USED CAR VALUE PREDICTION MODEL

PROBLEM STATEMENT:

Car Value Prediction Is Essential For Individuals and Companies

Each day, thousands of pre-owned cars are sold worldwide. Prediction of the second hand vehicle price provides an important benchmark to both private buyer and the seller as well as business professionals such as car dealers.

When bank gives out a loan, they have to check the value of the pre-owned vehicle. Since the bank will be holding the title, they need to make sure the the vehicle is priced accurately. Insurance companies alike need to be able to assess the value of the pre-owned vehicles, since they will be calculating premiums. Moreover, in case of a totaled car, an insurance company will be paying its value. It is mandatory for insurance companies that they predict second hand car values so they can make their risk assessments.

The used car market is also a large and strategically important market for car manufacturers since it is closely connected to the new car business. Trading-in used cars in new car retail sales and handling lease returns, repossessions and fleet returns from car rental companies necessitate car manufacturers to engage in the used car market. Therefore, car makers require sophisticated decision support systems to sustain the profitability of the used car business.

Edmunds, Kelley Blue Book, NADA Blue Book are popular car valuation companies, whose services are used by millions of professionals and individuals each year. These companies have statistical models, massive databases and they use machine learning algorithms to effectively predict the value of innumerable car brands and models.

PROBLEM AREAS:

Dataset is not professionally prepared, both continuous and binary variables, multi-dimensional

Databases of major automotive information companies are confidential, therefore for this capstone project, a publicly available dataset downloaded from a German used car sale website will be used.

Desired end state is to be able to predict the value of any brand and model, given the dataset comprising of current car prices, brand, model, year and condition information. Since the dataset is not prepared by professionals, but by private sellers and dealers, there are many arbitrary elements as well as missing and erroneous data.

Dataset needs precise cleaning because it has data points entered by users that are off scope, where prices, brand, model names can be wrong or misleading. Exploratory data analysis reveals numerous outliers on the price column because the advertiser entered monthly rental fee to a sale website, or entered arbitrary numbers to the price.

A filter stage needs to be developed to filter out the erroneous data in order to have a healthy prediction model. Data manipulation is also required to feed the prediction model with meaningful information. For example year of a specific vehicle may not be too much of a value for the model, but age may be.

Once data cleaning has been done, the correlation between factors affecting the pricing has to be measured. While some price influencing factors are continuous variables, such as age or mileage of the vehicle, some are discreet and binary For example the transmission type and fuel type are not continuous variables and they affect the price. This further complicates the problem and a classification algorithm may be required.

Since various methods of machine learning can be utilized, the most effective method must be identified and employed. A revisit to the dataset may be required if the results are not satisfactory, since truncation of meaningful information or another flawed approach may have caused problems.