Business Intelligence in Automobile Retail Industry

A bachelor thesis written in the Business Analytics specialization in Business Administration at NTNU Business School. By students Malin Gundersen and Andreas Torkildsen Hjertaker.

Abstract

The automobile retail industry is facing more growth than ever before, and new technologies in the machine learning and business intelligence domains contribute to decision-making and automating industries worldwide. This thesis uses the CRISP-DM methodology and looks into a hypothetical case company, "Automo inc," to increase its EBITDA. We supply domain aspects of Automo based on an actual interview from a local retailer. To reach the objective, we will utilize the possibilities of Python and the machine learning libraries to create a price prediction model. To this extent, we look at a data set retrieved from Kaggle for academic purposes to demonstrate the possibility of gaining business value by automating the price prediction process. This thesis emphasizes the limitations of both the model and methodology, and it concludes with recommendations and future work for the industry.

Table of contents

List of Figures

Preface

We make up the team of two students for this group project, and both study Business Administration at NTNU Business School. It is exciting how machine learning algorithms may improve business in various industries. Therefore, at the beginning of the project, we were determining which data set to use for the assignment. After some research, we found out that we wanted to write about price prediction of used cars, and agreed that the data set we chose had the most relevant information for the model we would develop. The first thing we did was to make a plan for how we would work with the project, and we made a structure for the whole assignment.

After deciding which data set to use, we gathered and pre-processed the data. Then we worked out the problem definition. We started to write the background part and business case, while making the model in Python. After finishing all the preparations of the testing and training sets, we started testing linear regression but ended up using random forest regression, as this had the highest model accuracy. Meanwhile, we started to write about approach and data strategy, as well as business understanding. We discussed how used cars sell, the limitations of using this model, and the recommendations as we collectively interpreted the results. This is a preliminary statement introducing the assignment to explain its scope and intention. We hope you have a pleasant reading.

- *Malin Gundersen and Andreas Hjertaker*

1 Introduction

1.1 Background

Today the used car market for passenger cars in Norway is more prominent than ever. According to an article on rb.no, the market for used cars is much larger than for new cars. In 2020, there were 519,288 such changes of ownership, while there were only 141,412 new cars registered. The corona pandemic is probably one of the reasons for this, like traveling with cars in Norway, closed factories, and tax relief. At the same time, Jan Petter Røssevold, marketing manager at the Road Traffic Information Council in Norway, says that the number of ownership changes continued to increase, although the supply of new cars was more normalized in the second half of 2020. This is probably because a newfound interest in used cars already had been created, both among buyers and dealers. (Abrahamsen, 2021).

New cars have a more significant decline in value, which may be one of the reasons why more people choose to buy used cars. In recent years the used cars market has become huge, and there is another reason for that. The pandemic has increased prices for many used cars in recent years. Knut Skogstad, editor of TV2 (2021), reports that the used cars market has behaved abnormally since the corona pandemic hit us in March 2020. Increased demand for used cars has led to higher prices. Nothing at this point indicates a decline in demand for used cars. (Skogstad, 2021).

The growth in the used car market leads to greater competition for the sellers, and it can also be more challenging to set prices. To mislead the incorrect price of the car, we want to create a model that predicts the prices of used cars based on variables such as the brand, the year or edition of the model, kilometers-driven, and fuel type. We want to look at how the variables affect what price one should take for the specific car.

1.2 Problem Definition

As the background information suggests that the car retail industry experiences growth, our problem definition is: **“How can retail automobile companies leverage machine learning to predict car prices for increased EBITDA, and should this technology be utilized?”**

2 Approach and Data Strategy

As we worked with applied data science, we utilized the popular CRISP-DM methodology throughout our thesis. It adapts to the business needs and has a reasonable approach to big data. Furthermore, we have a hypothetical business case partly based on an interview in the car retail industry and use available data to prove our concept. The phases of the CRISP-DM concept are described in detail in the next subsections.

2.1 CRISP - DM

Data mining is a creative process that requires different skills and knowledge. Currently, there is no standard framework to carry out data mining projects. This indicates that the success or failure of a data mining project depends on the particular team carrying it out, and success can not necessarily be repeated across the enterprise. Data mining needs a standard approach that can help translate business problems into data mining tasks, suggest appropriate data transformations and data mining techniques, and provide the means for considering the success of the results and documenting the experience. The CRISP-DM (Cross Industry Standard Process for Data Mining) project 1 addressed these problems by defining a process model that provides a framework for carrying out data mining projects of both the industry and the technology used. The CRISP-DM process model does significant mining projects, cheaper, reliable, repeatable, faster, and more manageable. (Wirth & Hipp, 2000).

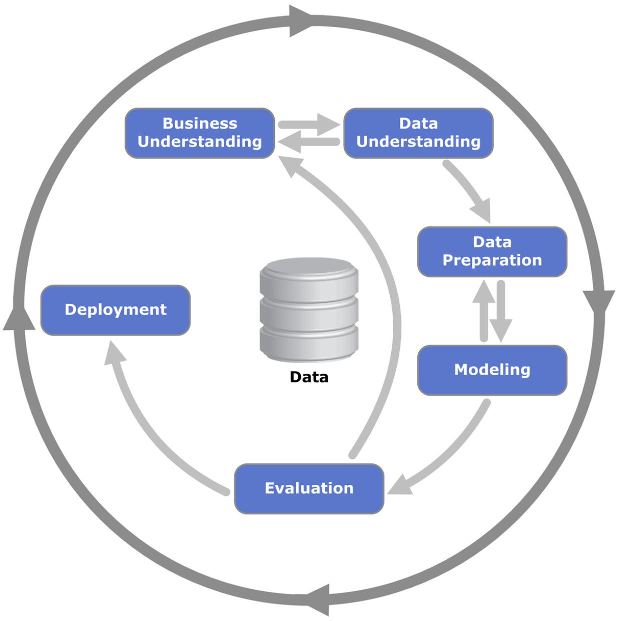


Figure 1: CRISP-DM.

2.1.1 Business Understanding

This initial phase focuses on understanding the project objectives and requirements from a business perspective and then converting this knowledge into a data mining problem definition and a preliminary project plan designed to achieve the objectives. (Wirth & Hipp, 2000).

2.1.2 Data Understanding

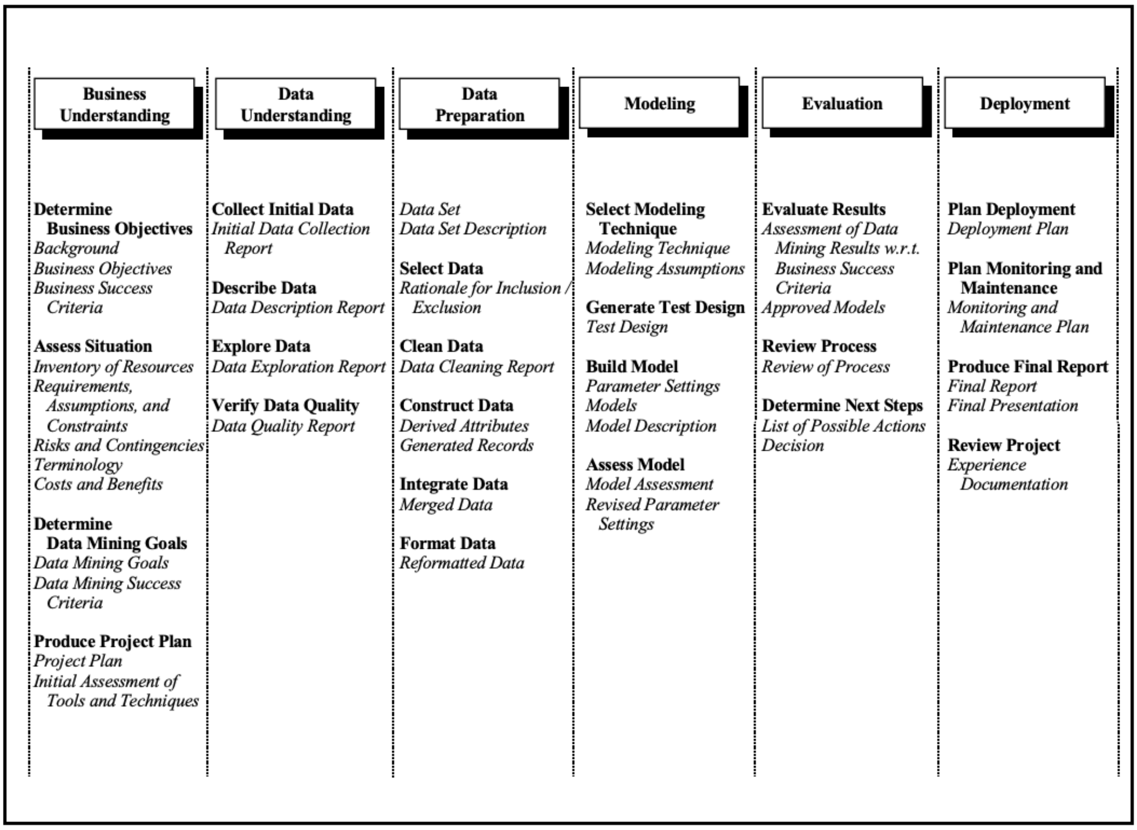
The data understanding phase starts with an initial data collection. It proceeds with activities to get familiar with the data, identify data quality problems, discover insights into the data, or detect interesting subsets to form hypotheses for confidential information. There is a close link between Business Understanding and Data Understanding. Formulating the data mining problem and the project plan requires at least some understanding of the available data. (Wirth & Hipp, 2000).

2.1.3 Data Preparation

The data preparation phase covers all activities to construct the final dataset from the initial raw data. Data preparation tasks are likely to be performed multiple times, not in any prescribed order. Tasks include table, record, and attribute selection, data cleaning, construction of new attributes, and data transformation for modeling tools. (Wirth & Hipp, 2000).

2.1.4 Modelling

In this phase, various modeling techniques are selected and applied, and their parameters get calibrated to optimal values. Typically, there are several techniques for the same data mining problem. Some techniques require specific data formats, and there is a close link between Data Preparation and Modeling. Often, one realizes data problems while modeling or gets ideas for constructing new data. (Wirth & Hipp, 2000).



Business Understanding

Determine Business Objectives
Background
Business Objectives
Business Success Criteria

Assess Situation
Inventory of Resources
Requirements, Assumptions, and Constraints
Risks and Contingencies
Terminology
Costs and Benefits

Determine Data Mining Goals
Data Mining Goals
Data Mining Success Criteria

Produce Project Plan
Project Plan
Initial Assessment of Tools and Techniques

Data Understanding

Collect Initial Data
Initial Data Collection Report

Describe Data
Data Description Report

Explore Data
Data Exploration Report

Verify Data Quality
Data Quality Report

Data Preparation

Data Set
Data Set Description

Select Data
Rationale for Inclusion / Exclusion

Clean Data
Data Cleaning Report

Construct Data
Derived Attributes
Generated Records

Integrate Data
Merged Data

Format Data
Reformatted Data

Modeling

Select Modeling Technique
Modeling Technique
Modeling Assumptions

Generate Test Design
Test Design

Build Model
Parameter Settings
Models
Model Description

Assess Model
Model Assessment
Revised Parameter Settings

Evaluation

Evaluate Results
Assessment of Data Mining Results w.r.t. Business Success Criteria
Approved Models

Review Process
Review of Process

Determine Next Steps
List of Possible Actions
Decision

Deployment

Plan Deployment
Deployment Plan

Plan Monitoring and Maintenance
Monitoring and Maintenance Plan

Produce Final Report
Final Report
Final Presentation

Review Project
Experience Documentation

Figure 2: Overview of the CRISP-DM tasks and their outputs.

2.1.5 Evaluation

The project team has built one or more models that appear to have high quality from a data analysis perspective at this stage in the project. Before proceeding to the final deployment of the model, it is essential to more thoroughly evaluate the model and review the steps executed to construct the model to be sure it properly achieves the business objectives. A key objective is to determine if the critical business issue was considered after all. At the end of this phase, one should use the data mining results. (Wirth & Hipp, 2000).

2.1.6 Deployment

The creation of the model is generally not the end of the project. Usually, the knowledge gained will need to be organized and presented so that the customer can use it. Depending on the requirements, the upcoming deployment phase can be as easy as generating a report or as complicated as implementing a repeatable data mining process. In many cases, the user, not the data analyst, will carry out the deployment steps. In any case, it is important to understand upfront what actions will need to be carried out to use the created models. (Wirth & Hipp, 2000).

2.2 Proof of Concept

Our goal for this project is to show how you can create a model that can predict the prices of used cars.

2.2.1 Hypothetical Case

We chose to solve this by creating a hypothetical case. The fictive company is Norwegian, and the dataset used to develop this model is based on Indian data. That means that the model we create will not be able to be used in reality, but it will be an example of how it can be developed.

2.2.2 Exemplified data

We use exemplified data to test how a price-estimating model can get developed using Python. The data is used as a placeholder for relevant data. We wanted to use a dataset with the most relevant variables possible and from which it was possible to create a model.

3. Business Case and Understanding

In our project, we will discuss a hypothetical automobile retail company that we will call "Automo Inc.". The case will hopefully increase readability and understanding throughout the assignment. Before we developed the business case, we had a phone interview with Kristian Rønning, team leader for used cars in a company in Trondheim called "Melhus Bil". We wanted the case to be as realistic as possible, so it has been partly based on this information.

Automo Inc. is a mature used car dealing company with ongoing operations for 18 years. This business case is about the store located in Trondheim, with additional locations in both Stjørdal and Oslo. The company in Trondheim is owned and driven by the CEO and administration and day-to-day management, different team leaders, sellers, and mechanics.

Their business model is a traditional automobile sales service for local customers in Trondheim. In 2020, the total turnover for Automo Inc. in Trondheim was 200.000.000 NOK. The store started as a smaller sole proprietorship in 2004, with only 40m^2 of facilities. Over the years, the concept has expanded to 60+ employees and 900m^2 over three operations locations. Automo Inc. has also developed custom digital solutions for its website and operations.

The used car market fluctuates at all times and for many reasons. The global situation concerning imports and emissions is an essential factor. Also, the demand for used cars is affected by the new car sale. Because of the pandemic in 2020, the new car sale decreased, and the demand for used cars increased. The increase is apparent in the company's turnover for 2020. All these factors make it challenging for Automo Inc. to set the optimal prices for the cars.

Automo Inc. has a CRM program based on a pricing tool with graphs, percentages, driveline graphs, estimated prices, and historical price statistics. Here they can see the average selling price of the car. The company mainly uses market fluctuations at Finn.no and their own experience to set the prices, but now they want an easier way. The company is growing and wants to grow even more. A crucial factor in reaching their goals is to set the correct prices. Therefore Automo Inc. wants us to make a model that can predict prices for used cars.

3.1 Objective

Automo Inc´s vision is to expand and proliferate with both velocity and direction as a company in growth. The company has an annual general assembly and established a goal for 2023 with a turnover of 250.000.000 NOK and EBITDA of 10.000.000 NOK. This goal demands growth in customers, marketing, and internal systems. The company has to adopt and facilitate a significant increase in supply and demand running through its operations. Compared to 2020, the initiative means increasing turnover by 25%. Increased EBITDA is vital because the stakeholders demand increased value proposition within the timeframe.

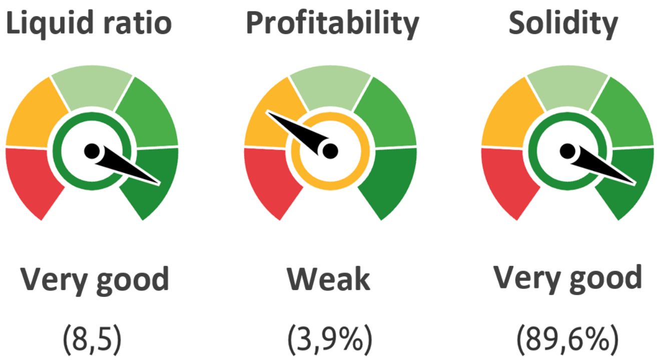


Figure 3: Financial summary of Automo Inc. 2020

Furthermore, Automo Inc. believes that increasing the turnover, in turn, will result in increased EBITDA. As of 2020, the EBITDA was 7.800.000 NOK. The business model facilitates increased profitability as the turnover increases because of the margin in used car dealing. The margin implies high variable costs but low fixed costs. It will take a lot of effort in different departments to reach their business initiative, and we will analyze their options and their implications. Applied data science might have insights for their direction.

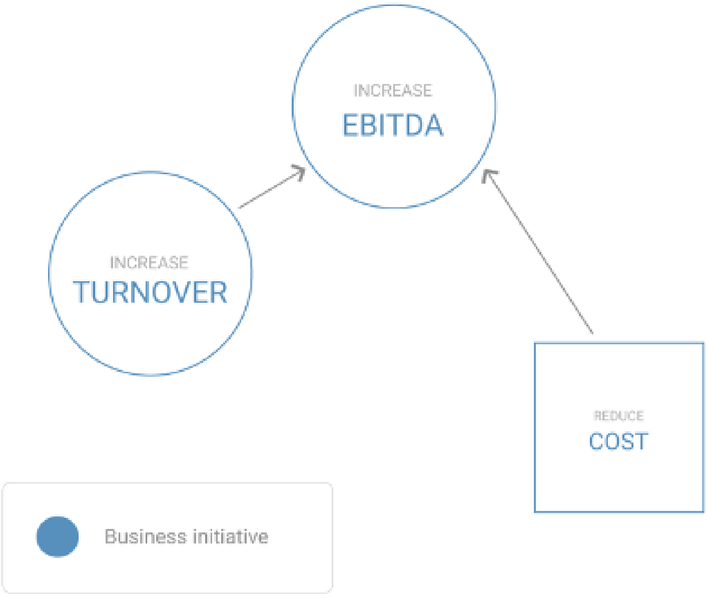


Figure 4: Business Initiative

Predicting the price of used cars is both a critical and exciting problem. According to data from our sources, the number of cars registered between 2003 and 2013 has experienced a significant increase of 234%. This number has now reached 160 701. With difficult economic conditions, second-hand imported cars and used cars will likely increase. We know that the sales of new cars have registered a decrease of 8% in 2013. (Pudaruth, 2014).

It is relatively common to lease a car rather than buy it outright in many developed countries. After the lease period, the buyer can buy the car at its residual value. Therefore, it is of commercial interest to sellers to predict the salvage value of cars with accuracy. If the value is under-estimated by the seller initially, the installments will be more significant for the clients, who will certainly then go for another seller. If the residual value is over-estimated, the installments will be lower for the clients, but then the seller may have much difficulty selling these high-priced used cars at this over-estimated residual value. (Pudaruth, 2014).

Therefore, we can see that estimating the price of used cars is of high commercial importance. German manufacturers lost 1 billion Euros in their USA market because of miscalculating the residual value of leased cars. Most people in Mauritius who buy new cars are also very apprehensive about the resale value after a certain number of years when they will possibly sell them in the used cars market. (Pudaruth, 2014).

3.2 Business Stakeholders

3.2.1 Customers

Customers are amongst the most important stakeholders because they trade value into the company. Therefore, reaching the business objective will imply bringing in more customers, increasing customer rebound rate, and increasing value with each purchase. All 25% of the increased turnover will need to result from increased customer activity. Automo Inc retails cars locally while having an in-house system and digital presence, being available to everyone interested at their website. The digital systems mean most activities get funneled through its systems. The system is then a hub for catching data from interaction behavior to demographics and preferences. We can also manage to differentiate the customers into B2B and B2C.

3.2.1.1 B2B

B2B is a more minor, more professional part of the company's customer base and is known for an increased need for tailoring and larger orders, and this drives a need for comprehensive manual work and capacity challenges.

3.2.1.2 B2C

The customer base is the primary driver of value but is more unpredictable and needy, especially regarding purchases, sales, customer service, and potential damage of received cars. These factors of manual work have to get managed with data-supported solutions.

3.2.2 Employees

Employees at Automo Inc are responsible for purchases and sales while delivering a service that satisfies the customer's needs. The service indicates the company's internal systems and price estimation administration, and a customer will not appreciate a faulty price estimate. The company also possesses personnel responsible for the administration of both accounting, marketing, customer service, and operations of the digital solution.

3.2.3 Partners

Today, several third parties are cooperating with Automo Inc. These range from mechanical services to premises for sales and marketplaces like Finn.no. These partnerships are essential to keep the functional quality that Automo Inc pursues and drive customers to the company in an affordable manner. They should get extended, more numerous, and dynamic given Automos business objective to drive more customer turnover.

3.2.4 Investors

Investors are essential in order to keep Automo Inc growing. They have invested money and work into the company, allowing it to grow. The investors have intentions to get a maximum return on their investment in the future. Today, Automo has three investors and equity of approximately 40.000.000 NOK.

3.2.5 Competitors

The used car rental market is competitive and exhausted by numerous competitors. The number of companies in the industry of trade-in retail has, with a minor exception in 2010, increased steadily since 2007. In 2015, there were 9,845 companies in this industry, 11.1 percent more than in 2007 (SSB, 2017).

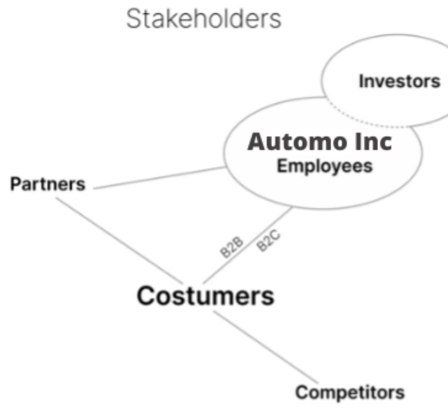


Figure 5: Automo Inc. stakeholders

3.3 Business Entities

3.3.1 Customers

Automo´s customers are also internet users, using Automo´s web services for information and ordering manual car estimates. Since Automo provides online services via web pages, data can get collected from the user in several ways. Upon request from an evaluation, some customers will provide information about themselves, such as demographics, age, or name, and share their experience with the company. Tracking the long-term loyalty of a customer is also helpful, as well as purchase history, payment method, the platform used, etc.

The level of customer satisfaction is vital to increase the rebound rate and increase the extent of verbal recommendations that drive more new customers. Customer satisfaction receives insights with Automo’s post-evaluation, which facilitates data collection for internal action and better experiences in the future.

3.3.2 Digital solution

The second entity of interest will get divided into interaction design and functions of the digital solution. Detailed data collection of these fields can turn into valuable insights that might help us make better business decisions. It should also get mentioned that the company's back-end systems include a database with valuable data such as sales and customer demographics.

3.3.2.1 Interaction design

When hosting a client-based service that represents the company to customers, keeping the interface user-friendly and gathering data on the site is essential. Data gathering might give insight into where to position elements or how car packages should get organized if a client fails to navigate, as examples. The optimization will have an impact on the churn rate of potential customers. Google Analytics might be a valuable integration that supports this type of data collection.

3.3.2.2 Automation, functions and integrations

Furthermore, it is crucial to keep a stable and sustainable digital solution that Automo can trust to be reliable. Data analysis regarding possibilities for automation and expansion of the autonomous system is vital for Automo´s competitiveness. Other data can be the quality of the platform with the function and service it supplies to both the customer and the administrative employees of the company. Analytics should ensure that the current on-site functions are valuable for customers, while parallel insights regarding new possibilities should get investigated. The functions will backbone its profitability as it relies heavily on its digital solutions and integrations.

3.3.3 Marketing

A primary driver for applied data science in this growth-oriented startup company should be the entity of marketing. The marketing orientation is because of the need for practical, targeted, and relatively affordable marketing, which fuels the expansion.

3.3.3.1 Traffic

To drive more traffic to the company's domain is a significant factor in receiving more new customers and reaching the business initiative. Automo accomplishes this with tools like Google Adwords, which track search terms, and techniques like on-site and off-site search engine optimization. Today, this optimization follows principles of try-and-error, while applied data science with collection and insight facilitates effective ways of reaching out to new customers.

3.3.3.2 Leads

Traffic leads us on to, well, leads! Because more relevant targeting methods will drive customers into the business domain. The content on-site must be relevant to drive more initiative to complete the sales process. The science behind handling leads while on-site resembles or mentions user-friendly interfaces, but the nature of this dimension grows further to what information the customer receives underway. A way of measuring this might be to collect data from different mixtures of information to create a perfect fit ultimately.

3.3.3.3 Sales

Other dimensions of marketing include sales with branding and rebound marketing. One can imagine that a strong brand generates more strong customer relations and influences the state of mind of their respective customers. A strong brand will get empowered with a clear identity and value proposition. The best way to do this is to investigate the most relevant target group and what drives value to those customers. This information can get gathered through a feedback system.

3.3.4 Employees

As a data-driven company interested in autonomous solutions, Automo should also collect data regarding their employee's needs, work hours, salary, and problems. A monthly evaluation might collect valuable data for insight and future action. The evaluation would also affect the employee's contentment and culture while on the clock.

3.3.5 Competitors

Collecting data from competitors and analyzing competitor reviews helps understand what Automo does better than the competition and improve. If a competitor has higher prices and still satisfies their customer base, then low price might not be the best strategy course.

Automo might also uncover patterns and answers of why a competitor performs in several areas, leading to increased insight for better business decisions.

3.3.6 Services

3.3.6.1 Main Sales and Purchases

Automo´s primary activities consist of different sales and purchases of used cars, and the delivery gets customized to the customer based on their needs. A customer can choose between smaller or larger cars, supporters, extra equipment, as examples. The sales database can keep track of order statistics, analyze the impact of different price strategies, and more. Car estimation quality will then possibly be optimized.

3.3.6.2 Add-ons

Today, many of Automos' pieces of equipment get offered as an optional add-on to customers. The mixture of available options with its information, price, type, and restrictions might also be analyzed through sales and interaction data while optimizing the mix.

3.3.7 Maintenance

Automo is a company that relies totally on its mechanics. The value drivers in the company are the functional cars and equipment delivered. That implies a potential for improvement as the cars should get estimated in terms of quality and status. A defect should be detected and funneled through an autonomous estimator. "Wear and tear" will demand restoration service from the company's employees and partners, and damaged equipment needs to be addressed and replaced. Measuring a product's lifecycle and detecting typical problems will increase the effectiveness of daily operations.

3.4 Company use cases with estimation software

A. Increase rebound rate

A machine learning-backed estimation increases trustworthiness amongst customers, which increases the possibility of later rebound. The rebound will significantly increase turnover as effort spent retaining an existing customer is significantly less than obtaining a new customer.

B. Reduce manual work per sale

As manual work is costly, automating more business areas would reduce work hours and future need for salary. Automation would significantly reduce costs and be a valuable case.

C. Increase website traffic and autonomous sales

Automo would need to gain more traffic to receive more new customers to expand as dramatically as the business initiative demonstrates. Traffic would drive more attention to the website and be valuable. The estimation software could help customers to order cars online.

D. Reduce marketing cost per acquisition

Marketing is one of the most significant cost factors, and reducing costs by implementing more effective and accurate marketing campaigns would be significant for the company's expansion. Estimation software allows for customized pricing in marketing efforts.

E. Increase prices

As more customers drive sales, slightly higher prices will multiply the company's total turnover. A correctly estimated price with a reasonable surcharge will drive income to Automo.

F. Expand with add-on options

More add-on options could grant better quality service in the form of customization and drive sales and turnover up because of the adjourning add-on to the overall price. However, it entails demands for the estimators' development and acquisition of new solutions and equipment.

G. Reduce extra maintenance cost risk per car

Less maintenance implies two things. Both less need for manual work for the company's employees and reduced risk of the unpredicted cost of restoration and replacement. A well made predictor increases the predictability of extra maintenance in acquisition.

The figure below illustrates interdependencies between the business initiative, entities, and use cases. It aims to point out some coherence between the objective and price predictor utilization.

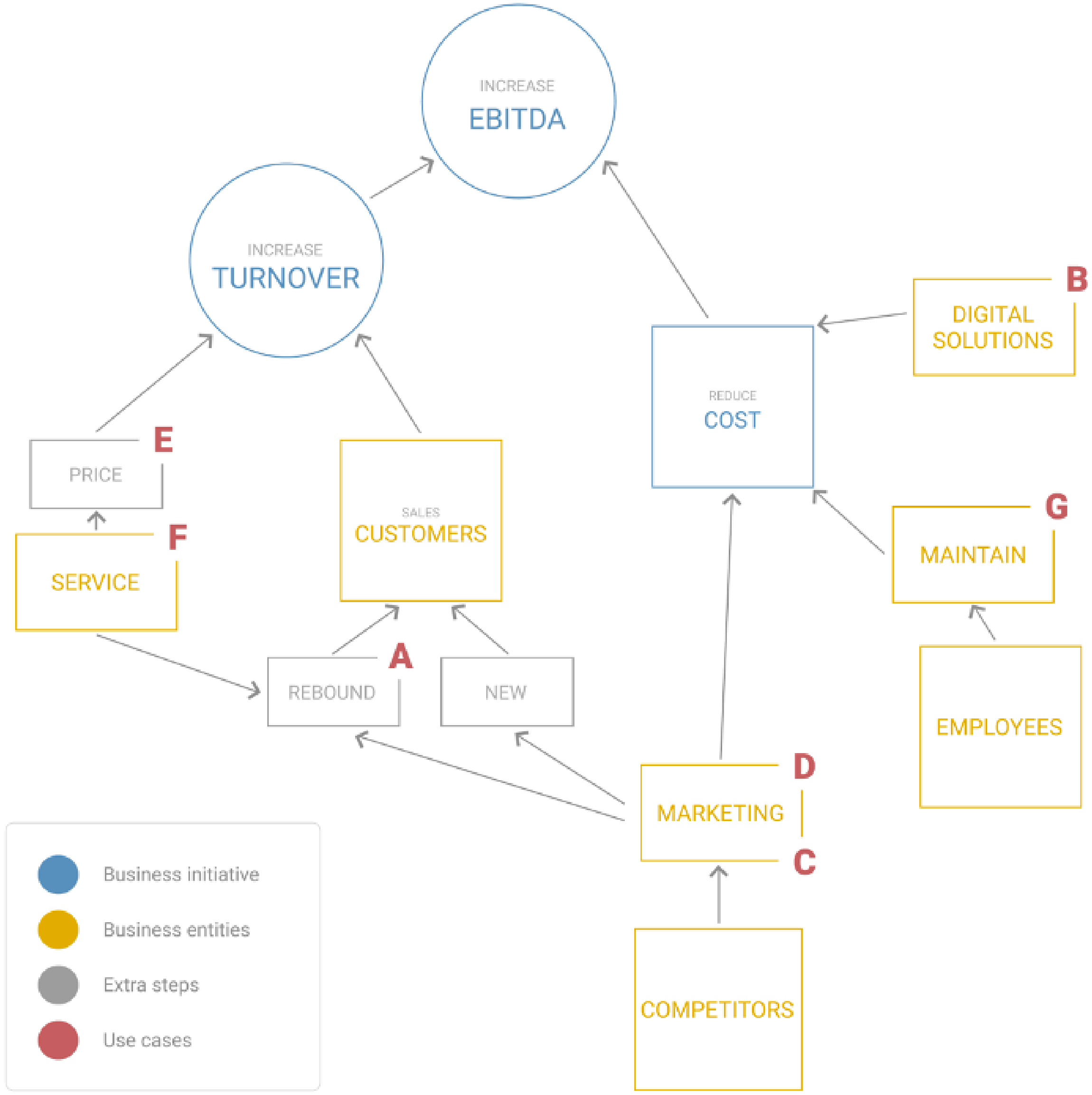


Figure 6: Interdependencies between the business initiative, entities, and use cases

4 Method

4.1 Data Understanding

We have received data from Kaggle to make a functional illustrative predictive model. This dataset is inaccurate because it does not contain actual data from our use case company but stands as an excellent example of how such an estimator can get created in practice. The dataset was published three years ago and got 182 upvotes and 78 creative contributors by the time of extraction. Furthermore, it has 14 columns and 7253 rows in total. The rows are already divided into training and testing as of extraction from Kaggle. (Kasliwal, 2019).

4.1. Why this dataset

As we were searching for opportunities for data backing our model creation, we noticed the hardship of acquiring relevant data from local retailers. Data is becoming a commodity of tremendous value for many domains, and this leads to a rapid increase in the number of data sources and public access data services, such as cloud-based data markets and data portals, that facilitate data collection, publishing, and trading. Data sources typically exhibit a wide variety and heterogeneity in the types or schemas of the data they provide, their quality, and their fees for accessing their data.

Users who want to build upon such publicly available data must first discover relevant sources then identify sources that collectively satisfy their applications' quality and budget requirements. This research gets done with only a few helpful clues about the quality of the sources and then repeatedly invests many person-hours in assessing the eventual use of data sources. All three steps require investigating the content of the sources manually, integrating them, and evaluating the actual benefit of the integration result for the desired application. Unfortunately, when the number of data sources is extensive, humans have little reasoning about the actual quality of sources and the tradeoffs between the benefits and costs of acquiring and integrating sources. This lack of reasoning is reflected in most of the local car retail companies in Trondheim today.

Systems enable the interactive exploration of different sources given a user's budget, allowing users to truly understand the quality and cost tradeoff between different integration options. We highlight the significant challenges in building such a system, such as supporting diverse integration tasks and multiple users, assessing the quality of data sources, and enabling the interactive exploration over different sets of sources. We believe that it is time for a new type of data portal that will allow data scientists and analysts to find the most valuable data sets for their tasks, even in the Norwegian car retail industry, and limit the person-hours spent in validating the quality of data. (Rekatsinas et al., 2015).

4.2 Data Set Attributes

The first step in building our predictive pricing model is to select the attributes that can explain the pricing of the given car. After listing each of the 14 attributes, each was then manually examined to see eligible attributes. Attributes that we want to investigate further are grouped into feature aspects to get a better overview of the different data types present in the listings data set. We further explored the different attributes by looking at the categorical and numerical distribution.

Car specifications	**Wear and tear**	**External factors**	**Price**
Name, Fuel type, Transmission, Mileage, Engine, Power, Seats	Year, Kilometers driven	Location, Owner type	New price, Price

Table 1: Attributes in data set

4.3 Data Preparation

The car retail data set must be processed before we can train the model on the data set. All the attributes were transformed into numerical values before being digested in the prediction model. Hence, all attributes got classified as either numerical or categorical features. The categorical features were hot-encoded, meaning they expanded into their numerical attributes describing whether or not the feature is present in a listing.

When looking closer at the distribution of price, we realized several issues. The target attribute price ranged from 0 - 160 000$, while most of the prices lay below 40$, indicating outliers of the price that could get removed to get a more accurate model.

4.3.1 Outliers of Attributes

There are a handful of outliers in the attributes of the cars. The distribution shown in the figure below shows that the kilometers driven for one car is 6.500.000, which is most likely a car that does not represent the specifications in which we are looking to predict prices. Extreme values like this might skew the model. The kilometers driven entry with several million kilometers, but it seems like the performance stayed intact after all.

A more thorough approach is to calculate outliers using “Z-Score” to find values that deviate too far automatically instead. Z-Score then quantifies the data outliers by the number of standard deviations above and below each value's mean. Nevertheless, one problem with this approach is that it relies on the normally distributed data, which leads us to our next problem.

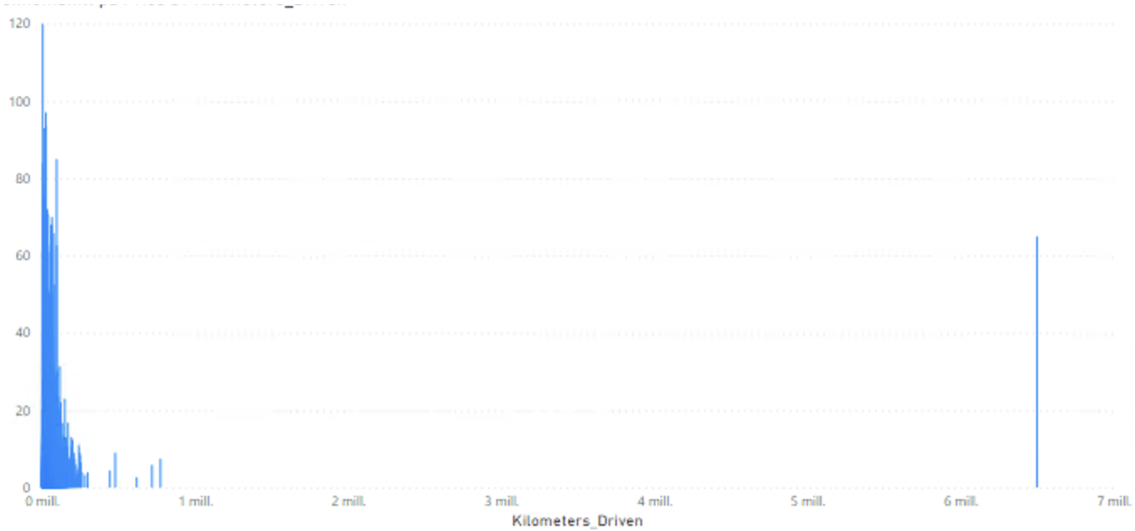


Figure 7: Plot of kilometers driven.

4.3.2 Skewness of Price

When looking at the price distribution figure below, it is apparent that the price distribution is not even. The price tends to skew in the leftmost direction, and the skew is then a problem as it reflects the prediction model's accuracy to predict price. The skew affects the precision of the model, and high values make the model skewed if they're not removed.

One way to reduce the skew is to transform the price by the logarithmic function. By doing this, the distribution curve would be more centered and evenly distributed, therefore increasing the prediction performance of our model. While this approach worked with the price, it did not reduce the skew sufficiently in the other numerical values and would not improve the initial results of the model.

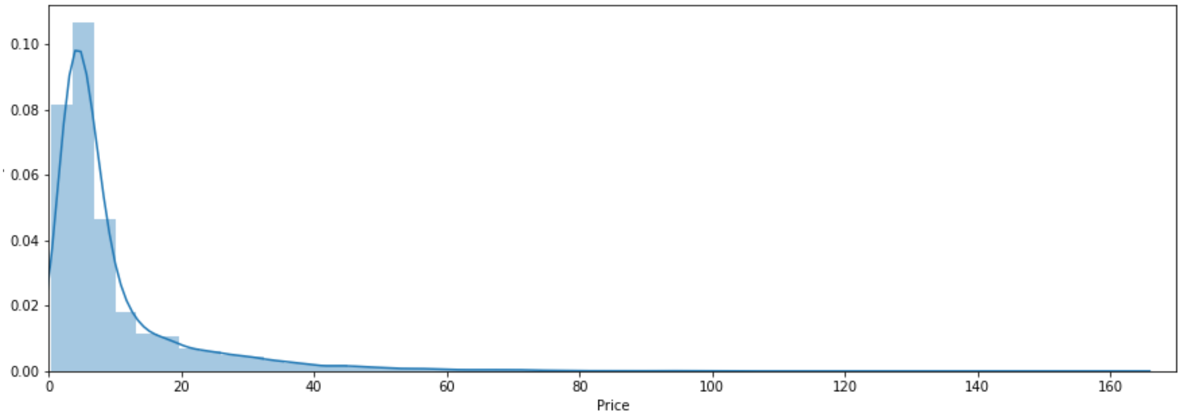
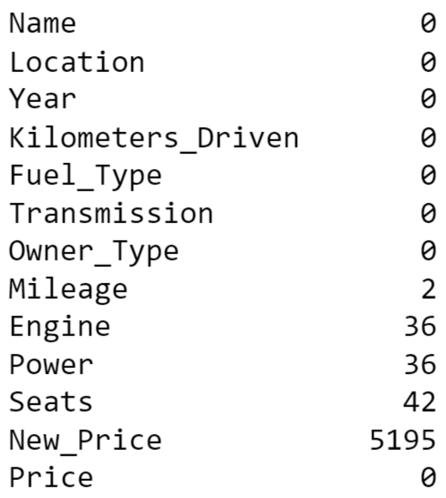


Figure 8: Plot of price

4.3.3 Null Values

Some attributes had a number of null values and could get replaced with the mean value of the said attribute or ignored. An example of this is the new price, which had mostly null values and was considered removed from the initial attribute selection process described above. The variable "New_Price" has 5195 null-values, but we chose to ignore this, because of the overall lack of variables. Also, we found it interesting to include this attribute for further analysis.

The other rows with attributes containing null values were dropped as part of the prepossessing of the data set, as these would contribute in a limited fashion and most likely only function as noise.



```
Name                    0
Location                0
Year                    0
Kilometers_Driven       0
Fuel_Type               0
Transmission            0
Owner_Type              0
Mileage                 2
Engine                 36
Power                  36
Seats                  42
New_Price            5195
Price                   0
```

Table 2: Attributes with null values

Because the new price had a considerable proportion of missing values, the attribute cannot be set with the mean value as it would decrease the model performance and, in general, be arbitrary to the corresponding actual rating. Another option is to leave the null values alone and use a training model that can handle null values such as SKLearn´s RandomForestRegressor as we did.

4.3.4 Encoding Categorical Attributes

When encoding categorical attributes to numerical representations, several possible paths are to take. We used label encoding if the data was in any order, and one-hot-encoding if it was not. First off, we used label encoding for Owner Type with values from 1-4. We initially decided to one-hot-encode the categorical attributes Location, Fuel and Transmission, meaning that each category within categorical attributes will expand into their attributes. The values of these new attributes are binary 1 or 0, indicating if the said attribute is present. One-hot encoding is generally a good approach, given that there are relatively few categories in the attributes. Because there were 30 companies in the data set, we chose to drop them for further work.

Another approach uses target encoding, where one uses the average target value for each category to represent each attribute value. Target encoding is suitable because it picks up values that explain the target price. For instance, each location value row in the data set would be encoded as the average price of all listings in that location. This encoding could potentially solve our attribute expansion seen in one-hot encoding. Although we will use hot encoding as a Proof of Concept, using target encoding could potentially get a better result.

4.4 Methods and Tools

4.4.1 Python, Pandas, and RFR

Pre-possessing and our prediction model got created using Python's robust programming language and the famous data-frame library Pandas. More data analysis libraries, such as Matplotlib and Seaborn, were used for visualization. While we extracted the machine learning model, or RandomForestRegressor, with SKLearn.

4.4.2 Power BI

Power BI is a tool primarily used for data visualization and analytics. We decided to use Power BI because of the dynamic data selection features. It allows a simple interface for viewing many data and visual objects. These objects consist of unique diagrams, matrices, and scripting blocks. We could easily connect, model, and visualize the data with this tool. (Microsoft, 2022).

4.4.3 Random Forest Regression

Random Forest Regression is a supervised learning algorithm that uses the ensemble learning method for regression. The ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model(Bakshi, 2020).

4.4.3.1 Ordinary decision trees

Decision trees are sensitive to specific data when they are trained. If the training data is changes, the resulting decision tree can get a different result, and the predictions can be quite different. Decision trees are computationally expensive to train, carries a significant risk of overfitting, and tend to find local optima because they cannot go back after they have done a split. To address the weaknesses, we turn to Random Forest which combines many decision trees into one model (Chakure, 2019).

4.4.3.2 Pros with Random Forest Regression

Random Forest Regression operates by constructing many decision trees at training time and outputs the class that is the mode of the classes, or mean prediction of the individual trees. A random forest is a meta-estimator as it combines the result of multiple predictions that aggregate many decision trees with some helpful modifications. The number of features that can get split at each node is limited to some percentage of the total. The limit ensures that the ensemble model does not rely heavily on any individual feature and makes good use of all potentially predictive features. When generating its splits, each tree draws a random sample from the original data set, adding a randomness element that prevents overfitting. The above modifications help prevent the trees from being highly correlated. A random forest classifier will handle the missing values and maintain the accuracy of a large proportion of data (Chakure, 2019).

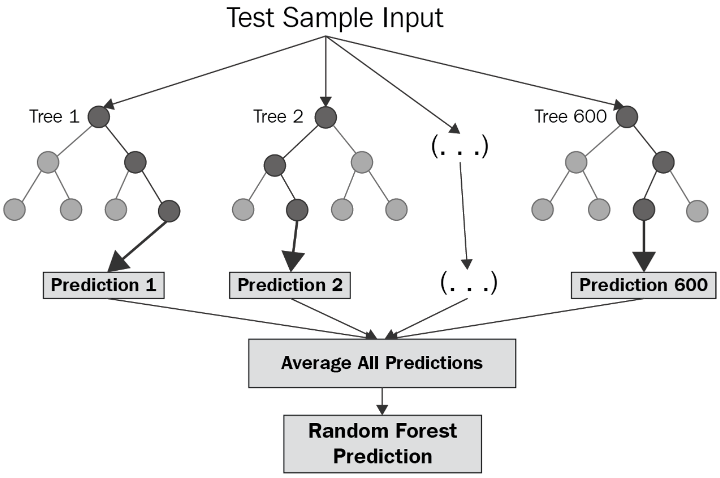


Figure 9: Random Forest Structure.

4.4.3.3 Shortcomings with Random Forest Regression

The prediction accuracy on complex problems is usually inferior to gradient-boosted trees. A forest is less interpretable than a single decision tree, and single trees may get visualized as a sequence of decisions.

Furthermore, the range of linear regression predictions in a Random Forest can make is bound by the highest and lowest labels in the training data and becomes problematic when the training and prediction inputs differ in their range. However, we chose to utilize the algorithm to get an accurate prediction (Thompson, 2019).

5 Analysis

This section will analyze the different numerical and categorical attributes of the data set and see how they correlate with the price in car sale entries.

5.1 Price and Numerical attributes

5.1.1 Attribute correlation

It is essential to get a sense of how the different attributes correlate to the price and research this early to get feedback on our initial hypothesis. Because of this, we point out one attribute that linearly correlates with price with the most certainty. The choice landed on the car's age, and then a plot investigating this guess was made.

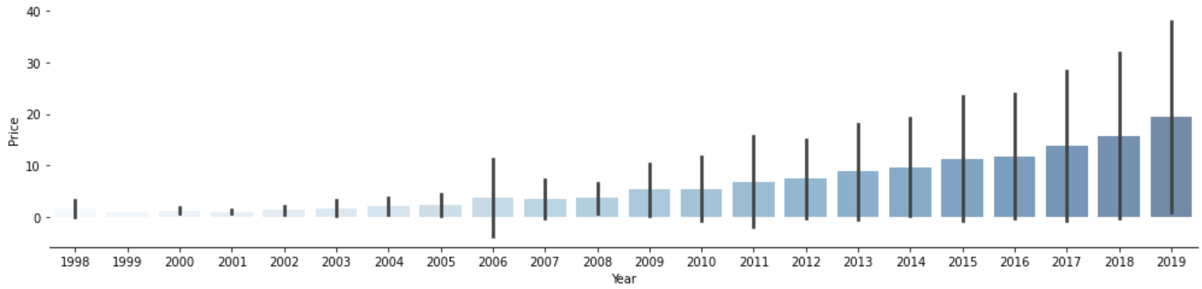


Figure 10: Barplot for year vs. price.

As shown by the plot, the car's age correlates with the price for this data set. An initial hypothesis was a high correlation between many of the original numerical attributes of the car with the price; Year, Kilometers driven, Owner type, seats, mileage, engine size, power, and new car price in the data set. To analyze this hypothesis, we plotted the attributes in a correlation heatmap.

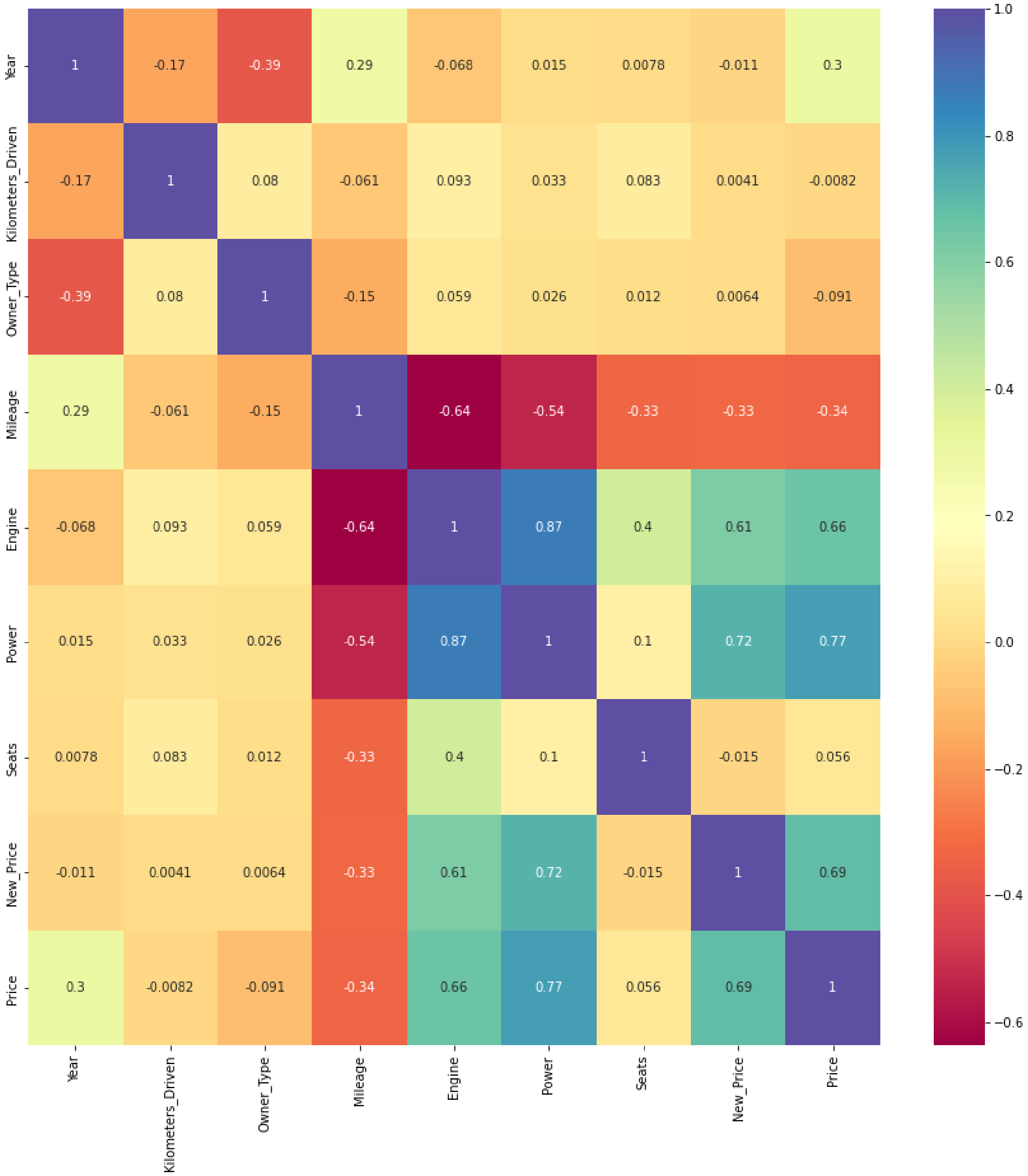


Figure 11: Correlation Heatmap of the original train data.

However, as we can see in the figure, this does not seem to be the case for the kilometers driven. The reason is that kilometers driven represent a vastly more significant number than, e.g., Year, which makes this visualization not that useful for that attribute.

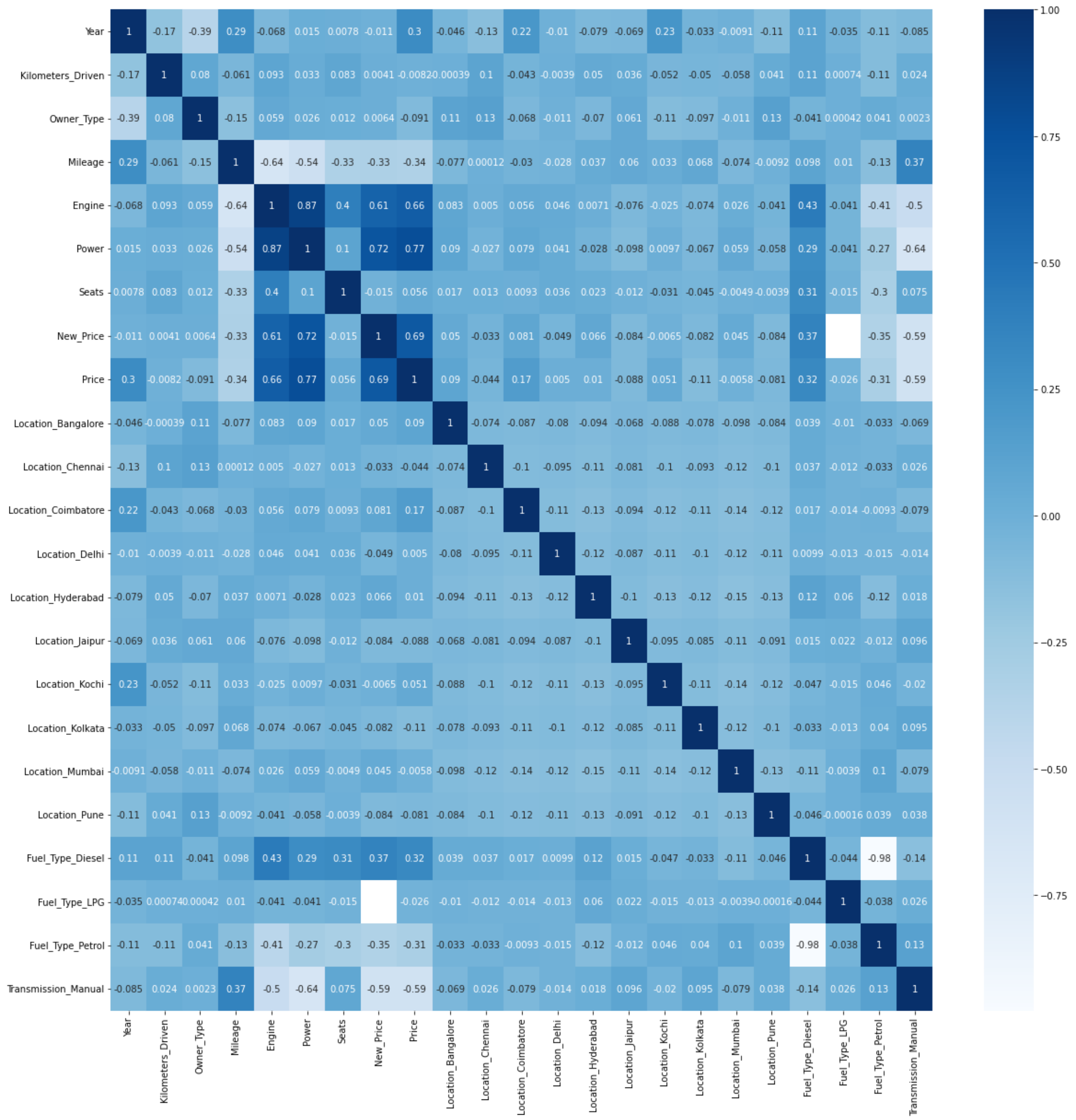
When looking at the figure, we can see that Mileage, Engine, Power, and new price are highly correlated to price and highly correlated to each other. The correlation makes sense as these attributes impose constraints on each other. When a larger engine, it also usually accommodates more power, while the fuel usage is generally higher. Larger and more powerful engines also drive new car prices as it takes more resources to build. Another note is that things are generally preferred when they are more spacious in many cases. Though interestingly, seats have a low correlation with the price of cars.

However, these highly correlated attributes can negatively influence the regressor model as highly correlated attributes can represent the same feature. When modeling highly correlated attributes in a model, the machine learning model does not necessarily recognize their similarity and the general feature they represent. That will cause it to get over-represented in the model causing high prediction volatility due to a bias towards engine size. This problem is referred to as multicollinearity. (Wikipedia, 2021).

5.2 Price and Categorical attributes

Another hypothesis was that the location and the price are correlated and influence each other. The result gets shown in the figure below. This figure shows that the listings prices correlate with location, and locations negatively correlate with each other. That makes sense because having a location works exclusively against all other locations. Then it makes sense that all correlations were about -0,1 when we got 11 locations. Furthermore, we can see in the figure that manual transmission correlates quite strongly negatively to the cars price. The fuel types also seem to matter in this data set.

Equipment is a set of different commodities present in a natural car listing. We hypothesize that the different equipment present in a car can drive the price upwards. The issue is that there is no regard to this in our data set, as there are many different equipment options and a few important ones. We can also imagine specific clustered equipment packages in real life. To avoid over-representation and multicollinearity, only one feature per such hypothetic cluster should get included.



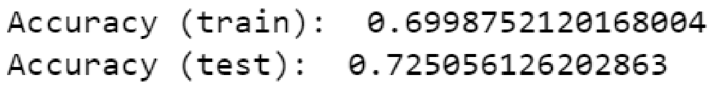
	Year	Kilometers_Driven	Owner_Type	Mileage	Engine	Power	Seats	New_Price	Price	Location_Bangalore	Location_Chennai	Location_Coimbatore	Location_Delhi	Location_Hyderabad	Location_Jaipur	Location_Kochi	Location_Kolkata	Location_Mumbai	Location_Pune	Fuel_Type_Diesel	Fuel_Type_LPG	Fuel_Type_Petrol	Transmission_Manual
Year	1	-0.17	-0.39	0.29	-0.068	0.015	0.0078	-0.011	0.3	-0.046	-0.13	0.22	-0.01	-0.079	-0.069	0.23	-0.033	-0.0091	-0.11	0.11	-0.035	-0.11	-0.085
Kilometers_Driven	-0.17	1	0.08	-0.061	0.093	0.033	0.083	0.0041	-0.0082	-0.00039	0.1	-0.043	-0.0039	0.05	0.036	-0.052	-0.05	-0.058	0.041	0.11	0.00074	-0.11	0.024
Owner_Type	-0.39	0.08	1	-0.15	0.059	0.026	0.012	0.0064	-0.091	0.11	0.13	-0.068	-0.011	-0.07	0.061	-0.11	-0.097	-0.011	0.13	-0.041	0.00042	0.041	0.0023
Mileage	0.29	-0.061	-0.15	1	-0.64	-0.54	-0.33	-0.33	-0.34	-0.077	0.00012	-0.03	-0.028	0.037	0.06	0.033	0.068	-0.074	-0.0092	0.098	0.01	-0.13	0.37
Engine	-0.068	0.093	0.059	-0.64	1	0.87	0.4	0.61	0.66	0.083	0.005	0.056	0.046	0.0071	-0.076	-0.025	-0.074	0.026	-0.041	0.43	-0.041	-0.41	-0.5
Power	0.015	0.033	0.026	-0.54	0.87	1	0.1	0.72	0.77	0.09	-0.027	0.079	0.041	-0.028	-0.098	0.0097	-0.067	0.059	-0.058	0.29	-0.041	-0.27	-0.64
Seats	0.0078	0.083	0.012	-0.33	0.4	0.1	1	-0.015	0.056	0.017	0.013	0.0093	0.036	0.023	-0.012	-0.031	-0.045	-0.0049	-0.0039	0.31	-0.015	-0.3	0.075
New_Price	-0.011	0.0041	0.0064	-0.33	0.61	0.72	-0.015	1	0.69	0.05	-0.033	0.081	-0.049	0.066	-0.084	-0.0065	-0.082	0.045	-0.084	0.37		-0.35	-0.59
Price	0.3	-0.0082	-0.091	-0.34	0.66	0.77	0.056	0.69	1	0.09	-0.044	0.17	0.005	0.01	-0.088	0.051	-0.11	-0.0058	-0.081	0.32	-0.026	-0.31	-0.59
Location_Bangalore	-0.046	-0.00039	0.11	-0.077	0.083	0.09	0.017	0.05	0.09	1	-0.074	-0.087	-0.08	-0.094	-0.068	-0.088	-0.078	-0.098	-0.084	0.039	-0.01	-0.033	-0.069
Location_Chennai	-0.13	0.1	0.13	0.00012	0.005	-0.027	0.013	-0.033	-0.044	-0.074	1	-0.1	-0.095	-0.11	-0.081	-0.1	-0.093	-0.12	-0.1	0.037	-0.012	-0.033	0.026
Location_Coimbatore	0.22	-0.043	-0.068	-0.03	0.056	0.079	0.0093	0.081	0.17	-0.087	-0.1	1	-0.11	-0.13	-0.094	-0.12	-0.11	-0.14	-0.12	0.017	-0.014	-0.0093	-0.079
Location_Delhi	-0.01	-0.0039	-0.011	-0.028	0.046	0.041	0.036	-0.049	0.005	-0.08	-0.095	-0.11	1	-0.12	-0.087	-0.11	-0.1	-0.12	-0.11	0.0099	-0.013	-0.015	-0.014
Location_Hyderabad	-0.079	0.05	-0.07	0.037	0.0071	-0.028	0.023	0.066	0.01	-0.094	-0.11	-0.13	-0.12	1	-0.1	-0.13	-0.12	-0.15	-0.13	0.12	0.06	-0.12	0.018
Location_Jaipur	-0.069	0.036	0.061	0.06	-0.076	-0.098	-0.012	-0.084	-0.088	-0.068	-0.081	-0.094	-0.087	-0.1	1	-0.095	-0.085	-0.11	-0.091	0.015	0.022	-0.012	0.096
Location_Kochi	0.23	-0.052	-0.11	0.033	-0.025	0.0097	-0.031	-0.0065	0.051	-0.088	-0.1	-0.12	-0.11	-0.13	-0.095	1	-0.11	-0.14	-0.12	-0.047	-0.015	0.046	-0.02
Location_Kolkata	-0.033	-0.05	-0.097	0.068	-0.074	-0.067	-0.045	-0.082	-0.11	-0.078	-0.093	-0.11	-0.1	-0.12	-0.085	-0.11	1	-0.12	-0.1	-0.033	-0.013	0.04	0.095
Location_Mumbai	-0.0091	-0.058	-0.011	-0.074	0.026	0.059	-0.0049	0.045	-0.0058	-0.098	-0.12	-0.14	-0.12	-0.15	-0.11	-0.14	-0.12	1	-0.13	-0.11	-0.0039	0.1	-0.079
Location_Pune	-0.11	0.041	0.13	-0.0092	-0.041	-0.058	-0.0039	-0.084	-0.081	-0.084	-0.1	-0.12	-0.11	-0.13	-0.091	-0.12	-0.1	-0.13	1	-0.046	-0.00016	0.039	0.038
Fuel_Type_Diesel	0.11	0.11	-0.041	0.098	0.43	0.29	0.31	0.37	0.32	0.039	0.037	0.017	0.0099	0.12	0.015	-0.047	-0.033	-0.11	-0.046	1	-0.044	-0.98	-0.14
Fuel_Type_LPG	-0.035	0.00074	0.00042	0.01	-0.041	-0.041	-0.015		-0.026	-0.01	-0.012	-0.014	-0.013	0.06	0.022	-0.015	-0.013	-0.0039	-0.00016	-0.044	1	-0.038	0.026
Fuel_Type_Petrol	-0.11	-0.11	0.041	-0.13	-0.41	-0.27	-0.3	-0.35	-0.31	-0.033	-0.033	-0.0093	-0.015	-0.12	-0.012	0.046	0.04	0.1	0.039	-0.98	-0.038	1	0.13
Transmission_Manual	-0.085	0.024	0.0023	0.37	-0.5	-0.64	0.075	-0.59	-0.59	-0.069	0.026	-0.079	-0.014	0.018	0.096	-0.02	0.095	-0.079	0.038	-0.14	0.026	0.13	1

Figure 12: Correlation Heatmap of the final train data.

5.3 Result of Pricing Model

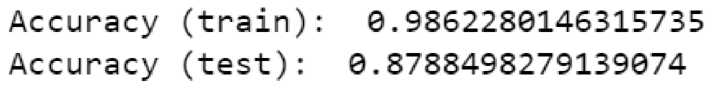
We used a linear regressor and a random forest regressor to create a regression model that can predict the prices of cars. The model and the pre-processing done are fitted to car retail specifically.

5.3.1 Accuracy



```
Accuracy (train):  0.6998752120168004
Accuracy (test):  0.725056126202863
```

Image 1: Linear Regression

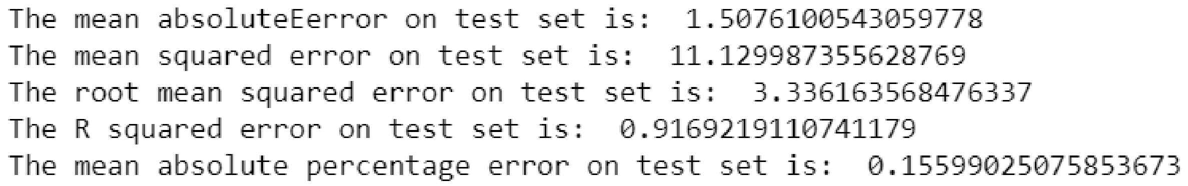


```
Accuracy (train):  0.9862280146315735
Accuracy (test):  0.8788498279139074
```

Image 2: Random Forest Regression

We can see that the linear regression model did not get outstanding results according to the accuracy tests of the train and test datasets. Because of this, we ran the random forest regression and achieved much better results. The good results show that modeling a predicting model for estimating the price of used cars is possible. We have successfully pre-processed and modeled using the tools mentioned above with our dataset. This averaging makes a Random Forest better than a single Decision Tree, improving its accuracy and reducing overfitting. Predictions from the Random Forest Regressor is an average of the predictions produced by the trees in the forest. (Mwiti, 2021).

5.3.1 Error



```
The mean absoluteEerror on test set is:  1.5076100543059778
The mean squared error on test set is:  11.129987355628769
The root mean squared error on test set is:  3.336163568476337
The R squared error on test set is:  0.9169219110741179
The mean absolute percentage error on test set is:  0.15599025075853673
```

Image 3: Error table

The MSE(Mean Squared Error) is relatively high. However, we argue that this is a pricing guideline and should consider what we are trying to suggest, which is most likely different from what previous car retailers have set as their price. The absolute percentage error is 15,6%.

6 Limitations

In retrospect to our understanding of the business case, understanding of the data set, pre-processing, and modeling, all this work is somewhat fraudulent due to the limitations of our chosen data set and limitations of the method itself. Some considerations for full real-life utilization of the model will now get reflected.

6.1 Data set Limitations

6.1.1 Indian data heritage

An apparent weakness with our dataset is that the data collected from India renders our solution useless in the Norwegian market. That conclusion is because of the significant differences in the markets and other factors affecting the retail business. Though, we needed a data set of a certain quality and availability. We discuss the shortcomings of using the Indian data set as an isolated topic. The rest of our thesis has a basis on the Norwegian case and the potential for Automo. According to the case, our model and statistics stand as a “Proof of Concept,” with most regards logic intact. We even have some valuable insights regarding the locations, which renders useful after all.

6.1.1.1 Differences

India is developing into an open-market economy, and policies have accelerated the country's growth, which has averaged under 7% per year since 1997. Slightly more than half of the workforce is in agriculture, but services are the primary source of economic growth, accounting for nearly two-thirds of India's output. The population size, needs, and resources decide what automobile demands the market is characterized by.

The Norwegian economy is prosperous, with a wealthy private sector, a large state sector, and an extensive social safety net. The government controls key areas, such as the vital petroleum sector, through extensive regulation and large-scale state-majority-owned enterprises. In anticipation of eventual declines in oil and gas production, Norway saves revenue from the petroleum sector in the world's second-largest sovereign wealth fund, valued at over $700 billion in January 2013, and uses the fund's return to help finance the public expenses. After solid growth in 2004-07, the economy slowed in 2008 and contracted in 2009 before turning to positive growth in 2010-12; however, the government budget remained surplus. Hence, the Norwegian market is subject to a large middle-class population with pricier car demands. (NationMaster, 2022).

6.1.1.2 Location insights

Even though the Indian heritage is a limitation, we can use the location-specific data to gain comprehensive insights into the car retail industry. We listed two models below, which first show the average price of a car in eleven locations and then the total sales numbers in each of these locations. We will not analyze this much further, but a location might affect the price of the car and what car segments might be possible to sell. This insight is, as mentioned, steadfast in the comparison between Norway and India. The effect is still valid and between the various cities in a country. It seems that the southern cities in India tend to sell more expensive cars on average than other cities. This variation can be valid for various reasons, but the socio-economical structure and level of wealth and needs may be leading indicators. Furthermore, this trend is not similar to the number of cars sold in each city. An explanation might be that India's southern cities have a higher level of wealth per citizen, while other cities contain many people but choose more affordable cars.

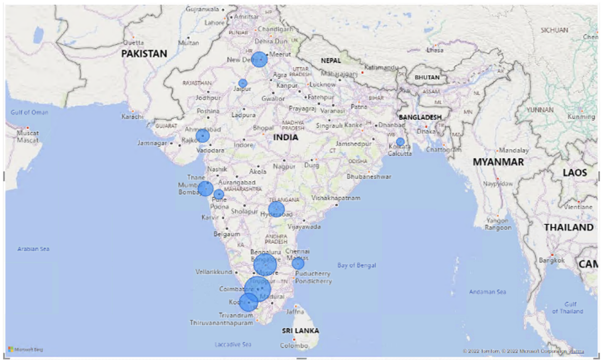


Image 4: Average price in locations

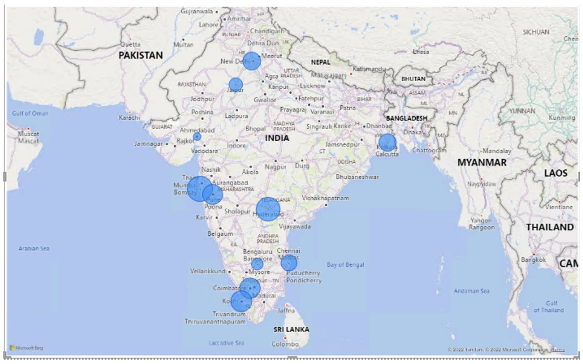


Image 5: Cars sold in locations

6.1.2 Limited data

Working with this dataset, we were limited to 7253 rows. With small data, one does not have the luxury to keep out many samples, and even when one does so, the number of observations in test data may be too small to give a meaningful performance estimate, and the number of observations in cross-validation data may be too few to guide parameter search optimally

(EduPristine, 2016). Representation learning with small labeled data has emerged in many problems since the success of prediction with machine learning often relies on the availability of a considerable amount of labeled data that is expensive to collect. Many efforts can change sophisticated training models with few labeled data in an unsupervised and semi-supervised fashion to address the limited data. (Qi & Luo, 2020).

6.1.3 Lack of Attributes

The data set is a proof of concept and does not contain enough attributes to make an accurate model. There are an incredible number of variables in the automobile industry that may affect price, and the count is ever increasing. For example, here are some variables that might improve the model: number of doors, technologies, color, mechanical and cosmetic reconditioning time, used-to-new ratio, appraisal-to-trade ratio. A realistic model might need hundreds or even thousands of attributes and millions of data entries to make real-time and accurate estimations while facilitating continuous growth of the attributes and entries.

6.2 Method Limitation

Now we reflect on some regards based on the methodological way to solve the business problems of our case. For full real-life utilization of the model, one needs to address the ever-changing macroeconomic environment, user adaptation and change management, and bias problems. Such considerations might render the estimation model invaluable.

6.2.1 Macroeconomic Factors

The macroeconomic aspect is a significant limitation of the model as it creates an unpredictable influence on pricing. We have chosen to use the PESTEL model to assess the macro factors, but Porter's five forces is an example of another model that could have been used. A PESTEL analysis is a tool used to gain a macro picture of an industry environment. PESTEL stands for Political, Economic, Social, Technological, Legal, and Environmental factors. It allows a company to form an impression of the factors that might impact a new business or industry.

Political
Legal
Economic
PESTEL
Environmental
Social
Technological

Figure 13: PESTEL

6.2.3.1 Political

The political factors are the stability of government, potential changes to legislation, and global influence. Those factors mean everything the government does with relevance to a business. (DigitalNorway, 2021). We have lived with a pandemic in recent years, which has affected many industries, including the car industry. The pandemic led to several restrictions by the government, for example, the travel ban. The ban resulted in people traveling around Norway, which led to people buying more cars, especially used cars. The increase in used car sales was also due to challenges in importing new cars. Because the demand has increased, sellers have increased their car prices. Now it looks like the pandemic is ending, and the future demand for used cars is uncertain. Here there are limitations because a price prediction model will not consider the government's regulations.

6.2.3.2 Economic

The economic factors are related to economic growth, employment rates, monetary policy, and consumer confidence. (DigitalNorway, 2021). There have been many economic fluctuations due to the pandemic in recent years, and this again results in fluctuations in the used car prices, which the model will not catch.

6.2.3.3 Social

Social factors include income distribution, demographic influence, and lifestyle factors. In the assessment of social factors, we form an image of the attitudes and patterns of the target group today and how the situation will be in the future. (DigitalNorway, 2021). Here there are limitations with the model. A model for predicting the price will maximize the profit and not include these social factors. Mainly it will include variables connected to the car and its functionality. In recent years the interest in the used car market has increased, primarily because of the pandemic, which has led to higher prices, but this can change at any time.

6.2.3.4 Technology

Technology is factors like international influences, changes in the information, and take up rates. In the phone interview with Melhus Bil, Kristian Rønning talked about constantly evolving technology. The car market includes new battery technologies, extended range, better heating, and charging speed. Also, more vans are becoming electric cars. (K. Rønning, personal communication, 4. mars 2022).

Further on, the technological factors include the impact of automation and data usage and how this can get transformed into value for the business. (DigitalNorway, 2021). When we talked about automated pricing, Rønning was skeptical because of the margin of error. He says the automatic assessment programs do not consider equipment and specific technologies, and they got a foundation on averages. The human conception is not taken into account. For example, if one considers the brand on the car, it can give a high risk of misjudgment. The system is working poorly with private resellers and companies. Factors such as experience, history, and relationships are essential. For example, a price estimation program will not describe how skilled the team leaders or salespeople are. (K. Rønning, personal communication, 4. mars 2022).

An automated solution for a price estimation or car sales will probably work better for new cars than used cars, according to Rønning. It is essentially what one physically sells, and there are many risk factors for the customer when buying a used car, such as wear and tear. The risks will not be easy to obtain on a computer. (K. Rønning, personal communication, 4. mars 2022).

6.2.3.5 Environment

Environment factors are, for example, regulations and restrictions and attitudes of customers. A price prediction model will try to be as effective as possible, designed to maximize profit, resulting in the model making decisions contrary to climate and environmental policy. (DigitalNorway, 2021).

6.2.3.6 Legal

The legal factors are connected to taxation policies, employment laws, industry regulations, and health and safety. Here we can mention GDPR. To develop the best possible model, we need as many variables as possible to get relevant information, and privacy information may be a limitation of the model. (DigitalNorway, 2021).

6.2.2 User adaptation and change

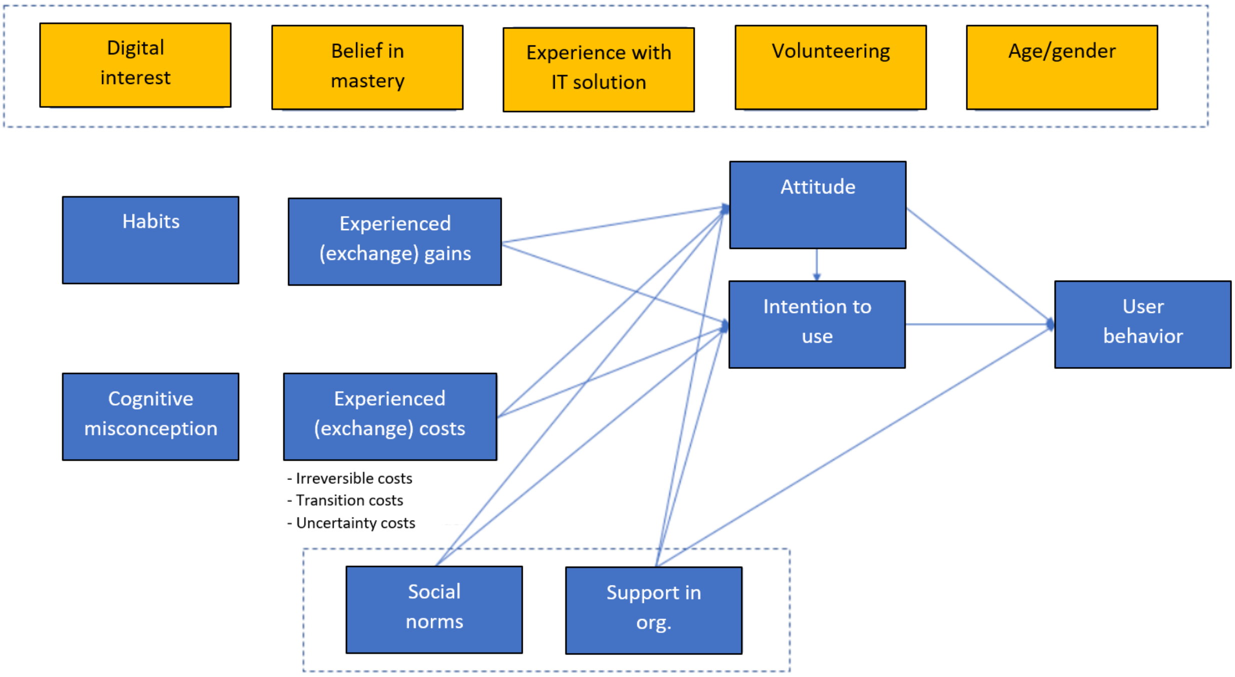


Figure 14: Model for technology acceptance.

The basis of any technology acceptance theory is that certain factors predict whether a user will have an intention to use a new technology, which will result in actual user behavior.

Change management is about influencing user behavior when digital solutions are introduced and used.

The two most important moments which affect the intention to use and the user behavior are perceived gains or benefits and experienced costs or effort. Perceived gains are about making assessments and creating expectations about what the use of a new IT solution will be of benefit or gains in the work situation. This experience will vary from group to group and individual to individual. Here are several things that affect, and then among other things, we must look at the other surrounding factors in the figure. Social norms, culture, and the way one talks about the new solution can, for example, affect what gains one makes envisions. In addition, as said, the yellow boxes' moderating variables will affect differently.

Perceived costs or efforts are about several aspects. Again, it is essential to note that it is a subjective experience, but at the same time that it can, of course, be about actual costs it some users will have to take (e.g., time to learn a new system). Three relevant points for our case got listed below this factor in the figure above. Irreversible costs ("sunk costs") are investments as a user has done, as with the transition to a new IT solution, new routines will get lost in Automo. Examples can be technical expertise around their old CRM system, which is not transferable to a new solution built on new technical standards.

Again, it is about experiencing loss and how one looks at having to leave something one has been working on for maybe many years for something new. Transition costs are about those costs and the extra effort a user must take in the preparation and the process of taking the new system into use. For example, there may be extra time spent training, and navigating a new solution. Many organizations experience much overtime in an initial phase after go-live with a new solution. The last point is uncertainty costs which is the burden related to uncertainty the user experiences when it comes to, for example, how much new one needs to learn, what tasks, and how one's work situation changes. Switching cost is essentially about relating the cost and effort to the current solution and way of working.

In the same way, we can also talk about “Exchange gain” because the gains measure against the status quo. “Cognitive Misconception” is drawn as a separate factor to show that people tend to give negative moments more attention than good moments. A disadvantage with a new solution may overshadow benefits and make technology acceptance or lack thereof partially irrational.

The assessment of perceived exchange cost/effort against perceived exchange gain/benefit will get concatenated to be able to predict whether the user tends towards acceptance of new technology or not. We have already seen that this cost-benefit assessment is only partially rational and that it is about an expectation created and influenced in a context of individual, technological, social, and organizational factors. Some of these are trapped in the other factors in yellow at the top of the figure and in the factors “social support” and “support in org.”. (Venkatesh, Morris, Davis & Davis, 2003).

6.2.3 Bias Exploration

Our model may be subject to bias, as earlier data entries might not be generalized for all future car predictions. If all the entries heir from the timeframe of COVID-19, then the model might be biased towards that macroeconomic structure. In machine learning, the term bias got introduced by Mitchell (1980) to mean “any basis for choosing one generalization over another, other than strict consistency with the observed training instances.” Examples of such biases include absolute biases and relative biases. An absolute bias is an assumption by the learning algorithm that the target function to be learned is a member of some designated set of functions, such as the set of linear discriminate functions or the set of boolean conjunctions. A relative bias assumes that the function to learn is more likely to be from one set of functions than from another. For example, the decision tree algorithms consider small trees before considering larger ones, and a larger one has not considered if these algorithms add a small tree that can correctly classify the training data. (Dietterich & Kong, 1995).

7 Recommendations and Further Work

7.1 Implementation and Recommendations

7.1.1 Possibilities

Use cases for Automo Inc
Increase rebound rate
Reduce manual work per sale
Increase website traffic and autonomous sales
Reduce marketing cost per acquisition
Increase prices
Expand with ad-on options
Reduce extra maintenance cost risk per car

Table 3: Use cases for Automo Inc.

A machine learning-backed estimation increases trustworthiness amongst customers, which increases the possibility of a later rebound; manual work is costly, automating more business areas would reduce work hours and, in turn, future need for salary. The estimation software could even help customers to order cars online. Estimation software also allows for customized pricing in marketing efforts. An estimated price with a reasonable surcharge will drive income to Automo. More add-on options could grant better quality service in the form of customization and drive sales and turnover up because of the adjourning ad-on to the overall price. However, it entails demands for the estimators' development and acquisition of new solutions and equipment.

7.1.2 Shortcomings

The limitation section mentioned several aspects that might get referred to as shortcomings in using the estimation software. These include; lack of relevant data, limited dataset sizes, lack of attributes, high level of macroeconomic complications, restrained user adaptation, and risk of bias. These aspects are highly considered shortcomings that prevent us from recommending the implementation of a prediction software in the variety of use cases in the table above. There is no sufficient reason to believe that the technology and data sources have evolved to a level where they can be utilized efficiently by local car retail companies. Even if they could benefit from such a well-developed software today, the resistance to change internally at companies like Automo would probably be too significant for implementation.

7.1.3 Conclusion

Even though there is obvious potential for business intelligence in the Automobile retail industry, the data set we have worked on and methodological limitations are too significant to recommend the usage of such a system for any car retailer. There are too many factors that need to get considered. The technology we just utilized does not have the necessary capabilities to apprehend the methodological shortcomings when we put the technology up against macroeconomic complexity, user adaptation, willingness to change in the retail business, equipment, specification-driven complexity, and biased models. However, the potential for our use cases will probably still be relevant when this technology and resources are required to catch up in the future.

7.2 Future Recommendation and Analysis

7.2.1 Data collection

Data science as a field and its potential for precious business intelligence is a driver for companies that will keep their competitive advantage in the future. To embrace the future possibilities, one must adapt to new technology in heart and culture and begin to utilize the new possibilities. One way to do this is by gathering and storing all seemingly relevant data and data with hidden potential. A car retail company should prepare by storing all sales data and all their access to in general. However, legal regard ensures that the business runs on clean sheets. By employing data storage, a company can get valuable business insights in the future that might save the company's existence in an ever-changing market.

7.2.1 Feature diversification

This thesis mentioned the vast number of attributes that one would have to combine to create a usable model. There are simply a lot of different features and variations that distinguish the cars and manufacturers from each other. However, we have found no clues that suggest car manufacturers want to standardize their cars and technologies. On the contrary, the industry seems interested in diversifying its cars and brands. Interests in diversifying may be explained by their branding strategies and the need to customize their products specifically to a small niche and identity. Unwillingness to conform amongst the manufacturers might render our mission to compare and estimate all cars efficiently in the future impossible.

References

Abrahamsen, M. (2021, march 3). *Bruktbilmarkedet er rekordstort - dieser er fortsatt kongen.* rb. https://www.rb.no/bruktbilmarkedet-er-rekordstort-diesel-er-fortsatt-kongen/s/5-43-1532717

Bakshi, C. (2020, june 8). *Random Forest Regression*. gitconnected. https://levelup.gitconnected.com/random-forest-regression-209c0f354c84

Dietterich, T. G. & Kong E. B. (1995). *Machine Learning Bias, Statistical Bias, and Statistical Variance of Decision Tree Algorithms*. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.38.2702&rep=rep1&type=pdf

DigitalNorway. (2021, january 20). *Hvilke eksterne drivkrefter påvirker bedriften? SLik gjør dere en PESTEL-analyse.* DigitalNorway. https://digitalnorway.com/pestel-analyse/

EduPristine. (2016, february 15). *Problems of Small Data and How to Handle Them.* https://www.edupristine.com/blog/managing-small-data

Kasliwal, A. (2019, june 25). *Used Cars Price Prediction.* Kaggle. *https://www.kaggle.com/avikasliwal/used-cars-price-prediction*

Microsoft. (2022). *Why Power Microsoft Power BI.* Microsoft. https://powerbi.microsoft.com/en-us/why-power-bi/

Mwiti, D. (2021, october 26). *Random Forest Regression: When Does It Fail and Why?* neptuneblog. https://neptune.ai/blog/random-forest-regression-when-does-it-fail-and-why

NationMaster. *Economy Stats: compare key data on India & Norway.* NationMaster. https://www.nationmaster.com/country-info/compare/India/Norway/Economy

Pudaruth, S. (2014). Predicting the Price of Used Cars using Machine Learning Techniques. *Academia.* *https://d1wqtxts1xzle7.cloudfront.net/54261672/2014_Predicting_the_Price_of_Used_Cars_using_Machine_Learning_Techniques-with-cover-page-v2.pdf?Expires=1646571715&Signature=b7M~03sHP6g0PXYDHTci3NHOp9L1cRylskHUxt-exLeJAhGBmeaUCo~KrsuKGKqt476~woZr~Ra3HiS2qDv0PLEkTKZSiPL1IzLqbT1CQToF99Gb4l0H5Hrmk~H98hv-~Z4wiH5PZ4qQKVEkpDjD8OGbPL3hXuk87*

Tiu9aI6~fmOZ3jb0CXbatmZ~Y0sWKNkBADDoHlRnvJf1q1ZsX~zccGYOKYsZ2DvHVkXwWFQAAb0zcJWsWRQG-LYZ48bxMlGtvTgeXZG9lPXqnumdeFxgAM~kWUnH~bFoHSvb4dAvI4f5sssvW~BNPMUie7i7u-m4FW~1AQKSTYjIuBhzJIegw__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA

Qi, G. J. & Luo J. (2020, october 19). *Small Data Challenges in Big Data Era: A Survey of Recent Progress on Unsupervised and Semi-Supervised Methods.* IEEE Xplore. https://arxiv.org/pdf/1903.11260.pdf

Rekatsinas, T., Dong X. L, Getoor, L. & Srivastava, D. (2015). *Finding Quality in Quantity: The Challenge of Discovering Valuable Sources for Integration.* https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.671.2958&rep=rep1&type=pdf

Skogstad, K. (2021, september 21). *Derfor har mange bruktbiler økt i pris det siste året.* TV2, broom. https://www.tv2.no/a/14235654/

SSB. (2017, june 1). *Sterkeste vekst på fem år for bilforhandlere og verksteder.* Statistisk sentralbyrå. https://www.ssb.no/varehandel-og-tjenesteyting/artikler-og-publikasjoner/sterkeste-vekst-pa-fem-ar-for-bilforhandlere-og-verksteder

Venkatesh, V., Morris, M., Davis, G. B., & Davis, F. D. (2003). *User acceptance of information technology: Toward a unified view.* MIS Quarterly, 27(3), 425–478. https://doi.org/10.2307/30036540

Wikipedia. (2021, september 24). *Multikollinearitet.* Wikipedia. https://no.wikipedia.org/wiki/Multikollinearitet

Wirth, R. & Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining. *unibo.* http://www.cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf

Chakure, Afroz. (2019). Random Forest Regression. https://medium.com/swlh/random-forest-and-its-implementation-71824ced454f

Thompson, Ben. (2019). A limitation of Random Forest Regression. https://towardsdatascience.com/a-limitation-of-random-forest-regression-db8ed7419e9f

Attachments

Price_Prediction_Modelling.ipynb - The Python programming in this thesis.