Capstone Project: Used Cars Price Prediction
Final Report

Executive Summary

An Extreme Gradient Boosting (XGBOOST) Model is proposed for the prediction of the prices of used cars in this project. There is a huge demand for used cars in the Indian market and this has resulted in a decrease in the sales of new cars. XGBOOST has quite a few features that improves its performance such as built support for regularization as well as giving additional hyperparameters to tune such as tree pruning and number of decision trees amongst others. However, it has its limitations such as the time taken in training the data. The most impactful features recommended by this model should be considered in bringing about an improvement in the number of used cars sold which will in turn result in an increase the profits from selling used cars.

Problem Summary

Unlike new cars where the price is deterministic and managed by Original Equipment Manufacturer (OEMs), such is not the case for used cars. Several factors such as mileage, brand, model etc. can influence the valuation of a used vehicle. It is important to solve this problem so both buyers and sellers of used cars have a better valuation of a used car. Having a better estimation of the value of a used car will help sellers not make a loss and buyers to not overpay for a car. It is important to identify which factors influence the price of used cars in order to know what steps and decision to take in order to improve your business and not to waste resources on the wrong departments.

The dataset had a lot of missing data which could highly bias the outcome of the models. The missing data was treated. The price and Kilometers_Driven variables were scaled down by applying log transformation to get rid of skewness and normalize the distribution.

Solution Design

A few regression models and ensemble methods were explored as part of the solution design. These includes models like linear regression, ridge, decision tree, random forest, hyperparameter tuned decision tree and random forest and Xgboost regressor.

The OLS linear regression had good variability and the r squared barely changed after using the wards test and VIF(variance inflator factor) to get rid of insignificant columns.

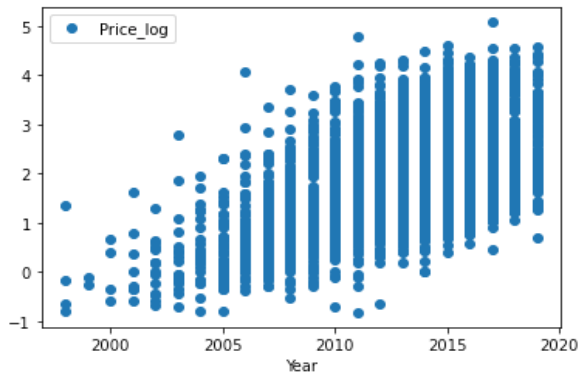

Figure 1: Scatter Plot Price vs Year

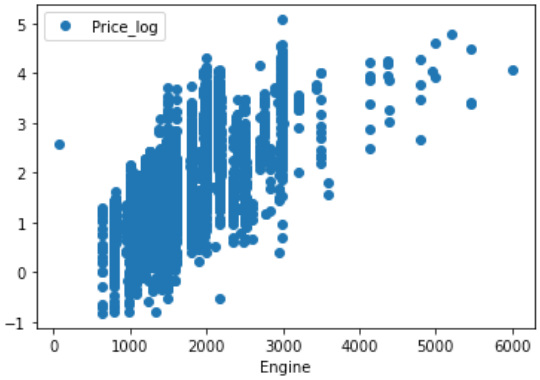

Figure 2: Scatter plot Price vs Engine

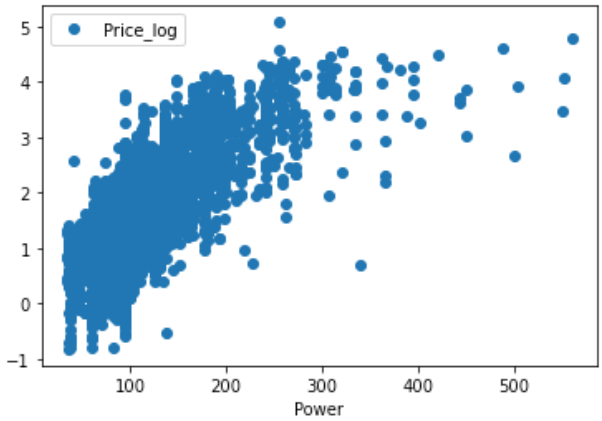

Figure 3: Scatter plot Price vs Power

These three features have them most correlation with the target variable.

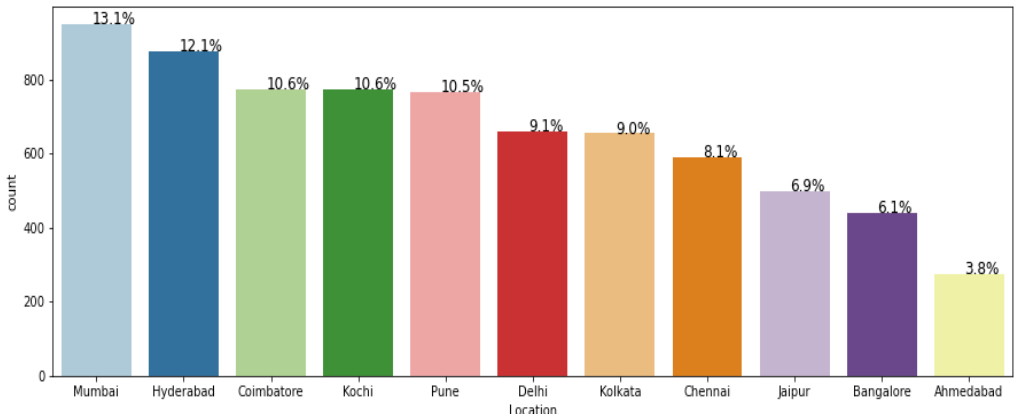

Figure 4: Bar Plot for Locatkion

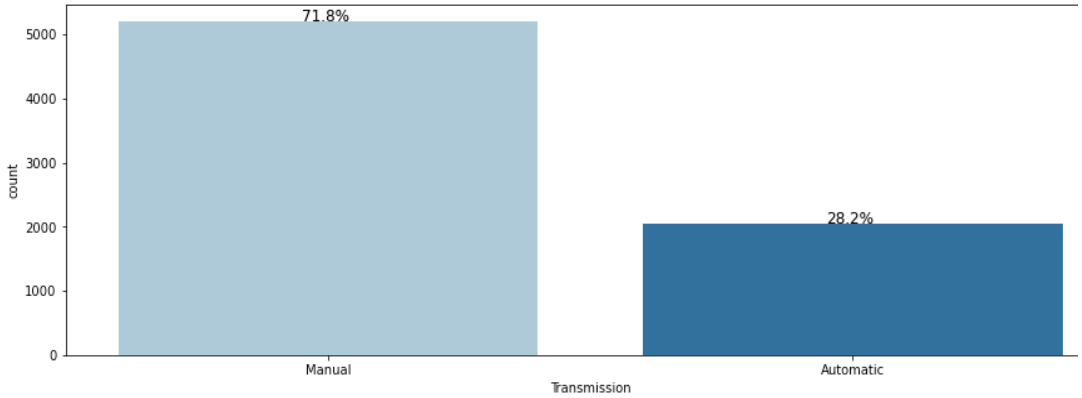

Figure 5:Bar Plot for Transmission

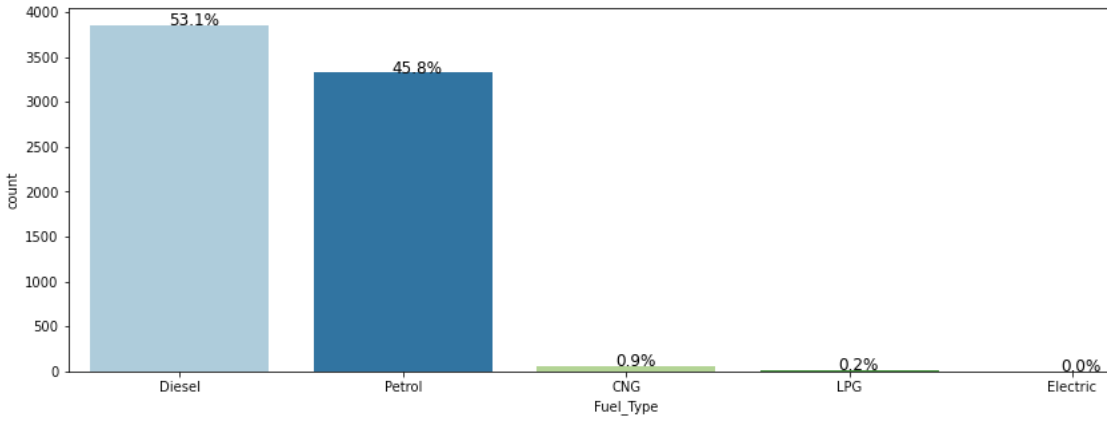

Figure 6:Bar Plot for Fuel Type

From the bar plot , it is observed that Mumbai, Hyderabad and Coimbatore are the cities where the most used cars are sold while

Ahmedabad is the city with the least. Newer cars were also sold significantly more than older cars. The fuel type used by most of the cars is Diesel closely followed by Petrol. It is difficult to state why this is so from just this data. About 70% of the cars sold have a manual transmission.

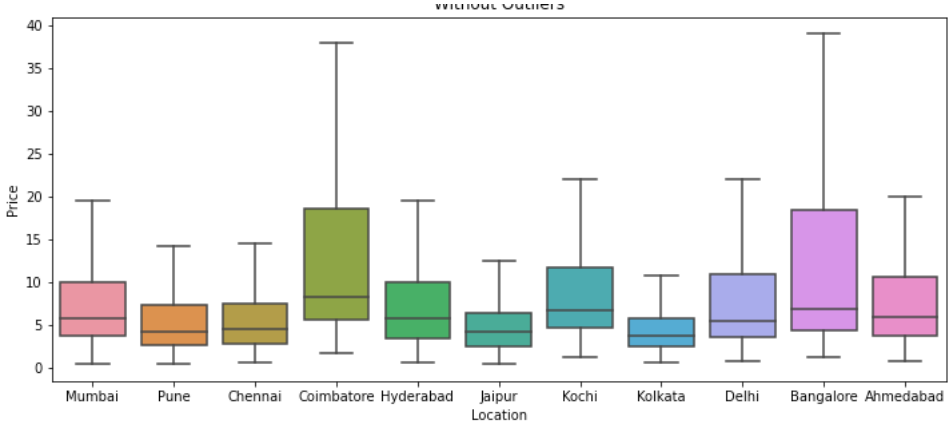

Figure 7: Box Plot-Price vs Location

From the above boxplot it is evident that Coimbatore makes the most money from selling used cars which probably means they sell the more expensive used cars while Kolkata makes the least money from selling used cars. This means Coimbatore is one of the best location to sell used cars.

After thoroughly analyzing the data and ridding it off missing data as well as dropping some features, the model building process commenced. Firstly, the necessary libraries were imported then the data was split into train and test data to enable the evaluation of the model's performance.

Decision tree and random forest overfit the data but after hypertuning the models using grid search cv, both perfomed relatively better. Linear regression and regularization technique Ridge have relatively the same performance.The coefficient of the determinant r squared , which is the variance in the target variable that is predictable from the independent variables tells us that the data bests fits the

XGBOOST Regressor model. This model also has the lowest Root Mean Squared Error.

Model	Train_r2	Test_r2	Train_RMSE	Test_RMSE
Linear Regression	0.861378	0.860809	4.159755	4.158004
Ridge	0.861341	0.860791	4.160308	4.158271
Decision Tree	0.999997	0.827663	0.020693	4.626672
Random Forest	0.974964	0.857122	1.767817	4.212711
Decision Tree Grid_Cv	0.836065	0.791020	4.523631	5.094850
Random Forest Grid_Cv	0.848269	0.810148	4.352002	4.856092
XGBOOST Regressor	0.988866	0.908525	1.178874	3.370788

Figure 8: Results of the perfomance of the models

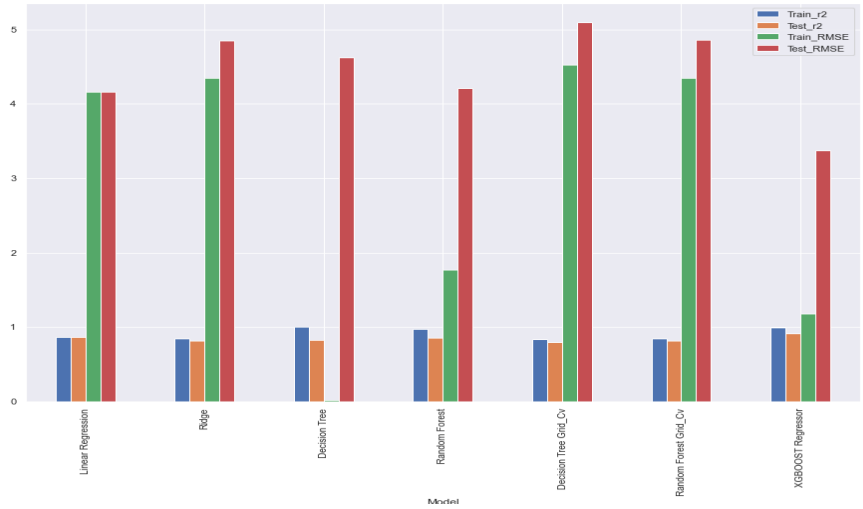

Figure 9: Bar Plot of the results

The final proposed solution is the Extreme Gradient Boosting(XGBOOST) which was optimized using the GridSearchCV parameter tuning to tune parameters such as the loss, learning rate, n_estimators and max_depth. Despite performing relatively better than the other models, the XGBOOST regressor does have its limitations. It takes a long time to train the data.

```
parameters = {
                "loss": ['squared_error', 'absolute_error','huber'],
                "learning_rate":[0.1,0.01,0.3],
                "n_estimators":[200,300,400,450],
                "max_depth": [3,4,5,6,7]

    }
Gbr= GradientBoostingRegressor()
grid_cv_gbr = GridSearchCV(Gbr, parameters, cv=5)
grid_cv_gbr.fit(X_train, y_train['Price_log'])
print("Best Hyperparameters::\n{}".format(grid_cv_gbr.best_params_))
```

```
Best Hyperparameters::
{'learning_rate': 0.3, 'loss': 'squared_error', 'max_depth': 3, 'n_estimators': 450}
```

Figure 10:Best Parameters for Gbr using GridSearchCV

The gradient boosted regression is an ensemble method which combines multiple decision trees to create a more powerful model. GridSearchCv is used to find the optimal parameter values from a given set of parameters in a grid. It took quite some time to obtain these parameters probably because GridsearchCV checks every single parameter value provided to find which one is optimal. A learning rate of 0.3, loss chosen is “squared_error”, max_depth of 3 and n_estimators of 450.

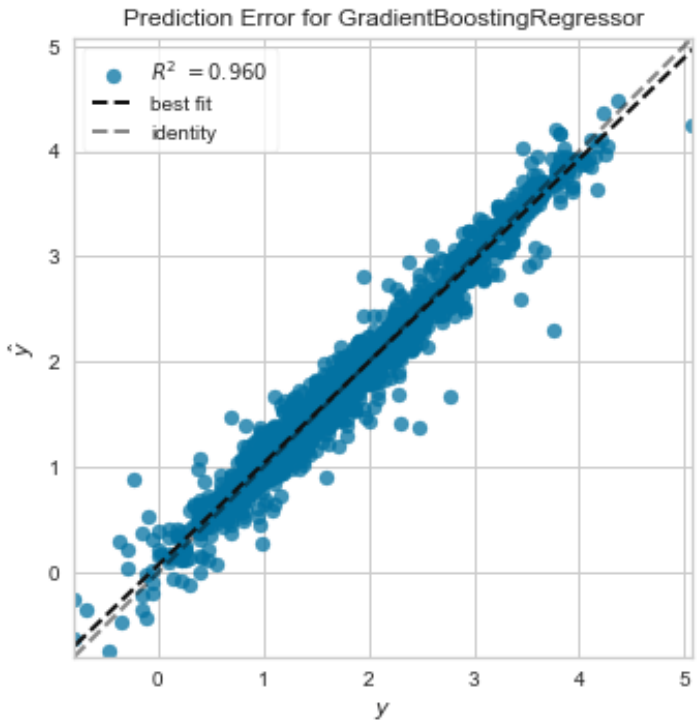

Figure 11:Prediction Error for GradientBoosting Regressor

Recommendations for Implementations

A lot was observed from the data explanatory phase. From our heatmap it was clear that Power, Engine and Year have the most positive correlation with the target variable Price_log. From the results obtained from the XGBOOST regressor, these three features had the most impact on the target variable.

Cars purchased by the business for sale should be cars of good engine, power and be as new as possible since these are the type of cars that bring in the most revenue.

Cities such as Coimbatore, Kochi and Pune and perhaps Delhi should be targeted as possible marketable destinations to sell used cars. Mumbai and Hyderabad market for used cars could already be saturated which would yield a low revenue output.

Cities like Ahmedabad , Bangalore and Jaipur have the lowest sales of used cars. It will be important to find out why. It could be from a lack of availability which would be an advantage to any used cars sales business because this would mean these cities are untapped revenue sources.

Just cars with Diesel and Petrol as their Fuel type should be sold. Getting a car of a different Fuel type will prove difficult to sell because buyers are not interested in vehicles of a different Fuel type.

Majority of the cars sold should have manual Transmission because majority of the people are interested in Manual cars but Automatic should not be completely overlooked as they are more expensive and will bring in more money.

Challenges

There are several other factors such as safety, comfort and entertainment features which buyers could consider while purchasing a used car. The lack of such features could or will prove to be a problem. Airbags, Air conditioning, tyre conditions, working sound system are all features buyers might consider when purchasing a used car.