Naveen Bagam

Bellevue university – Master of Science in Data Science

Course Name: DSC550-T302 Data Mining (2231-1)

# Project: Predicting used car prices

# Instructor: Brett M. Werner

Due Date: 11/19/2022

# Predicting used car prices using Regression:

**Introduction: -**

Approximately [40 million vehicles sold](https://static.ed.edmunds-media.com/unversioned/img/industry-center/insights/2019-used-vehicle-outlook-report-final.pdf) each year. Effective pricing strategies can help any company to efficiently sell its products in a competitive market and make a profit.

In the automotive sector, pricing analytics play an essential role for both companies and individuals to assess the market price of a vehicle before putting it on sale or buying it.

There are two main goals I want to achieve with this Data Science Project.

First, to estimate the price of used cars by considering a set of features, based on historical data.

Second, to get a better understanding of the most relevant features that help determine the price of a used vehicle.

The data that will be used for this project is accessible at [Kaggle](https://www.kaggle.com/austinreese/craigslist-carstrucks-data) and has been scraped from Craigslist, the world’s largest collection of used vehicles for sale.

The Database consists of 426,881 rows and 26 features, one of which will be the continuous dependent variable (“price”) that I want to predict.

**Organized and detailed summary of Milestones 1-3:**

The numerical features play a big role in this Regression model, so it is important to understand well how they are distributed in the Database.

As part of this project, I performed EDA, and found outliners and NaaN values/ missing values from the dataset.

I extracted Key information from “description” feature to fill NaaN, duplicates, and missing values.

there are values that have been added inconsistently. For example, we have these kinds of values in the same column: ‘awd’ and ‘ awd ’. To unify the strings, the strip() method is used.

Following shows outliners in the “Price” and “Odometer”

Chart

Description automatically generated

Since “price” had such a big difference between the minimum value and the 25% percentile as well as between the maximum value and the 75% percentile, I will be leaving out 10% of the values on each end.

“Odometer” feature has NaN values, so it is trickier to deal with. used scatterplot to visualize the outliers.

Chart, scatter chart

Description automatically generated

Since the outliers start at approximately 3,000,000; I have dropped the values that exceeds that. Also, it is a good idea to drop the minimum value (0) since it greatly differs from the 25% percentile. With this, the feature “odometer” has also been successfully adjusted.

“Year” Feature outliners Graphical user interface, application

Description automatically generated

From Above I see car data is available from 1900, which seems not a realistic, so I considered only data greater than or equal to 1950 for our model.

Filled Cylinders feature NaN values with drive feature.Graphical user interface, text, application

Description automatically generated

since "region" and "state" are connected, I have combined both of them, also dropped "posting\_date" feature, as it is not useful for model built.

Out of 13 features 10 are categorical, so applied Label encoder to each categorical feature to convert.

After data cleanup, I have applied "Random Forest Regression" model to get the "price" predictions, this model has the power to handle large data set with highest dimensionality and it won't allow over fitting to the model.

Later combined related columns “condition” and “title\_status”. since They are similar and basically tells us about the state of used cars.

Most of the title\_status feature data is clean, so we I have add "Clean" to missing values.

For condition feature missing values, I have divided them into bins and assigned the values “excellent” and “Good” values to the missing values.

Later dropped the “size” feature which has more missing values and not useful for the model training.

Also filled 25% NaN values features with 'ffill' ('drive','type','paint\_color').

In this project I have used "Random Forest Regression" model to get the "price" predictions, this model has the power to handle large data set with highest dimensionality and it won't allow

over fitting to the model

To start with, first I divided the data into Train and Test set, we can take 80:20 distribution.

First, I re-index the database and put the dependent variable "Price" as last column for a simpler splitter.

Then applied Feature scaling, as it is important to do it AFTER splitting the database to avoid data leakages in machine learning model.

To Fit the model in the Training Set using the Cross-Validation method to estimate the accuracy of the model more precisely. I utilized the R-squared metric to evaluate the performance of the model and transformed the values we get from R square into percentage.

We got 87.38%, which is not bad, but we tried to improve by tuning the hyper parameter, but the percentage didn’t get increased.

The score is even more accurate in its prediction in the Test Set (88.08%) is great.

**Conclusion:-**

For this project I have used a single model (Random Forest Regressor) to predict the price of used cars.it has shown an excellent performance in such a big dataset and it has performed consistently throughout the training and testing process, the test sets are even better than train set data, it gave 88.0% accuracy in predictions.

The goals of the project were to create a model that was able to estimate the price of used cars and we already achieved it.

With this project, we have built a model that can predict with a 88.00% of accuracy the price of used cars, given a set of features. This information can have an enormous value for both companies and individuals when trying to understand how to estimate the value of a vehicle and, more importantly, the key factors that determine its pricing.

As expected, the year of the vehicle is by far the main factor when calculating the price with almost a 43%, followed by odometer. I expected the state of the car and the odometer to be deeply related but there is a big gap in the difference of relevance between both measures.

That said, it seems the region also plays a part, which totally makes sense. There might be more general vehicles that are liked everywhere but specialized cars like sport or convertibles would be a better fit in warmer areas whilst bigger trucks and SUVs would play a better role in colder places.

Mastering the art of pricing is not an easy task, but with the study of historical data it is possible to find patterns that lead to accurate results. Acquiring this knowledge can provide you with a comparative advantage before putting a vehicle on sale or buying it on the market.

That said, it seems the region also plays a part, which totally makes sense. There might be more general vehicles that are liked everywhere but specialized cars like sport or convertibles would be a better fit in warmer areas whilst bigger trucks and SUVs would play a better role in colder places.

Mastering the art of pricing is not an easy task, but with the study of historical data it is possible to find patterns that lead to accurate results. Acquiring this knowledge can provide you with a comparative advantage before putting a vehicle on sale or buying it on the market.