Predicting the Resale Value of Used Cars: A Data-Driven Approach

ORIE 5751 Jiehan Wu (jw2568) , Di Wang (dw624)

1. Introduction

The goal of our project is to build a prediction model to predict the prices of used cars. The project has important real-life implications, focusing on two different aspects:

- For used car retailers, this can create more efficient and accurate pricing decisions, and more accurately predict the resale prices based on empirical historical data.
- For customers, there will be an easy-to-use tool to check the purchase prices provided by the retailers, and thus help them make better buying decisions.
- For research purposes, the model can help researchers gain better insights and understanding of used car markets.

Given the potential benefits of our project for both the enterprise and the wider research community, we believe that it is worthwhile for us to invest our time and resources into it.

2. Data

For this given project, we collected data from Kaggle in the form of csv files. The csv file will be uploaded into Jupyter Notebook using read_csv function for easier data processing. The file that we have chosen has the following entries: carID, brand, model, year, transmission, mileage, fuelType, tax, mpg (miles per gallon for fuel efficiency), engineSize, out of which we will be mainly focusing on the entries other than carID, brand and model.

The transmission and fuel type entry are in nominal values and categorical, for which we won't be converting them to other data types since there are only 3 different categotries in each entry. The other data are all in numerical values which are straightforward and ready to be used. We plan on doing the following process for data processing and visualization:

1) Data cleaning (e.g. clearing out the NaN value from the table)
2) Select the columns that we will use to predict process (as narrated above)
3) Visualize the data from the following aspects:
    a) A correlation table for all the entries we will be using as mentioned above
    b) A histogram of number of car sales grouped by their production year
    c) A box plot of year of the car (on x axis) and the pricing (on y axis)
    d) Two box plots for transmission vs pricing and fuel type vs pricing respectively

And after all these steps to help us and the audience better understand our dataset, we will be using SVM to build the model for prediction. We chose SVM because it works well with high-dimensional data and non-linear relationships. By using SVM, we can build a more accurate prediction model that takes into account multiple factors that affect the resale value of used cars.