Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

Gorm Finne Engelsen

Interpretable and Intelligble Statistical Models Applied to Data From Suicide Prevention Research

Master's thesis in Applied Physics and Mathematics
Supervisor: Mette Langaas
Co-supervisor: Linde Melby
June 2024

Gorm

NTNU
Norwegian University of
Science and Technology

Gorm Finne Engelsen

Interpretable and Intelligble Statistical Models Applied to Data From Suicide Prevention Research

Master's thesis in Applied Physics and Mathematics
Supervisor: Mette Langaas
Co-supervisor: Linde Melby
June 2024

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

NTNU
Norwegian University of
Science and Technology

ABSTRACT

In this thesis, we investigate the relation between the commencement of medication at an acute psychiatric department and the clinical assessment tool PANSS-EC. Explainable Boosting Machine (EBM) is used to model the probability of the commencement of medication and evaluate its performance by comparing the results to the two benchmark models Generalized Additive Model (GAM) and Generalized Linear Model (GLM). The performance measure used is the area under the receiver operating characteristic curve (AUC). Simulation studies are performed to analyze the model performance on known underlying relations to further provide a basis for the analyses of the real data.

Our results on simulated and real data indicate that the EBM model is not notably better than GAM and GLM in terms of general performance, but has clear strengths in inherently detecting interactions and discontinuities in the data. Nonlinear relations between the commencement of benzodiazepines, mood stabilizers, and PANSS-EC were identified.

The contributions of this thesis include a comprehensive presentation and comparison of models to analyze small binary classification datasets from medical research, an up-to-date explanation of the theory behind the EBM model, a thorough evaluation of the use of the EBM model in suicide research data and useful insight into the underlying relations between the commencement of medication and PANSS-EC at an acute psychiatric department.

SAMANDRAG

I denne avhandlinga undersøkjer vi samanhengen mellom oppstart av medisinar ved ein akutt psykiatrisk avdeling og det kliniske vurderingsverktøyet for uro, PANSS-EC. Vi bruker Explainable Boosting Machine (EBM) for å modellere sannsynet for oppstart av medisinering og vurderer prestasjonen til modellen ved å samanlikne resultata med referansemodellane Generaliserte Additive Modellar (GAM) og Generaliserte Lineære Modellar (GLM). Prestasjonen til modellane vert målt ved bruk er arealet under "receiver operating characteristic" kurva (AUC). Simuleringstudier er utført for å analysere modellprestasjonen på kjende underliggjande relasjonar. Dette vart gjort for å ytterlegare underbygge analysane av den verkelege dataa.

Resultata våre tilseier at EBM-modellen ikkje er vesentleg betre enn GAM og GLM i høve til generell prestasjon, men har klare styrker i å påvise interaksjonar og diskontinuitetar i dataa. Ikkje-lineære samanhengar mellom oppstart av benzodiazepinar og PANSS-EC, samt mellom oppstart av stemningsstabiliserande middel og PANSS-EC vart identifiserte og presentert.

Hovudbidraga frå denne avhandlinga inkluderer ei grundig framstilling og samanlikning av modellar for å analysere små binære klassifiseringsdatasett frå medisinsk forsking, ein grundig vurdering av bruken av EBM-modellen i denne forskinga knytt til sjølvmord og nyttig innsikt i dei underliggjande relasjonane mellom oppstart av medisinar og PANSS-EC ved ein akutt psykiatrisk avdeling.

PREFACE

This thesis is submitted to the Norwegian University of Science and Technology (NTNU) in the field of statistical learning and inference. It was written at the Department of Mathematical Sciences during the spring of 2024 in collaboration with St. Olav Hospital, Clinic for Mental Health Care. This master's thesis aims to investigate the commencement of medication at an acute psychiatric department using EBM, focusing on the relation to the clinical assessment tool PANSS-EC. Some parts of this thesis are extensions of a previous project thesis by the same author, written in collaboration with the same department and people. I want to thank Linde Melby (NTNU) for her insight in mental health and medicine. Thanks to $\int$-boys for making physics and mathematics endurable. I thank my friends, family and Emilie for being there for me throughout these five years. I hope you know how much it has meant to me. I would also like to express my utmost gratitude to my supervisor, Professor Mette Langaas (NTNU). Her help with this thesis has been invaluable, and I would not nearly have been able to complete it without her guidance and support. I feel truly lucky to have had the pleasure of getting to know her, working with her, and finishing five years of studies with her as my supervisor.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER

ONE

INTRODUCTION

Suicide is a significant public health issue worldwide. Despite this, a lack of knowledge of the root causes remains elusive. A more precise understanding of the relations of underlying factors of patients admitted to acute psychiatric departments is crucial in providing correct treatment for patients and enhancing important knowledge about general mental health issues.

New statistical models within machine learning allow for an unprecedented performance when it comes to predicting outcomes based on complex and high-dimensional data. In analysis within medicine, the understanding of the underlying processes in models is of importance. Therefore, the new wave of explainable artificial intelligence has proved very useful in this field.

In this thesis, we aim to investigate the relation between a clinical assessment tool measuring psychotic symptoms severity and the commencement of medication at an acute psychiatric department. Given our focus on understanding the relationship rather than merely its significance, interpretability and overall predictive performance are crucial. We will therefore use a relatively new machine learning model called Explainable Boosting Machine (EBM), which is known for its interpretability but has also shown to have as good performance, if not better, than other state-of-the-art full complexity machine learning methods (Nori et al. 2019*a*). The use of this model on the data presented in this thesis is evaluated and compared to the benchmark models, the Generalized Additive Model (GAM), and the Generalized Linear Model (GLM).

This thesis contributes to the field by enhancing our understanding of medication impacts within acute psychiatric departments, potentially leading to improved patient care and knowledge about mental health issues. This is in accordance with the United Nations sustainability goals, where the third goal is to "ensure healthy lives and promote well-being for all at all ages" (United Nations 2023).

This thesis is crafted with an emphasis on being interpretable by researchers at St. Olavs Hospital, Clinic for Mental Health Care, particularly in the presentation of the results, discussion and conclusions from our analyses. We have consciously

employed terminology and provided explanations accessible to a broad audience, not exclusively statisticians. This approach ensures that the findings are comprehensible and applicable clinically.

Parts of the work in this thesis are a continuation of previous work done in Engelsen (2024), a project thesis written in the winter of 2024 in collaboration with the same departments and people. Section 2.6, 2.7 and 2.8 in Chapter 2 are in large parts similar to the project thesis, with minor modifications to notation. The data presented in Chapter 3 is partly from the same studies as in the project thesis and, therefore, holds similarities in the presentation of the data used.

The thesis is structured in the following manner. We begin Chapter 2 by presenting the theoretical background that will be used in later analyses. This is mainly focused on the three models, GLM, GAM, and EBM, and the inference metrics used for the models. In Chapter 3, we present the data used in the analyses and the feature engineering done. Further, in Chapter 4, four simulation studies comparing the three models on synthetic datasets are presented. The synthetic datasets are meant to mimic variables of particular interest in the real dataset and will provide insight into the models. In Chapter 5, we present the results from the analyses using the data in Chapter 3. We end the thesis by discussing the results obtained and drawing conclusions in Chapter 6 and 7, respectively.

CHAPTER

TWO

THEORY

In this chapter, the theoretical background of the thesis analyses is presented. Pertinent terminology is specified before the problem is defined and presented, along with the relevant underlying theory of the models used. Lastly, the theory behind the performance measures and tests used for evaluating the models is thoroughly deduced.

2.1 Terminology

Interpretability, *explainability* and *intelligibility* are words that differ little in literal meaning but are often used to describe different characteristics of models in machine learning. Following the terminology of Nori et al. (2019*a*), one can divide machine learning models into two categories, glass box models and black box models. Glass box models are models that are directly interpretable due to their structure, while black box models are more difficult to understand due to their complexity. When using black box models, there is in general a need to use post-hoc methods to further explain the results, while in glass box models, the results are directly interpretable. The term *explainability* is often used while describing the ability to understand or explain black box models, while the term *intelligibility* is used when describing the ability to understand glass box models. Following Oxford Learner's Dictionaries *intelligibility* is defined as "the fact of being able to be easily understood" (Oxford University Press n.d.). In this thesis, we will use the term *intelligibility* when describing the ability to understand a model, as we are working with glass box models.

The terms *variable*, *feature* and *covariate* are often used interchangeably in statistics and machine learning. In this thesis, the term variable refers to the properties of the data and can be divided into the response variable and explanatory variables. The term covariate refers to the explanatory variables when used to model the response variable. Much of the literature on the subject also uses the term feature instead of covariate. The terms explanatory variable, covariate and feature can be used interchangeably without loss of meaning. To make matters worse, the

word "term" can also be used when referring to a variable used in a model. In this thesis, term will mostly be used when referring to interactions between two explanatory variables. On behalf of the statistical community, we apologize for the ambiguity.

2.2 Binary Classification Problem

A sample of size M, called a training set, and a sample of size N, called a test set, is randomly drawn from a population of interest. In these samples a response variable Y and a vector of covariates $X = (x_1, x_2, \ldots, x_p)$ are observed. Focusing on a binary response variable, noting $Y = 0$ as a negative and $Y = 1$ as a positive observation, two groups of observations in the sets, $\mathcal{D}_0$ and $\mathcal{D}_1$, can be inferred. For the training set, two groups of size $|\mathcal{D}_0^{train}| = m_0$ and $|\mathcal{D}_1^{train}| = m_1$, where $m_0 + m_1 = M$, are obtained. Similarly, for the test set, two groups of size $|\mathcal{D}_0^{test}| = n_0$ and $|\mathcal{D}_1^{test}| = n_1$, where $n_0 + n_1 = N$, are obtained.

A binary classification problem aims to construct a classifier that can separate the two groups $\mathcal{D}_0$ and $\mathcal{D}_1$ in the best way possible. The training set is used to construct the classifier, introducing an estimated prediction, $\hat{p}(X) = \hat{\mathbb{P}}(Y = 1|X)$, defined as the estimated probability of an observation being positive as a function of the covariate vector X. The goodness of the classifier can then be evaluated on new unseen data using the test set.

The training set of M observations is denoted as $\mathcal{D}^{train} = \{(X_i, Y_i)\}_1^M$, where the observed vector of covariates for observation i is denoted $X_i = (x_{i1}, x_{i2}, \ldots, x_{iq})$ and the corresponding response variable is denoted Y_i.

2.3 Generalized Linear Model

The GLM is a generalization of the standard linear regression model. The model can be characterized by three key components, the random component, the systematic component and the link function.

The random component defines the distribution of a response variable Y coming from the exponential family of probability distributions. The exponential family is a class of probability distributions where the probability or density functions can be written in the form

$$\mathcal{P}(Y_i|\theta_i) = \exp\left(\frac{Y_i\theta_i - b(\theta_i)}{\phi} w_i + c(Y_i, \phi, w_i)\right), \tag{2.1}$$

where θ_i is the canonical parameter, ϕ is the dispersion parameter, w_i is a weight function, $b(\theta_i)$ is a known function called the cumulant function and c is a known function (Dunn & Smyth 2018, p. 212). The exponential family include distributions such as the normal, Poisson, binomial and gamma distribution.

The systematic component for a GLM is a linear combination of the predictors $x_{i1}, x_{i2}, \ldots, x_{iq}$ and the parameters $\beta_0, \beta_1, \ldots, \beta_q$, written as $\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_q x_{iq}$.

The link function, denoted as $g(\cdot)$, is a known, monotonic, differentiable function that relates the expected value of Y_i, $E[Y_i] = \mu_i$, to the systematic component and thus binds the systematic component to the random component,

$$g(\mu_i) = \eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_q x_{iq}. \tag{2.2}$$

The link function is canonical if $\eta_i = \theta_i = g(\mu_i)$, which is highly practical as it simplifies the estimation of the parameters. The inverse of the link function is referred to as the response function, denoted $h(\eta_i)$. In Table 1, some examples of the canonical link and response functions for different random component distributions are shown.

Distribution	**Link function**	**Response function**	**Name**
Normal	$g(\mu) = \mu$	$h(\eta) = \eta$	Identity
Poisson	$g(\mu) = \log(\mu)$	$h(\eta) = e^{\eta}$	Log
Binomial	$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$	$h(\eta) = (1 + e^{-\eta})^{-1}$	Logit
Gamma	$g(\mu) = -\mu^{-1}$	$h(\eta) = -\eta^{-1}$	Negative inverse

Table 1: Link and response functions for typical distributions from the exponential family.

Dealing with a binary classification problem, this thesis focuses on responses from the binomial distribution and employs the logit link function. Denoting p_i as the probability of a positive for observation i with covariate vector X_i, it follows that $E[Y_i] = \mu_i = p_i$ and therefore

$$p_i = \mathbb{P}(Y_i = 1|X_i) = \frac{e^{\eta_i}}{1 + e^{\eta_i}} = \frac{1}{1 + e^{-\eta_i}}, \tag{2.3}$$

deduced from the logit link function and (2.2).

2.3.1 Parameter Estimation

The unknown parameters $\beta_0, \beta_1, \ldots, \beta_q$ can be estimated by using maximum likelihood estimation (MLE). The likelihood of the parameter vector β is defined as the probability of observing the training data given the parameters β, $L(\beta) = \prod_{i=1}^{m} f(Y_i|\beta)$. Since the natural logarithmic function is monotonic, it is practical to work with the log-likelihood function. Using the notation from (2.1), the log-likelihood function is defined as

$$l(\beta) = \sum_{i=1}^{m} l_i(\beta) = \sum_{i=1}^{m} \frac{1}{\phi} \left(Y_i \eta_i - b(\eta_i)\right) w_i + \sum_{i=1}^{m} c(Y_i, \phi, w_i).$$

The score function is defined as the derivative of the log-likelihood function with respect to the parameters β,

$$s(\beta) = \frac{\partial l(\beta)}{\partial \beta}.$$

This can be written on the matrix form $s(\beta^{(t)}) = \mathbf{X}^T\mathbf{D}\Sigma(\mathbf{Y} - \mu)$, where $\mathbf{X}$ is the design matrix, $\mathbf{D}$ is a diagonal matrix with the derivative of the response function on the diagonal, and $\mathbf{\Sigma}$ is a diagonal matrix with the variance of the response vector $\mathbf{Y}$ on the diagonal.

Further, the observed information matrix is defined as the negative derivative of the score function with regard to β,

$$\mathcal{J}(\beta) = -\frac{\partial s(\beta)}{\partial \beta^T} = -\frac{\partial^2 l(\beta)}{\partial \beta \partial \beta^T}.$$

The maximum likelihood estimator of β is denoted as $\hat{\beta}$ and is found as the solution to $s(\beta) = 0$. A solution can be found using an iterative technique such as the Newton-Raphson method. The formula for iteration $t + 1$ of this method is

$$\hat{\beta}^{(t+1)} = \hat{\beta}^{(t)} + \mathcal{J}(\hat{\beta}^{(t)})^{-1} s(\hat{\beta}^{(t)}). \tag{2.4}$$

Since the observed information matrix is often difficult to compute, the expected information matrix is used instead, also referred to as the Fisher information (Dunn & Smyth 2018, p. 186). The Fisher information is defined as the expected value of the observed information matrix, $\mathcal{F}(\beta) = E(\mathcal{J}(\beta))$. This can be written on the matrix form $\mathcal{F}(\beta) = \mathbf{X}^T\mathbf{W}(\beta)\mathbf{X}$, where $\mathbf{W}$ is a diagonal matrix with elements $h'(\eta_i)^2 \text{Var}(Y_i)^{-1}$ for $i = \{1, \cdots, M\}$, referred to as working weights. Using the Fisher information in (2.4), gives iteration $t + 1$ of the Fisher scoring method,

$$\hat{\beta}^{(t+1)} = \hat{\beta}^{(t)} + \mathcal{F}(\hat{\beta}^{(t)})^{-1} s(\hat{\beta}^{(t)}). \tag{2.5}$$

It is worth mentioning that for canonical link functions, the observed and expected information are equal, and the Newton-Rhapson and Fisher scoring methods are equivalent to each other (McCullagh & Nelder 1989, p. 43). The Fisher scoring iteration can be written in the matrix form as

$$\hat{\beta}^{(t+1)} = (\mathbf{X}^T\mathbf{W}(\hat{\beta}^{(t)})\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}(\hat{\beta}^{(t)})\tilde{\mathbf{Y}}^{(t)}, \tag{2.6}$$

where $\tilde{\mathbf{Y}}$ is referred to as the working response vector and has elements $\tilde{Y}_i^{(t)} = \eta_i + (Y_i - h(\eta_i))h'(\eta_i)^{-1}$ for $i = \{1, \cdots, M\}$ (Dunn & Smyth 2018, p. 246). $\hat{\beta}$ is then normally distributed with mean and variance, $\hat{\beta} \sim \mathcal{N}(\beta, \mathcal{J}^{-1})$.

Running Fisher scoring iterations while updating the working response vector and the working weights is called the iterated reweighted least squares method (IRLS). The IRLS method is a fast and efficient method for estimating the parameters of a GLM and is the method used in the `glm()` function in the `stats` package in R (R Core Team 2023), which is used in the analyses later.

Non-linear transformations and interaction terms can be modeled in GLM, though this needs to be explicitly specified. This often requires a priori knowledge and can be a cumbersome process.

2.4 Generalized Additive Model

GAM is a generalization of GLM created by Hastie and Tibshirani in Hastie & Tibshirani (1987) and allows for non-linear relationships in the systematic component while maintaining additivity (James et al. 2021, p. 307). This is based on that each covariate x_j is modelled using functions $f_j(x_j)$, resulting in the following model for the systematic component,

$$\eta = \beta_0 + f_1(x_1) + f_2(x_2) + \ldots + f_q(x_q). \tag{2.7}$$

Following Lou et al. (2012), these functions are called *shape functions*. GAM is considered more general than GLM since it, with shape functions, can without explicit specification model non-linear relationships that need to be manually specified in GLM. The main downside to GAM is that it is limited to maintaining additivity and can, for many covariates, miss important interactions (James et al. 2021, p. 309). Interaction terms can be added manually, though this can be cumbersome as there often are many possible interactions to consider for inclusion.

2.4.1 Shape Functions

The shape functions, $f_j(x_j)$, model a single covariate in the systematic component and can be functions of a variety of types, such as splines, univariate trees or univariate ensembles of trees. This flexibility contrasts with GLM, which only uses linear functions of the covariates by default. However, it is important to note that in GLM, the linearity refers to the model parameters β, and users have the option to incorporate non-linear transformations of the covariates.

The most common shape functions used in GAMs are splines and local regression, though other shape functions using decision trees have laid the foundation for the EBM model, presented later. The theory behind these shape functions will be presented in the following sections.

Smoothing Splines

Smoothing splines for regression problems are based on minimizing the residual sum of squares for some shape function, $f_j(x_j)$, with a constraint on the smoothness of the function to prevent overfitting. In the case of classification problems, the same constraint on smoothness is used but with a different likelihood function. The log-likelihood in a binomial GAM with a logit link function is

$$l(\beta) = \sum_{i=1}^{M} \left[Y_i \log(p_i) + (1 - Y_i) \log(1 - p(x_i)) \right],$$

where p_i is the probability from (2.3). For the smoothing spline for a binomial logit GAM, the log-likelihood to maximize is the penalized log-likelihood criterion

$$l_{pen}(\beta, \lambda) = l(\beta) - \frac{1}{2}\sum_{j=1}^{q} \lambda_j \int f_j''(t_j)^2 dt_j.$$

The optimal $f_j(x_j)$ is then a finite-dimensional natural spline with knots at the unique values of x_j, defined as $f_j(x_j) = \sum_{k=1}^{K} N_k(x_j)\theta_{kj}$ with K degrees of freedom, where N_k is the basis function for the i-th natural spline basis function and θ_{kj} are coefficients (Hastie et al. 2009, p. 127). In the analyses later, smoothing splines will be used as shape functions in the GAM model using the `gam` package in R (Hastie 2023). In this package, the degrees of freedom are set to $K = 4$ by default, which is the number used in later analyses.

2.4.1.1 Uncertainty

First, consider an additive model with a normal response. After a smoothing spline has been obtained, one can write the shape function as

$$\hat{f}(x) = \mathbf{SY},$$

where $\mathbf{S}$ is the smoothing matrix and $\mathbf{Y}$ is the response vector. If we assume that $\text{Cov}(Y) = \sigma^2 I$, where I is an identity matrix, the covariance matrix for the shape function is

$$\text{Cov}(\hat{f}(x)) = \mathbf{SS}^T\sigma^2. \tag{2.8}$$

The pointwise standard errors bands can be calculated as $\pm 1.96\sqrt{\text{diag}(\text{Cov}(\hat{f}(x)))}$ and results in a 95% confidence interval (Hastie & Tibshirani 1990, p. 60). This can be generalized to a logistic regression fitted by backfitting (see Section 2.4.2) as the final iteration can be seen as a weighted linear regression of an adjusted dependent variable, see Hastie & Tibshirani (1987) Section 4.4 for more details.

Local Regression

Local regression is based on fitting a regression function only at the nearby observations for a certain point and doing this over the entire range of observations. The observations are weighted so that the closest observations are weighted more than observations further away. One can use different regression functions, such as constant, linear or quadratic regression. For linear regression, the local regression function to minimize at x_0 is defined as

$$\text{argmin}_{\beta_0,\beta_1} \sum_{i=1}^{N} K_{i0}(y_i - \beta_0 - \beta_1 x_i)^2,$$

where K_{i0} are the weights for the neighboring observations around x_0 (James et al. 2021, p. 304). Local regression is the other possibility for shape function in the `gam` package in R (Hastie 2023). It will not be used in our analyses later but is included here for completeness.

Decision Tree

A decision tree is a non-parametric supervised learning method used for classification and regression. For regression problems, decision trees are often called regression trees, where the goal is to divide the covariate space into non-overlapping regions in order to estimate a response variable. Each leaf node in the regression tree corresponds to a numerical estimation of the response variable, often being calculated as the mean of all observations in the leaf node. The division of the covariate space is done by a splitting criterion in a greedy manner, such as variance reduction or minimizing the mean squared error (MSE) defined as

$$MSE = \frac{1}{M}\sum_{i=1}^{M}(Y_i - \hat{Y}_i)^2,$$

where Y_i is the observed response and $\hat{Y}_i$ is the estimated probability response for observation i. As mentioned above, all shape functions only relate a single attribute to the target, and decision trees are no exception, even though decision trees are usually built on multiple covariates.

In the case of a binary classification problem, a decision tree is called a classification tree. For classification trees, the goal is to divide the covariate space into non-overlapping subspaces in order to create a decision rule for predicting class membership. The division is done by a splitting criterion in a greedy manner, often using different measures of minimizing node impurity, such as the Gini index or cross-entropy. These are defined as

$$G = \sum_{k=1}^{K}\hat{p}_{lk}(1 - \hat{p}_{lk})$$
$$D = -\sum_{k=1}^{K}\hat{p}_{lk}\log(\hat{p}_{lk}),$$

where $\hat{p}_{lk}$ is the proportion of training observations in leaf node l that are from class k and $K = 2$ for binary classification (Hastie et al. 2009, p. 271). The overall misclassification rate remains stable across a reasonable range of splitting rules, although Gini is often preferred over cross-entropy (Breiman et al. 1984, p. 94, 111). The tree is expanded by adding binary splits until a stopping criterion is met, such as a minimum number of observations in each leaf node.

Bagged Decision Trees

An ensemble of decision trees can be used to improve the prediction accuracy of a single decision tree, which in combination with bagging (bootstrap aggregating) greatly reduces variance (Bauer & Kohavi 1999). Bagging is done by bootstrap sampling the training data into B bootstrap samples, which means sampling the data with replacement. A decision tree is then fitted to each sample on each of the q covariates. After fitting q trees for each of the B samples, we are left with

$B \times q$ trees and can estimate the final prediction for each covariate by taking the average of the predictions, $\hat{p}_j^b$, of all trees for the corresponding covariate,

$$\hat{f}_j(x_j) = \frac{1}{B} \sum_{b=1}^{B} \hat{p}_j^b(x_j).$$

Algorithm 1 Boosted Bagged Trees Algorithm

```
F(x) ← 0
f_j ← 0
for b ∈ 1 : B do                          ▷ Iterate over bootstrap samples
    for j ∈ 1 : q do                      ▷ Iterate over covariates
        Fit tree on variable j on the residuals
        Update f_j by adding new f_j
        Update F(x) with new f_j
    end for
end for
```

Boosted Decision Trees

Gradient boosted trees is a method that builds an ensemble of trees sequentially, where each tree is fitted to a function of the result of all previous trees. The standard gradient boosting algorithm was first presented in Friedman (2001), where the goal is to minimize the expected value of some loss function. The loss function can be the negative log-likelihood function, which, for our binary logit GAM, is

$$\sum_{i=1}^{M} L(Y_i, \hat{Y}_i) = \sum_{i=1}^{M} Y_i \log(\hat{Y}_i) + (1 - Y_i) \log(1 - \hat{Y}_i),$$

where $\hat{Y}_i$ represents the probability of belonging to class 1 and is obtained through the sigmoid function,

$$\hat{Y}_i = \frac{1}{1 + e^{-\hat{f}(X_i)}},$$

where $\hat{f}(X_i) = \sum_{b=1}^{B} \sum_{j=1}^{q} \delta \hat{f}_j^b(x_{ij})$ being the final model estimation defined as the sum of all earlier estimations with $\delta \in [0, 1]$ being the learning rate and $\hat{f}_j^b$ being the contribution to $\hat{f}(X_i)$ of tree b for covariate j (Midtfjord et al. 2022). After reaching a maximum number of trees or not improving the loss function, the final prediction is defined as the sum of all trees, and will result in the shape function.

Boosted Bagged Trees

Combining the methods of boosting and bagging results in a method called boosted bagged trees. This method is based on using a bagged ensemble of the training data

in each step of stochastic gradient boosting. This means that instead of sampling without replacement, as in stochastic gradient boosting, sampling is done with replacement using bagging. It is shown in Lou et al. (2012) that using gradient boosting of size-limited bagged trees for the shape functions yields better accuracy than other methods while still maintaining the intelligibility of GAM. This is, according to these authors, due to the limitations of spline-based methods, which tend to underfit the data and miss potentially crucial non-smooth tendencies.

2.4.2 Estimation

Local Scoring Algorithm

Two different methods for fitting additive models are presented. For a binomial logit GLM, it is shown in Section 2.3.1 how parameter estimation is done using the IRLS algorithm. The penalized logistic regression with smoothing splines can be maximized using a "backfitting algorithm within a Newton-Raphson procedure" (Hastie et al. 2009, p. 261), which is called a local scoring algorithm and is presented in Algorithm 2. This algorithm is based on Algorithm 9.1 and 9.2 in Hastie et al. (2009). An important feature of the backfitting algorithm is the existence and uniqueness of solutions for linear smoothers. A constraint on the shape function is necessary to ensure that the solution is unique, which is the sum-to-zero constraint. This constraint is achieved by mean centering the shape functions so that the expected value of each shape function is zero (Hastie & Tibshirani 1990, p. 115).

Algorithm 2 Backfitting

1: Compute $\hat{\beta}_0 = \log \frac{\bar{y}}{1-\bar{y}}$, where $\bar{y} = \frac{1}{M}\sum_{i=1}^{M} y_i$ and set $f_j(x_j) = 0, \forall j$.

2: Define $\hat{\eta}_i = \hat{\beta}_0 + \sum_{j=1}^{p} f_j(x_{ij})$ and $\hat{p}_i = \frac{e^{\hat{\eta}_i}}{1+e^{\hat{\eta}_i}}$.

a) Compute the working response

$$z_i = \hat{\eta}_i + \frac{y_i - \hat{p}_i}{\hat{p}_i(1-\hat{p}_i)}$$

b) Compute the working weights

$$w_i = \hat{p}_i(1-\hat{p}_i)$$

c) Use weighted backfitting to the working response, z_i, and weights, w_i, that is:

Iterate:

$\hat{f}_j \leftarrow$ Fit a smoothing spline with weights, w_i to $\{z_i - \hat{\beta}_0 - \sum_{k\neq j}^{q} f_k(x_{ik})\}$.

$\hat{f}_j \leftarrow \hat{f}_j - \frac{1}{m}\sum_{i=1}^{m} \hat{f}_j(x_{ij})$ (mean centering).

3: Continue step 2, until the change in the functions falls below a prespecified threshold

Gradient Boosting

Gradient boosting can also be used to fit additive models, but since shape functions are made for all covariates, the algorithm is modified to sequentially cycle through

all covariates for each boosting iteration (Lou et al. 2012). The algorithm for classification problems can be seen in Algorithm 3, which is similar to the one shown in Algorithm 2 in Lou et al. (2012).

Algorithm 3 Gradient Boosting for Classification

$f_j \leftarrow 0, \forall j$ and $F(x_i) = \sum_{j=1}^{q} f_j(x_{ij})$.
for $l = 1$ to L **do**
  for $j = 1$ to p **do**
    $\tilde{y}_i \leftarrow \frac{e^{F(x_i)}}{1+e^{F(x_i)}}$
    Learn $\{R_{kl}\}_{k=1}^{K}$, a tree with K leaf nodes using $\{(x_{ij}, -y_i + \tilde{y}_i)\}_{i=1}^{M}$.
    $\gamma_{kl} \leftarrow \frac{y_i - \tilde{y}_i}{\tilde{y}_i(1-\tilde{y}_i)}$ for $k = 1, \ldots, K$.
    $f_j \leftarrow f_j + \sum_{k=1}^{K} \gamma_{kl} I(x_{ij} \in R_{kl})$.
    Update $F(x_i) = \sum_{j=1}^{q} f_j(x_{ij})$.
  end for
end for
return $F(x_i) = \sum_{j=1}^{q} f_j(x_{ij})$.

2.5 Explainable Boosting Machine

A major limitation of standard GAM is that it is not able to handle interactions of predictors automatically. There are several approaches to detecting interactions in additive models, however, according to Lou et al. (2013), all these approaches fall short in terms of complexity, power and correctly identifying interactions compared to the approach they present. They introduce the GA^{2}M model that adds selected terms of interactions to the systematic component of the standard GAM model. This can be written as

$$\eta = \sum_{j=1}^{q} f_j(x_j) + \sum_{\forall j,k, j \neq k} f_{jk}(x_j, x_k).$$

Using the notation introduced in Lou et al. (2013), let $\mathcal{U}^1 = \{\{i\} | 1 \leq j \leq q\}$ be the set of all indices for all covariates, $\mathcal{U}^2 = \{\{j, l\} | 1 \leq j < l \leq q\}$ be the set of all indices for all pairs of covariates and $\mathcal{U} = \mathcal{U}^1 \cup \mathcal{U}^2$. Denote x_u as the set of all covariates whose indices are in $u \subseteq \{1, \ldots, q\}$. For any $u \in \mathcal{U}$, let $\mathcal{H}_u$ denote the Hilbert space of Lebesgue measurable functions $f_u(x_u)$, such that $E[f_u(x_u)] = 0$ and $E[f_u(x_u)^2] < \infty$. The Hilbert space is equipped with the inner product $\langle f_u, f'_u \rangle = E[f_u f'_u]$. Further, let $\mathcal{H}^1 = \sum_{u \in \mathcal{U}^1} \mathcal{H}_u$ denote the Hilbert space of shape functions on univariate covariates and $\mathcal{H} = \sum_{u \in \mathcal{U}} \mathcal{H}_u$ the Hilbert space of both univariate and bivariate interaction shape functions of the form $F(\mathbf{x}) = \sum_{u \in \mathcal{U}} f_u(x_u)$. Lou et al. (2013) now formulate the problem to be solved as

$$\min_{F \in \mathcal{H}} E[L(y, F(\mathbf{x}))], \tag{2.9}$$

where $L(y, F(\mathbf{x}))$ is some loss function.

GA^{2}M Framework

The goal is to solve (2.9) and find the best model $F \in \mathcal{H}$ that minimizes some expected loss function $E[L(y, F(\mathbf{x}))]$. For regression problems, the loss function is often the squared error loss, $L(y, F(\mathbf{x})) = (y - F(\mathbf{x}))^2$. The general algorithm for the regression setting in the GA^{2}M framework is seen in Algorithm 4, where S is the set of selected pairs of covariates, and Z is the set of remaining pairs of covariates. The algorithm iterates until convergence, where the best additive model is found by minimizing the expected loss function over $H^1 + \sum_{u \in \mathcal{S}} \mathcal{H}_u$. For each pair of covariates $u \in Z$, the shape function F_u is calculated. Then, the best interaction pair is selected, added to the set of selected pairs S, and removed from the set of remaining pairs Z. This is done until there is no gain in accuracy.

Algorithm 4 GA^{2}M Framework (Regression setting)

$S \leftarrow \emptyset$
$Z \leftarrow U^2$
while not converge **do**
    $F \leftarrow \text{argmin}_{F \in H^1 + \sum_{u \in S} H_u} E[L(y, F(\mathbf{x}))]$
    $R \leftarrow y - F(\mathbf{x})$
    for all $u \in Z$ **do**
        $F_u \leftarrow E[R|x_u]$
    end for
    $u^* \leftarrow \text{argmin}_{u \in Z} E[L(y, F(\mathbf{x}))]$
    $S \leftarrow S \cup \{u^*\}$
    $Z \leftarrow Z - \{u^*\}$
end while

Fast Interaction Detection

The general GA^{2}M algorithm above is very computationally expensive for large datasets, specifically datasets with many covariates, due to the large number of possible interactions. Lou et al. (2013) presents a computationally efficient method of ranking all possible pairs of covariate interactions for inclusion in the model. This method is referred to as Fast Interaction Detection (FAST) and is based on creating simple estimations of the interaction pairs and ranking the pairs based on the fit of the model. For regression problems, the residual sum of squares is used to rank the pairs. For the detailed algorithm and specifications regarding calculations in the FAST algorithm, see Lou et al. (2013).

Explainable Boosting Machine

The FAST implementation of the GA^{2}M algorithm is referred to as Explainable Boosting Machine (EBM) and is introduced in the InterpretML framework draft presented in Nori et al. (2019*a*). The Explainable Boosting Machine is a glass-box model and is found to outperform many state-of-the-art black-box models, such as Random Forest and Boosted Trees while still offering the interpretability of GAM. An implementation written in Python of the method is available at Nori et al. (2019*b*). Here, the two main functions are `ExplainableBoostingClassifier()`

and `ExplainableBoostingRegressor()` for classification and regression problems, respectively. From these, the corresponding feature importance and shape functions can be extracted using the `explain_global()` function.

Variable Importance

The variable importance presented in EBM is a measure of how much each covariate contributes to the final prediction. This is calculated by using the *weighted mean absolute score* (WMAS) defined as the average of the absolute contribution score of each covariate over all observations. It can be written as

$$\text{WMAS}(x_j) = \frac{1}{M} \sum_{i=1}^{M} |f_j(x_{ij})|. \tag{2.10}$$

Another possibility is to use the difference between the maximum and minimum contribution score. In this thesis, WMAS is used, which is also the default setting for EBM.

An equivalent constraint to the sum-to-zero constraint in GAM is used in EBM, which results in the shape functions being centered around zero (InterpretML Team 2020).

Uncertainty

The EBM model builds an ensemble of boosted bagged trees and takes the average of all bagged ensembles to get the predictions. Each boosting round can be referred to as an internal bagged EBM model. The uncertainty of the EBM model is presented as error bars for each interval of the covariate. This is calculated as the standard deviation of the predictions over all bagged models. This is not the same as a confidence interval but it gives a good general sense of the uncertainty within the given interval (InterpretML Team 2021).

2.6 Receiver Operating Characteristic

For binary classification problems, a practical way of evaluating classifiers is by using receiver operating characteristic curves, referred to as ROC curves. For a given classifier estimated using a training set, this curve is constructed using a test set and gives an evaluation of the classifier at all possible classification thresholds. This gives insight into how well the classifier performs on new unseen data. There are different ways of defining ROC curves, and two different approaches are presented.

2.6.1 Conventional Definition

The conventional definition of an ROC curve is based on estimating the sensitivity and specificity for all possible classification thresholds c. The following notation is introduced for this definition.

Sensitivity is the probability of a positive observation being predicted as positive. Consider $\hat{p}(X)$ as the estimated probability that an observation belongs to class 1, which is a random variable since it depends on the response variable. The sensitivity can then be defined as the probability function $\mathbb{P}(\hat{p}(X) \geq c|Y = 1)$, for some threshold $c \in [0, 1]$. An unbiased estimator for the sensitivity is the true positive rate (TPR) and is calculated from the proportion of positive observations in a test set correctly classified for a given threshold c,

$$\text{TPR}(c) = \frac{1}{n_1} \sum_{i \in \mathcal{D}_1^{test}} I(\hat{p}(X_i) \geq c),$$

where $I(\cdot)$ is the indicator function returning 1 if the condition is true and 0 otherwise (Hastie et al. 2009, p. 277).

Specificity is the probability of a negative observation being predicted as negative, which can be defined as the probability function $\mathbb{P}(\hat{p}(X) < c|Y = 0)$. An unbiased estimator for specificity is called the true negative rate (TNR), defined as the proportion of negative samples correctly classified for a given threshold c,

$$\text{TNR}(c) = \frac{1}{n_0} \sum_{i \in \mathcal{D}_0^{test}} I(\hat{p}(X_i) < c),$$

(Hastie et al. 2009, p. 277). The estimated ROC curve is then obtained by plotting the true positive rate on the y-axis against the false positive rate, $\text{FPR} = 1 - \text{TNR}$, on the x-axis for all possible c values ranging in $c \in [0, 1]$. Be aware that different axes can often be used when plotting the ROC curve. With this definition, the closer to the upper left corner an ROC curve is, the better the classifier performs. A perfect classifier will have an ROC curve that goes through the point $(0, 1)$, while a random guessing classifier will have an expected ROC curve along the diagonal line and has a 50-50 chance of classifying observations correctly.

2.6.2 Hand and Till Definition

Following the notation and procedure presented in Hand & Till (2001), let $\hat{q} = 1 - \hat{p}$ be a stochastic variable and the estimated probability of an observation being negative. Let $h(\hat{q}) = h(\hat{q}|Y = 0)$ be the probability function of the estimated probability of belonging to $\mathcal{D}_0$ for negative observations and let $g(\hat{q}) = g(\hat{q}|Y = 1)$ be the probability function of the estimated probability for belonging to $\mathcal{D}_0$ for positive observations. Further we let $H(\hat{q}) = H(\hat{q}|Y = 0)$, and $G(\hat{q}) = G(\hat{q}|Y = 1)$ be the cumulative distribution functions corresponding to $h(\hat{q})$ and $g(\hat{q})$ respectively. An ROC curve can then be defined as the plot of $G(\hat{q})$ on the y-axis versus $H(\hat{q})$ on the x-axis for all values of $\hat{q} \in [0, 1]$. As in the previous section, an ROC curve that is on the diagonal line, $G(\hat{q}) = H(\hat{q})$, corresponds to a random guessing classifier, while a good classifier will have an ROC curve above this line, $G(\hat{q}) > H(\hat{q})$.

It is not obvious that these two conventions are equivalent from the definitions above. The equivalence is shown by the following relations,

$$\text{Sensitivity}(c) = \mathbb{P}(\hat{p} \geq c|Y = 1) = \mathbb{P}(\hat{q} < c|Y = 1) = \int_0^c g(\hat{q})d\hat{p} = G(c)$$
$$1 - \text{Specificity}(c) = 1 - \mathbb{P}(\hat{p} < c|Y = 0) = \mathbb{P}(\hat{q} < c|Y = 0) = \int_0^c h(\hat{q})d\hat{q} = H(c)$$

It is then clear that plotting the estimated Sensitivity(c) versus $1 -$ Specificity(c) for all $c \in [0, 1]$ is equivalent to plotting the estimated $G(c)$ versus $H(c)$ for all $c \in [0, 1]$.

2.7 Area Under ROC Curve

Various classification methods may result in different predictions and, therefore, different ROC curves on the test set. Comparing the ROC curves of two classifiers on the same test set is a good way to evaluate the classifier's performance. One can also use the area under the ROC curve, AUC, to examine how well the classifier can separate observations from $\mathcal{D}_0$ and $\mathcal{D}_1$. The AUC is equivalent to the probability that a randomly chosen observation from $\mathcal{D}_0^{test}$ has a lower prediction, $\hat{p}$, than a randomly chosen observation from $\mathcal{D}_1^{test}$. This is shown by using the definition of the ROC curve from Section 2.6.2,

$$\text{AUC}(\hat{p}) = \int_0^1 G(c)dH(c) = \int_0^1 G(c)h(c)dc, \tag{2.11}$$

and the general definition of the area under parametric curves (Hand & Till 2001). We have that for a specific estimated prediction $\hat{q} = c$, the probability that a randomly chosen sample from $\mathcal{D}_1^{test}$ has smaller $\hat{q}$ than c is $G(c)$. It can then be deduced that the probability of a randomly chosen observation from $\mathcal{D}_1^{test}$ having a smaller $\hat{q}$ than a randomly chosen observation from $\mathcal{D}_0^{test}$ is $\int G(c)h(c)dc$, which is equivalent to the AUC in (2.11). Using $\hat{q} = 1 - \hat{p}$, it is clear that

$$\mathbb{P}(\hat{q}(X_j) \leq \hat{q}(X_i)) = \mathbb{P}(1 - \hat{p}(X_j) \leq 1 - \hat{p}(X_i)) = \mathbb{P}(\hat{p}(X_i) \leq \hat{p}(X_j)).$$

This is the probability that a randomly chosen observation from $\mathcal{D}_0^{test}$ will have smaller $\hat{p}$ than a random observation from $\mathcal{D}_1^{test}$ and is defined as $\theta = \mathbb{P}(\hat{p}(X_i) \leq \hat{p}(X_j))$, for some $X_i \in \mathcal{D}_0^{test}$ and $X_j \in \mathcal{D}_1^{test}$. According to the work of Bamber (1975), for this to be valid for both continuous and finitely discrete $\hat{p}$, the following definition is presented,

$$\theta = \mathbb{P}(\hat{p}(X_i) < \hat{p}(X_j)) + \tfrac{1}{2}\mathbb{P}(\hat{p}(X_i) = \hat{p}(X_j)).$$

An unbiased estimator for this probability is

$$\hat{\theta}(\hat{p}) = \frac{1}{n_0 n_1} \sum_{i \in \mathcal{D}_0^{test}} \sum_{j \in \mathcal{D}_1^{test}} \left(I(\hat{p}(X_i) < \hat{p}(X_j)) + \tfrac{1}{2} I(\hat{p}(X_i) = \hat{p}(X_j)) \right). \quad (2.12)$$

This shows a simple way of calculating the AUC for a given classifier. Another practical way of calculating the AUC is based on first ranking all samples from both groups by their prediction, i.e., probability of being positive. The AUC can then be estimated by summing up all ranks, r_i, of the negative observations and subtracting their rank position, i, with the formula

$$\hat{\theta}(\hat{p}) = \frac{1}{n_0 n_1} \sum_{i \in \mathcal{D}_0} (r_i - i) = \frac{1}{n_0 n_1} \left(\sum_{i \in \mathcal{D}_0} r_i - \tfrac{1}{2} n_0 (n_0 + 1) \right).$$

Here r_i is the mid-rank for the i-th negative observation defined as $r_i = k + (l+1)/2$, where k is the number of observations ranked higher than the i-th negative sample and l is the number of observations with the same prediction $\hat{p}$. If there are no ties present, then $l = 1$ and $r_i = k+1$, which is just the normal rank. This is equivalent to (2.12), and can be shown to be equivalent to the Wilcoxon-Mann-Whitney U test (Hand & Till 2001).

2.7.1 Confidence Interval for θ

From Hanley & McNeil (1982), it is shown that the standard deviation of the estimated AUC, $\hat{\theta}$, can be calculated as

$$SD(\hat{\theta}) = \sqrt{\frac{1}{n_0 n_1} (\theta(1-\theta) + (n_0 - 1)(Q_1 - \theta^2) + (n_1 - 1)(Q_2 - \theta^2))}, \quad (2.13)$$

where Q_1 is equivalent to the probability that two randomly chosen positive observations have higher predictions than a randomly chosen negative observation and Q_2 is equivalent to the probability that a randomly chosen positive observation has a higher prediction than two randomly chosen negative observations. This can be used to construct a confidence interval for θ using the fact that $\hat{\theta}$ is asymptotically normally distributed under H_0 (Mann & Whitney 1947).

Estimating (2.13) by replacing θ with $\hat{\theta}$ and estimating Q_1 and Q_2 by randomly sampling from the test data and calculating the proportion of times the condition is true, $\widehat{SD}(\hat{\theta})$ is obtained. A 95% confidence interval for θ can then be constructed defined as

$$\mathrm{CI}_{\theta,0.05} = \hat{\theta} \pm z_{0.05/2} \widehat{\mathrm{SD}}(\hat{\theta}),$$

where $z_{0.05/2}$ is the critical value from the standard normal distribution at a 5% level.

2.8 Comparing Two or More ROC Curves

As mentioned earlier, an ROC curve is a good tool for evaluating how well different classifiers work on unseen test data. A statistical way of comparing the AUC of two different classifiers on the same test set can be deduced following the procedure of DeLong et al. (1988).

Given two different models, A and B, used on the same set of test data, the two models produce two different estimated predictor values, $\hat{p}_A$ and $\hat{p}_B$, and will therefore result in two different ROC curves. To evaluate the difference between area under the ROC curves, the following test hypotheses are introduced,

$$H_0 : \theta_A = \theta_B,$$
$$H_1 : \theta_A \neq \theta_B.$$

These test hypotheses can be generalized using a linear contrast, $\mathbf{L}\theta^\top$, with the (1×2) row vectors $\mathbf{L} = [1 \quad -1]$ and $\theta = [\theta_A \quad \theta_B]$. Using this, the following test statistic is obtained,

$$T = \frac{\mathbf{L}\hat{\theta}^\top}{[\mathbf{L}\mathbf{S}\mathbf{L}^\top]^{\frac{1}{2}}} \sim N(0,1), \tag{2.14}$$

where S is the estimated covariance matrix for $\hat{\theta}$ and $\hat{\theta} = [\hat{\theta}_A \quad \hat{\theta}_B]$, deduced in DeLong et al. (1988). The test statistic in (2.14) can be used to test $\mathbf{L}\theta^\top = 0$ vs $\mathbf{L}\theta^\top \neq 0$ and a p-value of the test may be calculated using

$$(\hat{\theta} - \theta)\mathbf{L}^T\left[\mathbf{L}\mathbf{S}\mathbf{L}^T\right]^{-1}\mathbf{L}(\hat{\theta} - \theta)^T \sim \chi^2_l. \tag{2.15}$$

A p-value, first used in Arbuthnott (1710), is the probability of obtaining results at least as extreme as the measured results under the assumption that the null hypothesis is true. In (2.15), l is the rank of $\mathbf{L}\mathbf{S}\mathbf{L}^\top$ and χ^2_l is the chi-squared distribution with l degrees of freedom. From this, a confidence interval can also be calculated as

$$\mathbf{L}\hat{\theta}^\top \pm z_{\alpha/2}[\mathbf{L}\mathbf{S}\mathbf{L}^\top]^{\frac{1}{2}}.$$

CHAPTER

THREE

DATA

In this chapter, we present the data used in this thesis. The data is collected in two cohort studies conducted by the Department of Acute Psychiatry at Østmarka, St. Olav's Hospital. The first is the *Acute Agitation study* (AA) conducted from September 2011 until May 2012 (Prestmo et al. 2020) and the second is the *Genetic and Affective Prediction study* (GAP) conducted from January 2016 until June 2017 (Høyen et al. 2022). The two studies are aggregated into one dataset and regarded as one study consisting of 710 patients after 17 patients were found to be present in both datasets and were consequently removed from the GAP dataset.

The combined AA and GAP data were analyzed in Ludvigsen (2023), where the aim was to identify risk factors for a syndrome called the suicide crisis syndrome. The combined dataset is also analyzed in Melby (2024), with a focus on describing the patient data on the suicide crisis syndrome and the correlation to clinical variables. Data on medication usage at admission and during the stay at the psychiatric department have recently been collected from the electronic patient journals for the combined AA and GAP datasets and play a key role in our data analyses. Medication data were not part of the work done in Ludvigsen (2023) nor Melby (2024).

In this thesis, the variable names obtained directly from the datasets is used in figures and tables, as this is deemed more fitting for the further use of the findings in this thesis. If there is any ambiguity regarding what the variables mean, we encourage the reader to refer to Table 4, which will be elaborated in Section 3.2.

3.1 Medication Usage

Information about the patients' current use of medication and commencement of medication at the time of admission to the acute psychiatric department was collected from the patient journals. The medications of interest are divided into the four following categories: antipsychotics, benzodiazepines, hypnotics and mood stabilizers. For medication use up to the time of admission, each category is coded

as 1 for usage and 0 for non-usage, and for the commencement of medication at the time of admission, each category is coded as 1 for commencement and 0 for non-commencement. If a patient has registered usage of the specific medication category at the time of admission, then the patient cannot have registered commencement of the same drug. The variables for the commencement of medication obtained from the studies are listed with descriptions in Table 2.

In Appendix A, the frequencies of the medication variables in AA and in GAP are compared. For many of these variables, including the four variables presented in Table 2, there is a significant difference in usage between AA and GAP. It is therefore expected that the study indicator will be an important covariate in later analyses.

Variable	**Description**	**Type**
OPPST_ANTIPSYK	Commencement of Antipsychotics	Binary
OPPST_BENZO	Commencement of Benzodiazepines	Binary
OPPST_HYPNOTIKA	Commencement of Hypnotics	Binary
OPPST_STEMNINGSSTAB	Commencement of Mood stabilizers	Binary

Table 2: Description of the commencement of medication variables, after the time of admission.

3.1.1 Missing Data

In Figure 1, a quick overview of the missing data patterns of the commencement medication variables can be seen. Of the total 710 patients, 110 patients are missing all four medication commencement variables. The other patterns have much smaller numbers of missing patients, with a total of 13 patients. Since the commencement of medication is the response variable in this thesis, all patients with missing observation of these variables will not be a part of the analyses. The following numbers are shown in the bar chart in the lower left part of Figure 1. There are 111 patients with missing information on the commencement of antipsychotics, 118 on the commencement of benzodiazepines, 121 on the commencement of hypnotics and 112 on the commencement of mood stabilizers. Further information on the size of the four datasets will be given in Section 3.4.

3.2 Clinical Variables

The variables used in Ludvigsen (2023) and Melby (2024) are defined as clinical variables. These variables are presented in Table 4 reproduced with permission from Ludvigsen (2023). In this table, some descriptive specifications have been made, as well as some minor adjustments. The §3.2 and §3.3 categories for referral and specialist paragraph are merged into a new category of "forced hospitalizations". The variables associated with diagnosis are recategorized from 11 factors into the four categories *Affective Disorders*, *Substance Abuse Disorders*, *Psychosis Disorders* and *Other Disorders*, shown in Table 3. From a medical point of view, this is sufficient, and this results in fewer categories and more patients in each

category. This will potentially improve the fit in the statistical analyses where the diagnosis category will be a covariate.

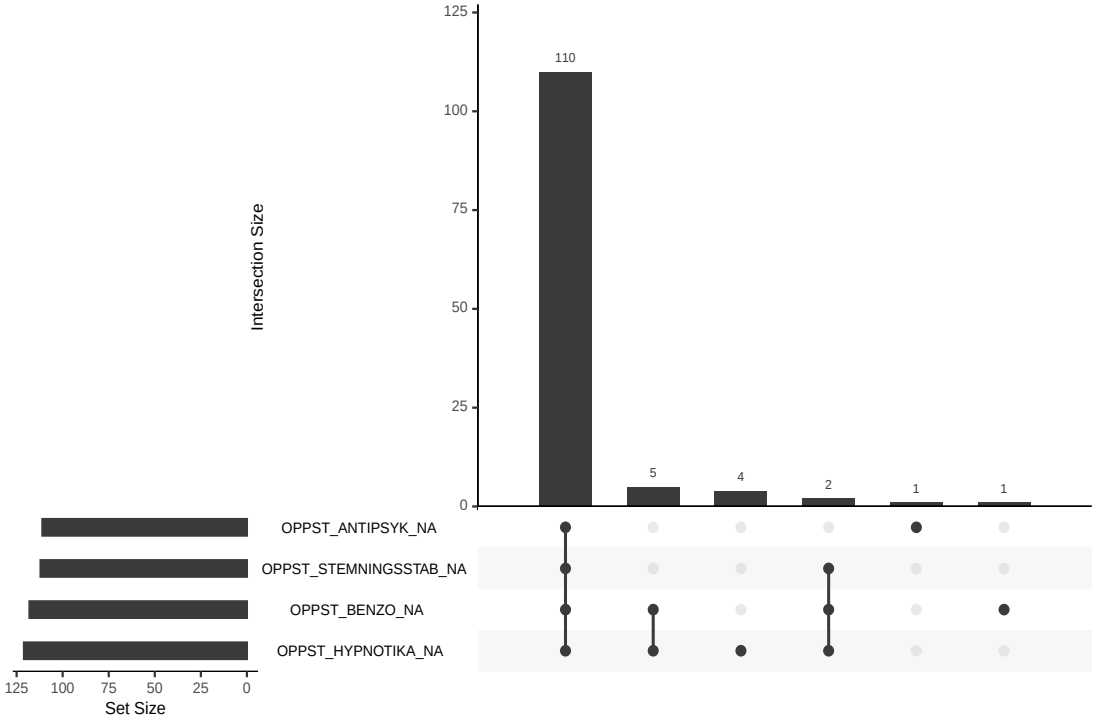


Figure 1: Missing data patterns in the medication data for the four commencement variables.

Diagnosis Categories	**Original Diagnosis Factors**
Affective Disorders	1. Behavioural and emotional disorders with onset usually occurring in childhood and adolescence
Substance Abuse Disorders	5. Mental and behavioural disorders due to psychoactive substance use
Psychosis Disorders	9. Organic, including symptomatic, mental disorders 10. Schizophrenia, schizotypal and delusional disorders
Other Disorders	2. Behavioural syndromes associated with physiological disturbances and physical factors 3. Disorders of adult personality and behaviour 4. Disorders of psychological development 6. Mental retardation 7. Mood [affective] disorders 8. Neurotic, stress-related and somatoform disorders 11. Uncertain mental disorder

Table 3: Recategorization of diagnosis factors into four diagnosis categories.

3.2.1 Missing Data

Figure 2 shows the number of missing observations for the clinical variables. From this, the variables associated with loss of self-esteem, lack of network, recent insomnia and if the patient has recently lost a relationship were concluded not to be suitable for further analysis due to high numbers of missing data. Assuming that the missing data is missing completely at random (MCAR), single imputation can be performed on all missing values. This means that all observations are assumed to have the same probability of being missing and that the missing mechanism is not dependent on other observed and unobserved variables. For categorical variables, this is done by taking the most frequently observed value and inserting it for the corresponding missing values, while for continuous variables, the mean value is used. The PANSS Exited Component (PANSS-EC) score, is defined as the sum of 5 other clinical variables and is therefore not necessary to impute after the other variables are imputed. More information on this variable is in the next section.

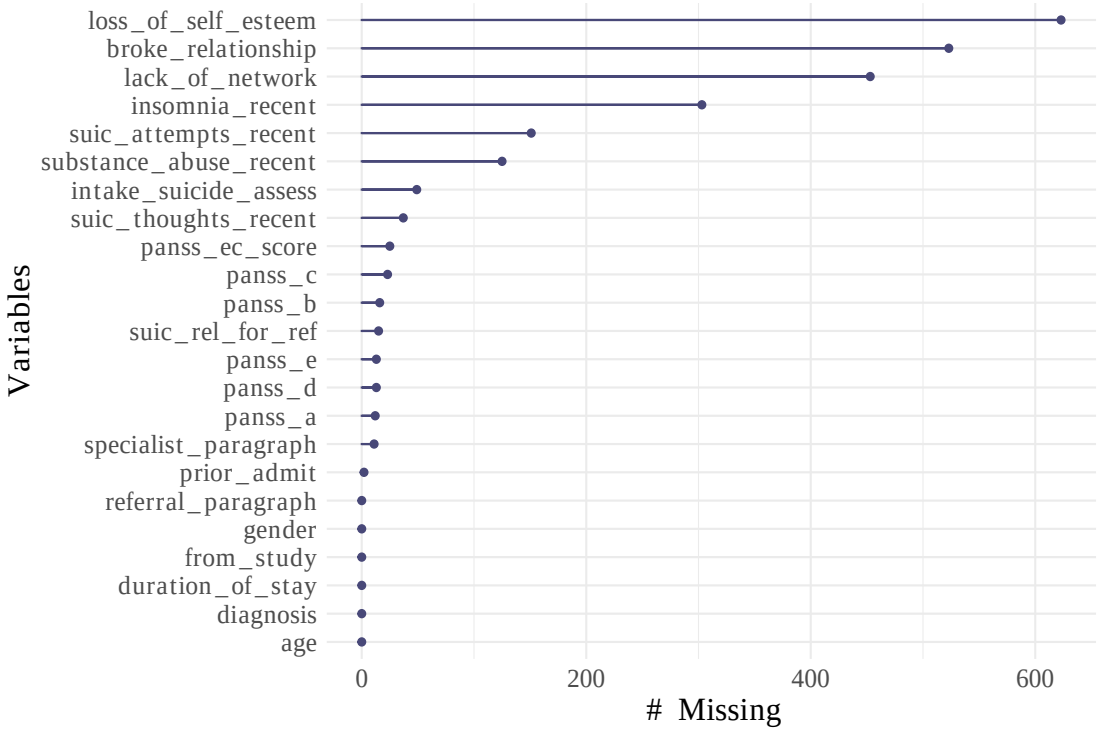


Figure 2: Number of missing observations in clinical data.

Variable	Description	Type of variable
from_study	which study the patient is from	binary (AA/GAP)
age	age of the patient	integer
gender	gender of the patient	binary (man/-woman)
prior_admit	has the patient been admitted to the psychiatric department before	binary (yes/no)
duration_of_stay	for how long was the patient admitted to the psychiatric department	integer
suic_rel_for_ref	was suicide relevant for the referral of the patient for the psychiatric department	binary (yes/no)
intake_suicide_assess	was the suicide risk high or low for the patient at admission as evaluated by the doctor on duty	binary (high/low)
referral_paragraph	the referral paragraph at admission given by the doctor that referred the patient. §2.1: voluntarily admission, §3.2: forced hospitalization without known diagnosis and §3.3: forced hospitalization with diagnosis	binary (§2.1 / §3.2 or §3.3)
specialist_paragraph	the specialist paragraph at admission given by the doctor on duty. §2.1: voluntarily admission, §3.2: forced hospitalization without known diagnosis and §3.3: forced hospitalization with diagnosis	binary (§2.1/ §3.2 or §3.3)
panss_ec	PANSS-EC, standardized questionnaire. Items A to E	ordinal(1-7)
panss_ec_score	sum of the PANSS-EC score	integer
suic_thoughts_recent	has the patient had suicidal thoughts recently (one month prior to the time of admission)	binary (yes/no)
suic_attempts_recent	has the patient had any suicidal attempts recently	binary (yes/no)
insomnia_recent	has the patient had insomnia recently	binary (yes/no)
broke_relationship	had the patient lost a relationship	binary (yes/no)
lack_of_network	has the patient a lack of network	binary (yes/no)
loss_of_self_esteem	has the patient lost their self-esteem	binary (yes/no)
substance_abuse_recent	has the patient abused some substance recently	binary (yes/no)
diagnosis_category	the patient main diagnosis category at discharge	factor (4 different categories)

Table 4: Description of clinical variables adapted from Ludvigsen (2023).

3.3 Positive and Negative Syndrome Scale

The Positive and Negative Syndrome Scale (PANSS) is a standardized questionnaire used to assess the severity of symptoms in patients with schizophrenia and was published in Kay et al. (1987). PANSS is also widely used to assess general psychopathology in patients. The most common structure to PANSS is the five-factor solution consisting of the components *Positive*, *Negative*, *Disorganized*, *Excited* and *Anxiety/Depression*.

The exited factor in PANSS is referred to as the PANSS Excited Component, referred to as PANSS-EC, and is considered to be one of the most simple and intuitive scales to assess agitation in patients (Montoya et al. 2011). Only the PANSS-EC is used in this thesis. It consists of 5 the subfactors excitement, tension, hostility, uncooperativeness and poor impulse control, which all are given a score ranging from 1 (not present) to 7 (extremely severe), seen as the variable `panss_ec` in Table 4. The evaluation is performed by trained medical doctors. The total score of a PANSS-EC evaluation ranges between 5 and 35, and can be seen in Table 4 as `panss_ec_score`.

3.4 The Datasets

The clinical and medication data are combined into one dataset. This is done using the unique patient numbers as key. The combined dataset consists of 710 patients and 17 variables. Since the commencement of medication will be the response variable and the inference between the commencement of different medications is not of interest, the other commencement variables are consequently removed when analyzing each of them separately. This results in four different datasets, one for each of the four commencement variables, which each consists of 13 clinical variables, as well as the relevant commencement variable. For each commencement variable, all patients with registered usage (at admission) of the corresponding drug are not included. This is due to the fact that only the patients who have the possibility to commence on a new drug are of interest. The sample sizes of the final datasets are presented in Table 5. The column Usage_0 shows the number of patients who don't have registered usage of the medication at admission and are then the possible candidates for the commencement of the medication. The column $\text{Commencement}_{\text{NA}}$ shows the number of patients with missing values for the commencement of the medication for the possible candidates. Removing the number of patients with missing medication commencement from the possible candidates gives the total number of patients in each dataset in the column Patients. The column Case Ratio shows the ratio between the number of patients who commence a medication and the total number of patients in the dataset. This is used later when splitting the datasets into training and test sets.

From the initial 710 patients, the dataset has been refined to four distinct datasets seen in Table 5. Each dataset consists of 13 variables, where ten are binary, two are continuous, and one is nominal. The variables can be seen in Table 4, where all except the variables associated with PANSS-EC (not PANSS-EC score), recent loss of a relationship, lack of network, loss of self-esteem and recent insomnia are included in the datasets.

	Usage_0	$\text{Commencement}_{\text{NA}}$	Patients	Case Ratio
Antipsychotics	529	76	453	0.296
Benzodiazepines	641	99	542	0.382
Hypnotics	638	102	536	0.267
Mood Stabilizers	620	95	525	0.065

Table 5: The number of possible candidates for the commencement of medication (Usage_0), i.e., patients who don't have registered usage of the medication at admission, the number of patients with missing values for the commencement of the medication for the possible candidates ($\text{Commencement}_{\text{NA}}$), the total number of patients and the case ratio for each of the four datasets.

3.4.1 Correlation Analysis

Correlations between the clinical variables on the dataset before imputations in each of the four datasets are calculated using the Pearson correlation, both for binary and continuous variables. The nominal variable for diagnosis category is removed from the correlation analysis since it is not suitable to include with our correlation measure. It should be noted that a correlation is expected between the diagnosis category *Substance Abuse Disorders* and the variable associated with recent substance abuse; however, this has not been further investigated.

Correlation plots for each of the datasets are shown in Figure 3. It is clear that there is little difference between the datasets, with the correlation coefficients not differing more than ± 0.1 for any of the variables. The correlation plots can then be discussed jointly. Most variables are weakly correlated (less or equal to $|0.3|$), with the exception of one moderate correlation and two strong correlations. The correlation between PANSS-EC and the referral paragraph is 0.4 in the Hypnotics dataset, but is less than 0.3 in the other datasets and is therefore considered a weak correlation. The moderate correlations are between the PANSS-EC score and the specialist paragraph, with a correlation of 0.4 and 0.5. The two strong correlations are between the variables for recent suicide thoughts and if suicide was relevant for the referral, and between the specialist paragraph and referral paragraph, with a correlation 0.7. Looking at Table 4, it is understandable that these two correlations are present. The first strong correlation may be due to the fact that if a patient's referral to the psychiatric department was due to suicidal relevance, this would then naturally correlate with recent suicidal thoughts. It should be noted that for suicidal thoughts, a lot of patients tend to hold back information, which may be the reason for difference in the two variables. The last strong correlation is due to the fact that the paragraph given by the doctor who referred the patient often will be the same as the paragraph created by the doctor on duty at admission.

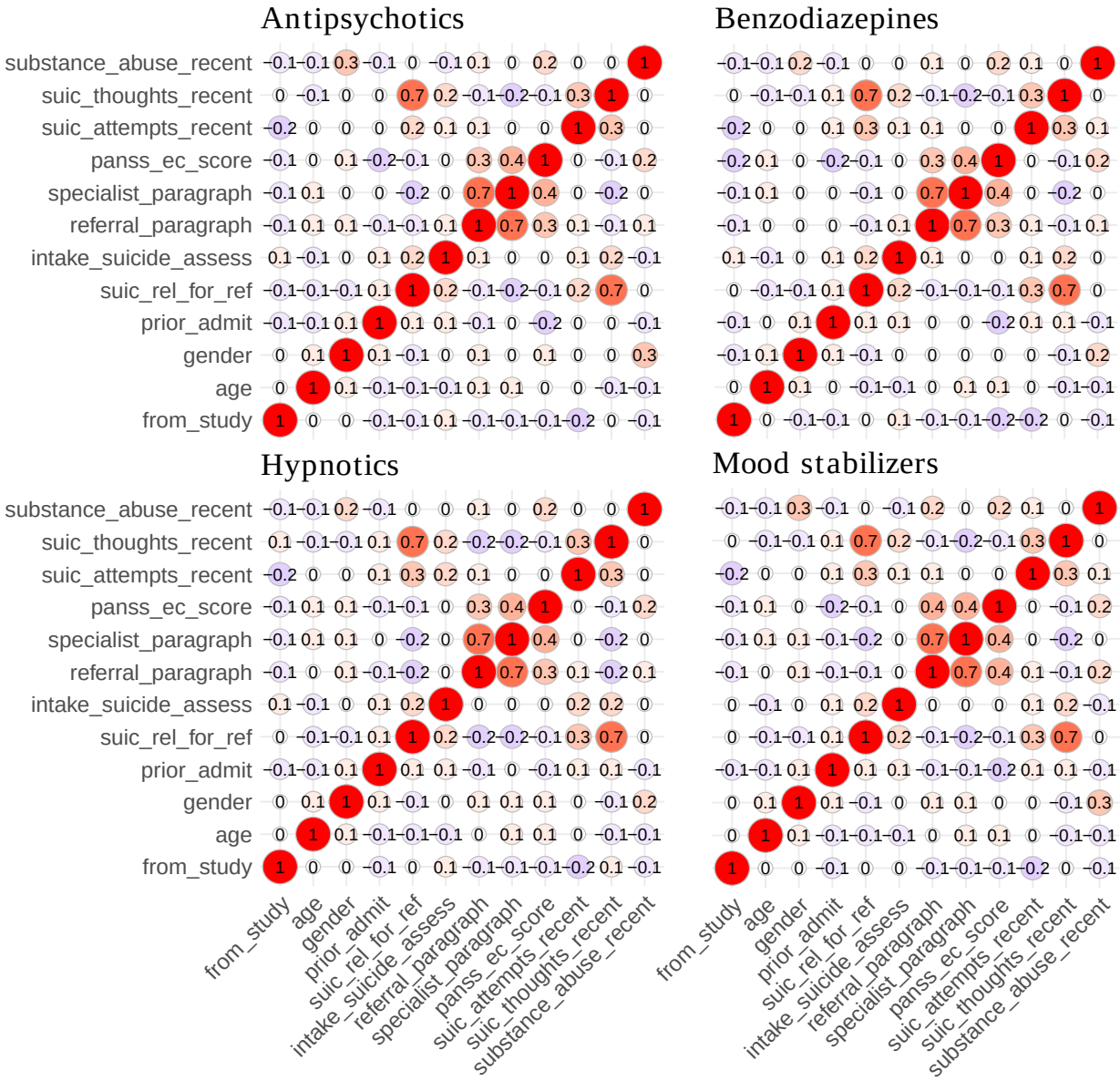


Figure 3: Plots of the correlation coefficients between pairs of clinical variables, expect the variable for diagnosis categories, before imputation in each of the four datasets.

CHAPTER

FOUR

SIMULATION STUDIES

In this section, simulation studies are conducted to better understand how the three models GLM, GAM, and EBM, described in Chapter 2, perform on data resembling the data presented in Chapter 3. This is done by generating data from known distributions, splitting the data into a training and test set and then fitting the models to the training set. The models are evaluated on the test set and compared knowing the underlying true model. The results from the simulation studies can then give insight into the models' performance, which can be of use when analyzing the real data from Chapter 3, later in Chapter 5.

4.1 Generating Data and Workflow

The simulation studies are based on modeling a binary response variable, Y, using two predictors, X_1 and X_2. X_1 is a binary variable drawn from a binomial distribution with probability 0.5, meant to mimic a typical binary variable presented in Chapter 3. X_2 is a continuous variable drawn from a truncated normal distribution meant to mimic the PANSS-EC score from Section 3.3.

For each simulation study, 500 observations are drawn for each variable, with X_1 and X_2 being drawn independently of each other. The data generating is seeded to make the results reproducible. The models' performance is evaluated when the true underlying model has linear, non-linear and interaction effects in the covariates and noise variables are present. Four different true systematic components, η, are then defined. The formulas for the true η's are presented at the start of each simulation study.

For each simulation study, the observed systematic components are calculated from the generated X_1, X_2, and the corresponding η formula. The η's can be transformed to probabilities, p, using the inverse logit function, also called the sigmoid function, seen in (2.3),

$$\log \frac{p}{1-p} = \eta \iff p = \frac{1}{1+e^{-\eta}}.$$

The probabilities are used to draw the binary response variable Y from a binomial distribution. The data is then split into a training and a test set of size 350 (70%) and 150 (30%), respectively, stratified on the response variable using the `createDataPartition()` function from the `caret` package in `R` (Kuhn 2023). This ensures the case ratio is approximately the same in the training and test set. The training set is used to fit models (presented in the next section), and the test set is used to evaluate the models. The size of the datasets is chosen to mimic the size of the datasets presented in Chapter 3.

This process is repeated 1000 times, and for each repetition, referred to as a simulation, the models are fitted to the training set and evaluated on the test set. The resulting predictions from the models on the test set are used to evaluate the models. The evaluation of performance is done using the AUC from Section 2.7, using the `pROC` package in `R` (Robin et al. 2023). The AUC of each model prediction from each simulation is saved, and the AUC from all simulations is used to compare the three models using violin plots with the `ggplot2` package in Wickham (2016). A violin plot is a combination of a boxplot and a kernel density plot, which shows the distribution of the data over all quantiles, not just the quartiles as in boxplot. Bland-Altman plots are used to evaluate the agreement between the models, where the difference between the AUCs of the two models is plotted against the average of the AUCs for each of the 1000 train/test splits. The blue dots each represent a train-test split, the purple dotted line is the mean difference between the AUCs and the black dotted lines are the 95% limits of agreement defined as the mean difference $\pm 1.96 \times$SD (Bland & Altman 1986).

For the last of the 1000 simulations, the corresponding shape functions for the three models are extracted and presented. For GLM, $\hat{\beta}_2 X_2$ is simply plotted over the range of X_2, which is the information that corresponds to the shape functions of GAM and EBM. The uncertainty of the shape functions is also plotted along with the shape functions. For GLM, this is the 95% confidence interval of the estimated coefficient multiplied by X_2, over the range of X_2. For GAM, this is the confidence interval calculated from the smoothing matrix similar to what seen in (2.8), and for EBM, this is the standard deviation of the predictions from the bagged models explained at the end of Section 2.5. The flowchart for each simulation study is similar to the one seen in the left part of Figure 12, in Chapter 5.

4.2 Models

Generalized Linear Model

The GLM model is run using the `glm()` function from the `stats` package in `R` (R Core Team 2023). The family parameter is set to `binomial`, and the rest of the parameters are set to default values. Confidence intervals for the coefficients are calculated using the `confint()` function, and the model predictions are obtained by using the `predict()` function from the same `R` package, `stats`. Regression coefficients with standard errors and test results are obtained from the `summary.glm()` function from the same package.

For Simulation Study 1-3, the model is only given X_1 and X_2, while for Simulation Study 4, linear terms in the noise variables $\text{Noise}_1 - \text{Noise}_{10}$ (see below) are also present in the model formula. Interaction terms and non-linear transformations are not specified to the GLM model formula in any of the simulation studies.

Generalized Additive Model

The GAM model is run using the `gam()` function from the `gam` package in `R` (Hastie 2023). This package was chosen as it is the original GAM implementation. Please note that the `R` package `mgcv` is more common and, in most cases, more practical to use (Wood 2023). The model is run with smoothing splines with 4 degrees of freedom for the continuous variables and the family parameter is set to `binomial`. The rest of the parameters are set to default values. The model predictions are obtained by using the `predict()` function from the `stats` package in `R` (R Core Team 2023). Regression coefficients with standard errors and test results are obtained from the `summary.glm()` function from the same package.

For Simulation Study 1-3, the model is given X_1 and $s(X_2, 4)$, while for Simulation Study 4, linear terms in the noise variables $\text{Noise}_1 - \text{Noise}_{10}$ (see below) are also present in the model formula. Interaction terms and discontinuities are not specified to the GAM model formula in any of the simulations.

Explainable Boosting Machine

The EBM model is run from the `interpret` library in `Python` with the `ExplainableBoostingClassifier()` function (Nori et al. 2019*b*). The model is run with default parameters, the most notable being the total number of boosting rounds, set to 5000, and the learning rate, set to 0.01. Default early stopping is employed so that if no improvement in the loss function is obtained within 50 rounds, then the model fitting algorithm stops. The training data is fitted using the `fit()` function and the test data predictions are obtained by using the `predict_proba()` function, both from the same library.

For all four simulation studies, all training data are presented to the EBM model. For Simulation Study 1-3, the model is only given X_1 and X_2, while for Simulation Study 4, the noise variables $\text{Noise}_1 - \text{Noise}_{10}$ are also added. Non-linearity and interactions are not specified in the model formula, as the model inherently captures these.

4.3 Results from the Simulation Studies

The results from the four simulation studies will now be presented. For each simulation study, the shape functions for X_2 is presented for GLM, GAM and EBM with the corresponding barplot of the distribution of X_2 for different intervals from EBM. For Simulation Study 3 and 4 the shape function for the interaction term is included. Violin plots showing the distribution of AUCs for the three models are presented for each simulation study in Figure 4, and Bland-Altman plots can be seen in Figure 15. The average AUCs over all 1000 simulations for the four simulation studies are presented along with corresponding standard deviations in Table 7. The average coefficient estimates with standard deviations for the three models are calculated for each simulation study and presented in Table 6. For EBM, this is only the coefficient for the variable X_1, which is centered around 0, in contrast to the estimates from the GAM and GLM.

Following the terminology used in the EBM model, the word score is used to describe the contribution to the probability of a positive response variable. This is not to be confused with the score function or the PANSS-EC score presented in earlier chapters.

	GLM			GAM		
	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
Sim1	-2.34±2.17	5.41±2.12	-0.20±0.04	-2.37±1.88	5.41±1.84	-0.20±0.04
Sim2	-2.45±0.56	4.92±0.46	-0.05±0.04	-2.54±0.63	5.12±0.50	-0.06±0.05
Sim3	-1.45±0.56	-2.18±2.47	-0.06±0.06	-1.47±0.54	-2.15±2.19	-0.05±0.05
Sim4	-1.58±0.66	-1.60±0.66	-2.24±2.19	-2.24±1.98	-0.06±0.07	0.01±0.25

	EBM	
	Score_0	Score_1
Sim1	-2.08 ± 0.32	2.09 ± 0.33
Sim2	-2.12 ± 0.25	2.13 ± 0.25
Sim3	0.59 ± 0.24	-0.60 ± 0.25
Sim4	0.40 ± 0.18	-0.39 ± 0.18

Table 6: Average coefficients and standard deviations over the 1000 simulations for the three models in the four simulation studies.

	GLM	GAM	EBM
Sim1	0.920 ± 0.022	0.919 ± 0.023	0.916 ± 0.024
Sim2	0.922 ± 0.026	0.929 ± 0.023	0.928 ± 0.023
Sim3	0.707 ± 0.073	0.705 ± 0.070	0.716 ± 0.063
Sim4	0.670 ± 0.081	0.672 ± 0.079	0.655 ± 0.085

Table 7: Average AUC with standard deviations over the 1000 simulations for the three models in the four simulation studies.

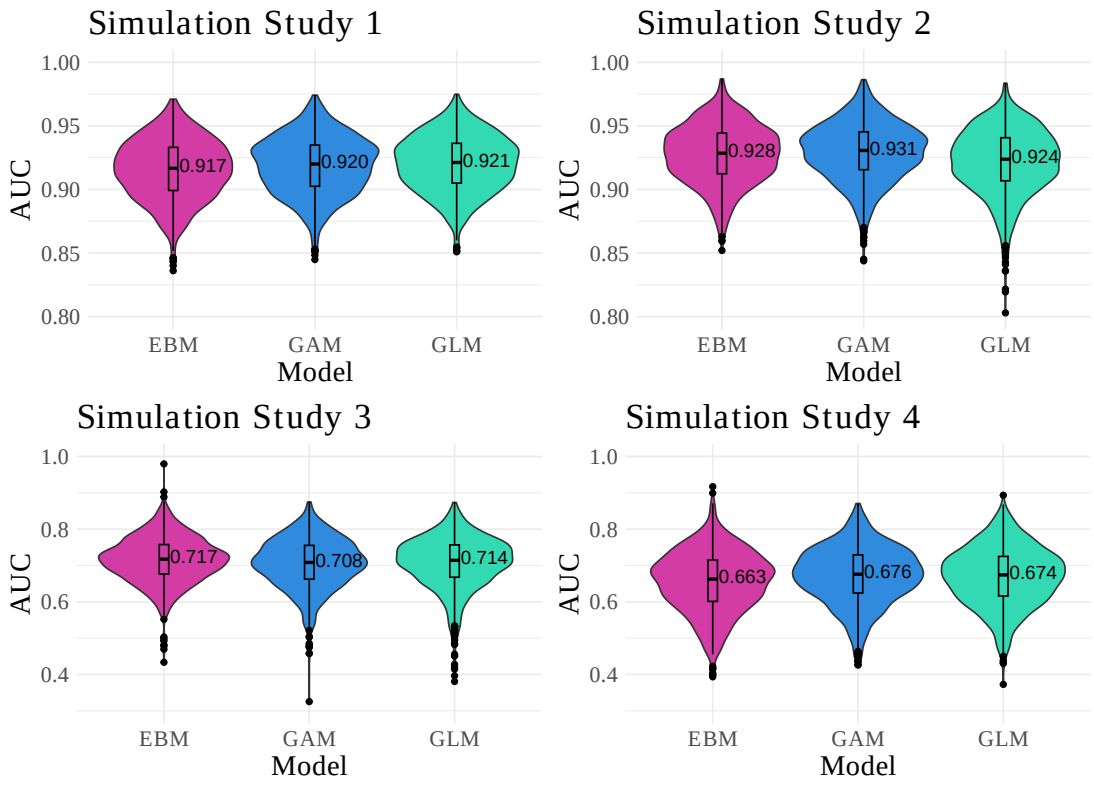


Figure 4: Violin plot for the AUCs on the test set for the 1000 simulations for each of the four simulation studies.

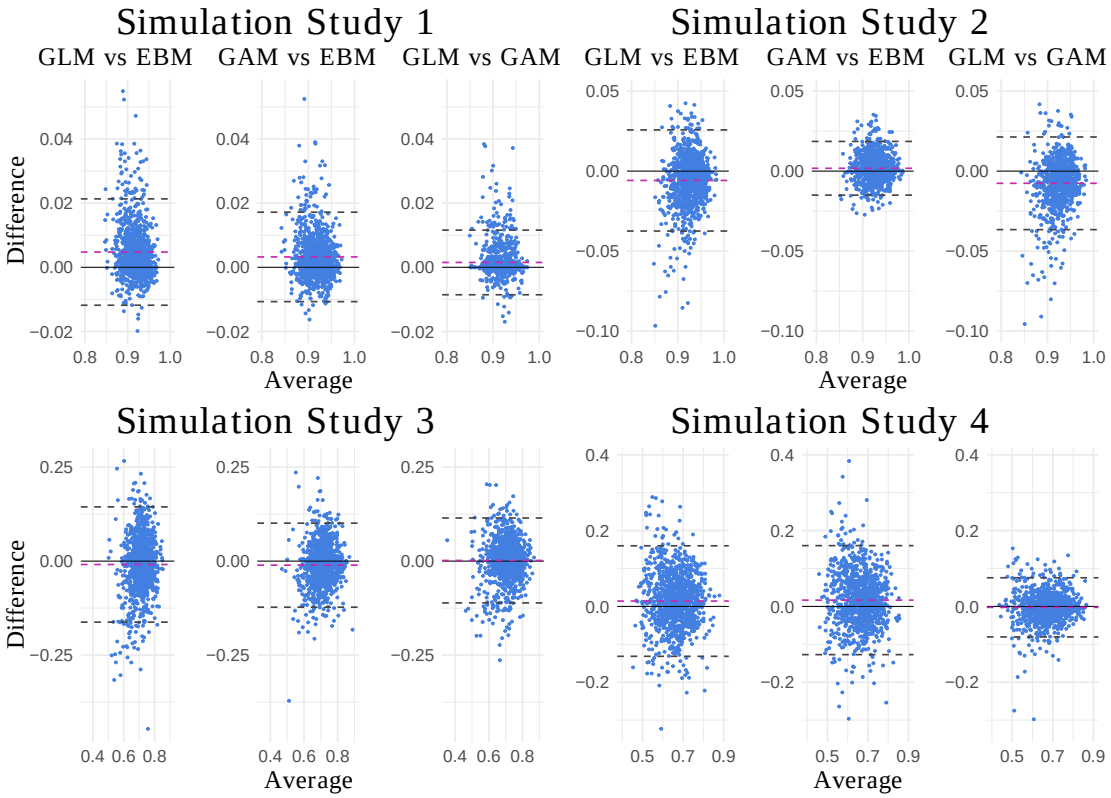


Figure 5: Bland-Altman plot for the AUC from EBM, GAM and GLM from the 1000 simulations for Simulation Study 1-4. Note that the axes are different between the simulation studies.

4.3.1 Simulation Study 1

True Model

In the first simulation study, the true systematic component is

$$\eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2,$$

where $\beta_0 = -2$, $\beta_1 = 5$ and $\beta_2 = -0.2$. This is a simple linear GLM with a logit link function and a binary response. The resulting shape functions can be seen in Figure 6. Since the true model is linear all three models are expected to perform well, but it is of interest to investigate to which degree the flexible GAM and EBM models may overfit the data.

Results

Based on the estimated shape function from X_2 in Figure 6, it is clear that both GAM and EBM fit a non-linear function and seem to capture an increase in score for X_2 larger than 20. This is likely due to outliers since there are very few observations in this interval, as seen in the barplot in Figure 6. GAM is forced to use 4 degrees of freedom in the smoothing spline, which may be too flexible. The EBM model seems to be more optimistic of the variance in the data for high values of X_2 compared to GAM.

From Table 8, it can be seen that the EBM adds significant importance to the interaction term in the last train-test split, which is not in accordance with the average importance over all 1000 simulations. Since the true model doesn't include any interaction, this shows that the EBM model overfits the data. The interaction term was included in all simulations and, therefore, deemed to increase the performance of the model in all simulations.

The violin plot in Figure 4 shows that all models perform excellent, according to the classification of AUC in Hosmer et al. (2013, p. 177), with all median AUCs above 0.9. The EBM performs slightly worse than the other two models, but the difference is small. The AUC results of Table 7 complement the violin plot. The Bland-Altman plot in Figure 15 doesn't show any trends in the agreement between the models.

Variables	Imp	Avg Imp	SD
X1	1.923	2.081	0.301
X2	0.517	0.571	0.162
X1 & X2	1.248	0.313	0.257

Table 8: Variable importance from the last simulation (Imp) with the corresponding average importance (Avg Imp) and standard deviation (SD) over the 1000 simulations from the EBM model in Simulation Study 1.

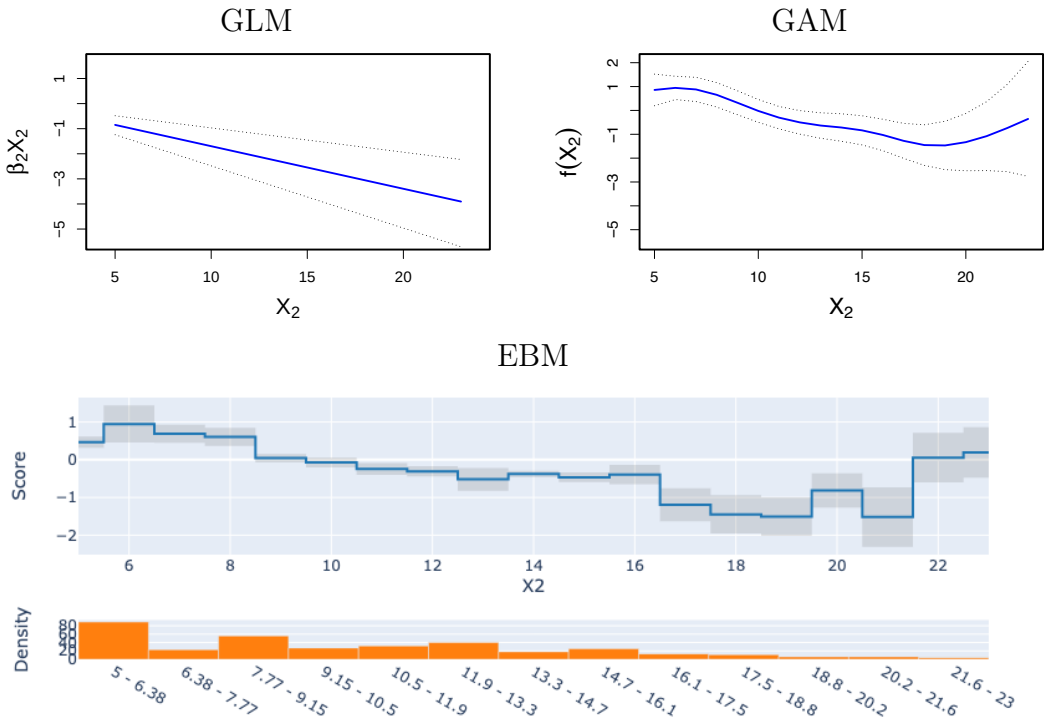


Figure 6: Comparison of the shape functions of GLM, GAM and EBM in Simulation Study 1, with a barplot showing the distribution of X_2 across various intervals.

4.3.2 Simulation Study 2

True Model

In the second simulation study, the true systematic component is

$$\eta = \beta_0 + \beta_1 X_1 + \beta_2 f_2(X_2),$$

where $f_2(X)$ is a non-linear step function seen in Figure 7 and β_0, β_1 and β_2 have the same values as in Simulation Study 1. We expect the GLM to perform poorly since it will only use a linear term in X_2, while GAM and EBM should perform well due to their ability to capture non-linear effects. Since the step function used is non-smooth, we expect the EBM to perform better than the GAM model due to the GAM model's smoothing characteristic. Note that $\beta_2 = -0.2$, which means that the true shape function for X_2 is an inverted and scaled version of the step function.

Results

The shape functions from Simulation Study 2 can be seen in Figure 8. It can be seen that the GAM and EBM models capture the non-linearity of the true model, while the GLM model fails to do so. The GAM model reports a much larger uncertainty in the upper tail of the X_2 variable than the EBM model. The EBM model gives the same uncertainty for X_2 values around 7, 13 and 22 in contrast to the GAM model. This indicates that the EBM model is not just more optimistic

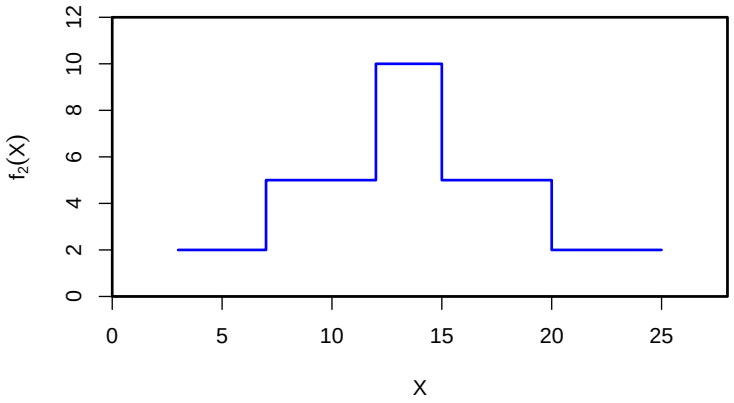


Figure 7: Step function, f_2, used in Simulation Study 2.

than GAM in general but gives different uncertainty estimates as well.

From Table 9 we see that the EBM importance from the last simulation is in line with the average importances over all 1000 simulations. The most important variable is by far X_1, with an average importance of 2.117. The variables X_2 and the interaction term are less important, with average importances of 0.300 and 0.172, respectively.

In the violin plot for Simulation Study 2 in Figure 4, EBM and GAM seem to perform equally well, while the GLM model performs slightly worse and has a larger span than the other two models. The AUC results of Table 7 complement the results in the violin plot. The Bland-Altman plot for Simulation Study 2 in Figure 15 shows no trends in the agreement between the models, other than that for low AUCs, the EBM outperforms GLM, which is tobe expectedt looking at the violin plot.

Variables	Imp	Avg Imp	SD
X1	2.076	2.117	0.219
X2	0.355	0.300	0.126
X1 & X2	0.101	0.172	0.135

Table 9: Variable importance from the last simulation (Imp) with the corresponding average importance (Avg Imp) and standard deviation (SD) from the EBM model in Simulation Study 2.

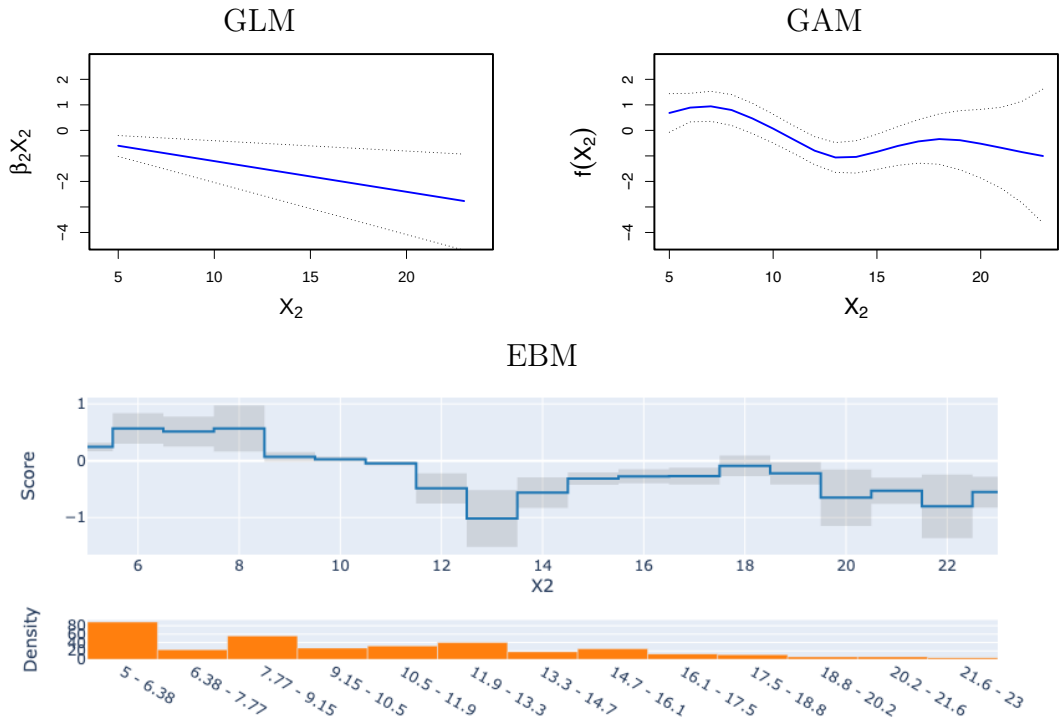


Figure 8: Comparison of the shape functions of GLM, GAM and EBM in Simulation Study 2, with a barplot showing the distribution of X_2 across various intervals.

4.3.3 Simulation Study 3

True Model

In the third simulation study, the true systematic component is

$$\eta = \beta_0 + \beta_1 X_1 + \beta_2 f_3(X_1, X_2),$$

where $f_3(X_1, X_2) = X_1 X_2^2$ and β_0, β_1 and β_2 have the same values as in Simulation Study 1 and 2. Note that there is no term for X_2 alone. $f_3(X_1, X_2)$ can be seen in Figure 9. In this simulation study, we expect EBM to perform better than the other models due to the interaction term, which is not explicitly specified in GAM and GLM. We also expect the EBM to have a shape function for X_2 that is approximately constant since the interaction term should pick up the non-linear effect of X_2. It is unknown how GLM and GAM will estimate the shape function for X_2. As in Simulation Study 2, note that $\beta_2 = -0.2$, which means that the true shape function for X_2 is an inverted and scaled version of Figure 9.

Results

The shape function for the three models can be seen in Figure 10, this time including the interaction shape function. The shape functions from GAM and EBM have the same shape only with GAM not capturing the same drastic drop in score for X_2 values around 18 due to its smoothing characteristic. Both models seem to have an approximately constant shape around zero, but not for large

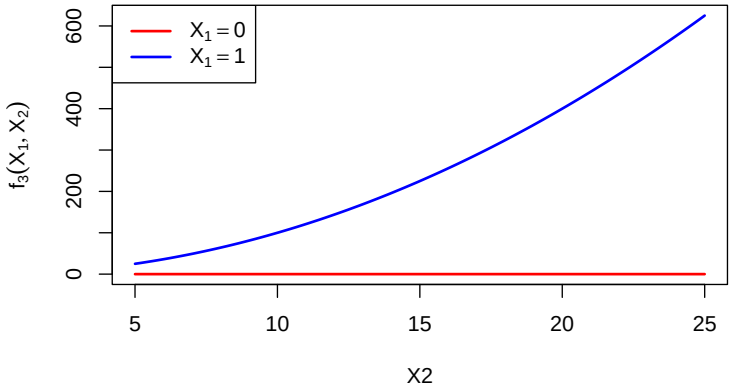


Figure 9: Interaction function, f_3, used in Simulation Study 3.

values of X_2. The uncertainty in both models also increases drastically for these values of X_2, though the GAM model seem to estimate a much larger uncertainty than EBM. In the shape function for the interaction term we see that the EBM model captures the interaction term well.

From Table 10, it is clear that the interaction term is by far the most important variable in the last simulation. This is in line with the average importance over all 1000 simulations, though the interaction does it a little better and X_1 a little worse in the last simulation compared to the average importance.

The violin plot in Figure 4 shows that all models now perform considerably worse than in the previous simulation studies. All models have a median AUC of around 0.7, with the EBM model performing slightly better than the other two models. The AUC results of Table 7 complement the violin plot. The Bland-Altman plot in Figure 15 shows no clear trends in the agreement between the models.

Variables	Imp	Avg Imp	SD
X1 & X2	1.878	1.230	0.557
X1	0.281	0.592	0.241
X2	0.326	0.375	0.313

Table 10: Variable importance from the last simulation with the corresponding average importance and standard deviation from the EBM model in Simulation 3.

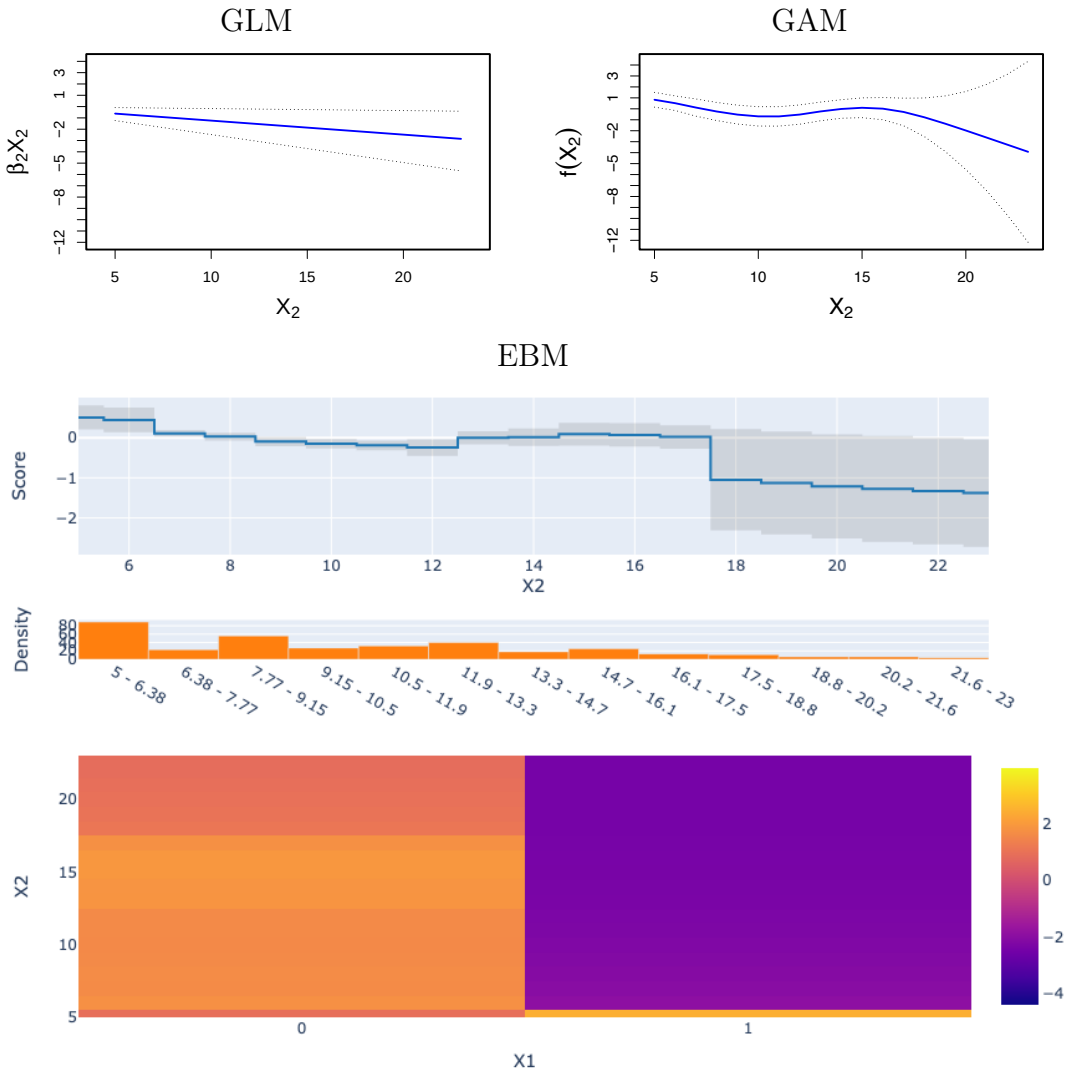


Figure 10: Comparison of the shape functions of GLM, GAM and EBM in Simulation Study 3, with a barplot showing the distribution of X_2 across various intervals. The shape function for the interaction term from EBM is seen at the bottom.

4.3.4 Simulation Study 4

True Model

In the fourth simulation study, the true systematic component is the same as in Simulation Study 3. Ten noise variables with mean 0 and standard deviation 1 are added to the data so that each model now has ten noise variables they are using when predicting Y in addition to X_1 and X_2. Note that the ten noise variables are not included in the true systematic component. This is done to see how the models handles additional variables that are not related to the response variable. We expect the EBM model to overfit the noise variables with interaction terms, while GAM and GLM should perform similarly to Simulation Study 3, only a little worse due to the noise variables.

Results

The resulting shape functions can be seen in Figure 11, also including the shape function for the interaction term between X_1 and X_2 in EBM. The shape function for GLM is approximately zero, which is to be expected since the true model doesn't include X_2. The shape functions of GAM and EBM are similar in shape, and both seem to have low importance around zero, though both increase for high values of X_2. The uncertainty for both GAM and EBM grows drastically for high values of X_2. The shape function for the interaction term in EBM shows that the model is far less capable of capturing the interaction term than it was in Simulation Study 3.

In Table 11, the importance of all variables in the EBM model, including the noise variables, can be seen. X_1 is the most important variable, followed by the three noise variables before the interaction between X_1 and X_2. We see that although some noise variables are deemed important in the last simulation, all of them has an average importance over the 1000 simulations considerably lower, around 0.086. Most notably is that the true interaction between X_1 and X_2 is only included in the model in 567 of the 1000 simulations.

In Table 12, we have included the p-values for the coefficients in the GLM and GAM models for the last simulation, along with the number of times the p-value of the coefficient is less than 0.05 over the 1000 simulations. We see that the p-values for the coefficients for the noise variables, on average, have a count of around 50. If we repeat an experiment where the null hypothesis is true 1000 times and count the number of times the p-value is below 0.05, we would assume on average that the count is around 50 for an exact test.

The violin plot for Simulation Study 4 in Figure 4 shows that the span of the AUC of all models is large, with AUCs ranging from 0.4 to 0.9. The EBM model performs slightly worse than the other two models considering the medians in the violin plot. Table 7 shows that this difference is even greater when looking at the average AUCs over all 1000 simulations.

The Bland-Altman plot in Figure 15 shows no clear trends in the differences between the models.

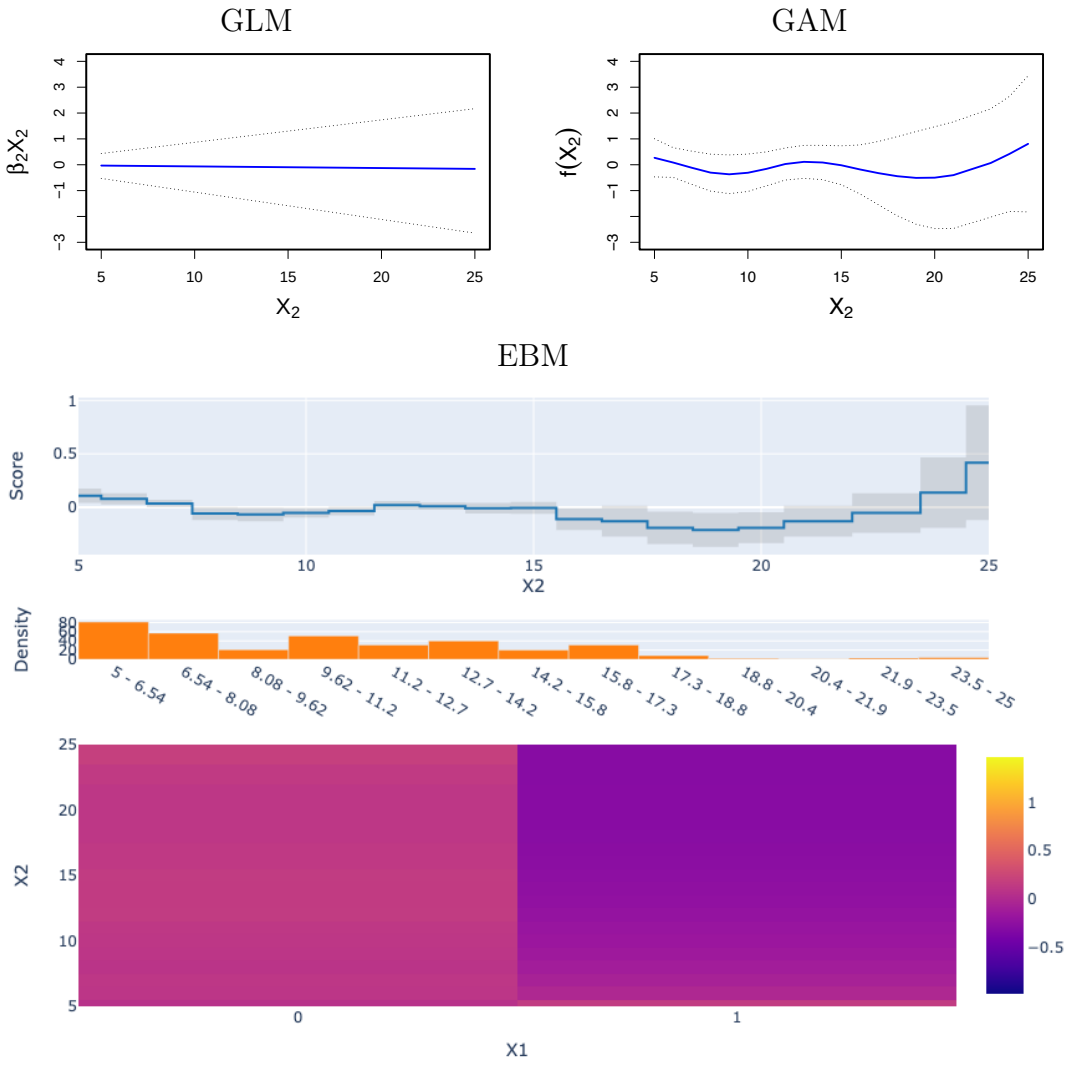


Figure 11: Comparison of the shape functions of GLM, GAM and EBM in Simulation Study 4, including the shape function for the interaction term from EBM.

Variables	Imp	Avg Imp	SD	Inclusion Count
X1	0.560	0.393	0.181	1000
Noise_2	0.203	0.086	0.054	1000
Noise_8	0.176	0.086	0.052	1000
Noise_6	0.174	0.087	0.052	1000
X1 & X2	0.145	0.212	0.108	567
Noise_5	0.142	0.085	0.054	1000
X1 & Noise_5	0.130	0.142	0.067	62
Noise_4	0.119	0.085	0.051	1000
X1 & Noise_6	0.109	0.115	0.058	63
Noise_9	0.105	0.086	0.056	1000

Table 11: Importance (Imp) from the last simulation, average importance (Avg Imp), standard deviation (SD) and inclusion count in over the 1000 simulations from EBM in Simulation Study 4.

	GLM		GAM	
	p-value	Count < 0.05	p-value	Count < 0.05
(Intercept)	0.001	741	0.237	769
X1	0.000	939	0.003	938
X2	0.896	206	0.010	176
Noise_1	0.956	57	0.158	50
Noise_2	0.054	49	0.733	62
Noise_3	0.997	49	0.327	53
Noise_4	0.927	40	0.584	46
Noise_5	0.368	38	0.723	43
Noise_6	0.109	60	0.380	66
Noise_7	0.904	49	0.520	52
Noise_8	0.157	49	0.464	54
Noise_9	0.145	53	0.735	54
Noise_{10}	0.291	53	0.883	50

Table 12: P-values for the last simulation and count of p-values below 0.05 over the 1000 simulations for GLM and GAM in Simulation Study 4.

4.4 Discussion

From the results above, we have observed that the EBM and GAM models both tend to overfit the data for regions with few observations. This is seen clearly in Simulation Study 1, where both models fail to capture the linearity of the true model for high values of X_2 where there are few observations. Through all simulation studies, the EBM model seems to be more optimistic than the GAM model, which is seen in the uncertainty estimates. This probably comes from the fact that EBM presents error bars for each interval used for the variable, calculated from the standard deviation of all bagged models. The low number of observations will then most likely not affect the uncertainty estimates as much as in the GAM model, where the uncertainty is calculated from a smoothing matrix similar to what is seen in (2.8). This is seen in all simulation studies but especially in Simulation Study 3 and 4, where the uncertainty for high values of X_2 is much larger for GAM than for EBM.

When only one possible interaction is present, EBM will always include the interaction term in its prediction. This is also the case even when it is not part of the true model, as seen in Simulation Study 1 and 2. Although the EBM model over the 1000 simulations, in general, sets a low importance to the interaction term in these simulation studies, it also sometimes gives a relatively high importance, which is seen in the last simulation presented in Simulation Study 1. The inclusion of an interaction term in the model, when it is not part of the true systematic component, indicates overfitting.

Comparing the estimation of coefficients from GLM and GAM in Table 6. We see that both models on average give the same estimates with approximately the same variability between the simulations. The scores for variable X_1 from the EBM

model can be seen in the same table. Here, the scores are centered around 0, which means that we have to look at the difference between the scores to compare them to the coefficients from GLM and GAM. The scores for X_1 are in the same range as the coefficients from GLM and GAM, though in general a little lower. This may be due to the interaction term that is included in the EBM model, which potentially makes the model less dependent on the variable X_1.

When adding noise variables to the data, the general performance of all models decreases, which is as expected. The span of the AUCs is also much larger when the noise variables are added. This can be seen in Figure 14 and in Table 7. We see that the EBM model is not able to capture the true interaction between the variables X_1 and X_2 in Simulation Study 4, as it does very well in Simulation Study 3. It only includes the interaction term in 567 of the 1000 simulations, which is a clear sign of fitting the wrong model. This shows that adding noise variables to the dataset can have a large impact on the EBM model's performance. This is contrary to the GLM and GAM, where the noise variables are not found to be significant (on the level expected for true null hypotheses).

Another train-test split is chosen randomly from the 1000 simulations, and the corresponding shape functions are presented in Appendix C. This is to compare the shape functions from the last simulation with another randomly chosen train-test split.

Comparing the shape functions in this chapter to the shape functions seen in Appendix C, we see that for a different train-test split the shape functions are similar, but with some clear differences. For Simulation Study 1, we see that the shape functions for both GAM and EBM in the appendix catch the linearity better due to the data not having the same outliers as in the last train-test split. In Simulation Study 2, we see a little difference in the shape of the functions, most notably for EBM with regard to shape and uncertainty. First comparing the univariate shape function, for Simulation Study 3, the uncertainty for GAM is much lower in the appendix, but the shape is similar. For the EBM model, the uncertainty is much larger in the appendix and the shape is also different. In Simulation Study 4, the shape of the functions are similar, but the uncertainty is smaller for the GAM model in the appendix. The EBM model results in similar shapes and relatively similar uncertainty. Comparing the shape functions for the interaction term, we see that the EBM in Simulation Study 3 fails to capture the interaction in the appendix, while the EBM in Simulation Study 4 is able to capture it. This shows how sensitive the shape functions are to change in the data, which is important to keep in mind when analyzing the real data later.

This is useful insight that will be taken into consideration when analyzing the real data in Chapter 5.

CHAPTER

FIVE

RESULTS

This chapter now turns to the data presented in Chapter 3 for statistical analyses. The aim of the analyses is presented with an overview of the process along with specifications regarding the three models used. The results from a single train-test split is presented, followed by the results from 1000 train-test splits. The results from the Benzodiazepines and Mood Stabilizers datasets are further investigated with regard to variable importance, shape functions and coefficients. The results for the Antipsychotics and Hypnotics datasets are presented in Appendix B.

5.1 Aim of the Analyses

The aim of the analysis is to identify the importance of clinical variables in predicting the commencement of medication at the time of admission. This is done by evaluating the three different models, EBM, GAM and GLM, on the four datasets *Antipsychotics*, *Benzodiazepines*, *Hypnotics* and *Mood Stabilizers* from Section 3.4. The models are evaluated by their predictive performance, measured by AUC. The relation between the variable for PANSS-EC score and the commencement of medication is of particular interest.

5.2 Overview of the Analyses

A flowchart of the data analysis for each dataset is presented in Figure 12. The dataset is divided into a training and a test set of size 70% and 30%, respectively stratified on the response with the `createDataPartition()` function from the `caret` package in `R` (Kuhn 2023). All data splitting is seeded to ensure reproducibility. The training set is used to fit the models GLM, GAM and EBM, which allows for the shape functions, variable importances and coefficients to be extracted for each of the models. The fitted model is then used on the new unknown data in the test set, which then allows for the ROC curves for each model to be plotted using the `pROC` package in `R` (Robin et al. 2023). The DeLong test from Section 2.8, is used to compare the three ROC curves pairwise in each dataset

with the `roc.test()` function from the same package. From the ROC curves, the AUC is obtained. This process is repeated 1000 times with different train-test splits, and the AUCs for each model are saved for each repetition. In the same way as in the simulation studies in Chapter 4, the AUC densities for each model can be plotted, as well as the Bland-Altman plots of the AUC values to compare the models performance. For EBM, the average and standard deviation of the 1000 variable importances for each variable are presented.

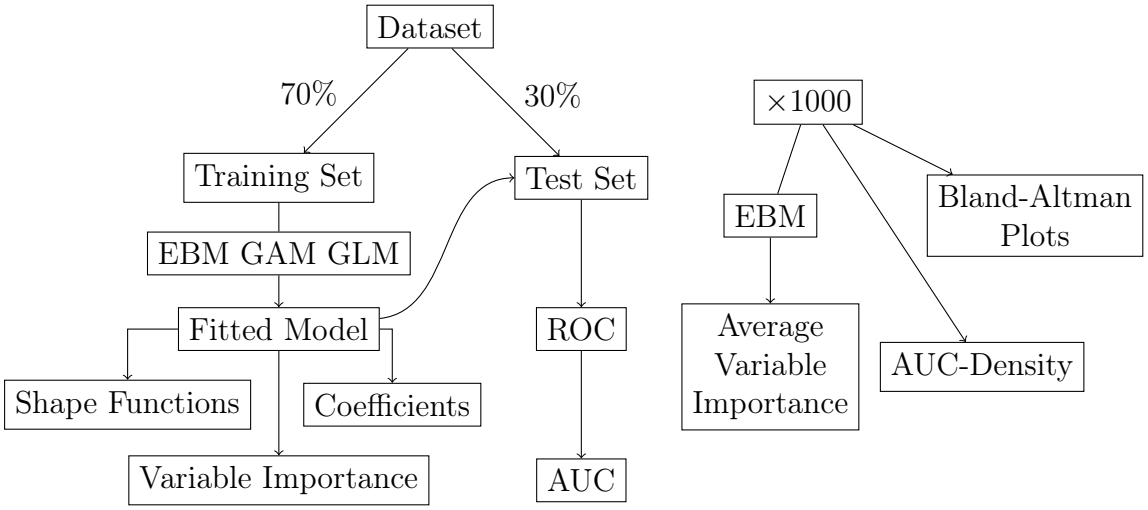


Figure 12: Flowchart of the analysis for each of the four datasets.

5.3 Models

Details about the models fitted will now be presented. All available covariates, ten binary, two continuous and one nominal with four categories, are included in the systematic component of the models. Looking at Table 4 these are all variables except the ones associated with PANSS-EC (not the total PANSS-EC score), broken relationship, lack of network, loss of self-esteem and recent insomnia. The models are used with the same functions, packages and settings as presented in Chapter 4, using data presented in Chapter 3. See Section 4.2 for further details on the models.

For the GLM, no non-linear transformations are specified and neither GLM nor GAM have any interaction terms included in the model. For the covariates in GAM where a smoothing spline is used, the `anova.GAM()` function from the `gam` package is used to test the null hypothesis that a linear term in the covariate is sufficient. This is done using a score-type test, according to the documentation of `anovaGam`. The p-value from this test is then a measure of the non-linearity of the variable, while the estimate from `summary.glm()` measures the effect the spline term has on the response.

5.4 Results

5.4.1 ROC Curves and AUCs for All Analyses

The ROC curves for the test set of three models on the four training datasets may be seen in Figure 13. This will be referred to as the original train-test split.

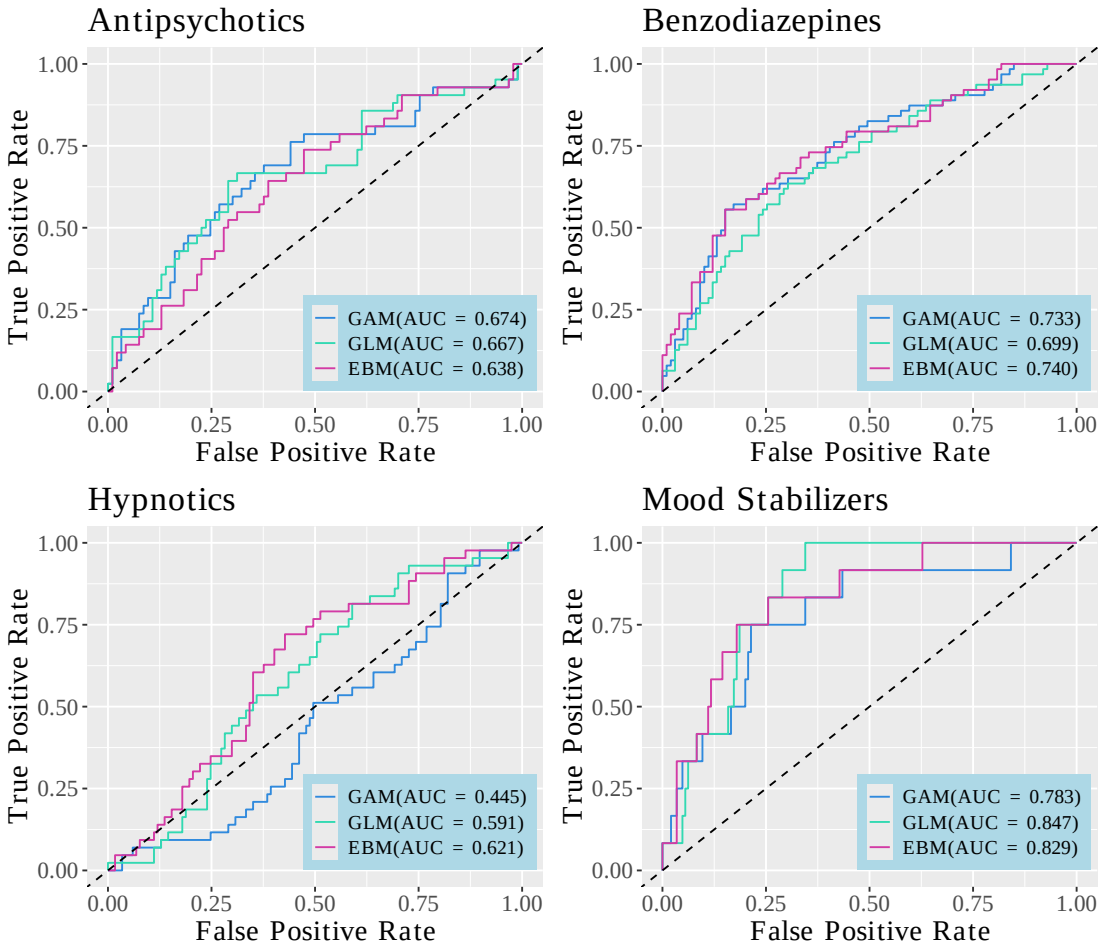


Figure 13: ROC Plots of GLM, GAM and EBM for the original train-test split for the four datasets.

In the plot for the Antipsychotics dataset in Figure 13, the GAM model performs better than GLM, which in turn outperforms EBM, but the general performance is poor with AUCs below 0.7. For the Benzodiazepines dataset, the EBM and GAM model outperform the GLM model, with an acceptable performance of AUCs above 0.7 (0.699 for GLM). For the Hypnotics dataset, EBM outperforms the other two models with an AUC of 0.621, while the GLM model results in an AUC of 0.591. This is not a good performance, but much better than the GAM model, which resulted in an AUC of 0.445, which is considered worse than a random guessing classifier. For the Mood Stabilizers dataset, the three models perform well with AUCs around 0.8. In this dataset, the GLM model has the highest AUC, ahead of the EBM and GAM models.

Using the DeLong test introduced in Section 2.8, the ROC curves of two models can be compared. The null hypothesis of the test is that the two models have

the same AUC. The p-values from the DeLong test for the three models on the four datasets are presented in Table 13. The only two models that are deemed significantly different at a level of 95% are the EBM and the GLM model and the GAM vs GLM model in the Benzodiazepines dataset.

	Antipsychotics	Benzodiazepines	Hypnotics	Mood Stabilizers
EBM - GAM	0.143	0.637	0.056	0.422
EBM - GLM	0.256	0.047	0.293	0.628
GAM - GLM	0.765	0.021	0.121	0.252

Table 13: P-values from the DeLong test on the different models for all four datasets.

5.4.2 1000 Train-Test Splits

As explained in the flowchart in Figure 12, the process above is repeated 1000 times with different train-test splits, and the AUC for each split is saved for all three models. The distribution of the AUCs for each model is plotted in a violin plot, along with boxplots showing quartiles and the median value using ggplot2 (Wickham 2016). The violin plots are found in Figure 14. For the Antipsychotics dataset, the EBM model has the lowest median AUC of 0.595, while GAM and GLM have 0.633 and 0.635, respectively. The span of the AUCs seems to be similar between the models, although GLM has some outliers with very low AUC compared to its AUC density. For the Benzodiazepines dataset the three models seem to perform equally well with similar AUC densities, though having a difference in AUC density in accordance to the difference in medians. For the Hypnotics dataset, the three models seem to perform poorly with an AUC of around 0.56, with the GLM model seeming to outperform the other two models slightly. For the Mood Stabilizers dataset, the three models perform well with a median AUC of around 0.8. The GLM has a slightly higher median AUC than the other two models, but has a substantially larger span with one outlier scoring an AUC as low as 0.4. The GAM model also has a large span but with the lowest value slightly below 0.5. The EBM model has the smallest span and seems to outperform the two other models based on this.

Although the AUC density plots are a nice way of comparing the three models, we are also interested in the agreement between the models. The Bland-Altman plot is used to show the agreement and can visualize trends in the differences in AUC between two models. The results may be seen in Figure 15. The zero-difference line is outlined in black. When assessing a Bland-Altman plot, we want to evaluate the agreement between the two models, that is, to observe if the differences between the AUC of the two models are consistent across the range of train-test splits and to identify any systematic bias or trends. The mean difference between the models, referred to as the bias, is also evaluated. If one model's AUC is consistently higher or lower than the other, the bias will be different from zero. Trends in the data, as well as outliers, will also be assessed.

For the Antipsychotics dataset, we see that the agreement between the EBM vs the the other models, is much lower than for GLM vs GAM. This is in line with

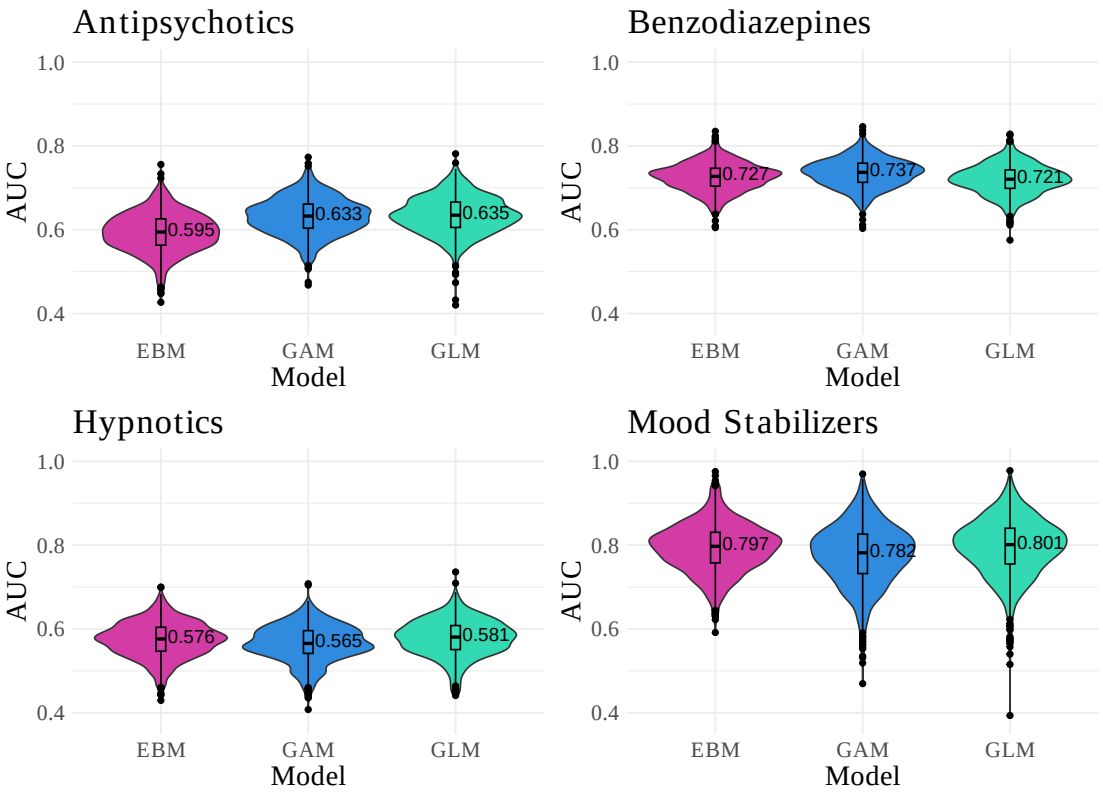


Figure 14: AUC density plots for the 1000 train-test splits for all three models on the four datasets.

the results seen in Figure 14, as the EBM model has an AUC density lower than the two others. This indicates that the AUC of EBM model is consistently lower than the two others. Considering GLM vs GAM, there doesn't seem to be any notable trends, except for a few outliers.

For the Benzodiazepines dataset, the agreement between the models is high compared to the other datasets. There are no notable signs of trends or outliers.

For the Hypnotics dataset, the agreement between the models is relatively low and of a similar degree to what we see in the Antipsychotics dataset. There is a tendency in the data that for an average AUC around 0.5, that the models differ substantially more than for the other values of average AUC. This may be due to the low AUCs for the models in the dataset in general and the fact that for train/test splits where the AUC is higher, the models differ less.

For the Mood Stabilizers dataset, the agreement is by far the worst, which is to be expected from the large span in AUC densities seen in Figure 14. In the plot for GLM vs EBM and GAM vs EBM, there is a clear trend in the data, where for low average AUCs, the EBM model outperforms the other two models. This is seen by the multiple outliers in the lower left corner of the plots. For the higher average AUCs, the agreement between the models is far better. This is in line with the results seen in Figure 14, where the EBM model has a lower span in AUCs, but a similar median value compared to the other models.

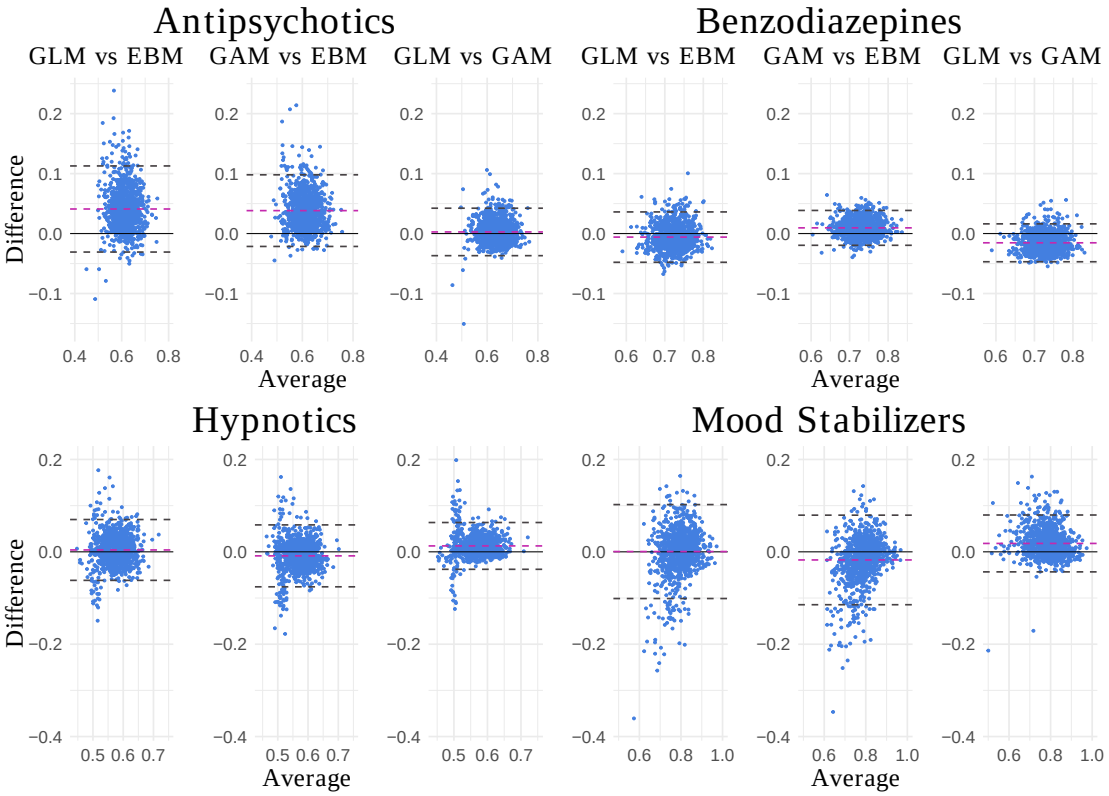


Figure 15: Bland-Altman plot for the AUC from EBM, GAM and GLM from the 1000 train-test splits for all 4 datasets. Note that the axes are not the same for all plots.

To summarize the results from Figure 14 and Figure 15, the Mood Stabilizer dataset gave the highest AUC of the four datasets, with a median AUC over the 1000 train/test splits of 0.797 for EBM, 0.782 for GAM and 0.801 for GLM. Following the rule of thumb, presented in Hosmer et al. (2013, p. 177), the performance of a classifier is considered poor if the AUC is between 0.5 and 0.7, acceptable if the AUC is between 0.7 and 0.8, excellent if the AUC is between 0.8 and 0.9 and outstanding if the AUC is above 0.9. The Mood Stabilizers dataset results in an acceptable/excellent AUC. For the Benzodiazepine dataset, the median AUCs for the model were all around 0.73 which is considered acceptable. The two remaining datasets, Antipsychotics and Hypnotics, both had median values below 0.7 for all models, which is considered a poor performance, not much better than random guessing. The Bland-Altman plots in Figure 15, further supports these results, where there is no clear trends that contradicts the inference made from the AUC densities.

We wish to further investigate the results for the datasets that perform well. The focus is then on the Benzodiazepines and Mood Stabilizers datasets since these are the two datasets with the highest overall performance with regard to AUC. The Mood Stabilizers dataset has a very low case ratio compared to the other three datasets, which can be a challenge for the models as the number of cases compared to the number of covariates can be low. With a low case ratio, AUC may not be the best measure of performance since it is sensitive to class imbalance. The results for the Antipsychotics and Hypnotics datasets are presented in Appendix B.

5.5 Benzodiazepines

We will now do a deeper dive into the results for the Benzodiazepines dataset and investigate the variable importance, shape functions and coefficients for the three models.

5.5.1 Variable Importance

From the EBM model, the variable importance for the 10 most important variables (including interactions) is extracted. The variable importance is defined as the weighted mean absolute score, seen in (2.10). We wish to compare these results with the 1000 other different train-test splits, so the average importance over the 1000 fitted models is calculated, along with the corresponding standard deviations for each variable. Since EBM estimates the interaction between all variables and then only includes those that increase the performance of the model, the number of times a variable is included in the model out of the 1000 train-test splits is counted. This is due to the fact that an interaction can be included with relatively high importance but only in a few fitted models. This will result in a high average importance, which is misleading without the insight of the inclusion count. The results are shown in Table 14.

All variables display a consistency in importance between the original train-test split and the average importance from the 1000 train-test splits. There are no interaction terms present among the top 10 variables, which makes the inclusion count redundant since all univariate terms are automatically included in the EBM model. The variable for study is clearly the most important variable, which is to expect following the analysis in Appendix A, where we see that the commencement of Benzodiazepines is significantly different between the two studies.

Variable	Imp	Avg Imp	SD	Inclusion Count
from_study	0.509	0.536	0.069	1000
diagnosis_category	0.281	0.235	0.044	1000
suic_rel_for_ref	0.209	0.256	0.045	1000
intake_suicide_assess	0.186	0.127	0.044	1000
panss_ec_score	0.154	0.201	0.049	1000
age	0.141	0.218	0.056	1000
substance_abuse_recent	0.101	0.073	0.042	1000
gender	0.093	0.082	0.046	1000
suic_thoughts_recent	0.085	0.069	0.037	1000
suic_attempts_recent	0.057	0.037	0.026	1000

Table 14: Variable importance (Imp), from EBM for our training set from the original train-test split, including average importance (Avg Imp), standard deviation (SD) and inclusion count from the 1000 train-test splits of the Benzodiazepines dataset.

5.5.2 Shape Functions and Coefficients

The shape functions for the two continuous variables for PANSS-EC score and age are extracted from the EBM model. These are presented with a barplot showing the distribution of observations in the range of the variable for different intervals. The shape functions are shown in Figure 16. Following the terminology used in EBM, the contribution of the shape function to the systematic component is referred to as the score, not to be confused with the PANSS-EC score. Note that score on the y-axis of the plots is here the contribution to the systematic component, not to be confused with the PANSS-EC score. The shape function for PANSS-EC score begins at -0.27 before increasing to a score around 0.30 for PANSS-EC scores of 10 to 20. For PANSS-EC scores from 20 to 25, the score decreases to around 0 before increasing to 0.80 at around 30. The uncertainty increases with the value of the PANSS-EC score. The shape function for age starts at a little below -0.5 and steadily increases to a peak at around 0.3 for ages between 40 and 50. After this, the score decreases to around 0 for ages 55 to 70 before decreasing drastically to a score of -1 for age 85. The uncertainty increases considerably for those of age above 70.

For the categorical variables, the variable scores are extracted from the EBM model and shown in Table 15. For the binary variables, the Score_0 is the contribution to the systematic component of the model when the variable is equal to 0 and the Score_1 is the contribution when the variable is equal to 1. When it comes to the variable for diagnosis category, the scores are the contribution to the systematic component of the model for each of the four categories.

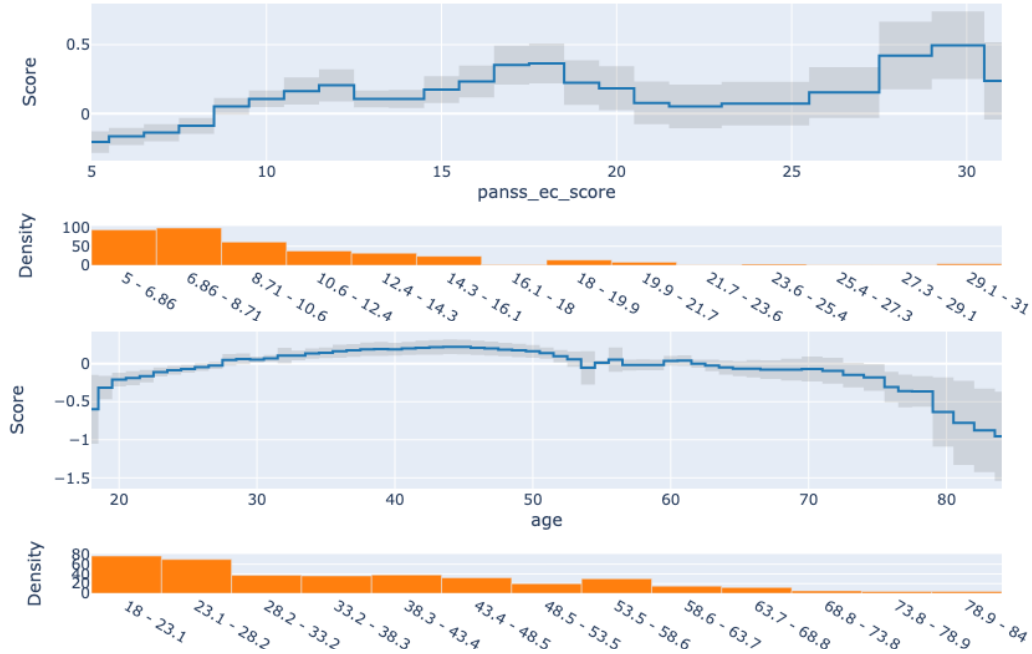


Figure 16: Shape functions for variables for PANSS-EC score and age from EBM for the Benzodiazepines dataset.

Table 15 shows the variables that contribute to an increased probability of commencing benzodiazepines. These include being part of the GAP study, being male, having prior admissions, being referred for reasons not related to suicide, being

Variable	Score_0	Score_1
from_study	-0.532	0.489
gender	0.096	-0.090
prior_admit	-0.069	0.044
suic_rel_for_ref	0.436	-0.137
intake_suicide_assess	-0.128	0.339
referral_paragraph	0.022	-0.096
specialist_paragraph	-0.029	0.288
suic_attempts_recent	-0.034	0.169
suic_thoughts_recent	0.162	-0.057
substance_abuse_recent	-0.077	0.148

Variable	Affective	Other	Psychosis	Substance Abuse
diagnosis_category	0.131	-0.503	0.216	0.294

Table 15: Scores for binary variables (top) and diagnosis category (bottom) from EBM for the Benzodiazepines dataset.

assessed as high suicide risk, and having a referral for voluntary hospitalization. Other contributing factors are forced hospitalization, as noted by the specialist, recent suicidal attempts, absence of recent suicidal thoughts, and recent substance abuse.

5.5.3 GAM and GLM

The shape functions for the two continuous variables for PANSS-EC score and age are extracted from the GAM model and shown in Figure 17. The shape functions are similar to the ones from EBM seen in Figure 16, though the GAM model doesn't seem to catch the same patterns as EBM does as it is smoothed. This is to be expected as GAM uses smoothing splines and EBM uses a maximum of 5000 decision trees for each shape function. As seen in Chapter 4, the EBM claims to have a smaller uncertainty in the shape function than GAM, which is more pessimistic.

The p-values, testing if each of the model coefficients is different from 0 for the variables in the GLM and GAM models on the benzodiazepine dataset, are extracted and sorted from lowest to highest. The results are shown in Table 16 and Table 17. The p-values are not a direct measure of importance, but can nevertheless be used when comparing the models. For the continuous variables in the GAM model, the p-value presented in the top table in Table 17 is the p-value from `summary.glm()`, while the p-value in the bottom table is the p-value from the ANOVA test. See Section 5.3 for further details. Comparing the results from the GLM and GAM models, we see that all variables have the same direction of the estimate for both models, except for the variable regarding suicidal thoughts recently, which also has the highest p-value in the GLM and GAM. Comparing to the direction of the estimates from the EBM model, which is the difference between the estimates for Score_1 and Score_0 in Table 15, we see that the direction is the same for all variables except for the variable for suicide thought recently where it has the same

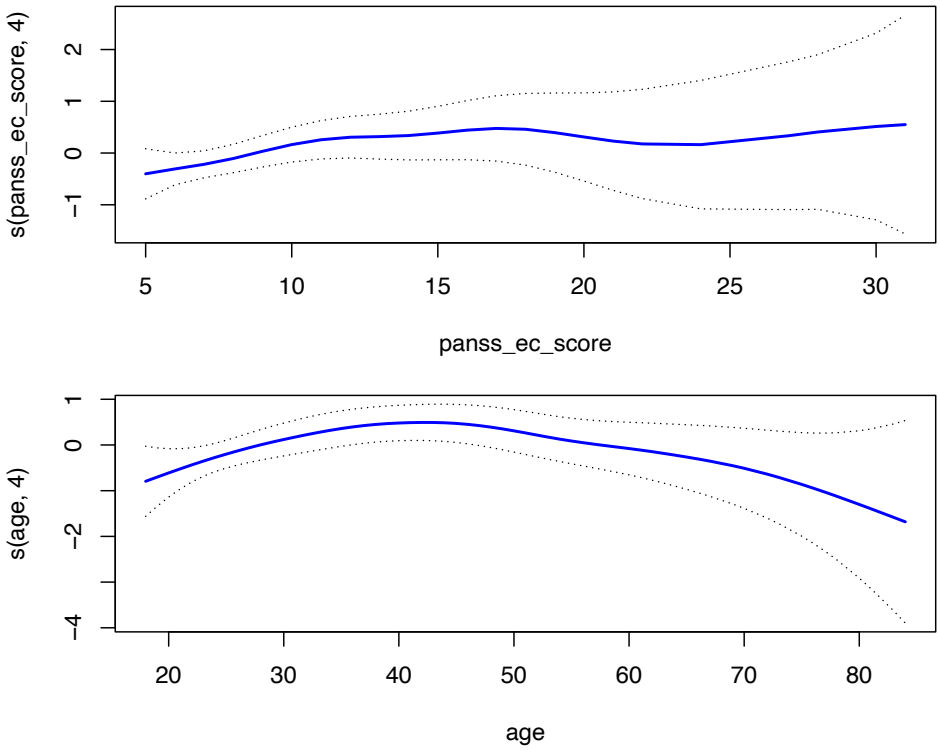


Figure 17: Shape functions for variables for PANSS-EC score and age from GAM for Benzodiazepines dataset.

direction as the GLM model. The p-values are similar for GAM and GLM, with little difference in the ranking of the variables between them. Comparing these two to the EBM model, we see that there is little difference, other than the GLM ranking the variable age low compared to the two other models.

5.5.4 PANSS-EC Score

In the EBM model, the PANSS-EC score is ranked as the 5th most important variable with an importance of 0.154, seen in Table 14. This is a little below, but in line with, the average importance from the 1000 train-test splits which is 0.201. This indicates that the train-test split used in Figure 13 is representative and implies that inference can be made from the shape function of the PANSS-EC score in Figure 16. Comparing Figure 16 and Figure 17, the shape functions for PANSS-EC score are similar with the greatest difference being larger uncertainty in GAM. This also implies that the inference made from the EBM model is reliable, though conclusions regarding the high values of PANSS-EC score should be made with caution. It should be noted that the GAM model does not consider the spline for PANSS-EC score to be significant as a 4-degree spline and the ANOVA test shows that the non-parametric additions to an additive term are not significant. This might imply that a linear term in PANSS-EC would be more suitable. For the GLM model the PANSS-EC score is deemed significant with a positive coefficient. The shape function in Figure 16 implies that patients with a PANSS-EC score between 10 and 20, and over 25, have a higher probability of commencing benzodiazepines than those with a score lower than 10 or between 20 and 25.

	Estimate	Std. Error	z value	Pr(>|z|)
from_studyGAP	1.579	0.267	5.922	3.19e-09
diagnosis_category_Other	-1.309	0.335	-3.914	9.09e-05
intake_suicide_assess	0.724	0.287	2.528	0.0115
suic_rel_for_refyes	-0.848	0.386	-2.194	0.0282
panss_ec_score	0.055	0.027	2.055	0.0399
referral_paragraph	-0.612	0.488	-1.255	0.209
(Intercept)	-1.446	1.179	-1.227	0.220
suic_attempts_recentyes	0.420	0.349	1.202	0.229
diagnosis_category_Psychosis	0.411	0.386	1.063	0.288
specialist_paragraph	0.611	0.637	0.960	0.337
substance_abuse_recentyes	0.251	0.317	0.794	0.427
genderWoman	-0.203	0.257	-0.791	0.429
diagnosis_category_Substance_Abuse	0.219	0.397	0.553	0.580
age	0.004	0.008	0.457	0.647
prior_admityes	0.097	0.262	0.369	0.712
suic_thoughts_recentyes	-0.069	0.391	-0.177	0.859

Table 16: Summary of GLM results for the Benzodiazepines dataset.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.900	1.207	-1.574	0.116
from_studyGAP	1.692	0.270	6.256	3.95e-10
diagnosis_category_Other	-1.300	0.339	-3.840	0.000123
intake_suicide_assess	0.778	0.294	2.647	0.00812
suic_rel_for_refyes	-0.942	0.393	-2.398	0.0165
s(panss_ec_score, 4)	0.048	0.027	1.791	0.0733
suic_attempts_recentyes	0.467	0.354	1.320	0.187
specialist_paragraph	0.841	0.648	1.298	0.194
referral_paragraph	-0.627	0.495	-1.267	0.205
diagnosis_category_Psychosis	0.397	0.396	1.003	0.316
substance_abuse_recentyes	0.246	0.324	0.760	0.447
genderWoman	-0.167	0.260	-0.641	0.521
s(age, 4)	0.004	0.009	0.481	0.630
diagnosis_category_Substance_Abuse	0.188	0.405	0.465	0.642
prior_admityes	0.034	0.267	0.127	0.899
suic_thoughts_recentyes	0.018	0.398	0.046	0.964

	Npar Df	Npar Chisq	P(Chi)
s(panss_ec_score, 4)	3.000	3.346	0.341
s(age, 4)	3.000	11.457	0.009

Table 17: Summary of GAM results (top) and ANOVA (bottom) from the Benzodiazepines dataset.

5.6 Mood Stabilizers

We perform the same inference as in 5.5 for the Mood Stabilizers dataset. This dataset resulted in the highest AUC for all three models, but it stands out from the other datasets as the case rate is only 0.065.

5.6.1 Variable Importance

As done in Section 5.5, the variable importance for the ten most important regression terms in the EBM model is extracted along with the corresponding average importance, standard deviation and inclusion count over the 1000 train-test splits. The results are presented in Table 18. There are five univariate terms and five interactions in the top ten terms. The univariate terms all have an average importance in agreement with the average importance from the 1000 train-test splits, and is included in all 1000 models. The term with the highest average importance, of 0.483, is the variable for diagnosis category. The interaction term with the highest importance is between prior admits and diagnosis category and has an importance of 0.227. It is included in 375 of the 1000 models, which implies that it is not a consistently important term in the EBM model. The same goes for the two other interactions ranked 9th and 10th in the table. These are only included in 269 and 133 of the 1000 models. The two remaining interactions are between age and the PANSS-EC score, and age and diagnosis category. These have an importance of 0.182 and 0.160, and are included in 999 and 969 of the 1000 models. This implies that these interactions are consistently important terms in the EBM model over the 1000 train-test splits.

Variable	Imp	Avg Imp	SD	Inclusion Count
from_study	0.373	0.339	0.099	1000
diagnosis_category	0.353	0.483	0.162	1000
substance_abuse_recent	0.290	0.313	0.089	1000
prior_admit&diagnosis_category	0.227	0.162	0.058	375
specialist_paragraph	0.199	0.160	0.040	1000
age & panss_ec_score	0.182	0.118	0.085	999
intake_suicide_assess	0.179	0.139	0.090	1000
age & diagnosis_category	0.160	0.113	0.056	969
from_study & age	0.159	0.090	0.038	269
from_study&suic_thoughts_recent	0.159	0.116	0.039	133

Table 18: Variable importance (Imp) from EBM for our training set in the original train-test split, including average importance (Avg Imp), standard deviation (SD), and inclusion count, derived from the 1000 train-test splits of the Mood Stabilizers dataset.

Due to the multiple non-consistent interaction terms in the top ten terms in Table 18, we wish to investigate the average importance further and extract the 15 terms with the highest average importance. The results are shown in Table 19. All interaction terms that have an inclusion count below 500 are deemed not consistently important and are excluded from this table. This number is chosen

from the analysis in Section 4.3.4. We see that when only considering the average importance, the top ten variables change considerably, as expected due to the interaction terms with low inclusion count in Table 18. The only interaction term that qualifies for the top ten is then the interaction between age and gender, which has an inclusion count of 784. The interaction between age and PANSS-EC score is only ranked 14th in the table, which is a big difference from the 6th place in Table 18. Most notably the PANSS-EC score didn't make the table in Table 18 and is ranked 5th in Table 19 with an average importance of 0.245, however, the interaction between PANSS-EC score and age made the list.

Variable	Avg Imp	SD	Inclusion Count
diagnosis_category	0.483	0.162	1000
from_study	0.339	0.099	1000
substance_abuse_recent	0.313	0.089	1000
age	0.263	0.130	1000
panss_ec_score	0.245	0.132	1000
gender	0.207	0.097	1000
prior_admit	0.192	0.102	1000
suic_rel_for_ref	0.172	0.066	1000
age&gender	0.162	0.101	784
specialist_paragraph	0.160	0.040	1000
suic_rel_for_ref&diagnosis_category	0.140	0.044	910
intake_suicide_assess	0.139	0.090	1000
suic_thoughts_recent&diagnosis_category	0.124	0.034	806
age&panss_ec_score	0.118	0.085	999
age&diagnosis_category	0.113	0.056	969

Table 19: Top 15 terms with regard to average importance from EBM for the Mood Stabilizers dataset.

5.6.2 Shape Functions and Coefficients

As shown in Figure 18, the shape functions for the two continuous variables, PANSS-EC score and age, are extracted from the EBM model. We see that the PANSS-EC score has a shape function that starts negative before increasing to two small peaks at a PANSS-EC score of about 10 and 20 and then increasing to a peak of a score of 1 with a PANSS-EC score of above 30.

The age variable has a shape function that starts at a slightly negative score and increases steadily to a peak at age around 60, with a score of a little under 0.5. After this the score decreases to a score of approximately -0.3 around age 80. For both shape functions the uncertainty increases with the value of the variable.

The shape functions of the interaction between age and gender, and age and PANSS-EC score is also of interest to investigate due to their relatively high average importance in Table 19. The shape functions are shown in Figure 19. These shape functions imply that the probability of commencing mood stabilizers is the opposite for men and women, with the score increasing with age for women, and

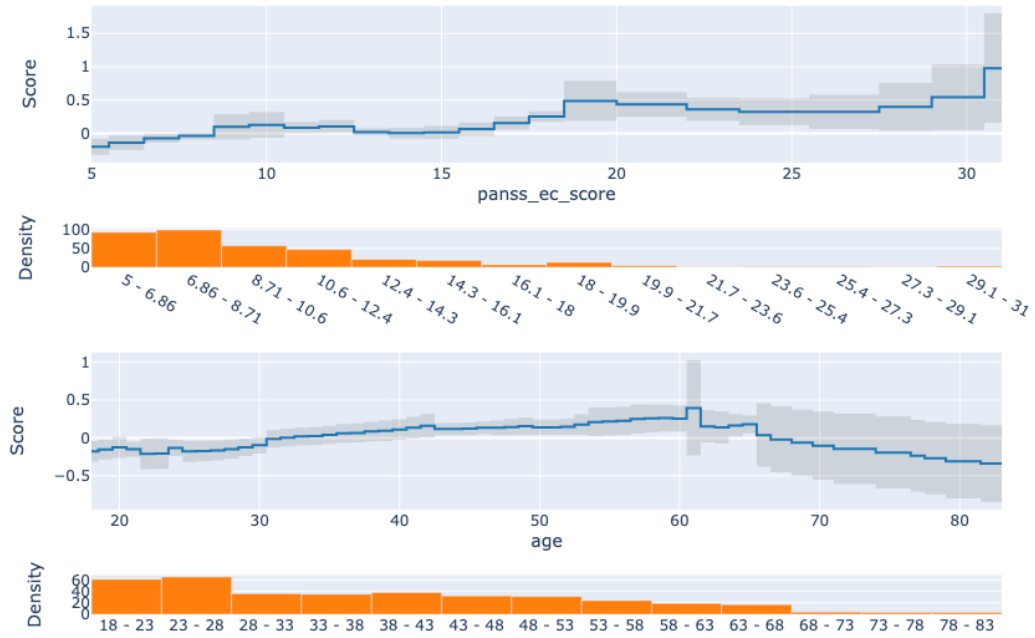


Figure 18: Shape functions for the variables for PANSS-EC score and age from EBM for the Mood Stabilizers dataset.

decreasing for men. The interaction between age and PANSS-EC score is more complex, with PANSS-EC scores over 10 the score is high for age around 20, and then decreasing as the age approaches 80. For PANSS-EC scores below 10 the score is negative for ages around 20, and then increases with age. The contributions in these should be valuated in light of the variable importances from Table 18 and Table 19, and the fact that these plots are presented without uncertainty.

Additionally, the variable scores for the categorical variables are shown in Table 20. This shows the variables that contribute to an increased probability of commencing mood stabilizers. These include the estimates with positive direction as being part of the GAP study, having prior admissions, having a referral for forced hospitalization, having a specialist referral for forced hospitalization, recent suicidal attempts and recent substance abuse. The estimates with negative direction are being a man, suicide was not relevant for the referral, being assessed at low suicide risk and no recent suicidal thoughts.

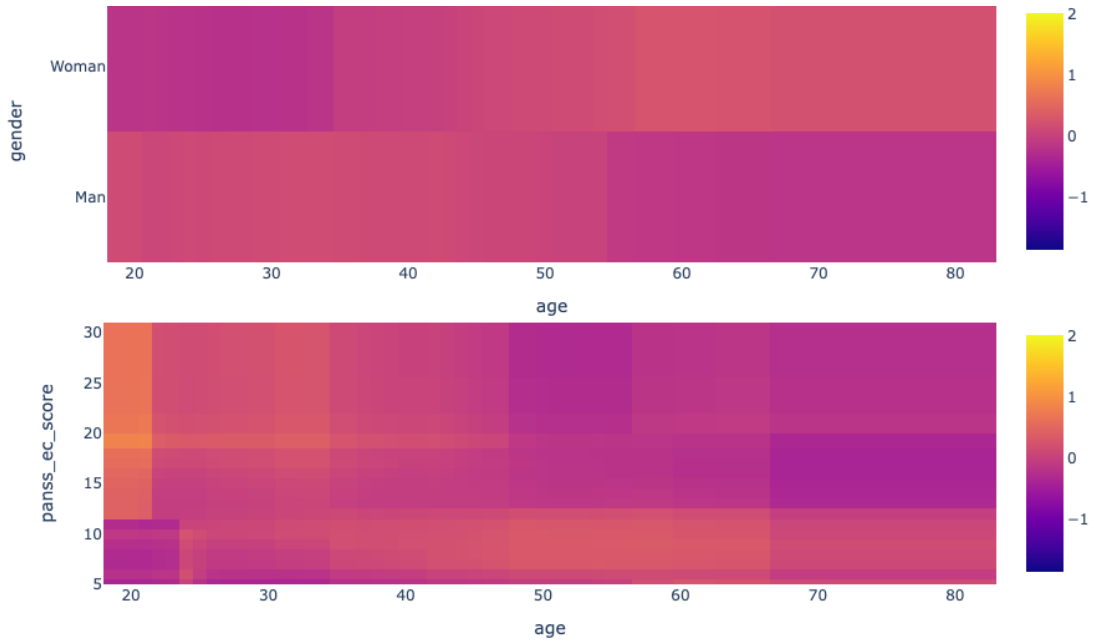


Figure 19: Shape functions for the interaction between age and gender, and age and PANSS-EC score from EBM for the Mood Stabilizers dataset.

Variable	Score_0	Score_1
from_study	-0.381	0.365
gender	0.116	-0.111
prior_admit	-0.130	0.082
suic_rel_for_ref	0.268	-0.096
intake_suicide_assess	0.124	-0.323
referral_paragraph	-0.073	0.304
specialist_paragraph	-0.113	0.830
suic_attempts_recent	-0.088	0.451
suic_thoughts_recent	0.159	-0.060
substance_abuse_recent	-0.226	0.404

Variable	Affective	Other	Psychosis	Substance Abuse
diagnosis_category	0.155	-0.415	-0.420	0.584

Table 20: Scores for binary variables (top) and the diagnosis categories (bottom) from EBM for the Mood Stabilizers dataset.

5.6.3 GAM and GLM

The shape functions for the continuous variables from GAM are extracted and shown in Figure 20. The shape functions are similar to the ones from EBM seen in Figure 18, though the GAM model doesn't seem to catch the same local patterns as EBM does, which is to expect from the smoothing splines used in GAM. As seen in Chapter 4 and in Section 5.5, the EBM claims to have a smaller uncertainty in the shape function than GAM, which is more pessimistic.

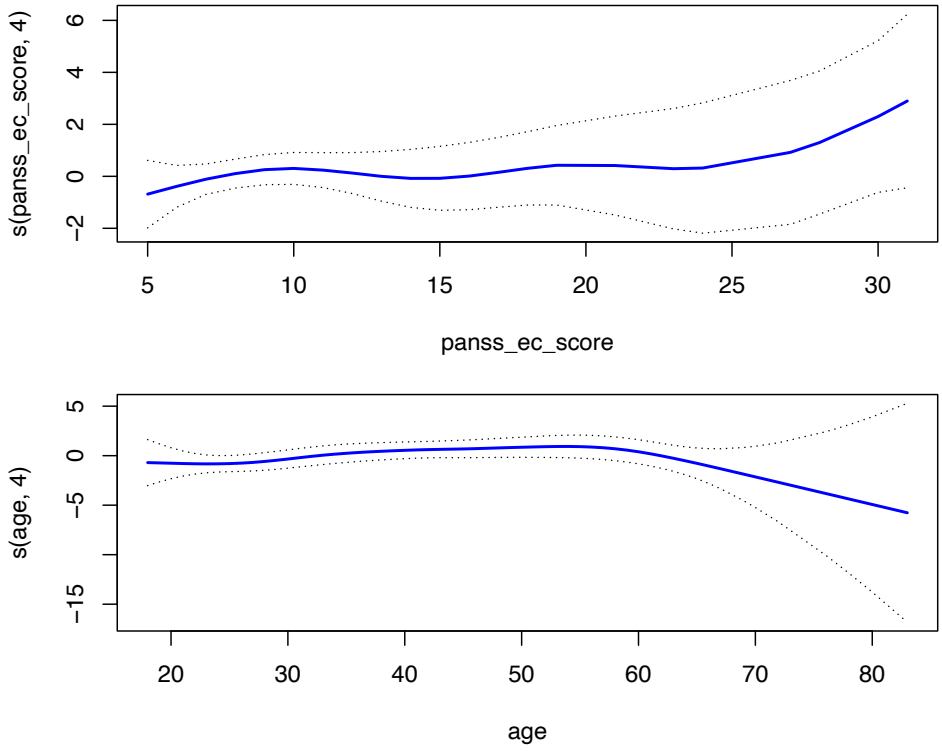


Figure 20: Shape functions for PANSS-EC score and age from GAM for Mood Stabilizers dataset.

The p-values for the variables in the GLM and GAM models on the Mood Stabilizers dataset are extracted and sorted from most significant to least significant. The results are shown in Table 21 and Table 22.

All variables have the same direction of the estimates in all three models. Looking at the p-values from the GLM and GAM model, we see that the seven most significant variables, not counting the intercept, are the same for both models. Most notably, the variable for which study is the most significant variable in GAM and GLM and is also the variable with the highest importance in the EBM model, though the variable for diagnosis category has a higher average importance in the EBM model. It is difficult to compare the lower ranks of the variables, as the EBM model has interaction terms that are not present in the GLM and GAM models.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-10.490	2.279	-4.603	4.17e-06
from_studyGAP	1.691	0.619	2.731	0.00631
specialist_paragraph	3.680	1.444	2.549	0.0108
diagnosis_category_Psychosis	-2.453	1.191	-2.060	0.0394
suic_attempts_recentyes	1.413	0.736	1.919	0.055
intake_suicide_assess	-1.274	0.855	-1.491	0.136
diagnosis_category_Substance_Abuse	0.940	0.762	1.233	0.218
panss_ec_score	0.065	0.055	1.199	0.230
diagnosis_category_Other	-1.347	1.128	-1.194	0.232
substance_abuse_recentyes	0.945	0.814	1.161	0.246
referral_paragraph	-1.211	1.299	-0.932	0.351
suic_rel_for_refyes	-0.631	0.777	-0.813	0.416
genderWoman	-0.447	0.581	-0.770	0.441
age	0.013	0.018	0.742	0.458
suic_thoughts_recentyes	-0.512	0.811	-0.631	0.528
prior_admityes	0.364	0.601	0.605	0.545

Table 21: Summary of GLM results for Mood Stabilizers dataset.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-12.103	2.421	-4.998	5.78e-07
from_studyGAP	2.225	0.675	3.295	0.000984
specialist_paragraph	4.804	1.550	3.100	0.00194
diagnosis_category_Psychosis	-2.789	1.228	-2.271	0.0231
suic_attempts_recentyes	1.763	0.802	2.200	0.0278
intake_suicide_assess	-1.481	0.922	-1.606	0.108
diagnosis_category_Substance_Abuse	1.212	0.801	1.514	0.13
s(panss_ec_score, 4)	0.076	0.056	1.346	0.178
referral_paragraph	-1.749	1.382	-1.265	0.206
suic_thoughts_recentyes	-0.999	0.828	-1.206	0.228
diagnosis_category_Other	-1.237	1.150	-1.076	0.282
s(age, 4)	0.020	0.021	0.957	0.338
genderWoman	-0.515	0.620	-0.832	0.406
suic_rel_for_refyes	-0.666	0.808	-0.825	0.41
substance_abuse_recentyes	0.523	0.822	0.636	0.525
prior_admityes	0.390	0.630	0.620	0.536

	Npar Df	Npar Chisq	P(Chi)
s(panss_ec_score, 4)	3.000	4.101	0.251
s(age, 4)	3.000	6.165	0.104

Table 22: Summary of GAM results (top) and ANOVA (bottom) from the Mood Stabilizers dataset.

5.6.4 PANSS-EC Score

In the EBM model, the interaction between age and PANSS-EC score is ranked as the 6th most important variable, with an importance of 0.182 seen in Table 18 and an average importance of 0.118 seen in Table 19. This difference should be taken into account when interpreting the corresponding shape function in Figure 19. This shape function implies most notably that the probability of commencing mood stabilizers is higher for patients with a PANSS-EC score over 10 and an age around 20, than for those with a PANSS-EC score over 10 and an older age. The opposite applies for the PANSS-EC score below 10. The PANSS-EC score is itself not a part of the most important variables in this table, despite it having a considerably higher average importance seen in Table 19. This should be taken into consideration when making inferences from the shape function for PANSS-EC score in Figure 18. This shape function is similar to the one from GAM seen in Figure 20, though as we have seen before, the GAM model is more pessimistic in the uncertainty of the shape function. The shape function from EBM implies, in short, that patients with a low PANSS-EC score have a lower probability of commencing mood stabilizers than those with a higher score.

CHAPTER

SIX

DISCUSSION & FURTHER WORK

We will in this section discuss the results from Chapter 5. The performance of EBM is evaluated in comparison to the benchmark models GLM and GAM, across the four datasets. Before performing any of the analyses, we would expect the EBM to outperform GAM and GLM. This is partly due to the results presented in (Nori et al. 2019*a*) and Lou et al. (2013), where EBM is shown to outperform other methods for a variety of datasets. In addition to this, EBM will inherently consider all possible interactions, while no interactions are explicitly included in the other two models. We would also expect the EBM and GAM to outperform GLM due to allowing for non-linear relations in the continuous variables. We will also discuss the PANSS-EC score's relation to the different medications and how the shape functions can be used to interpret this relationship.

6.1 Performance of the Models

In Chapter 5, we saw that the only two datasets that resulted in acceptable performance with regard to AUC are the Benzodiazepines and Mood Stabilizers datasets.

The Benzodiazepines dataset resulted in the second-best performance where there is little difference between the models, though GAM has a slightly higher AUC than EBM, which in turn has a slightly higher AUC than GLM. Looking at the variable importances, there doesn't seem to be any interaction terms of notable importance in the EBM model. This is likely the reason for EBM not outperforming GAM in this dataset, and may also be the reason the EBM model performs slightly worse than the GAM model, since EBM is prone to overfitting interaction terms as seen in Chapter 4. From the results seen in 5.5, we see that the two continuous variables are not given a particularly high importance in the EBM model, ranking in as number 5 and 6. For GAM, we see that the age variable is deemed significant in the ANOVA test, indicating a non-linear relationship between age and the commencement of benzodiazepines, though the coefficient seen from the `summary.glm()` function shows that the age variable is not significant. This is consistent with the GLM model, where the linear coefficient for the age variable is not deemed significant either. Due to the two continuous variables not showing

clear non-linear properties as well as not being of particular importance, it is not surprising that the GLM model is not considerably outperformed by the EBM and GAM models.

The Mood Stabilizers dataset had the best performance of the datasets, where even though the median performance is close between the models, the EBM model seems to outperform the other two models due to the more consistent results seen in the small span of the AUCs in the violin plots in Figure 14. For the EBM model, there are several interaction terms of consistent importance, which likely explains the better performance of the EBM model. The variables for age and PANSS-EC both have high importance in the EBM model, ranking as number 4 and 5 over the 1000 train-test splits. These variables are deemed much less important in the GAM and GLM models, since they are not considered to significantly contribute to the model, although this is compared to the original train-test split where these variables also were not of significant importance in the EBM model. It would be of interest to study more closely the results from GAM and GLM for all 1000 train-test splits and compare them to EBM. As seen in the results from Simulation Study 4 in Chapter 4, the EBM model is very prone to overfitting, especially when there are numerous variables in the model that don't have an underlying relation to the response. This may result in true underlying relations being missed and not being included in the model consistently. This is likely the reason why EBM is not outperforming GAM and GLM more clearly.

The Antipsychotics and Hypnotics dataset resulted in all models having an average AUC well below 0.7 and have, out of the 1000 train-test splits, both a considerable number of splits where the AUC is less than 0.5. From this we conclude that reliable inference cannot be made regarding the underlying relations in these datasets. Despite all models performing badly for these datasets, it is of interest that EBM performs in general worse than the GAM and GLM in the Antipsychotics dataset. This stands in contrast to our expectation that EBM should outperform GAM and especially GLM.

Regarding performance, there is little from our analyses that implies that the EBM model should be preferred over the GAM and GLM models for our data. Both GAM and GLM are here only used as benchmark models, and have not been optimized in any way with regard to variable selection, interaction terms or hyperparameters. It is likely that the performance of the GAM and GLM models could be improved by including interaction terms and non-linear transformations of the continuous variables. In the `gam` package used for modeling GAM, one needs to specify constant degrees of freedom for the spline terms for the continuous variables. It would be interesting to estimate this for each variable, as this would avoid overfitting. This can be done in the `mgcv` package in `R` (Wood 2023), which would be of interest to use instead of the `gam` package. This gives the possibility to, by default, let all continuous variables have smoothing spline terms, since the degrees of freedom are estimated.

It should be noted that with the small datasets used in this thesis, there are large variations in the performance of all models, and it is therefore questionable whether the results from a single train-test split are suitable for reliable inference. Especially in the case of the Hypnotics dataset is this the case, where the original

train-test split resulted in a terrifying AUC of 0.445 for GAM. This split is an outlier in the general performance of the model, and using another train-test split would likely result in more representative results. We are not aware of to which extent it is common in publications within the field of medicine to present results from several train-test splits, although we know that many analyses in the field of medicine do not involve a test set at all. More examples of the shape functions and variable importances from different train-test splits would be of interest to see, but have not been included in this thesis.

When it comes to intelligibility, both EBM and GAM can catch non-linearity in the data and visualize these. The EBM tends to be more optimistic in its uncertainty estimates, which is a feature that one should be aware of when using the model. A clear advantage that EBM has over GAM is that it can inherently handle discontinuities in the shape function, while this needs to be explicitly specified in GAM. Especially in the case of intelligibility is this a clear advantage, as this could reveal important patterns in the data that would otherwise be missed using GAM and GLM. It should be noted that the `mgcv` package in `R` provides different possibilities for splines that can be used to more easily model discontinuities in the data.

With regard to interaction, it is an advantage that EBM inherently handles all possible interactions, while these need to be explicitly specified in GAM and GLM. Although when specified, can be used in GAM similarly as in EBM. The `mgcv` package in `R` supplies numerous ways to visualize these interactions, which is limited in the current version of EBM.

A possible usage of the EBM model could be to use it as a first step in a modeling process. We then propose that the EBM model could be used to search for interactions and discontinuities, which then could be included in a GAM or a GLM model. This would be a way to combine the advantages of EBM with the advantages of the GAM and GLM models. With small datasets like the ones used in this thesis, it may be important to perform many train-test splits. For larger datasets, one may divide the data into three, where one part is used for model selection, one part is used for model training, and one part is used for model testing. The model selection may then include the EBM model, and GAM may be used with the discovered interactions and discontinuities for the model training.

6.2 PANS-EC Score

For the Benzodiazepine dataset, the evaluation of the PANSS-EC score's variable importance and the performance of the EBM model implies that inference can be made from the shape function for the PANSS-EC score. The shape function shows as mentioned a clear increase in probability of commencement of benzodiazepines for PANSS-EC scores around 17 and over 25, with a decrease in between. The prediction of the shape functions in areas with few observations is highly uncertain and should be interpreted with caution, as seen in Chapter 4.

The same goes for the viability of the Mood Stabilizers dataset, where the shape function of the PANSS-EC score and the interaction between PANSS-EC score and age is of interest. The shape function for PANSS-EC score shows an increase in

the probability of commencement of mood stabilizers with PANSS-EC scores of up to 19, where the probability remains stable for higher levels. As mentioned for the Benzodiazepines dataset, there are very few observations for the higher PANSS-EC scores, making the prediction of the shape function highly uncertain. The shape function for the interaction term shows an interesting relationship between the PANSS-EC score, age and the probability of commencement of mood stabilizers. For patients around age 20, the probability of commencement of mood stabilizers is much higher for patients with PANSS-EC scores above 12, than for patients with lower PANSS-EC scores. This relation changes with age and seems to be the opposite for patients older than 50, though not as obvious. This is a relationship that should be further investigated.

For the Antipsychotics and the Hypnotics dataset, we choose not to make any inference from the shape function of the PANSS-EC score, due to the poor performance of the models, and relatively low variable importance of the PANSS-EC score.

6.3 Theory

The EBM model is a relatively new model, and the published theory behind the model in several areas is scarce or difficult to find detailed descriptions of. Many theoretical aspects of the model could not be found in the documentation provided in publications, but had to be inferred from the code and issues on the GitHub repository. It should be mentioned that the creators of the model are very active on the GitHub repository and seem to have been very helpful in answering questions and providing guidance for users of the EBM, although these questions could be avoided with more thorough documentation. We have tried to contact the authors directly to confer about specific parts of the model, but have not succeeded in getting a response.

6.4 Contributions and Further Work

This thesis's contributions consist of a thorough evaluation of the usage of the EBM model in small medical datasets. We reveal strengths and weaknesses with EBM in comparison to the benchmark models GAM and GLM, which is of interest for the continuation of research on the topic of suicide prevention research. The up-to-date presentation of the EBM model is also a clear contribution of this thesis.

Information regarding the relation between the PANSS-EC score and the commencement of the relevant medication is obtained, seen in light of the viability of the models used on the datasets. This will hopefully prove useful in future research on the suicide prevention data used in this thesis.

In the future it would be of interest to examine the performance of other state-of-the-art machine learning models on the datasets to see if better performance can be obtained. The relations between the PANSS-EC score and the different medications should be further investigated, as there may be interesting underlying reasons for the relationships found.

CHAPTER

SEVEN

CONCLUSIONS

This thesis aimed to model the probability of the commencement of medication at an acute psychiatric department, using clinical data from suicide prevention research, and provide insights into the relation between the commencement of medication and the clinical assessment tool PANSS-EC. This was done using the Explainable Boosting Machine model (EBM), a new highly interpretable machine learning model. The EBM model was evaluated with regard to performance measured in area under the receiver operating characteristic, compared to the benchmark models Generalized Additive Models (GAM) and Generalized Linear Models (GLM). Simulation studies were performed to analyze the model performance on known underlying relations and further provide a basis for the analyses of the real data.

Our analyses found that the performance of the EBM model was not considerably better than that of GAM and GLM, although the model proves useful in interaction detection and uncovering of discontinuities in the data. The performance of all three models was varying between the medication studied. Meaningful relations between the commencement of medication and PANSS-EC was found for the medications where the performance of the model were deemed adequate.

Theoretical contributions of this thesis consist of a comprehensive presentation and comparison of models to analyze small binary classification data sets from medical research, a detailed up-to-date description of the existing literature on the subject and insight into the relations between the commencement of medication and PANSS-EC.

BIBLIOGRAPHY

Arbuthnott, J. (1710), 'An argument for divine providence, taken from the constant regularity observ'd in the births of both sexes', *Philosophical Transactions of the Royal Society* .

Bamber, D. (1975), 'The area above the ordinal dominance graph and the area below the receiver operating characteristic graph', *Journal of Mathematical Psychology* **12**(4), 387–415.
URL: https://www.sciencedirect.com/science/article/pii/0022249675900012

Bauer, E. & Kohavi, R. (1999), 'An empirical comparison of voting classification algorithms: Bagging, boosting, and variants', *Machine Learning* **36**, 105–139.
URL: https://doi.org/10.1023/A:1007515423169

Bland, J. M. & Altman, D. G. (1986), 'Statistical methods for assessing agreement between two methods of clinical measurement', *The Lancet* **327**(8476), 307–310. Originally published as Volume 1, Issue 8476.
URL: https://www.sciencedirect.com/science/article/pii/S0140673686908378

Breiman, L., Friedman, J., Olshen, R. A. & Stone, C. J. (1984), *Classification and Regression Trees*, 1 edn, Chapman and Hall/CRC.
URL: https://doi.org/10.1201/9781315139470

DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. (1988), 'Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach', *Biometrics* **44**(3), 837–845.

Dunn, P. K. & Smyth, G. K. (2018), *Generalized Linear Models with Examples in R*, Springer.

Engelsen, G. F. (2024), 'Insights into area under receiver operating characteristic curves and evaluating medication variables for suicidal crisis syndrome prediction', *TMA4500 Specialization Project* . Available upon request.

Friedman, J. H. (2001), 'Greedy function approximation: A gradient boosting machine', *The Annals of Statistics* **29**(5), 1189–1232.
URL: http://www.jstor.org/stable/2699986

Hand, D. J. & Till, R. J. (2001), 'A simple generalisation of the area under the roc curve for multiple class classification problems', *Machine Learning* **45**(1), 171–186.

Hanley, J. A. & McNeil, B. J. (1982), 'The meaning and use of the area under a receiver operating characteristic(roc) curve', *Radiology* **143**(1), 29–36.

Hastie, T. (2023), *gam: Generalized Additive Models*. R package version 1.22-3.
URL: `https://rdocumentation.org/packages/gam/versions/1.22-3`

Hastie, T. & Tibshirani, R. (1987), 'Generalized additive models: Some applications', *Journal of the American Statistical Association* **82**(398), 371–386. Accessed 9 June 2024.
URL: `https://doi.org/10.2307/2289439`

Hastie, T. & Tibshirani, R. (1990), *Generalized Additive Models*, 1 edn, Chapman and Hall, London, UK.

Hastie, T., Tibshirani, R. & Friedman, J. (2009), *The Elements of Statistical Learning*, Springer Series in Statistics, New York, NY.

Hosmer, D. W., Lemeshow, S. & Sturdivant, R. X. (2013), *Applied Logistic Regression*, 3 edn, John Wiley and Sons.
URL: `https://doi.org/10.1002/9781118548387`

Høyen, K. S., Solem, S., Cohen, L. J., Prestmo, A., Hjemdal, O., Vaaler, A. E. & Torgersen, T. (2022), 'Non-disclosure of suicidal ideation in psychiatric inpatients: Rates and correlates', *Death Studies* **46**(8), 1823–1831.
URL: `https://doi.org/10.1080/07481187.2021.1879317`

InterpretML Team (2020), 'Issue 207: Discussion on feature contribution calculation methods in ebm', `https://github.com/interpretml/interpret/issues/207`. Accessed: 2024-06-01.

InterpretML Team (2021), 'Issue 409: Explanation of error bars and model behavior in ebm', `https://github.com/interpretml/interpret/issues/409`. Accessed: 2024-06-01.

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2021), *An Introduction to Statistical Learning : With Applications in R*, Springer texts in statistics, second edition edn, Springer, New York.

Kay, S. R., Fiszbein, A. & Opler, L. A. (1987), 'The Positive and Negative Syndrome Scale (PANSS) for Schizophrenia', *Schizophrenia Bulletin* **13**(2), 261–276.
URL: `https://doi.org/10.1093/schbul/13.2.261`

Kuhn, M. (2023), *caret: Classification and Regression Training*. R package version 6.0-94.
URL: `https://rdocumentation.org/packages/caret/versions/6.0-94`

Lou, Y., Caruana, R. & Gehrke, J. (2012), 'Intelligible models for classification and regression', p. 150–158.
URL: `https://doi.org/10.1145/2339530.2339556`

Lou, Y., Caruana, R., Gehrke, J. & Hooker, G. (2013), 'Accurate intelligible models with pairwise interactions', p. 623–631.
URL: https://doi.org/10.1145/2487575.2487579

Ludvigsen, M. B. (2023), Suicide crisis syndrome in a norwegian acute psychiatric unit: Exploring risk factors using statistical learning and inference, Master's thesis, Norwegian University of Science and Technology. Available upon request.

Mann, H. B. & Whitney, D. R. (1947), 'On a test of whether one of two random variables is stochastically larger than the other', *The Annals of Mathematical Statistics* **18**(1), 50–60.

McCullagh, P. & Nelder, J. A. (1989), *Generalized Linear Models*, Vol. 37 of *Monographs on statistics and applied probability*, 2nd edn, Chapman and Hall, London. Bibliography: p. 479-499.

Melby, L. (2024), 'Symptoms of suicide crisis syndrome and associated risk factors in an acute psychiatric population, a cross sectional cohort study'. Available upon request.

Midtfjord, A. D., De Bin, R. & Huseby, A. B. (2022), 'A decision support system for safer airplane landings: Predicting runway conditions using xgboost and explainable ai', *Cold Regions Science and Technology* **199**, 103556.
URL: https://www.sciencedirect.com/science/article/pii/S0165232X22000751

Montoya, A., Valladares, A., Lizán, L., San, L., Escobar, R. & Paz, S. (2011), 'Validation of the excited component of the positive and negative syndrome scale (panss-ec) in a naturalistic sample of 278 patients with acute psychosis and agitation in a psychiatric emergency room', *Health Qual Life Outcomes* **9**, 18.
URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3078838/

Nori, H., Jenkins, S., Koch, P. & Caruana, R. (2019*a*), 'Interpretml: A unified framework for machine learning interpretability'. Preprint, arXiv:1909.09223v1 [cs.LG].
URL: https://doi.org/10.48550/arXiv.1909.09223

Nori, H., Jenkins, S., Koch, P. & Caruana, R. (2019*b*), 'Interpretml: A unified framework for machine learning interpretability', GitHub.
URL: https://github.com/interpretml/interpret

Oxford University Press (n.d.), 'Intelligibility'. Accessed on: 2024-05-12.
URL: https://www.oxfordlearnersdictionaries.com/definition/english/intelligibility?q=intelligibility

Prestmo, A., Høyen, K., Vaaler, A. E., Torgersen, T. & Drange, O. K. (2020), 'Mortality among patients discharged from an acute psychiatric department: A 5-year prospective study', *Frontiers in Psychiatry* **11**(Article 816).
URL: https://doi.org/10.3389/fpsyt.2020.00816

R Core Team (2023), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: `https://www.R-project.org/`

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C. & Müller, M. (2023), *pROC: An R Package for Displaying and Analyzing ROC Curves*. R package version 1.18.5.
URL: `https://CRAN.R-project.org/package=pROC`

Sprent, P. & Smeeton, N. C. (2012), *Applied Nonparametric Statistical Methods*, 4th edn, Springer.

United Nations (2023), 'Goal 3: Ensure healthy lives and promote well-being for all at all ages'.
URL: `https://sdgs.un.org/goals/goal3`

Wickham, H. (2016), *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag New York.
URL: `https://ggplot2.tidyverse.org`

Wood, S. N. (2023), *mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation*. R package version 1.9-1.
URL: `https://rdocumentation.org/packages/mgcv/versions/1.9-1`

APPENDIX

A

MEDICATION DATA COMPARISON STUDY OF AA AND GAP

A.1 Data

In Melby (2024), the AA and GAP studies are compared with respect to their clinical variables. Here, we will compare the two studies in terms of the medication variables. It is of special interest to see if the usage and prescription of medication are different between the two studies since they were conducted five years apart. The number of 0, 1 and missing values for all medication usage and commencement variables in the two studies are presented in Table 23. There are seven medication usage variables and five medication commencement variables in the dataset. The corresponding usage and commencement variables have the same name, but with the prefix `INN` for usage and `OPPST` for commencement. The variables `INN_SSRI` and `INN_ANTIDEP` have no corresponding commencement variables and are together with `INN_OPIOIDER` and `OPPST_OPIOIDER`, the only variables not being used in the rest of the thesis.

A.2 Methods

A.2.1 Odds Ratio

The odds ratio is the ratio between two odds. In our case, this is the odds of medication usage in study AA compared to the odds of medication usage in study GAP. The odds ratio is calculated as

$$\text{OR} = \frac{\text{Odds}_{\text{AA}}}{\text{Odds}_{\text{GAP}}} = \frac{p_{\text{AA}}/(1-p_{\text{AA}})}{p_{\text{GAP}}/(1-p_{\text{GAP}})}, \tag{A.1}$$

denoting p_{AA} as the probability of medication usage in AA and p_{GAP} as the probability of medication usage in GAP. An odds ratio of less than one indicates that the probability of medication usage for a patient in study AA is lower than the

	AA			GAP		
Variable	0	1	NA	0	1	NA
INN_ANTIDEP	274	103	3	235	95	0
INN_SSRI	317	60	3	274	56	0
INN_ANTIPSYK	279	97	4	250	80	0
INN_STEMNINGSSTAB	329	47	4	291	39	0
INN_BENZO	341	36	3	300	29	1
INN_HYPNOTIKA	349	28	3	289	39	2
INN_OPIOIDER	370	7	3	309	19	2
OPPST_ANTIPSYK	223	63	94	208	105	17
OPPST_STEMNINGSSTAB	275	10	95	285	28	17
OPPST_BENZO	213	67	100	160	152	18
OPPST_HYPNOTIKA	221	57	102	222	89	19
OPPST_OPIOIDER	276	4	100	302	8	20

Table 23: Number of patients with 0, 1 and missing value (NA) for all medication usage and commencement variables in the two studies.

probability of medication usage for a patient in study GAP. The opposite applies for an odds ratio greater than one, while an odds ratio equal to one indicates that the odds of medication usage are equal between the two studies. The intuition behind this is that if the two study groups can be considered as random samples from the same population, the odds ratio should be equal to one.

A.2.2 Fisher's Exact Test

Fisher's exact is a non-parametric test used to test, in our case, the odds of a medication variable being different between the two study groups AA and GAP. Removing all missing values (NA) for each medication variable in Table 23, a contingency table is created for each medication variable. The contingency table for medication variable X_i has the form

	AA	GAP
$X_i = 1$	a	b
$X_i = 0$	c	d

Table 24: Contingency table for medication variable X_i.

From this table, the odds ratio is calculated as

$$\text{OR} = \frac{ad}{bc}, \tag{A.2}$$

where a and b are the number of patients from study AA and GAP respectively with medication usage, while c and d are the number of patients from study AA and GAP respectively without medication usage (Sprent & Smeeton 2012, p. 172). The function `fisher.test()` in the `stats` package in R (R Core Team 2023) is used to run the Fisher's exact test and calculate the odds ratio. The corresponding

confidence intervals and p-value are obtained directly from the same function using the central hypergeometric distribution. See Section A.3 for the results.

A.2.3 Logistic Regression

One can also use GLM to test if the medication usage is different between the two studies. This is done by using the medication variable as the response variable and using the variable `from_study` as the only covariate. `from_study` is a binary variable indicating if the patient is from study GAP ($X_{\texttt{from\_study}} = 1$) or AA ($X_{\texttt{from\_study}} = 0$). The logistic regression model is then

$$g(\mu) = \beta_0 + \beta_1 X_{\texttt{from\_study}}, \tag{A.3}$$

where we have that $g(\mu) = \log(\frac{\mu}{1-\mu})$ and μ is equal to the probability of medication usage. The odds of a patient with medication, $\texttt{Med}_i$, from study GAP is then

$$\text{Odds}(Y_{\texttt{Med}_i} | X_{\texttt{from\_study}} = 1) = e^{g(\mu)} = e^{\beta_0 + \beta_1}, \tag{A.4}$$

and for a patient from study AA the odds is

$$\text{Odds}(Y_{\texttt{Med}_i} | X_{\texttt{from\_study}} = 0) = e^{\beta_0}, \tag{A.5}$$

due to the fact that $X_{\texttt{from\_study}}$ is binary. The odds ratio can then be calculated from

$$\text{OR}_i = \frac{\text{Odds}(Y_{\texttt{Med}_i} | X_{\texttt{from\_study}} = 1)}{\text{Odds}(Y_{\texttt{Med}_i} | X_{\texttt{from\_study}} = 0)} = \frac{e^{\beta_0 + \beta_1}}{e^{\beta_0}} = e^{\beta_1}, \tag{A.6}$$

where β_0 is the intercept and β_1 is the coefficient for $X_{\texttt{from\_study}}$. The coefficients β_0 and β_1 are obtained from the `glm()` function in R along with the corresponding p-values. The odds ratio is then calculated from A.6. The corresponding 95% confidence intervals are obtained using the `confint()` function in the `stats` package in R (R Core Team 2023). The confidence intervals for the odds ratio are calculated by taking the exponential of the confidence intervals for the coefficients. The results are presented in the following section.

A.3 Results

In Table 25, the corresponding odds ratios, 95% confidence intervals and p-values from Fisher's exact test and logistic regression are presented. As we see, the two tests give almost the same results. The variables `INN_HYPNOTIKA`, `INN_OPIOIDER`, `OPPST_ANTIPSYK`, `OPPST_STEMNINGSSTAB`, `OPPST_BENZO` and `OPPST_HYPNOTIKA` all have a p-value less than 0.05, indicating that the medication usage is different between the two studies. The remaining variables have a p-value that indicates that the medication usage is not different between the two studies. When analyzing the medication commencement in Chapter 5 the results from this analysis imply that we would expect the variable `from_study` to be an important covariate.

	Fisher's Exact Test			Logistic Regression		
Variable	OR	CI	p-value	OR	CI	p-value
INN_ANTIDEP	1.075	0.763-1.514	0.675	1.075	0.774-1.494	0.665
INN_SSRI	1.080	0.710-1.640	0.760	1.080	0.724-1.609	0.706
INN_ANTIPSYK	0.921	0.644-1.314	0.664	0.920	0.653-1.295	0.634
INN_STEMNINGSSTAB	0.938	0.580-1.512	0.818	0.938	0.594-1.474	0.782
INN_BENZO	0.916	0.528-1.577	0.795	0.916	0.545-1.527	0.736
INN_HYPNOTIKA	1.681	0.981-2.911	0.053	1.682	1.014-2.823	0.046
INN_OPIOIDER	3.245	1.285-9.257	0.008	3.250	1.408-8.413	0.009
OPPST_ANTIPSYK	1.785	1.222-2.622	0.002	1.787	1.244-2.583	0.002
OPPST_STEMNINGSSTAB	2.698	1.243-6.347	0.007	2.702	1.329-5.945	0.009
OPPST_BENZO	3.014	2.092-4.373	<0.001	3.020	2.129-4.317	<0.001
OPPST_HYPNOTIKA	1.553	1.044-2.323	0.028	1.554	1.064-2.284	0.023
OPPST_OPIOIDER	1.826	0.483-8.380	0.390	1.828	0.569-6.910	0.329

Table 25: Odds ratios, confidence intervals and p-values from Fisher's exact test and logistic regression.

APPENDIX

B

RESULTS FOR THE ANTIPSYCHOTICS AND HYPNOTICS DATASETS

B.1 Antipsychotics

We wish to perform the same inference as in 5.5 and 5.6 for the Antipsychotics dataset. This dataset resulted in a median AUC of 0.595 for EBM, 0.633 for GAM and 0.635 for GLM. All three models have similar AUC densities, though the GLM model seems to have lower outliers than the other two models.

B.1.1 Variable Importance

The variable importance for the 10 most important variables in the EBM model is extracted along with the corresponding average importance, standard deviation and inclusion count over the 1000 train-test splits. The results are presented in Table 26. All variables in the top ten are univariate variables, which makes the inclusion count redundant. The variable with the highest importance is the variable regarding recent substance abuse, with an importance of 0.058. This is in line with the corresponding average importance of 0.075, but significantly lower than the average importance of the diagnosis categories, which is 0.153. The variables regarding age and study also seem to be misrepresented in our model, as they also have a much higher average importance. The general low importance of the variables in the EBM model is in line with the low AUC of the model, which indicates that the model is not able to capture any clear underlying patterns in the data.

B.1.2 Shape Functions and Coefficients

The shape functions for the two continuous variables PANSS-EC and age are extracted from the EBM model. The shape functions are shown in Figure 21. The shape function for PANSS-EC score begins slightly negative and increases to a Score barely above 0 for PANSS-EC scores from 15 to 20. The Score then

Variable	Imp	Avg Imp	SD	Inclusion Count
substance_abuse_recent	0.058	0.075	0.039	1000
intake_suicide_assess	0.048	0.085	0.038	1000
prior_admit	0.044	0.084	0.041	1000
diagnosis_category	0.041	0.153	0.059	1000
age	0.030	0.102	0.056	1000
from_study	0.028	0.121	0.050	1000
suic_rel_for_ref	0.026	0.090	0.041	1000
referral_paragraph	0.023	0.076	0.035	1000
panss_ec_score	0.020	0.063	0.033	1000
specialist_paragraph	0.012	0.029	0.022	1000

Table 26: Variable importance (Imp), from EBM for our training set from the original train-test split, including average importance (Avg Imp), standard deviation (SD) and inclusion count from the 1000 train-test splits of the Antipsychotics dataset.

decreases to around -0.1 for PANSS-EC scores from 25 to 30. The uncertainty increases with the value of the PANSS-EC score, but is especially large compared to the other datasets. The shape function for age starts at 0 before dropping to a score of -0.05 for ages around 20. The Score then increases to 0.05 for ages 40 to 60 before decreasing to a score slightly negative for ages over 70. The uncertainty is very large for all ages except for ages 30 and 65.

For the categorical variables, the variable scores are extracted from the EBM model and shown in Table 15. For the binary variables, the Score_0 is the contribution to the systematic component of the model when the variable is equal to 0, and the Score_1 is the contribution when the variable is equal to 1. When it comes to the variable for diagnosis categories, the scores are the contribution to the systematic component of the model for each of the four categories.

The PANSS-EC score is ranked 9th in both importance and average importance in Table 26. Together with the low AUC of the EBM model, it is questionable whether reliable inferences can be made from the shape function for the PANSS-EC score.

B.1.3 GAM and GLM

The shape functions for the two continuous variables PANSS-EC and age are extracted from the GAM model and shown in Figure 22. The shape functions are similar to the ones from EBM seen in Figure 16, though the GAM model doesn't seem to catch the same patterns as EBM does as it is smoothed, as expected. The uncertainty seems to be more similar between the two models than what we have seen for the Benzodiazepines and Mood Stabilizers datasets, even though the uncertainty is still larger for the GAM model.

The p-values for the variables in the GLM and GAM models on the Antipsychotics dataset are extracted and sorted from most significant to least significant. The results are shown in Table 28 and Table 29. For the continuous variables in the

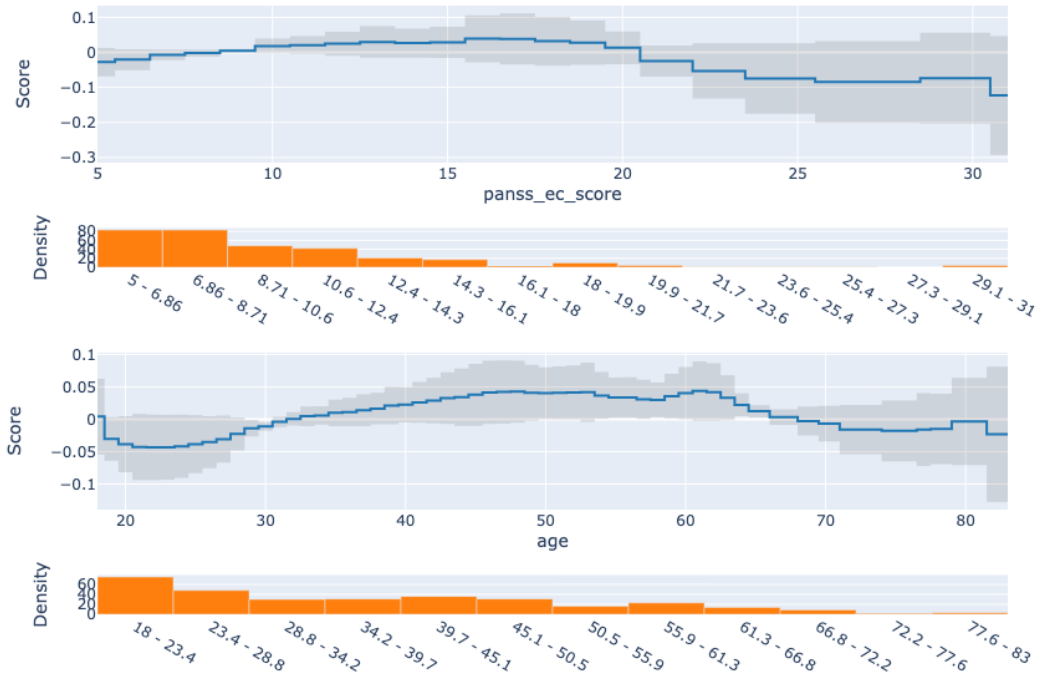


Figure 21: Shape functions for PANSS-EC score and age from EBM for Antipsychotics dataset.

GAM model, the p-value presented in the top table in Table 17 is the p-value from `summary.glm()`, while the p-value in the bottom table is the p-value from the ANOVA test. Comparing the results from the GLM and GAM models, we see that all variables have the same direction of the estimate, except for the variable for suicide attempt, which is ranked amongst the last anyway. The variable for the diagnosis categories is the most significant variable in both models. Most notably, the study variable is only ranked 6th as is not considered to be significant, which stands in contrast to the results seen in Appendix A, where the commencement of antipsychotics was found to be significantly different between the two studies. Comparing to the direction of the estimates from the EBM model, which is the difference between the estimates for Score_1 and Score_0 in Table 15, we see that the direction is the same for all variables except for the variable for suicide attempts recently, where the direction is the same as GAM. The ranking of variables between EBM and GAM/GLM is similar for many variables but differs enough with regard to both importance and average importance to conclude that they are considerably different.

Variable	Score_0	Score_1
from_study	-0.029	0.026
gender	0.002	-0.002
prior_admit	0.048	-0.040
suic_rel_for_ref	0.055	-0.017
intake_suicide_assess	-0.032	0.095
referral_paragraph	0.014	-0.060
specialist_paragraph	0.007	-0.053
suic_attempts_recent	-0.002	0.011
suic_thoughts_recent	-0.019	0.007
substance_abuse_recent	0.046	-0.081

Variable	Affective	Other	Psychosis	Substance Abuse
diagnosis_category	0.043	-0.074	0.031	0.002

Table 27: Scores for binary variables (top) and diagnosis category (bottom) from EBM for the Antipsychotics dataset.

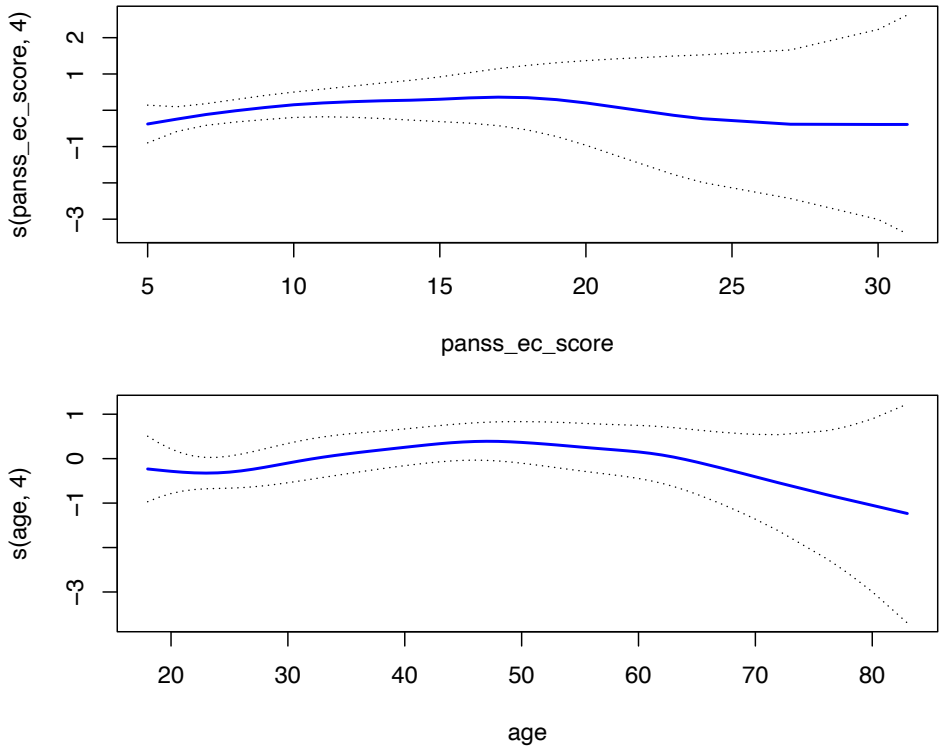


Figure 22: Shape functions for PANSS-EC score and age from GAM for Antipsychotics dataset.

	Estimate	Std. Error	z value	Pr(>|z|)
diagnosis_category_Other	-0.935	0.349	-2.681	0.00734
intake_suicide_assess	0.716	0.312	2.292	0.0219
substance_abuse_recentyes	-0.814	0.390	-2.089	0.0367
suic_rel_for_refyes	-0.798	0.417	-1.913	0.0558
prior_admityes	-0.389	0.274	-1.420	0.156
suic_thoughts_recentyes	0.554	0.413	1.342	0.18
from_studyGAP	0.343	0.273	1.254	0.21
panss_ec_score	0.039	0.034	1.162	0.245
(Intercept)	1.238	1.287	0.962	0.336
referral_paragraph	-0.547	0.625	-0.876	0.381
age	0.007	0.009	0.838	0.402
specialist_paragraph	-0.486	0.792	-0.614	0.539
diagnosis_category_Substance_Abuse	0.197	0.438	0.450	0.652
diagnosis_category_Psychosis	0.163	0.483	0.338	0.735
genderWoman	-0.074	0.280	-0.266	0.79
suic_attempts_recentyes	-0.044	0.365	-0.121	0.904

Table 28: Summary of GLM results for Antipsychotics dataset.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.911	1.298	0.702	0.483
diagnosis_category_Other	-0.903	0.352	-2.563	0.0104
intake_suicide_assess	0.733	0.315	2.327	0.0199
substance_abuse_recentyes	-0.857	0.392	-2.188	0.0286
suic_rel_for_refyes	-0.828	0.419	-1.978	0.0479
prior_admityes	-0.412	0.277	-1.488	0.137
from_studyGAP	0.385	0.276	1.395	0.163
suic_thoughts_recentyes	0.558	0.414	1.349	0.177
s(panss_ec_score, 4)	0.036	0.035	1.038	0.299
referral_paragraph	-0.593	0.625	-0.948	0.343
s(age, 4)	0.008	0.009	0.879	0.379
diagnosis_category_Psychosis	0.280	0.492	0.569	0.57
diagnosis_category_Substance_Abuse	0.165	0.438	0.377	0.706
specialist_paragraph	-0.292	0.787	-0.371	0.711
suic_attempts_recentyes	0.077	0.367	0.211	0.833
genderWoman	-0.057	0.284	-0.201	0.841

	Npar Df	Npar Chisq	P(Chi)
s(panss_ec_score, 4)	3.000	2.369	0.499
s(age, 4)	3.000	5.852	0.119

Table 29: Summary of GAM results (top) and ANOVA (bottom) from the Antipsychotics dataset.

B.2 Hypnotics

Following the procedure used in the other datasets, we wish to perform inference for the Hypnotics dataset. This dataset resulted in the lowest AUCs for all models, with a median AUC of below 0.6. The train-test split used in 13 and in our analysis resulted in the horrendous result of an AUC of 0.445 for the GAM model, while the EBM and GLM model had an AUC of 0.621 and 0.591, respectively.

B.2.1 Variable Importance

The variable importance for the ten most important variables in the EBM model is extracted along with the corresponding average importance, standard deviation and inclusion count over the 1000 train-test splits. The results are presented in Table 30. The variable importance is, in general, very low, which is in line with the low AUCs and poor performance of the model. The most essential variable in terms of importance and average importance is the variable regarding study, with an importance of 0.098. This is in agreement with the results from appendix A, where the commencement of hypnotics was found to be significantly different between the two studies. There is little difference in the importance of the other variables, indicating that the model is not able to capture any clear underlying patterns in the data. There are four interaction terms in the top ten, where only one can be considered to be consistently included in the model, which is the interaction between gender and PANSS-EC score included 929 times.

Variable	Imp	Avg Imp	SD	Inclusion Count
from_study	0.098	0.081	0.036	1000
intake_suicide_assess	0.060	0.027	0.019	1000
diagnosis_category	0.048	0.051	0.026	1000
suic_attempts_recent	0.047	0.020	0.016	1000
from_study&gender	0.045	0.054	0.021	135
intake_suicide_assess&diagnosis_category	0.039	0.056	0.021	564
age	0.038	0.064	0.035	1000
prior_admit	0.037	0.050	0.029	1000
gender&panss_ec_score	0.035	0.050	0.027	929
from_study&diagnosis_category	0.034	0.042	0.016	171

Table 30: Variable importance (Imp), from EBM for our training set from the original train-test split, including average importance (Avg Imp), standard deviation (SD) and inclusion count from the 1000 train-test splits of the Hypnotics dataset.

B.2.2 Shape Functions and Coefficients

The shape functions for the two continuous variables PANSS-EC and age are extracted from the EBM model. The shape functions are shown in Figure 23. The shape function for PANSS-EC score has a Score of 0 for PANSS-EC scores from 5 to 20, before decreasing to a Score of -0.5 for PANSS-EC scores around 30. The uncertainty is very low up to PANSS-EC scores of 20, where it drastically

increases as the PANSS-EC score increases. The shape function for age has a similar shape, with a Score of 0 for ages 20 to 55 before decreasing to a Score of -0.2 for ages around 80. The uncertainty is similar to the PANSS-EC score, with a low uncertainty for ages up to 55, before increasing as the age increases. The uncertainty for the patients of ages 18 to 20 is also considerably larger than for the other ages, but not close to as large as for ages above 55.

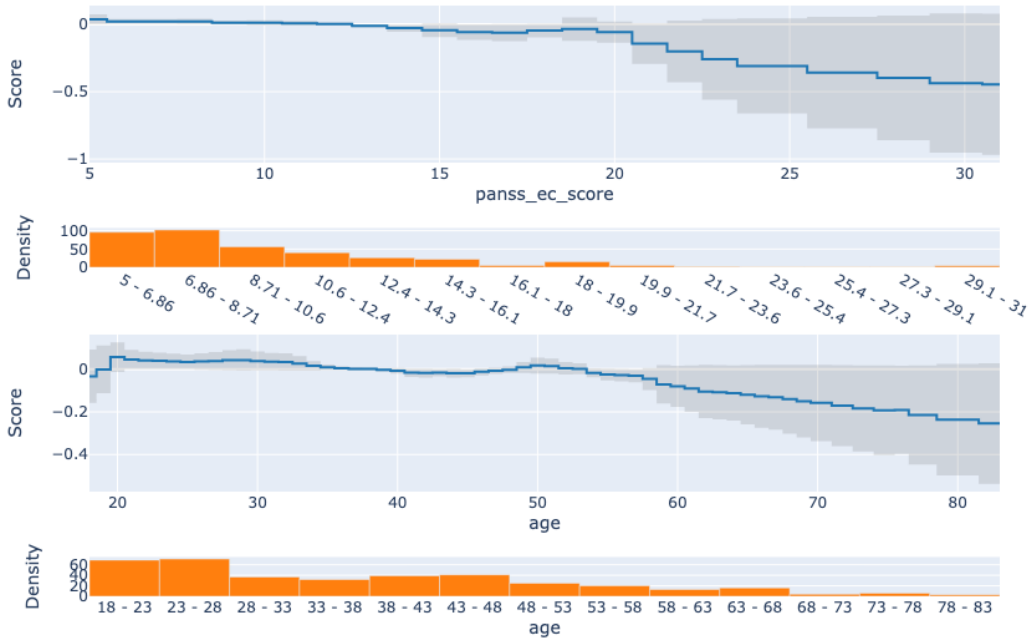


Figure 23: Shape functions for PANSS-EC score and age from EBM for Hypnotics dataset.

For the categorical variables, the variable scores are extracted from the EBM model and shown in Table 31. For the binary variables, the Score_0 is the contribution to the systematic component of the model when the variable is equal to 0 and the Score_1 is the contribution when the variable is equal to 1. When it comes to the variable for diagnosis category, the scores are the contribution to the systematic component of the model for each of the four categories.

The PANSS-EC score is not amongst the top 10 in Table 26. Together with the low performance of the model, it is questionable whether reliable inference can be made from the shape function for PANSS-EC score.

Variable	Score_0	Score_1
from_study	-0.100	0.096
gender	-0.021	0.021
prior_admit	0.047	-0.031
suic_rel_for_ref	0.035	-0.013
intake_suicide_assess	-0.041	0.113
referral_paragraph	0.008	-0.040
specialist_paragraph	0.011	-0.111
suic_attempts_recent	-0.029	0.132
suic_thoughts_recent	-0.022	0.008
substance_abuse_recent	-0.015	0.031

Variable	Affective	Other	Psychosis	Substance Abuse
diagnosis_category	0.067	-0.014	-0.043	-0.068

Table 31: Scores for binary variables (top) and diagnosis category (bottom) from EBM for the Hypnotics dataset.

B.2.3 GAM and GLM

The shape functions for the two continuous variables PANSS-EC and age are extracted from the GAM model and shown in Figure 24. The shape functions are similar to the ones from EBM seen in Figure 23, though the GAM model doesn't seem to catch the same patterns as EBM does as it is smoothed, as expected. The uncertainty seems to be more similar between the two models than what we have seen for the Benzodiazepines and Mood Stabilizers datasets, even though the uncertainty is still larger for the GAM model.

P-values for the variables in the GLM and GAM models on the Hypnotics dataset are extracted and sorted from lowest to highest value. The results are shown in Table 32 and Table 33. For the continuous variables in the GAM model, the *p*-value presented in the top table in Table 17 is the *p*-value from `summary.glm()`, while the *p*-value in the bottom table is the *p*-value from the ANOVA test. Comparing the results from the GLM and GAM models, we see that all variables have the same direction of the estimate and that there is little difference in the ranking of the variables. The variable for study is the most significant variable in both models, which is in agreement with the results from appendix A. Compared to the direction of the estimates from the EBM model, which is the difference between the estimates for Score_1 and Score_0 in Table 15, we see that the direction is the same for all variables. The ranking of variables between EBM and GAM/GLM is similar with regard to the univariate variables.

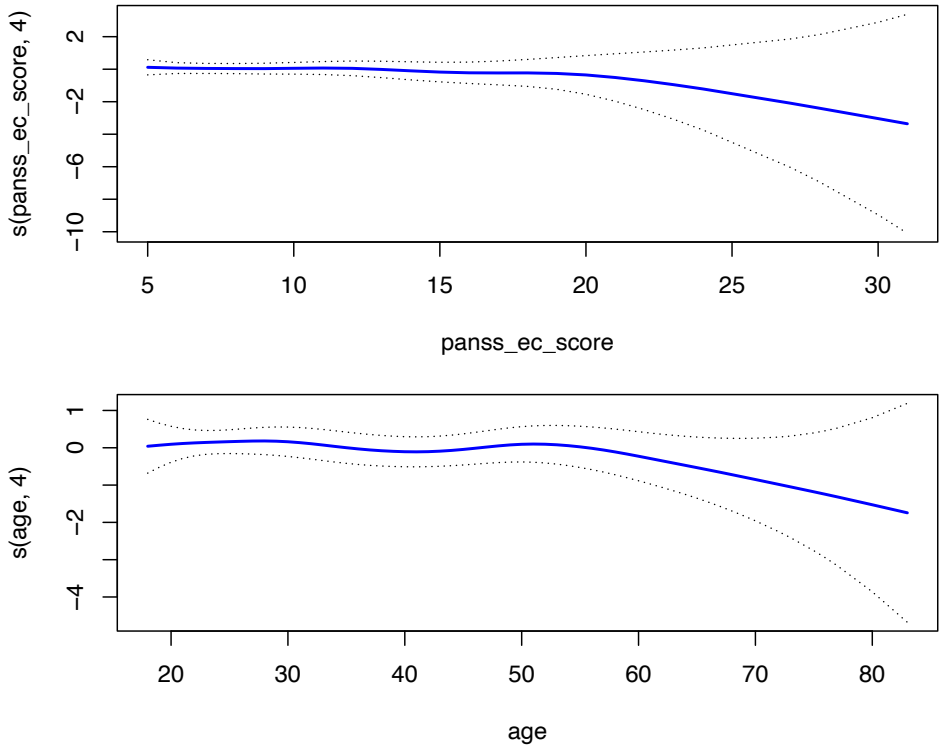


Figure 24: Shape functions for the variables for PANSS-EC score and age from GAM for Hypnotics dataset.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.771	1.416	0.545	0.586
from_studyGAP	0.550	0.253	2.170	0.030
suic_attempts_recentyes	0.599	0.321	1.868	0.0618
diagnosis_category_Substance_Abuse	-0.700	0.428	-1.636	0.102
intake_suicide_assess	0.395	0.273	1.449	0.147
diagnosis_category_Other	-0.423	0.303	-1.398	0.162
age	-0.012	0.009	-1.369	0.171
substance_abuse_recentyes	0.377	0.326	1.158	0.247
panss_ec_score	-0.035	0.033	-1.081	0.280
prior_admityes	-0.238	0.261	-0.913	0.361
suic_rel_for_refyes	-0.292	0.371	-0.787	0.431
specialist_paragraph	-0.527	0.761	-0.692	0.489
diagnosis_category_Psychosis	-0.248	0.408	-0.607	0.544
genderWoman	0.130	0.260	0.500	0.617
suic_thoughts_recentyes	0.116	0.384	0.301	0.764
referral_paragraph	-0.015	0.515	-0.029	0.977

Table 32: Summary of GLM results for the Hypnotics dataset.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.449	1.436	0.312	0.755
from_studyGAP	0.562	0.253	2.219	0.0265
suic_attempts_recentyes	0.619	0.322	1.924	0.0543
diagnosis_category_Substance_Abuse	-0.722	0.431	-1.673	0.0943
intake_suicide_assess	0.399	0.273	1.461	0.144
diagnosis_category_Other	-0.421	0.303	-1.392	0.164
s(age, 4)	-0.011	0.009	-1.230	0.219
substance_abuse_recentyes	0.386	0.329	1.175	0.240
prior_admityes	-0.249	0.262	-0.954	0.340
s(panss_ec_score, 4)	-0.030	0.035	-0.853	0.394
suic_rel_for_refyes	-0.265	0.367	-0.723	0.470
diagnosis_category_Psychosis	-0.256	0.410	-0.623	0.533
genderWoman	0.144	0.262	0.552	0.581
specialist_paragraph	-0.409	0.764	-0.535	0.592
suic_thoughts_recentyes	0.090	0.381	0.236	0.813
referral_paragraph	-0.019	0.515	-0.038	0.970

	Npar Df	Npar Chisq	P(Chi)
s(panss_ec_score, 4)	3.000	1.748	0.626
s(age, 4)	3.000	3.661	0.301

Table 33: Summary of GAM results (top) and ANOVA (bottom) from the Hypnotics dataset.

APPENDIX

C

EXAMPLES FROM SIMULATION STUDIES

In this appendix, the results from a different iteration of the simulation studies are presented for Simulation Study 1-4, than the one presented in Chapter 4. The results are seen in Figures 25 - 28. The resulting shape functions are similar to the ones seen in 4, but with less variance on the right edge of the shape functions, which might imply that outliers are present in the datasets used in 4.

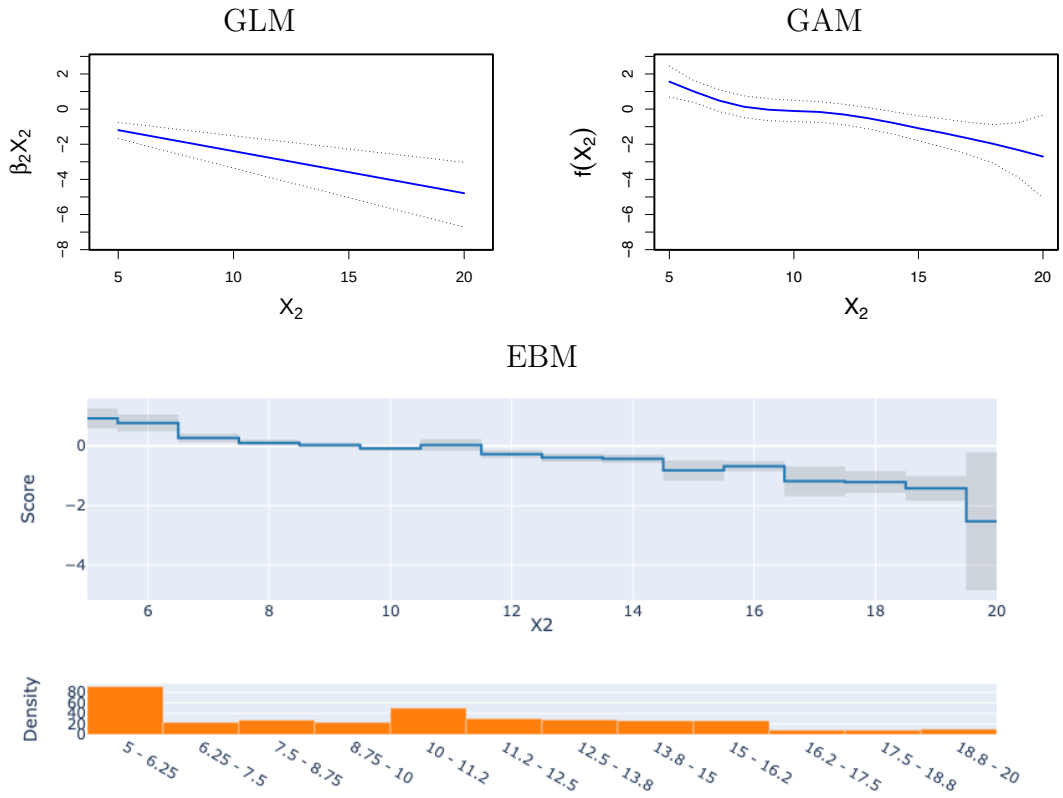


Figure 25: Comparison of the shape functions of GLM, GAM and EBM in Simulation Study 1 for another train-test split.

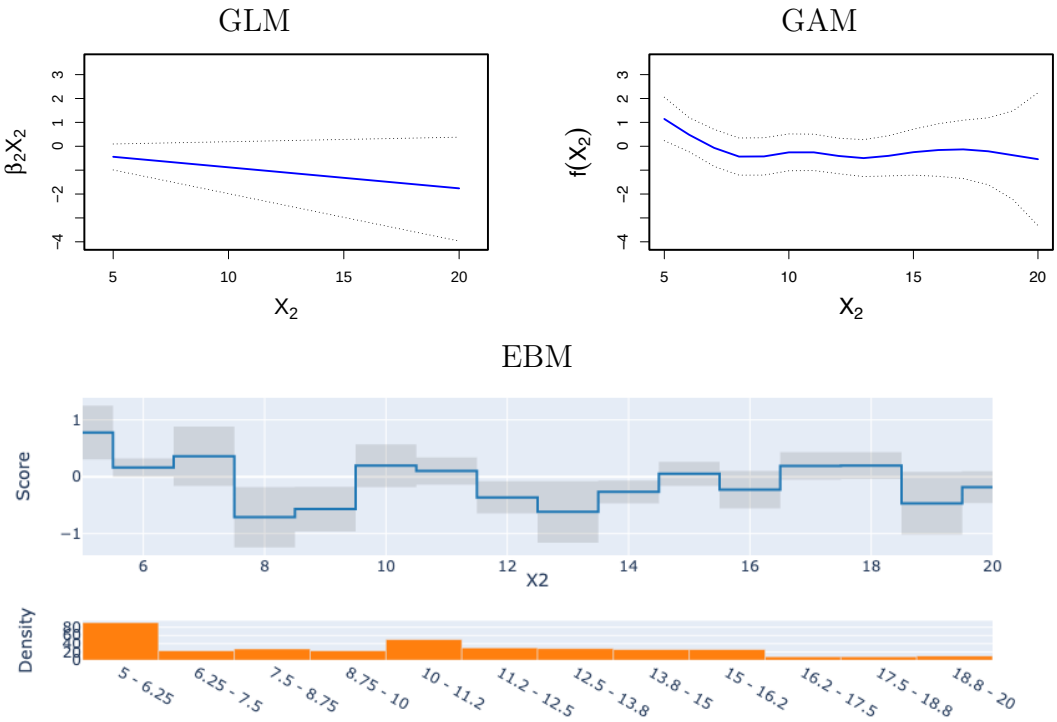


Figure 26: Comparison of the shape functions of GLM, GAM and EBM in Simulation Study 2 for another train-test split.

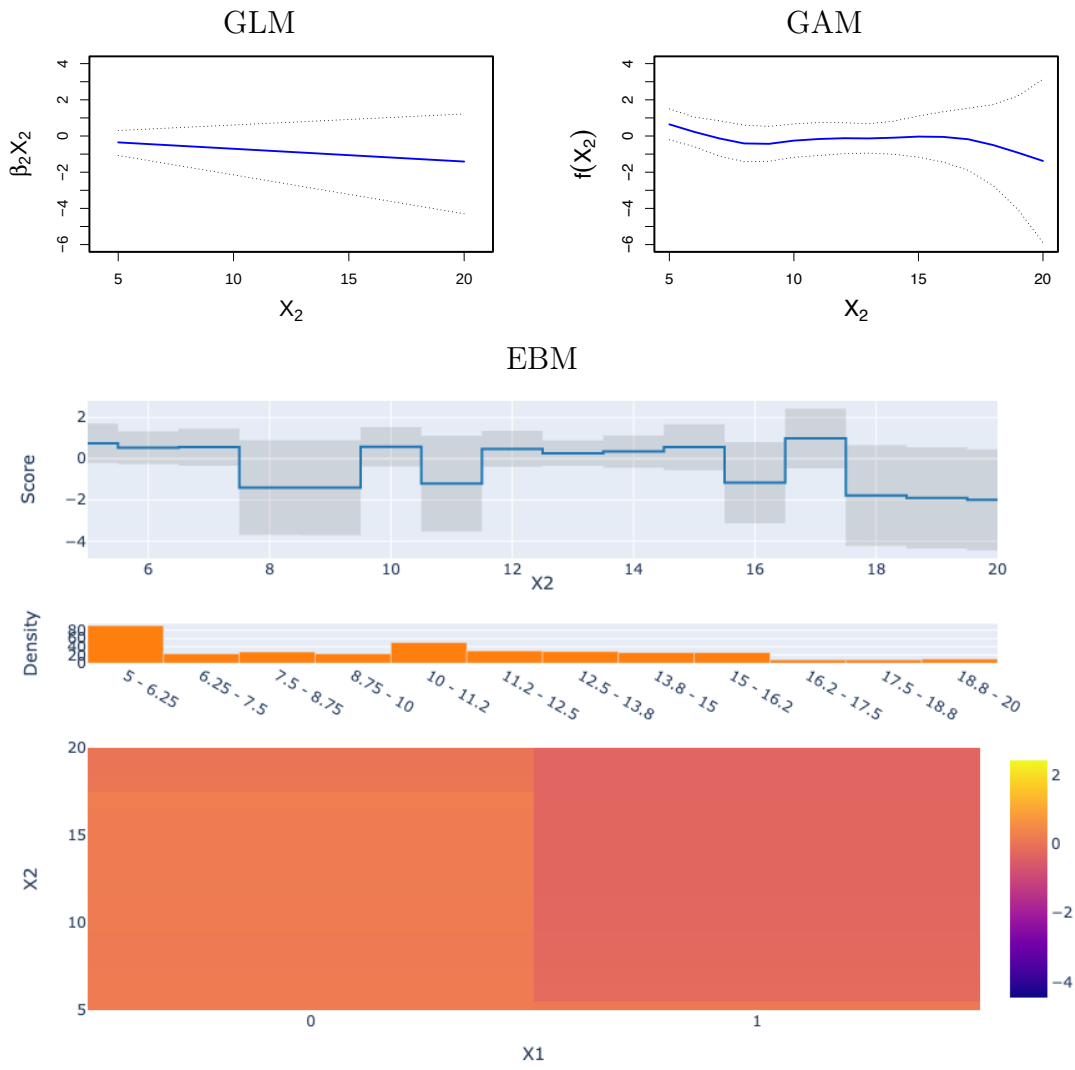


Figure 27: Comparison of the shape functions of GLM, GAM and EBM in Simulation Study 3, including the shape function for the interaction term from EBM, for another train-test split.

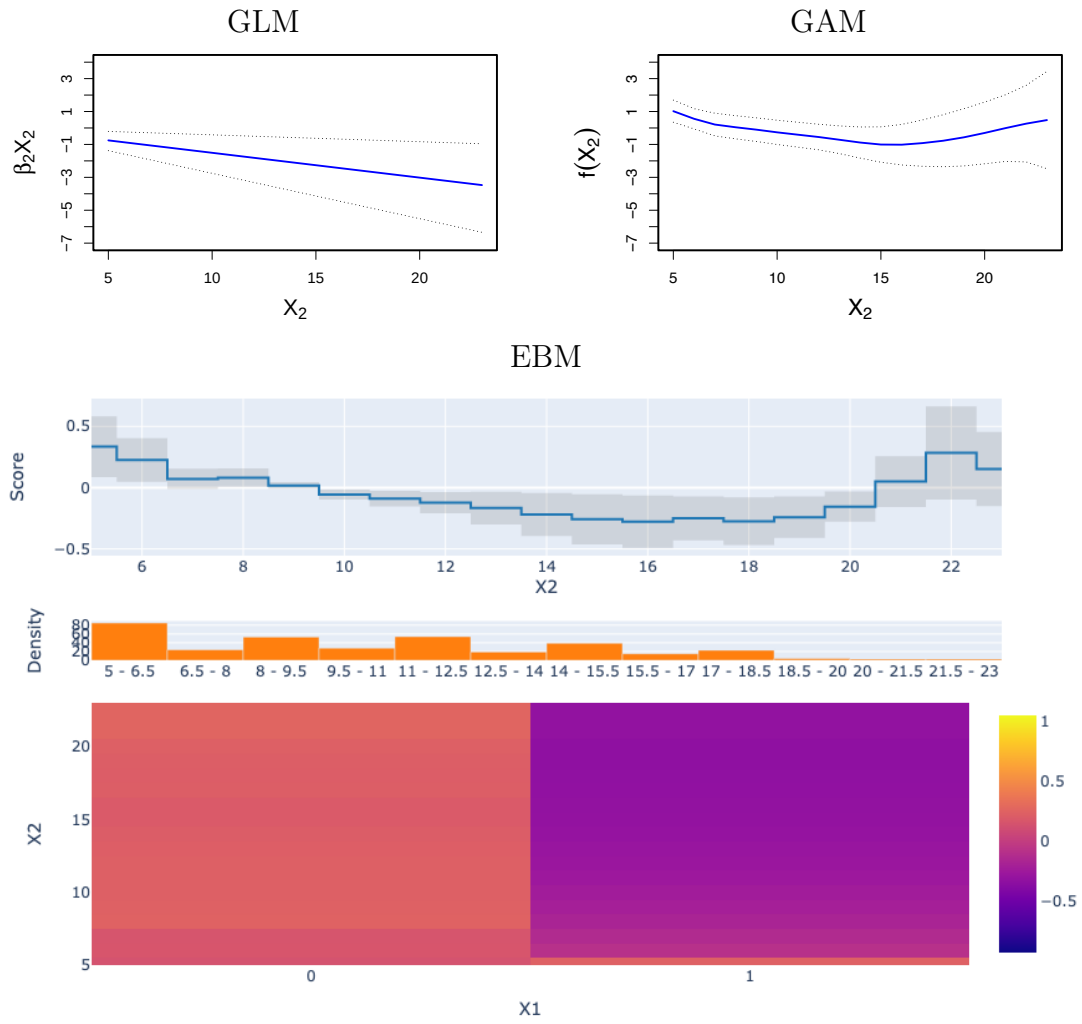
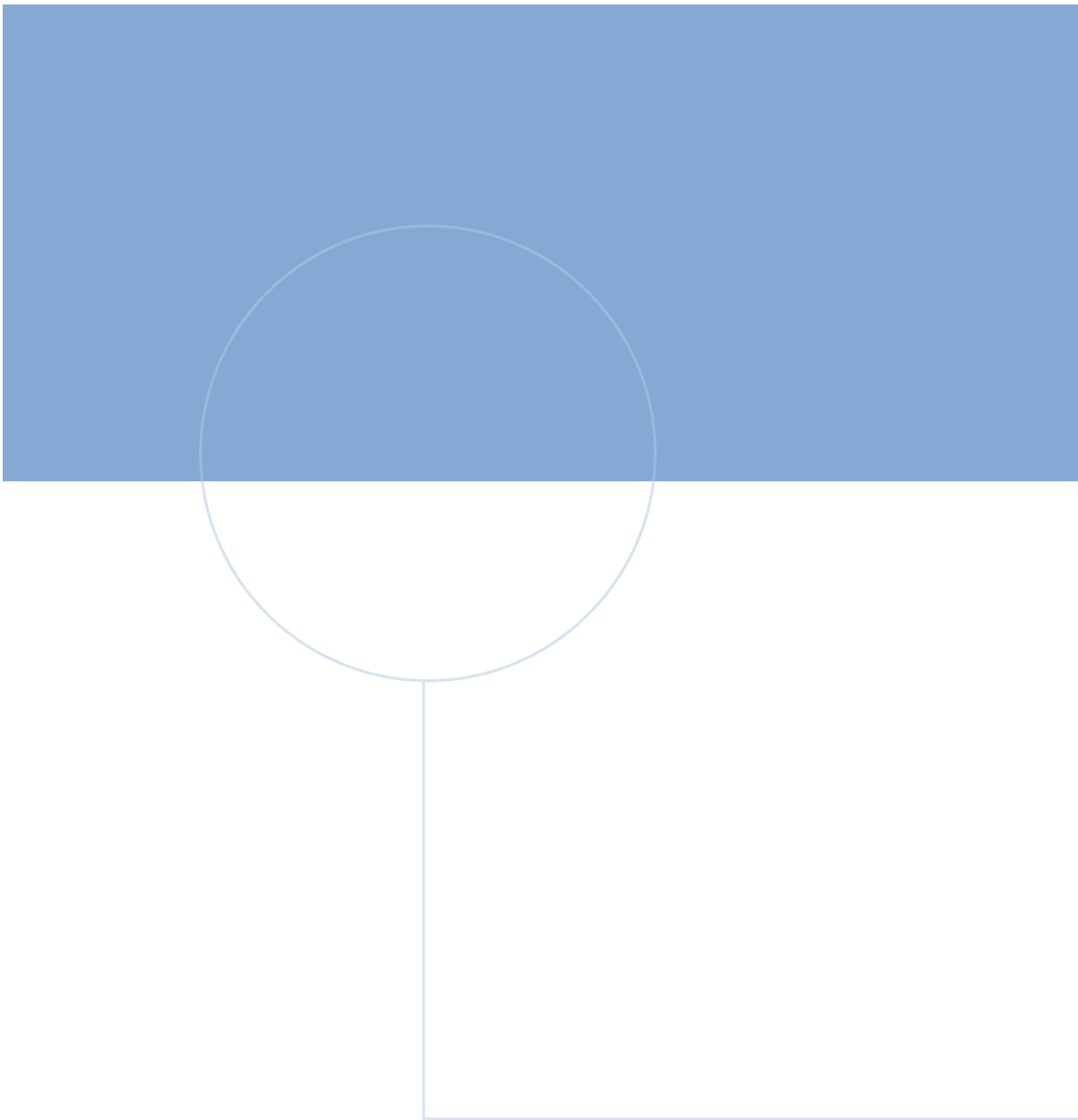


Figure 28: Comparison of the shape functions of GLM, GAM and EBM in Simulation Study 4, including the shape function for the interaction term from EBM, for another train-test split.



NTNU

Norwegian University of
Science and Technology