**Predicting the car resale price using machine learning techniques**

**Sharmila.J**

**Contents**

**1. Title of the project**

**1.1 Introduction**

**1.2 Objective**

**1.3 Methodology**

**1.3.1 Retrieving data**

**1.3.2 Data preprocessing and data cleaning**

**1.3.3 Data visualization**

**1.3.4 Data exploration**

**1.3.5 Data modeling**

**2. Architecture**

**3. Related work**

**4. Conclusion**

**5. References**

Abstract

A car price prediction has been a high interest research area, as it requires noticeable effort and knowledge of the field expert. Considerable number of distinct attributes are examined for the reliable and accurate prediction. Vehicle price prediction especially when the vehicle is used and not coming direct from the factory, is both a critical and important task. With increase in demand for used cars more and more vehicle buyers are finding alternatives of buying new cars. There is a need of accurate price prediction mechanism for the used cars. Prediction techniques of machine learning can be helpful in this regard. It is common to lease a car in many countries rather than buying a new car. After the lease period is over, the buyer has the possibility to buy the car at its residual value, i.e. its expected resale value.

Objective

The objective of the project is to find the possible resale value of a car based on its gearbox, model, brand, vehicle type, fuel type and whether repaired or not.

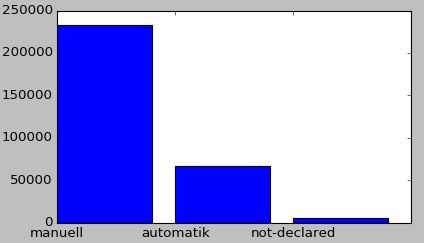
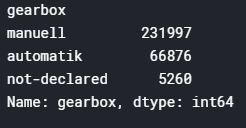
Methodology

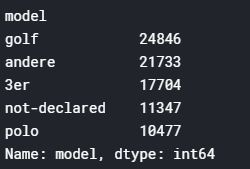
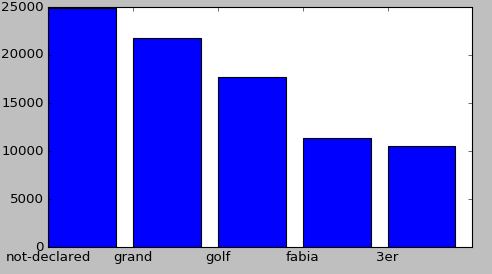
**Retrieving data**

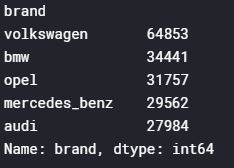
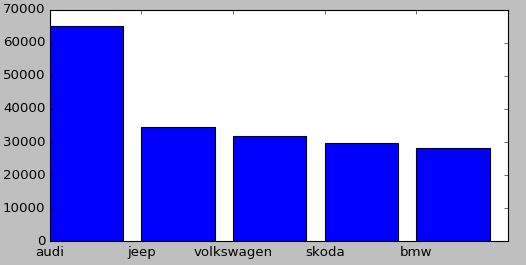
Retrieving Data is the important step that comes after setting the research goal, for this purpose and used online data source was used.

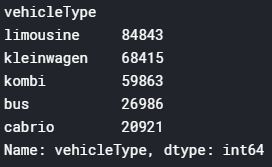
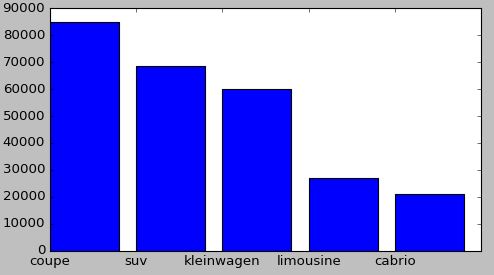
**Data preprocessing and data cleaning**

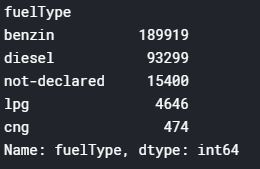
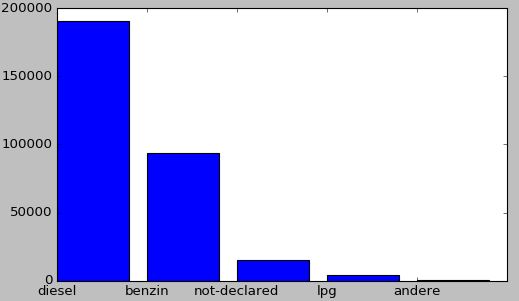
The reason behind the data preprocessing is to transform raw data to a useful one and also to reduce the data size so that it becomes easy to analyses. The name column is encoded as each element is different from each other. Name is used an integer instead of it as a string value add this as a column and drop ‘name’ for easy analysis. There are two columns of the same type and most of the given data belongs to only one type. Hence this column can be dropped. Similarly there are two types of offers and one of them is in majority so it can be dropped. There are null values as well in the dataset instead of deleting the rows a dictionary can be defined to put the same data set with the models as keys. The year of registration is considered only when the range is distributed highly.

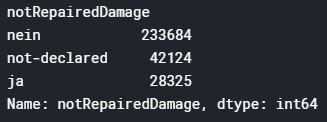
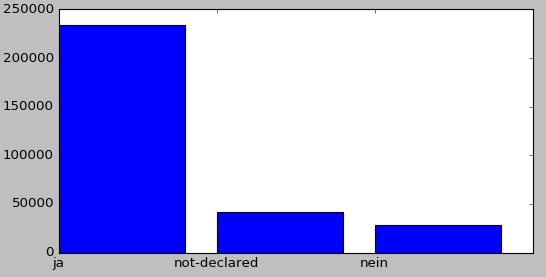


** **

** **

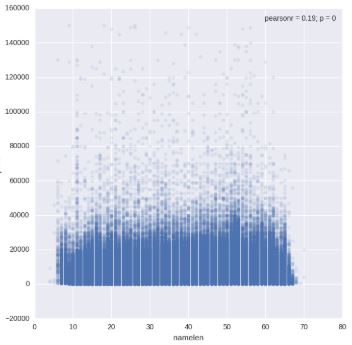
** **

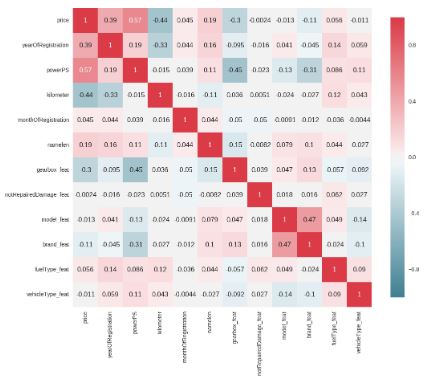
** **

** **

**Data visualization**

After the data extraction and data preprocessing steps, the dataset should now be visualized to have an insight of what is happening under the hood and how the data is distributed. The number of null values in the given dataset are visualized using a heat map. The count of different categories is checked to see which categories influence the prediction the most. Then gearbox, fuel type, not repaired damage are encoded and unused data columns are dropped i.e. number of pictures posted, postal code, last seen, date crawled.



1

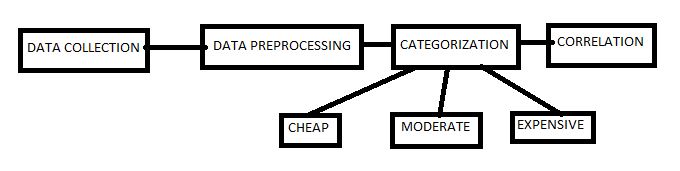
**Data exploration**

After completing all the pre-processing a correlation between the different variables can be observed.

**Data modeling**

The main task is data modelling, the dataset is divided into train and test sets. Here the data set is into train and validation data and tune the right-skewed sale price column.

Architecture



Related work

In her MSc thesis, Listiani showed that the regression mode build using support vector machines (SVM) can estimate the residual price of leased cars with higher accuracy than simple multiple regression or multivariate regression.

In another university thesis, Richardson working on the hypothesis that car manufacturers are more willing to produce vehicles which do not depreciate rapidly. In particular, by using a multiple regression analysis, he showed that hybrid cars (cars which use two different power sources to propel the car

Gonggi proposed a new model based on artificial neural networks to forecast the residual value of private used cars. The main features used in this study were: mileage, manufacturer and estimate useful life.

Conclusion

Used vehicle market involves many factors when it comes to predicting the fast-selling vehicles that maintain profit and reduce inventory cost for the retailers. The main aim of the project is to predict the price of second-hand reconditioned and second-hand used cars the average residual value was reasonably low for all the approaches. Thus, we conclude that predicting the price of second-hand cars is a very risky enterprise but which is feasible. This system will be very useful to car dealers and car owners who need to assess the value of their cars. In future research we can explore other factors that influence the sales period of a used vehicle. For example, the level of fuel-efficiency, whether the vehicle is electric or hybrid, level of discount from the original price. Incorporating these factors in the analysis can improve the accuracy to choose non-overage vehicles and have a positive impact on profit.

References

[1] Agency za statistiku BiH. (n.d.), retrieved from: http://www.bhas.ba. [Accessed July 18, 2018.]

[2] Listiani, M. (2009). Support vector regression analysis for price prediction in a car leasing application (Doctoral dissertation, Master thesis, TU Hamburg-Harburg).

[3] Richardson, M. S. (2009). Determinants of used car resale value. Retrieved from: https://digitalcc.coloradocollege.edu/islandora/object /coccc%3A1346 [accessed: August 1, 2018.]

[4] Ho, T. K. (1995, August). Random decision forests. In Document analysis and recognition, 1995. proceedings of the third intern

[5] Russell, S. (2015). Artificial Intelligence: A Modern Approach (3rd edition). PE.

[6] Aizerman, M. A. (1964). Theoretical foundations of the potential function method in pattern recognition learning. Automation and remote control, 25, 821- 837.