# Introduction

This paper is part of final report of academic project on Predictive Analytics. In this paper we will discuss the steps and techniques used to train a predictive model and conclude by evaluating model accuracy and feature importance.

Dataset used in this project is a consolidated record of used car advertisement in Craigslist. Dataset is part of monthly web scraping project hosted in GitHub and Kaggle, it contains every used vehicle entry within the United States on craigslist. As of now, dataset contains 423,857 records with 25 feature columns. Some of the important feature columns includes price of vehicle, year of manufacture, odometer reading, manufacturer, model, condition, number of cylinders, fuel efficiency, transmission type, etc. Dataset is quite large and includes cars as old as ones manufactured in 1965, to make sure the relevance of prices in current market, only the relevant records for the price prediction model will be selected.

Price of car is a continuous variable; hence we will be using regression techniques to fit the model. To evaluate and compare the performance of multiple regression model, we will model the same set of data to three model.

## Model Selection

We will train 3 models and the reason for selection of these models is discussed below:

**Multivariate linear regression**: Variable like odometer and year of manufacture have linear relationship with target variable. Although, rest of the feature may not have similar linearity, we will fit a multi variate linear regression model to evaluate performance difference among other sophisticated models.

**KNN regressor**: KNN regressor can be effective in this case because we are going to take reference of user’s price estimates of each car advertisement in craigslist. Using KNN regressor in this case practically allows us to take average of n nearest neighbor selling similar car in past to estimate value of new car. Since the dataset is significantly large, it can be a little show as it tries to reference all the record from training set. We will, experiment the accuracy at different value of K.

**Gradient Boosting**: Considering the complexity and dimension of dataset, an ensemble learning technique is a good idea. XGBOOST uses boosted tree algorithm with the ability to parallelly construct multiple trees utilizing multiple CPU cores during training. It uses gradient descent to minimize the residual of prior model and to achieve close to accurate result in final prediction at optimal cost. Ideally, this model is expected to have higher accuracy compared to multivariate linear regression and KNN regression.

# Implementation and Analysis

### Missing Value

As the dataset is direct result of scrapping craigslist’s used vehicle marketplace, there are many missing values, around 32% of data is missing. The reason for huge missing value is because few parameters were added later in vehicle advertisement form by craigslist and rest of the missing value must be the case where user chose to submit blank field in the form.

As part of cleaning the dataset, records missing all three important feature Year, model and odometer were dropped. The missing odometer reading was replaced by the average odometer reading for the respective Year of manufacture, with this we were able to fill missing value on odometer column with most accurate odometer reading.

Since the Model of car is a free text field in advertisement form, there are more than 126 thousand unique model, using this in the model will be extremely ineffective. As an attempt of standardizing model of car, we will filter out car model which have less than 200 frequency in dataset. Now, since model column is standardized, we will use it to fill the other missing value too. To fill the remaining missing value, a lambda function was used which replace missing value with mode or mean of respective model. This function works by grouping models and calculating mode for categorical column and mean for numerical column. The missing value are replaced with reference to the grouped table.

### Outliers

There are three numerical variables in dataset: price, odometer, and year. Outliers in any of these variables can be a nuance in model and can significantly reduce model’s ability to predict the target. In case of Price and Odometer reading, typos with one extra zero can be an outlier. To remove such instance, the advertisement priced below $200 and above $120,000 has been removed. Similarly, odometer reading above 750000 has been removed. Also, to make the model predicts the relevant price of car in modern market, all the records of cars manufacture before 2000 has been removed.

## Exploratory Data analysis (EDA)

EDA was performed on dataset to understand the distribution of data. EDA on this dataset can help us understand the market share of brand and which models were more famous and have good resale value in preowned vehicle marketplace.

Few of the average finding from EDA were, the average price of vehicle sold in craigslist is around $12000, Ford has biggest market share in used car market with their F-150 being most popular followed by explorer.

### Encoding

Before training the model, all the categorical variable must be changed to numerical variable for the compatibility of modeling algorithm. When encoded using one hot encoding technique, number of columns goes as big as 256. This can slow down model, training and testing can be very time consuming. Hence, in this case, we will use label encoder to change categorical variable to numerical variable.

### Scaling

Since the distribution of data on different columns is in different range, one feature can act dominant over others. To avoid this, we will scale the data using standard scaler. Using standard scaler, values has been transformed in the range of -1 to 1.

## Modeling and evaluation

Before fitting the model, we will split the dataset into training and testing set in 7:3 ratio. The model with be training using training test and we will evaluate fitted model using testing set to ensure the models are overfitted. We will use two metrics, Root Mean squared error (RMSE) and R squared score to evaluate model.

### Multivariate Linear Regression

Evaluation of Multivariate linear regression yielded **RMSE value of 5936.17 and R-squared score of 0.7176**. This indicates, model has accuracy of around 70% which is not ideally great. However, considering the complexity of dataset and nonlinear relationship of few feature with target, accuracy is ambitiously good.

### KNN regression

As expected, KNN regression performed better than multivariate linear regression. Evaluation of KNN regression model at different K values are listed below:

RMSE of Model is : 5067.319 when value of k is 1

R square of Model is : 0.794 when value of k is 1

RMSE of Model is : 4746.426 when value of k is 2

R square of Model is : 0.819 when value of k is 2

RMSE of Model is : 4622.365 when value of k is 5

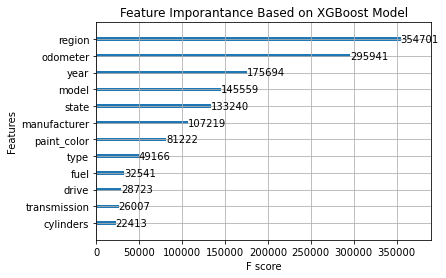
R square of Model is : 0.829 when value of k is 5

RMSE of Model is : 4638.722 when value of k is 10

R square of Model is : 0.828 when value of k is 10

Accuracy of the model was found optimal at K=5, model yielded RMSE value of 4622.36 and R-squared value of 0.829. KNN regressor worked well in this model because, we have huge sample of training which KNN use as reference for each prediction. Prediction is made based on average target value of K-nearest neighbor, in this case 5 nearest neighbor records.

### Gradient Boosting

 Gradient booting models were configured with following hyperparameter after multiple iteration of testing, learning rate was set to 0.4, max depth allowed was 25 and alpha (regularization) value was set to 8, number of estimators were set to 200. With this configuration in XGBOOST function, we were able to fit the model with good accuracy. Gradient boosting model, when tested with testing set yielded RMSE value of 3753.50 and R-squared score of 0.8871.

Feature importance plot based on fitted model indicates region where the advertisement was posted. This indicates, the same car can have drastically different value in different regional marketplaces. As expected, the following important features are odometer reading, year of manufacture and model of the car.

# Conclusion

From above three predictive model, we can conclude that ensemble learning algorithm XGBOOST based on gradient boosting is better compared to KNN regression and multivariate linear regression when working with data with high dimension and with complex relationship between feature variable and target variable.

As we found region of sales is very important, this indicates the scope of improvement of model by utilizing another geographical feature indicating latitude and longitude of advertisement posted. We can also work on standardizing model column more which will for sure improve the cleanliness of dataset and influence the filling of missing values in a positive manner.

The application of this model or similar predictive model can be a webapp. For this model, a webapp can be utilized by public to estimate the price of their vehicle if they wish to sell. This can also be used by craigslist themselves to suggest the price estimation to user trying to post a new add to sell their vehicle in craigslist.

References:

Dataset: <https://www.kaggle.com/austinreese/craigslist-carstrucks-data>

Scrapping project: <https://github.com/AustinReese/UsedVehicleSearch>