

**Used Cars Data Mining**

|  |  |
| --- | --- |
| **Anamika Avinash**  **Vikas Chhillar** | Name: |
| **16210480**  **16212887** | Student ID: |
| **anamika.avinash2@mail.dcu.ie**  **vikas.chhillar2@mail.dcu.ie** | Email: |
| **MCM (Data Analytics)** | Programme: |
| **CA683** | Module code: |
| **Dr. Andrew McCarren** | Lecturer: |
| **18-April-2017** | Submission date: |

**Declaration**

In submitting this project, we declare that the project material, which we now submit, is our own work. Any assistance received by way of borrowing from the work of others has been cited and acknowledged within the work. We make this declaration in the knowledge that a breach of the rules pertaining to project submission may carry serious consequences.

**Contents**

**1 Questions to be Answered**

**2 Introduction to the Dataset**

**3 Project Plan**

**3.1 Data Collection and Understanding**

**4 Dataset Description**

**4.1 Cleaning Dataset**

**5 Modelling**

**6. Conclusion**

**Abstract**

According to recent articles, sales have improved in the European Union since the recession that hit in 2008 but the auto industry is still recovering and trying to maintain a balance. Q3 2016 closed with 9.8M vehicles sold in the used market, an increase of 3.3% over the previous year[1] . This increase in overall values is heightened by the younger age of vehicles sold at the dealerships. These changes in new car buying behaviour is consequential to an increase in the used car market sales.

Our dataset called the Used Car Database is obtained from Ebay-Kleinanzeigen in German and consists of multiple attributes such as price, year of registration, powerPS, kilometer, fueltype, brand and so on. Predicting values for a used car based on the numerous attributes present in the dataset is the main objective of this project.

**1. Questions to be answered**

* Which is the most common car company considered when buying an used car?
* Relationship between price and age of the car?
* Estimate the cost of the car model considering the year of registration?
* Does the length of description influence the price?
* Price prediction analysis

**2. Introduction to the Dataset**

Dataset in consideration is obtained from Ebay-Kleinanzeigen in German and consists of 371528 rows and 20 attributes. The attributes include price, year of registration, powerPS, kilometers, fueltype, etc. Since the dataset is in german, there are a few values which needed to be altered. The data contains ads that are aired between March 2015 and April 2016.

Objective is to understand the growing used car market. It is a well known universal fact that price is the main factor when it comes to deciding on which car to buy but there are a few other factors also that effect the decision. To successfully explore the data , we have removed a few columns , added a few and cleaned and refined the remaining

The aim is to understand this growing used car market. When we look at used cars, price is the most important factor that influences opinions, but there are also few other facts that affect the purchase decision. So we tried to find out which variables affects the price most and how they do it.

.

**3. Project plan:**

The project is divided into 5 phases and the plan of the project was to understand the second hand cars market and give the predicted values for cars.

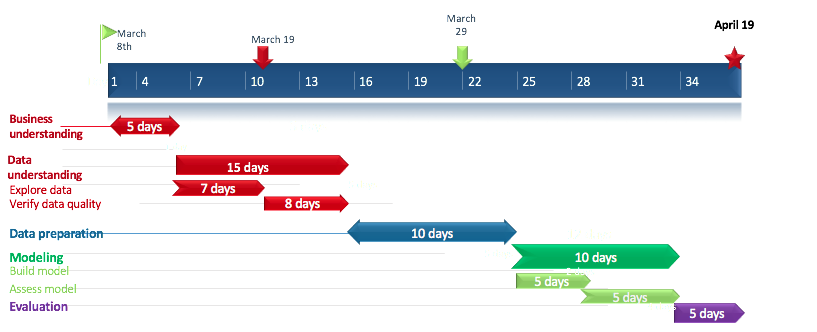


Fig 2.1 :Project Plan

**3.1 Data collection and understanding**

The dataset of second hand cars is collected for the project. The dataset contains prices and attributes for approximately 370,000 second hand cars of 40 unique brands. The dataset contains several variables for exploratory data analysis. Few attributes were dropped as they were of no use in the model.

We plotted various graphs to understand the data and the various features and patterns we could obtain from it. A simple bar graph that shows the brand distribution in the entire dataset was first done and the top 10 brands with the highest count was then plotted as shown below

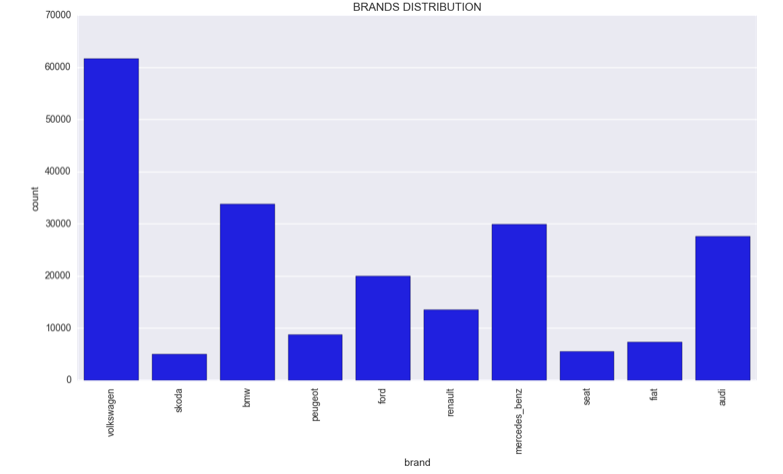
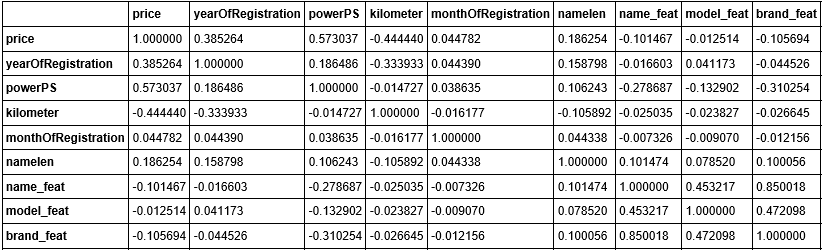
fig 3.1 Brand Distribution

Fig 3.2 Correlation matrix

The above table shows the correlation between the various attributes and features. It is evident that the correlation between price and powerPS , year of registration and namelen is something worth looking into.

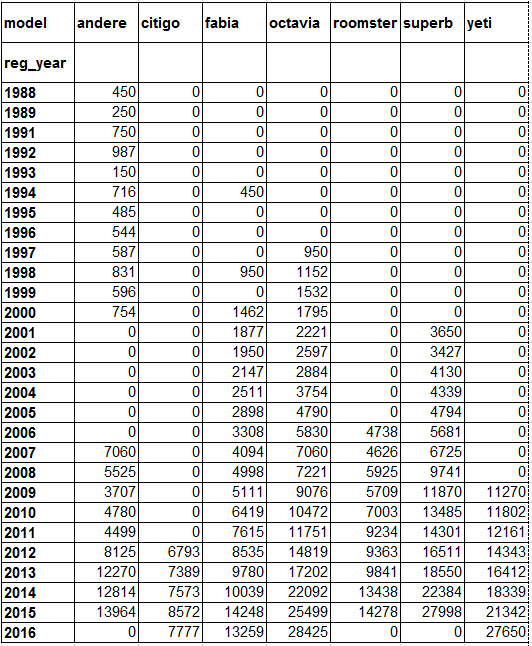


fig 3.3 : pivot table (skoda)

The pivot table above shows an estimate of the prices for the respective skoda car models depending on the year of registration. This is calculated using the mean of all the values.



Fig 3.4 Effect of length of description on price

The above plot shows the relationship between the price and the length of the description. The x axis consists of namelen and price on the y-axis. From the above plotted graph we can see that the having a description which ranges between 15 to 30 characters is nominal and gets a good price.

**4. Dataset description:**

The dataset consist of 20 attributes and the top 10 brands(BMW, Volkswagen, Opel, Mercedes, Audi, Ford, Renault, Peugeot, Seat and Feat) contributes to 80 percent of the dataset

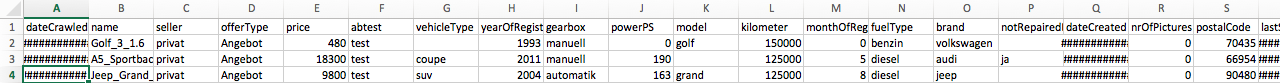


Fig 4.1 Dataset description

**4.1 Cleaning dataset:**

* The dataset contains some German names which are changed to English.
* Engine power cleaning : We observed that the engine power (PowerPS) column included extreme outliers, therefore after our first look we corrected them with quantile(). The sample value of engine power is set as NA when the value is lower than 40.
* Dropping some columns: Columns such as no. of pictures in the advertisement contains only one numerical values i.e “0” so it is of no use for analysis. The offertype column is also dropped because there are only 12 records out of 371528 having a value different from “Angebot”.

**5. Modelling**

**Audi**

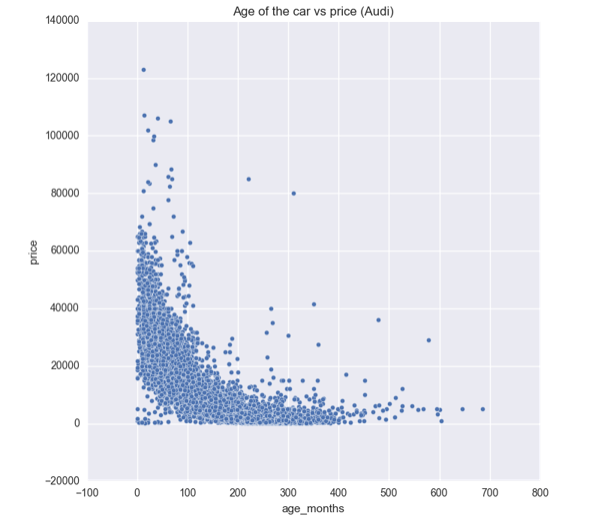
To develop a prediction model, we have considered one brand and one model of another brand. The brand considered in Audi because it has a considerably high count of cars 

Fig 5.1 : Age of the car vs price for Audi.

The above plot shows the relationship between the age of the car and the price. This is only for the brand Audi. The polynomial transformation is then taken into account with a degree of 2 and the factors considered are age of the car, kilometers and powerPS because these attributes had a higher value in the correlation table that was shown before. The linear regression for the above was then done using scikit learn library and the plot produced is shown below.



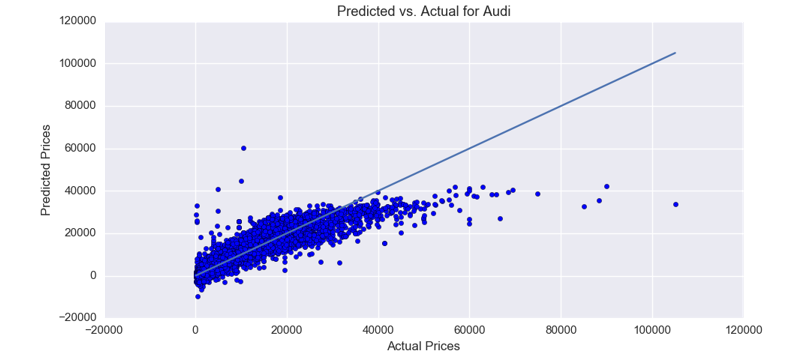


Fig 5.2 : Predicted vs Actual for Audi

Audi Q7 is a high end car by the company. The predicted vs actual plot is shown below and so is the R2 score.



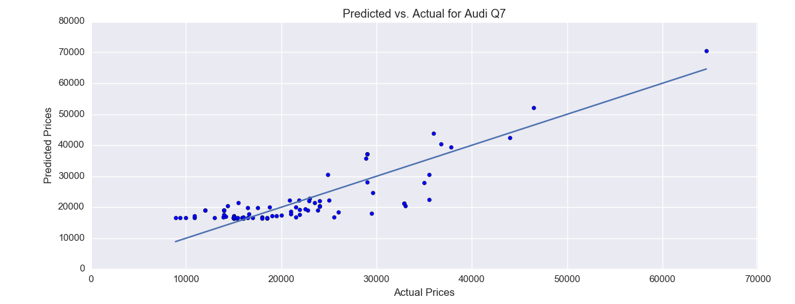


Fig 5.3 Predicted vs actual for Audi Q7

**SKODA**

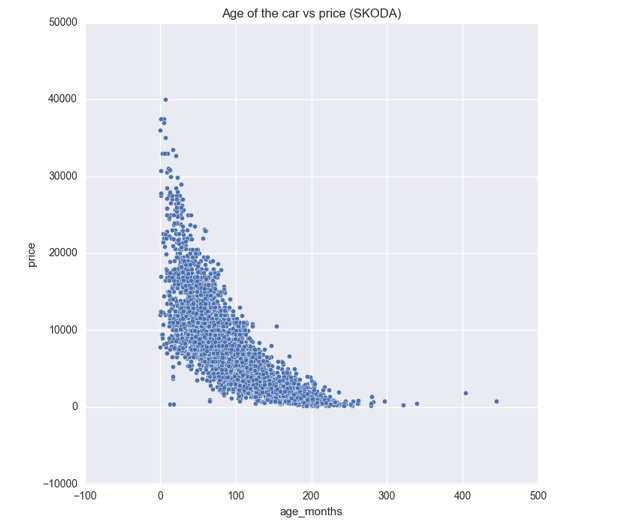
****

Fig 5.4 : Age of the car vs price for Skoda

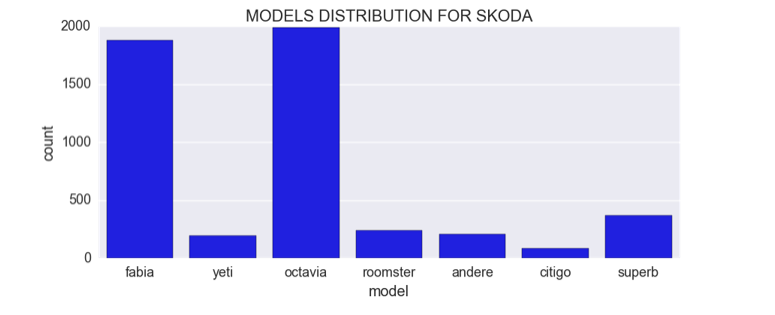


Fig 5.5 : Model distribution for skoda

Taking skoda into consideration, we plotted the model distribution for it and the age of the car vs price.

Since Octavia has a high count of cars, we have used it for our prediction model.



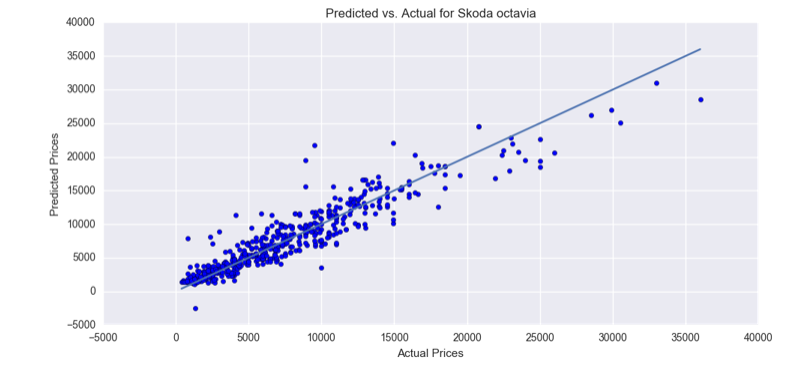


Fig 5.6 : Predicted vs actual for Skoda Octavia

**6. Conclusion**

* The cars with name length between 15 and 30 seems to have higher price for selling the used cars. One of the reason could be that the longer name means that there are more optional features and accessories which makes the price of the car higher.
* The top 10 popular brands are BMW, Volkswagen, Opel, Mercedes, Audi, Ford, Renault, Peugeot, Seat and Feat and contributes to around 80 percent of the cars(there are 40 brands in the dataset) and the most common brand is Volkswagen.
* The price and age of the car is related and is shown above using scatter plots in a brand wise manner.
* The cost of the car can be predicted if the year of registration is known with the help of a pivot table.
* The price of the car is continuously decreasing for its first 20 years but slightly increases after. 20+ years car are considered as antique by some people.

**7. References**

[1] http://dealers.edmunds.com/static/assets/articles/2016\_Q3\_Used\_Market\_Report.pdf

[2] https://www.kaggle.com/orgesleka/used-cars-database