DSA 6000 FINAL PROJECT REPORT

PREDICTING USED CAR PRICES USING STATISTICAL LEARNING METHODS

Members: Aseel Faddah, Rajratna Patil, Shivani Tayal, Waijeh Salman, Yousuf Qadri

Abstract

This project aims to use statistical learning techniques to accurately predict the price of used cars in the German market. A comprehensive historical dataset was scraped from Ebay Germany and linear regression models were applied to predict the car price accurately. The report compares the performance of different models that were used to fit the dataset. Prior to modeling, the data was cleaned and visualized using Tableau to gain insights about different parameters that affected the car price. The models are based on predictor variables such as vehicle type, fuel type, kilometers driven, and car age. Based on the results from the model, the predicted car prices were compared with the actual car prices in dataset and accuracy of the models were evaluated. The Square Root transformation model was found to generate best results with a R^2 of 0.81. The analysis also helped determine the relationship between the car's price and its features. The models created in the project are therefore useful for car buyers/sellers as it enables them to know what a good price is to buy/or sell a car based on the characteristics/features of that car.

Background

E-commerce continues to spread and is becoming the more popular method people use to buy and sell products and services. Thus, many industries and markets are adopting it, and the used car market is no different. Used car advertisements are posted on various online platforms around the world every day to promote online sales. The posted selling price on each ad is determined by the car's features and other attributes, like the car's age for instance. However, the relationship between these features and the price may not be clear. Both buyers and sellers want to have a clear understanding of the relation, as the seller needs to know the appropriate selling price to post without under-pricing a car, while customers would not want to overpay for a car.

This report focuses on the German used car market. Data collected from eBay-Kleinanzeigen, an online web-crawling platform using e-Bay in Germany, is used to make the analysis throughout the report. There are over 370,000 data samples that were collected regarding scraped used car pricings against over 20 factors, including mileage, brand and model, and time period that each car remained before it was sold. Given this information, tracking the factor's relation relative to the price of the car is possible and can be utilized by both the seller and customer. The data set can be referred from the following link: https://www.kaggle.com/orgesleka/used-cars-database#autos.csv

Methodology

For the purpose of this project, it's necessary to identify a series of steps that help achieve the objectives. These steps are first, cleaning the data to removing any outliers or variables that do not have an effect on the car's price; second, visualizing the data to observe any trends that may be present; third, using the cleaned and graphically visualized data to start building multiple models that predict the price of a car; and fourth, calculate the accuracy and error rates for each of the models and identify the one with the highest prediction accuracy. Each step will be further discussed in the next section.

1. Data Cleaning

The dataset originally has 371,521 observations and 20 variables from the used car data scraped from eBay - Germany between March and April of 2016. The data is in German and it contains 20 categorical and numerical variables. The main response variable in the dataset is the price of the used car. Data cleansing began by removing unwanted columns such as seller name, number of pictures shown in the ad, offer type, date crawled, and month of registration. Through further investigation on the pivot tables and in R, we found outliers in the data which were removed by assigning limiting ranges as follows:

- Adjusting the price range from $200-$150000
- Changing the year of registration from 1990-2016
- Adjust power PS from 25-600.

In the dataset, the conversion of data columns from German into English was done. Some columns like vehicle type and fuel type had blank entries and these were assigned as “others”. The “year of registration” variable was converted to car age in years. After cleaning, the dataset had 280,081 observations and 10 variables.

2. Data Visualization

Tableau was used to visualize data and identify key trends. This was useful in determining which variables were required to be chosen for model fitting and helped identify any further errors left in the data. Figure 1 shows that the German market is dominated by 5 major brands that account for the 70% of the total sales. Based on this visualization, it was considered adequate to use these brands to fit the model going forward. Figure 2 shows that an inverse correlation exists between car age and price. Similar trend was observed for kilometers driven vs car price. As the kilometers driven by the car increased, its price decreased. On the other hand, the car price increased with increasing power of the vehicle.

It is observed from Figure 2 that all the vehicle brands tend to reach similar price levels after 15 years due to depreciation. Using a second-degree polynomial fit, the trend line was created for each brand. By looking at the R-squared values and p-value, there is confidence to say that there is a strong relationship between the car age predictor and the response price.

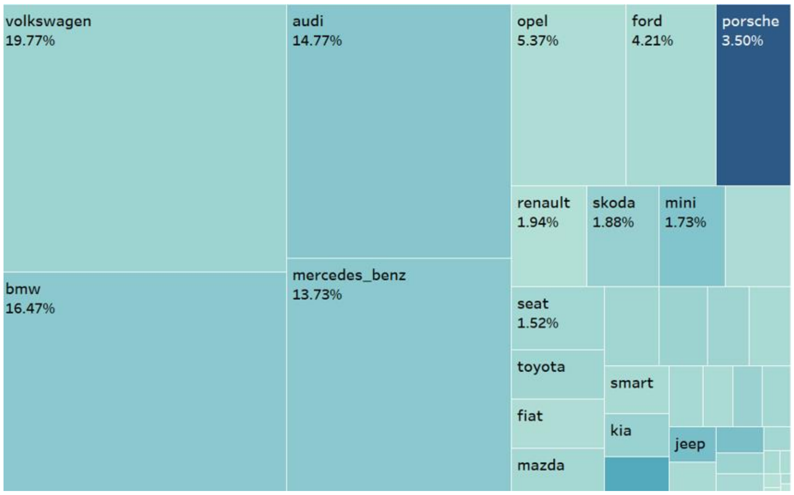

Figure 1: Vehicle brand distribution in German car market

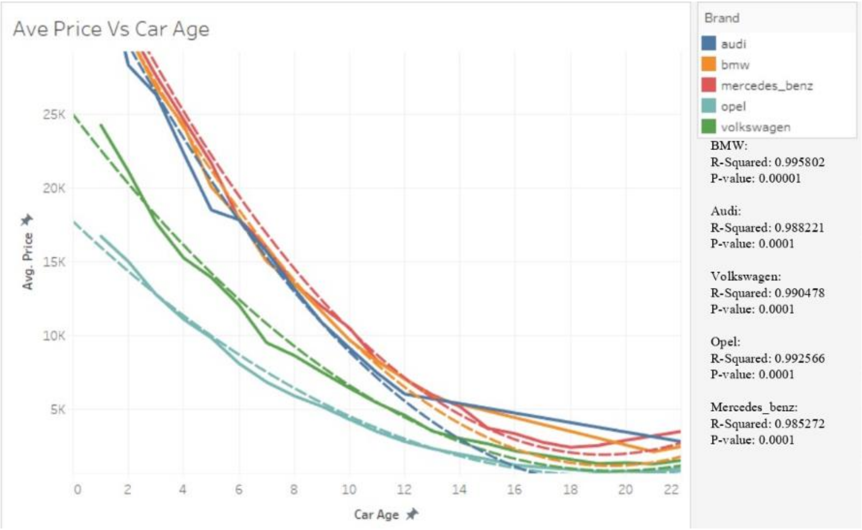

Figure 2: Vehicle price depreciation for all brands

3. **Model Building**

With the data cleaned and preliminary trends identified, the next step was to build predictive models for the dataset using R. Figure 3 shows the correlation plot between different numerical variables. Since correlations are not significantly higher (for example 0.52 between car age and kilometer driven), all the numerical variables were retained in the model. If the degree of correlation between variables is higher, it can affect the accuracy of the model.

In this project, Linear Regression was used. Linear Regression attempts to model the relationship between two or more variables by fitting a linear equation to observed data. It tries to find the line that best fits the data points available, so that we can use it to predict output values for input that are not present in the data set we have, with the belief that those outputs would fall on the line. One of the limitations of regression is that it can be used only for linear relationships.

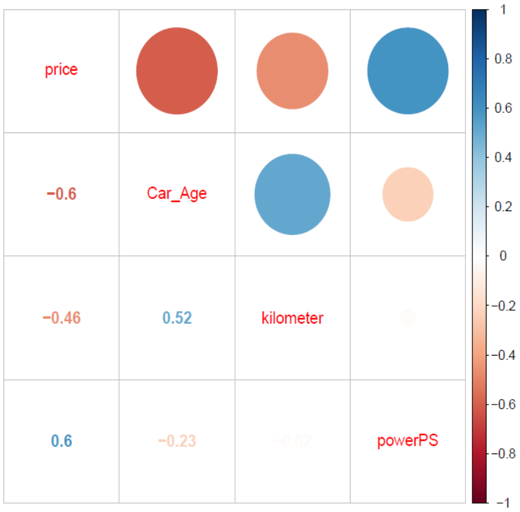

Figure 3: Correlation plot between numerical variables

The data was split into 70% training and 30% test data. To obtain high accuracy, several models were attempted in this study and R^2 and Residual Standard Error (RSE) for each was compared to assess the model performance. Table 1 shows the performance comparison for three main models that were tried. The first model is a simple linear model between car price as the response and vehicle type, gearbox, power, kilometers driven, fuel type, brand, car age as the predictors. These predictor variables remained constant across different models. The log and sqrt models were built by taking the log and sqrt transformation of the response variable. It is evident that the sqrt model performed best among the three since its R^2 is highest and RSE is minimum. Hence the sqrt model was selected as the final model for this project. Other methods including cube-root transformation and higher degree polynomials were also tested but showed negligible improvement over sqrt model.

The details and code samples for the Models are shown below:

Model 1: Linear Model

```
Model_simple<-lm(price~vehicleType+gearbox+powerPS+Kilometer+fuelType+brand+CarAge,
data=train)
```

Summary was used to calculate R^2 and RSE

Model 2: Logarithmic Model

```
Model_log<-lm(log(price)~vehicleType+gearbox+powerPS+Kilometer+fuelType+brand+CarAge,
data=train)
```

Since log(price) was used as a response variable, we used below calculations to calculate R^2 and RSE

```
pred_log<-exp(predict(Model_log,test))
RSS<-sum((test$price-pred_log)^2)
TSS<- sum((test$price-mean(test$price))^2)
RSE<- sqrt(RSS/(nrow(test)-16-1))
(R2<- 1-RSS/TSS)
```

Model 3: Square-Root Model

```
Model_Sqrt<-lm(sqrt(price)~vehicleType+gearbox+powerPS+Kilometer+fuelType+brand+CarAge,
data=train)
```

Since sqrt(price) was used as a response variable, we used below calculations to calculate R^2 and RSE

```
pred_sqrt<-predict(Model_sqrt,test)^2
RSS<-sum((test$price-pred_sqrt)^2)
TSS<- sum((test$price-mean(test$price))^2)
RSE<- sqrt(RSS/(nrow(test)-16-1))
(R2<- 1-RSS/TSS)
```

Table 1: Performance comparison of main models

	Simple Linear Model	Log Model	Sqrt Model
R^2	0.73	0.71	0.81
RSE	3950	4042.02	3273.87

The data was also checked for multicollinearity between independent variables using VIF (Variance Inflation Factor). There was a clear sign of no multicollinearity since all VIF values were less than 5.

Results

Using the sqrt model which performed best, car prices were predicted. Figure 4 shows the comparison of the predicted prices vs the actual car prices. The left plot shows that for most datapoints, predicted price values are quite close to the actual prices. The plot on the right shows true price on the x-axis and predicted price on the y-axis and we can notice a linear (y=x) correlation between the two for each vehicle brand.

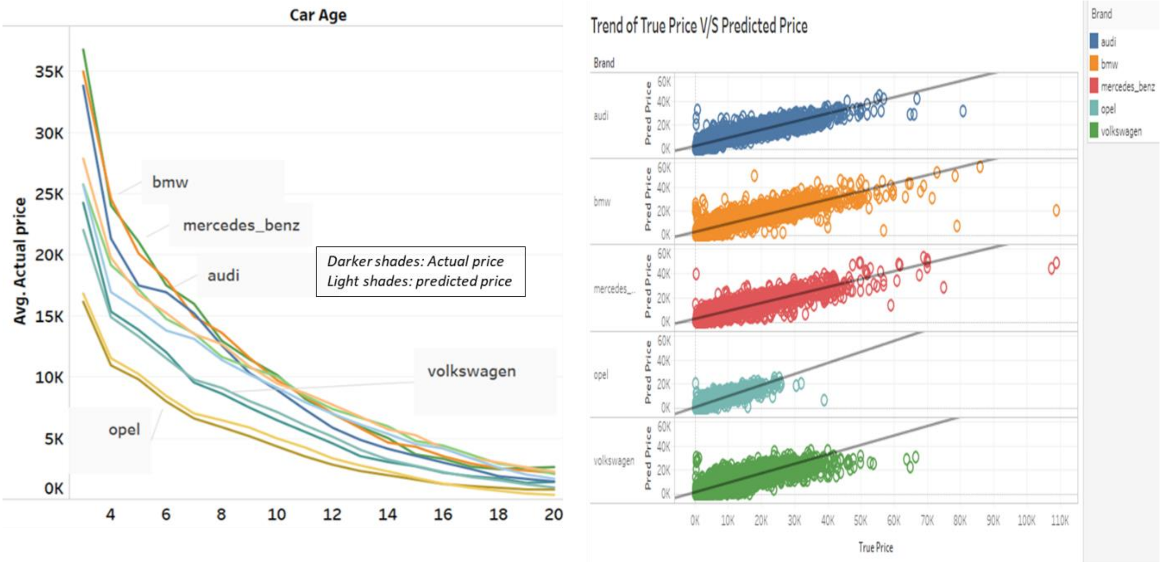

Figure 4: Comparison of predicted price vs actual car price

Conclusions

It was concluded in this project that the SQRT model gives the best fit for the current dataset. An R^2 of 0.81 and RSE of 3273.87 is achieved using the prediction model. Volkswagen, BMW, Audi, Mercedes and Opel were found to be the most popular brands in the German Market. The significant variables in the predicting the price of a used car are vehicle type, gearbox, power, kilometers driven, brand and car age. This project helped us understand the theory behind predictive analytics and its practical implementation on a real-world data set. By applying these principles, we were able to get results which seem quite promising. Also, this project strengthened our skills on R programming and tableau visualization.

References

1. http://faculty.marshall.usc.edu/gareth-james/ISL/ISLR%20Seventh%20Printing.pdf