Ryan Paik

Final Report:
Used Cars Price Prediction

Problem Statement

It is difficult to know if the value of a used car is listed at a fair value since there are a variety of features that affect vehicle price. As a buyer, it is troublesome to know what features and to what degree do these affect the price of a vehicle. With this data and model, the purpose is to help buyer and seller's have an estimated price to know if they is at a fair value.

Data Wrangling & Pre-Processing

There were 2 datasets that were used from Kaggle, brand new vehicles and used vehicles. The new car dataset contained the MSRP, which identified the starting price of a vehicle at perfect condition. The new car price dataset contained 32,316 rows and 57 columns. The dataset was trimmed down to the model of the vehicle and the MSRP. There were many other features that were listed with the vehicle, but those features were unnecessary and removed.

In the used car price dataset, the dataset contained 458,213 rows and 25 columns. Since there were too many features, I had to reduce the number to lower the dimensionality. The number of features was reduced from 25 to 12. All null values were removed given I did not want to skew the data with the mean or median. The new shape of the dataset is 90,817 and 12 columns.

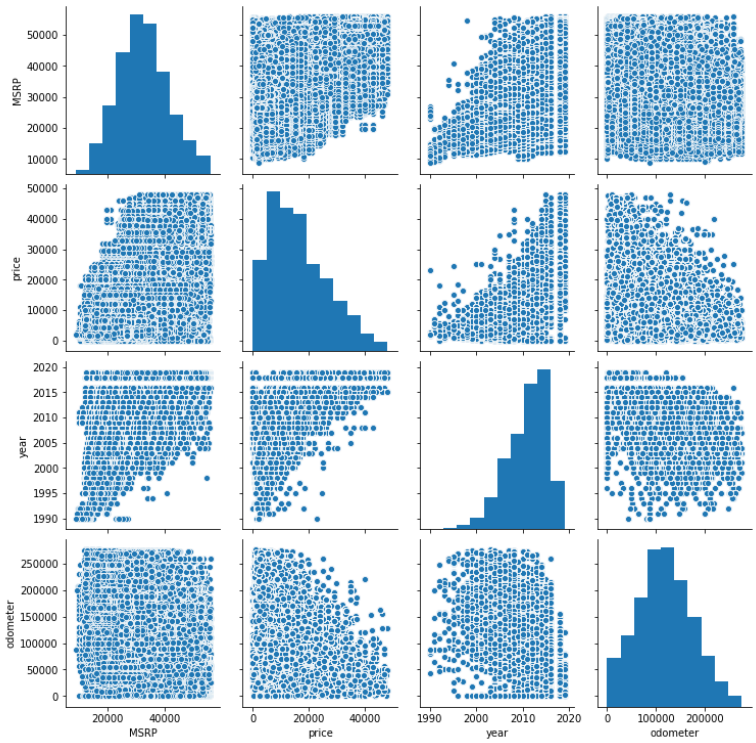
The 2 datasets were merged based on the model of the vehicle and the final dataset shape was 304,070 and 12 columns. The increase in rows is due to the increase of the same type of model vehicle. IQR was used to identify any outliers and the outliers were removed to limit skewness of the data. The final shape of the data used was 267,067 and 12 columns.

Exploratory Data Analysis

The following 12 columns within the merged dataset:

- Model
- MSRP
- Price
- Year
- Condition
- Cylinders
- Fuel
- Odometer
- Transmission
- Drive
- Size
- Type

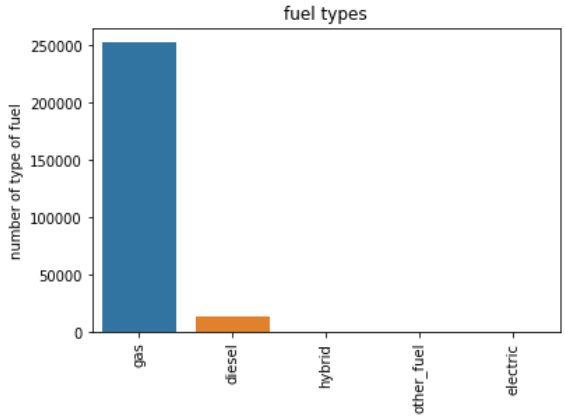
Within the 12 columns, only 4 of the columns contained numerical values: MSRP, price, year, and odometer. The 4 columns were placed in a pairplot to have a visualization of the data.

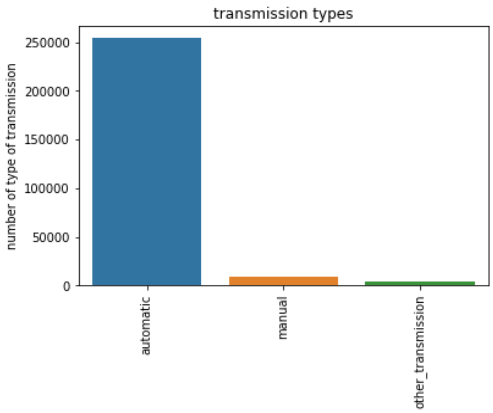


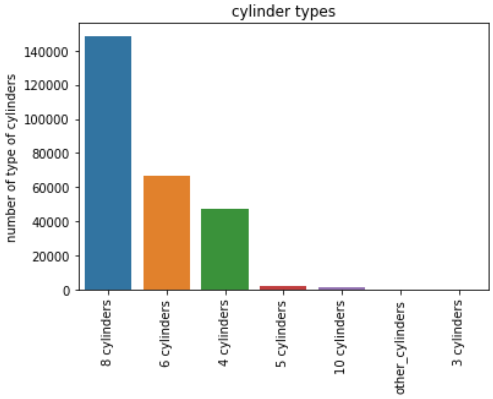
The odometer and MSRP values can be seen to have a normal distribution whereas the used car price and the year of the vehicle are skewed to the right and left, respectively. To see if there are any correlations, a heatmap was made.

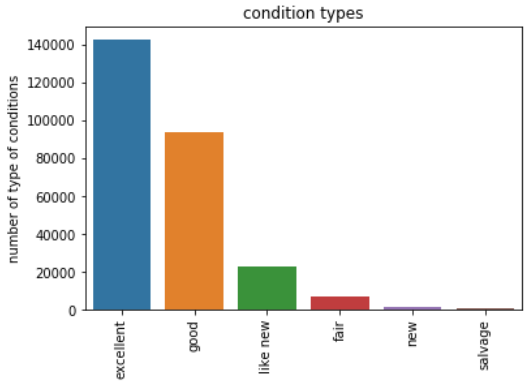
The heatmap helps visualize the correlations among the numerical values in the data. There is a positive correlation with the MSRP and the year of the vehicle. Depending on the model of the vehicle, the starting price (MSRP) depreciates to the used car price value over time. The MSRP also had a positive correlation with the years, since the more recent years will have the higher priced vehicle and as the vehicle gets older, the price will deprecate in value. Similarly, this also can be seen between the used car prices and the years. There is a negative correlation between the odometer with year and price. As the vehicle has more mileage, the price of the value with depreciate. Also with the year, it can be assumed that newer vehicles will have a lower mileage than a vehicle that has been used for more years. The MSRP and odometer has a -0.2 correlation, nearly 0, given that the MSRP is based on a vehicle with 0 mileage.

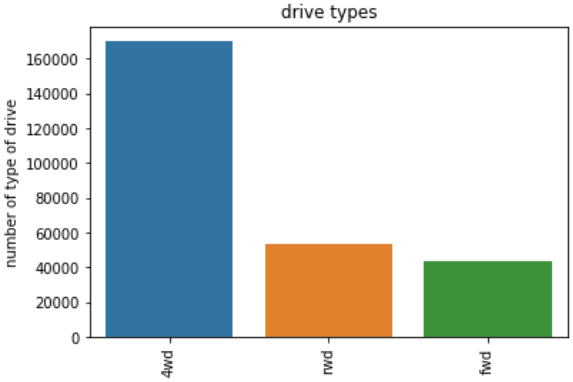
The other columns had categorical data and a bar plot were made for these features to have a quick visualization. The Model column bar plot was not made.

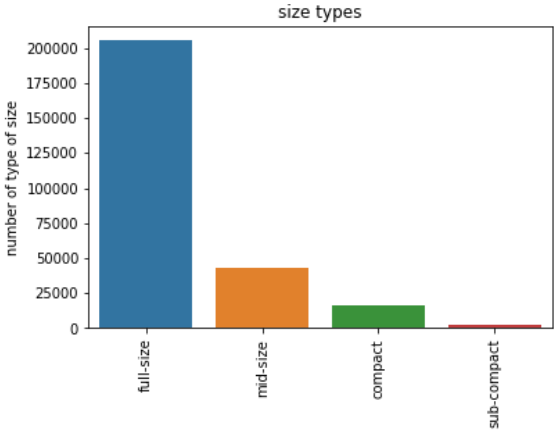


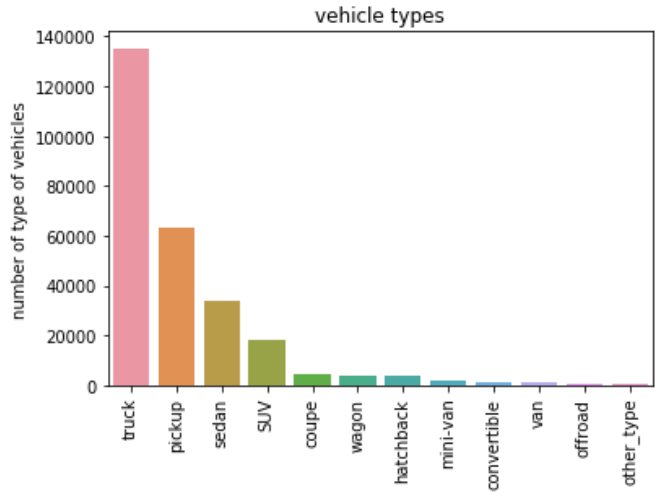












Model Selection

I tested 4 different machine learning regression models: Linear Regression, Decision Tree Regression, Random Forest Regression, and Gradient Boosting Regression. The model would take the different features and predict the value of the used price vehicle. In consideration to avoid overfitting, the type of vehicle model was not included in the machine learning model.

Prior to building the models, since there are 7 features with different categories, they needed to be turned into a numerical value to be compared to one another. This process was done using the get_dummies feature. After using the get_dummies feature, this led to the data having 43 columns. Once the data was scaled to have numerical values, the data was scaled to have consistency units across all the data. This was done using the Standard Scaler. The model was now ready to be created to be made into a train and test split. The test size was 20%. Without any tuning, Decision Tree, and Random Forest had an accuracy rate of 0.99% with a R2 score of 0.99. There were some fears that the model was overfitting. Since Random Forest had the highest percentage at 99.4%, a quick sanity test was done to compare the values that the model predicted.

The prices predicted were close; however, I wanted to see if there can be any adjustments made on all models with hyperparameter tuning. A quick cross validation was done on al models and similarly to our initial model results, the Decision Tree and Random Forest model had average scores of 0.99. This was to get a quick glimpse to see if there were any minute changes but hyperparamter tuning was still done on all models except for linear regression. GridSearchCV was used to do the hyerparamter tuning and this produced interesting results. The Gradient Boost model produced the highest validation score with a 98% however, the Decision Tree had an average validation score of 72% and the Random Forest had an average validation score of 88%. This means that the Decision Tree and Random Forest model were overfitting. This can conclude that the model that will be used will be the gradient boost model.

Future Improvements

There are some things to consider, and many improvements can be made. One concerning feature is that the used car prices are listed with their conditions based on the seller's opinion. There is not a standard scaling measurement on the car's features such as the condition of the vehicle. When there is a similar grading across all vehicles, then there can be a higher consistency among the prices. Furthermore, the gradient boost model can be made to now include future data. For example, a buyer can input their desired vehicle with the desired specs to output a list of vehicles within the buyer's parameters. Overall, I believe with this model and the dataset, this is a great start to identifying and helping buyer is get a value of a used vehicle.