**Used Car Price Prediction Model**

**Final Semester Project (STA-6704) – SUMMER 2021**

**Syed Muhammad Sabih**

**Final Project Executive Summary**

**PROJECT STATEMENT**For the US Cars Dataset, the analysis for this project pertains to build a Machine Learning Model that predicts the price of the car based on the imperative features.

**THE DATASET**The dataset used for the analysis in the project is from the Kaggle repository and was scraped from auctionreport.com. The dataset includes information about 28 brands of clean and used vehicles for sale un the US. The dataset comprises of 11 features: *price, year, brand, model, color, state, mileage, VIN, title\_status, lot,* and *condition.* The target variable for the analysis is *price*.

**EXPLORATORY DATA ANALYSIS**For the exploratory data analysis:

* Any irrelevant features were removed and the only features that were retained for the analysis were: *price, brand, model, year, title\_status, mileage, color, and state.*
* Further in the analysis, the features were investigated to have a distinction between categorical and numeric features. The decision was made to modify the *year* feature to reflect the age of the car as a numeric variable.
* The dataset was checked for duplicate records and four duplicate rows were removed.
* The dataset was investigated for any missing values and no missing values were found.
* The brands of the vehicles in the dataset were investigated, which resulted in removing 181 records, which included the top 5 brands in the dataset.
* Upon visualizing the numeric variables, the results showed that the data was skewed to the left, and a strong correlation between *year* and *mileage* was seen.
* The categorical variables were visualized which gave insight into the distribution of *brand, title\_status, model, color,* and *state.*

**DATA PREPARATION**After the exploratory data analysis, the data needed to be prepared for modeling. The process for preparation is as follows:

* The data was split into two dataframes with categorical and numeric variables, which enabled to OneHot Encode the categorical variable to produce binary columns for each level of categorical predictors.
* The numeric and encoded dataframes were concatenated and the input variables were scaled using StandardScaler().

**PREDICTIVE MODELING**The data prepared for modeling was split in to 70% *training* and 30% *testing* dataset. The analysis for modeling is as follows:

* Using K-Fold cross-validation on the fitting on the *training* dataset, ***Random Forest Regressor, XGBoosting Regressor, K-Nearest Neighbors Regressor, Decision Tree Regressor,*** *and* ***Gradient Boosting Regressor*** were evaluated using the scoring measure: r-square. The best performing algorithms were *RandomForestRegressor() and XGBRegressor()* with an r-square value of 66% and 68% respectively*.*
* *Random Forest and XGBoosting* were evaluated on making predictions for the *testing* dataset with the scoring measure of r-square. The best performing algorithm was *XGBRegressor()* with an r-square value of 73.9%.
* *XGBRegressor()* was enhanced with hyperparameter tuning and was evaluated on making predictions for the *testing* dataset with the scoring measure of r-square. The algorithm improved and the r-square value increased to 75.4%.
* A substantially fit model was produced to predict the *price* for used cars.