MEF UNIVERSITY

BIG DATA ANALYTICS ON USED CAR INFORMATION

Capstone Project

Efe Demir

İSTANBUL, 2021

MEF UNIVERSITY

BIG DATA ANALYTICS ON USED CAR INFORMATION

Capstone Project

Efe Demir

Advisor: Asst. Prof. Dr. Utku KOÇ

ISTANBUL, 2021

MEF UNIVERSITY

Name of the project: BIG DATA ANALYTICS ON USED CAR INFORMATION
Name/Last Name of the Student: Efe Demir
Date of Thesis Defense: 01/02/2021

I hereby state that the graduation project prepared by Efe Demir has been completed under my supervision. I accept this work as a "Graduation Project".

01/02/2021

Asst. Prof. Dr. Utku Koç

I hereby state that I have examined this graduation project by Efe Demir which is accepted by his supervisor. This work is acceptable as a graduation project and the student is eligible to take the graduation project examination.

01/02/2021

Prof. Dr. Özgür Özlük

Director of Big Data Analytics Program

We hereby state that we have held the graduation examination of _____ and agree that the student has satisfied all requirements.

THE EXAMINATION COMMITTEE

Committee Member	Signature
1. Prof. Dr. Özgür ÖZLÜK	……..
2. Asst. Prof. Dr. Utku Koç	……..

Academic Honesty Pledge

I promise not to collaborate with anyone, not to seek or accept any outside help, and not to give any help to others.

I understand that all resources in print or on the web must be explicitly cited.

In keeping with MEF University's ideals, I pledge that this work is my own and that I have neither given nor received inappropriate assistance in preparing it.

Name	Date	Signature
Efe Demir	01/02/2021	

EXECUTIVE SUMMARY

BIG DATA ANALYTICS ON USED CAR INFORMATION

Efe Demir

Advisor: Asst. Prof. Dr. Utku KOÇ

JANUARY,2021, 38 Pages

In this research, a decision support system is implemented on a used car dataset. The main purpose is to predict the price information and reveal the related features. The price prediction problem is classified as a regression problem. The key point is to find the best-fitting model and obtain the best accurate prediction outcomes. Should we buy this car, or at what price may I sell my car? This work is about to answer these questions. Various regression models are compared, and detailed results are explained correspondingly.

The constructed models will help customers to know about their car price and salability. And they can identify the buying opportunities. The percentage error approach which is detailed in the results section will be a guideline for customers/firms to make a market analysis or detect fraudulent listing information.

Key Words: Regression Model, Price prediction, Decision Support.

ÖZET

KULLANILMIŞ ARABA BİLGİLERİ ÜZERİNDEN BÜYÜK VERİ ANALİTİĞİ

Efe Demir

Proje Danışmanı: Dr. Öğr Üyesi Utku KOÇ

Ocak,2021, 38 Sayfa

Bu araştırmada ikinci el araba verisi üzerinde bir karar destek sistemi hayata geçirilmiştir. Çalışmanın temel amacı bir fiyat bilgisi tahmini yapabilmek ve bu tahmin kapsamındaki veri ilişkisini tespit edebilmektir. Fiyat tahmini problemi bir regresyon problemi olarak tanımlanmıştır. Veriyi en iyi açıklayan modeli bulup bu modeller ile en isabetli fiyat tahminlerini yapabilmek çalışmanın en kritik ögesidir. Bu arabayı almalı mıyız, ya da kendi arabamı bu fiyata satabilir miyim? Çalışma bu soruları cevaplayabilmek ile ilgilidir. Çeşitli regresyon modelleri kıyaslanmış ve detaylı çıktıları ifade edilmiştir.

Oluşturulan modeller arabalarının satılabilirliği ve fiyatı konusunda müşterilere yardımcı olacaktır. Ayrıca alım fırsatlarını da tespit edebilmeleri sağlanmaktadır. Araştırmanın sonuçlar kısmında detaylı olarak açıklanan yüzdelik hata yaklaşımı, müşterilere ve firmalara market araştırması yapmaları ve sahtecilik kapsamına girebilecek ilanları tespit edebilmeleri noktasında kılavuzluk edebilecektir.

Key Words: Regresyon Analizi, Fiyat Tahminlemesi, Karar Destek

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

1.INTRODUCTION

The data is everything to obtain information and make decisions. Any information is ready to check whether is valid or not with the use of big data technologies in every aspect of life. The mathematically describable nature of life is boosting the analyzing and modeling efforts in a scientific way.

Craigslist is one of the biggest platforms that a customer can purchase or sell an asset. A huge amount of transaction takes place every day in more than 70 countries and 700 different cities. Because of the subjective nature of the trade market, especially on second-hand sales, intelligent decision-making systems give an upper hand to customers or businesses which are buying or selling in this sector. In this work, car sales and purchases are focused.

Everybody can make predictions about car prices hence there are many regression models. It can be said for the different regression models for price prediction, the methodology to estimate the residual value is remarkably similar and traditional (such as usage of model, mileage in km, and year of manufacture) (Gegic, et al, 2019, pg.168)

The importance and success of work lie beneath the accurate and proper model selection. The research is going to find the best model by comparing the accuracies of the most popular fifteen of them. For the price prediction, more complex ensemble algorithms are recommended. (Kuiper, Shonda., 2018). With this perspective not only light-weight models that can be trained in a short period are used, but also more complex and more time-consuming models are built with detailed hyper-parameter tuning efforts.

1.1. Price Prediction: Literature survey

Price prediction is a common curiosity in market analysis. There is an upward trend in the numbers of big data applications in this subject. There are many related works about car price prediction of vehicles and houses. Both types of research have a common property: Comparison of the regression models.

One of the most comprehensive work uses several classification methods (Support Vector Machine, Extra Trees, Random Forest, and Logistic Regression) The significant features are highlighted as the brand, rounded price, vehicle age, and mileage. The best accuracy is achieved by the Random Forest technique with a %78 score on the test data. The author indicated that it would be possible to achieve better results with more features considered. (Zhang, et al, 2019) The idea of feature engineering about price and year columns provides insight into this work.

In another research, a multivariate regression model as a classification problem is implemented by using the 2005 General Motors dataset. The author conducts detailed work to encourage the people which are new to this area (Kuiper, Shonda., 2018). The selection technique of the proper features and the methodology is explained by cross-checking the regression models. First, simple models like Linear Regression and Support Vector Machines are used to identify feature importance. Then this feature selection is applied to more complex models to achieve better results. In this work, the author emphasizes that price prediction is a complex business, and the feature selection of each dataset should be cross-checked by the outcomes of different models.

Another model comparison-based work indicates that the usage of all features in a model could result in low accuracy on the test set. The author uses a car dataset of 2000 records within a short period. (Noor, Kanwal, and Sadaqat Jan 2017) The collected data includes not-so-relevant features like the advertisement, color, and advertisement company. That sort of evaluation results in a low correspondence of the features. By this knowledge, the feature selection is considered as a significant basis in this research.

Pudaruth (2014) set up a model to estimate the prices of used cars in Mauritius (an island country in Africa) using machine learning techniques. 58 percent accuracy is achieved. The main problem in the work is the limited number of observations. One thousand observations are selected as a train set with different characteristics. So, the built model remains incapable of expressing the test dataset. In this research, a huge number of observations are taken into account. This provides data diversity to build a better model.

In another university thesis, Richardson (2019) working on the hypothesis that car manufacturers are more willing to produce vehicles that do not depreciate rapidly. The engine property has particular importance in the prediction accuracy. He shows that cars that have hybrid engines are more able to keep their value. The relation between the accuracy and engine property is likely because of the fuel efficiency. The environmental concerns may cause that. The other

important features are selected as the car age, fuel efficiency, mileage, and the car model. Feature importance is highly supporting our research outcomes.

1.2. Linear Regression: Literature survey

The methodology aims to find out the best-fitting line as the line that minimizes the sum of squared errors (SSE) or mean squared error (MSE) between our target variable (y) and our predicted output overall samples i in our dataset of size n. The Algorithm aims to minimize the function in Equation 1.

$$MSE = \frac{1}{N} \sum_{(x,y) \in D} (y - prediction(x))^2$$

Equation 1: Mean Squared Error Equation

Multiple Linear Regression (MLR) is a supervised technique used to estimate the relationship between one dependent variable and more than one independent variable. Identifying the correlation and its cause-effect helps to make predictions by using these relations (Shim, Joo Yong, and Chang Ha Hwang, 2011, pg.166). To estimate these relationships, the prediction accuracy of the model is essential; the complexity of the model is of more interest. However, Multiple Linear Regression is prone to many problems such as multicollinearity, noises, and overfitting, which affect the prediction accuracy.

1.3. Support Vector Machine versus Support Vector Regression: Literature survey

Support Vector Machine (SVM) algorithm is a well-known classification algorithm. When the algorithm is applied to a regression model it is named Support Vector Regression (SVR). In general, both algorithms try to find a hyperplane that separates the classes by minimizing the error. The method is to maximize the distance margin between classes. Although the SVM works with regression problems, finding the optimal decision boundary may be tricky in regression problems. SVR has the same principles but minor differences to deal with the regression/prediction problems.

SVR formulates this function approximation problem as an optimization problem that attempts to find the narrowest tube centered around the surface while minimizing the prediction error, that is, the distance between the predicted and the desired outputs. (Shim, Joo Yong, and Chang Ha Hwang, 2011, pg.168). Both algorithms' goal is to maximize the minimum distance which is formulated in Equation 2.

$$w^* = arg_w max\ [min_n\ d_H(\phi(x_n))]$$

Equation 2: SVM Mathematical Equation

1.4. XGBoost and LightGBM: Literature survey

XGBoost is a scalable machine learning system for tree boosting. The system is available as an open-source package. The system has generated a significant impact and has been widely recognized in various machine learning and data mining challenges (Bo, et al., 2020). The algorithm has a huge advantage over scaling. In distributed systems, the algorithm is capable of processing millions of examples in a short period.

LightGBM is also a tree boosting algorithm like XGBoost and serves as a gradient boosting framework. The main difference between algorithms is that while XGBoost is tree level-wise, LightGBM is tree leaf-wise. (Bo, et al., 2020). Parallel learning is supported in LightGBM when dealing with large datasets. This gives an advantage to LightGBM but in much research, XGBoost is labeled as more reliable. This is stated as an open discussion.

First an initial model denoted as Fzero to predict the target variable y. This model is going to be associated with (y – Fzero) residual (Equation 3). Model 2 is created after modeling the residuals of the first model. (Equation 4). The general function is constructed in Equation 5 that m is the iteration. In this research square loss is selected as loss function, the model targets to minimize this loss function (Equation 6)

$$F_1(x) <- F_0(x) + h_1(x)$$

Equation 3: First Model Equation

$$F_2(x) <- F_1(x) + h_2(x)$$

Equation 4: Second Model Equation

$$F_m(x) <- F_{m-1}(x) + h_m(x)$$

Equation 5: The Model Equation of "m" th Iteration

$$F_0(x) = argmin_\gamma \sum_{i=1}^{n} L(y_i, \gamma)$$

Equation 6: The Loss Function of Model

1.5. Gradient Boosting vs Random Forest: Literature survey

Both algorithms are used to solve supervised learning problems. They are ensemble learning procedures that generate learners to build more robust and accurate models. Although they look common because both are decision-tree based algorithms, the way that the construction of the trees is different. GBT builds trees one at a time, where each tree helps to improve the error function made by previously trained ones. Random Forest trains independent trees and builds a model by combining them.

If the data is not clean, GBT is more sensitive to overfitting, independent trees make Random Forest more robust and hardly overfit. Although the training is relatively slower in GBT than Random Forest, in real-time problems GBT performs better. The Random Forest algorithm in huge datasets performs slowly because of the number of independent trees. (Prakash, et al.,2019)

1.6. Ridge Regressor: Literature survey

Ridge regression is the regularized form of linear regression. The analysis method estimates the relationship variables termed Ordinary least squares (OLS) regression. The difference from the

other linear regression techniques is minimizing the sum of the squares in the difference between the observed and predicted values of the dependent variable. As mentioned in Equation 7, the mathematical approach is to minimize the cost function.

Equation 7: Ridge Regressor Formula

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2=\mathrm{RSS}+\lambda\sum_{j=1}^{p}\beta_j^2,$$

The regularization concept here is the L2 regularization. "Squared magnitude" of the coefficient that is shown above as a penalty term is added to the loss function by ridge regression. (Roger W. Hoerl, 2020, pg.224)

2. ABOUT THE DATA

In this research, one of the leading American classified advertisements website Craigslist' car sale data is used. The data is retrieved from Kaggle.com. The data both consist of categorical features like model, region, condition, and numerical features like odometer (mileage information), year, and price. The target variable will be the price information.

2.1. Features

This dataset contains 458K used vehicle listing information from CraigsList.org. Each data is recorded as 26 features. Feature details are explained in Table 1.

Table 1: Feature Details

Features	**Description**
Id	Unique id of the listing
Url	Url of the record.
Region	Region information of the car sale
Region_Url	Region's url information
Price	Price of the vehicle. Will be target variable in this research
Year	Production year of the vehicle
Manufacturer	Manufacturer company of the vehicle
Model	Model of the vehicle in free text format
Condition	Condition of the vehicle
Cylinders	Cylinder count of the vehicle
Fuel	Fuel category of the vehicle
Odometer	The mileage information of the vehicle
Title_Status	An additional conditional information. Free-text user input
Transmisson	Gearbox property of the vehicle
VIN	A unique number of the vehicle listing
Drive	Wheel actuator information of the vehicle

Size	Size information. A categorical user input
Type	Type information of the feature. Sedan, SUV etc.
Paint_Color	The color of the vehicle
Image_Url	The url of the vehicle image
Description	A text description of the vehicle listing
State	The state information of the vehicle
Lat	Latitude information of the listing
Long	Longtitude information of the listing
Posting_Date	The posting date of the vehicle listing

Id, Url, Region_Url, Image_Url, Description, Region_Url, VIN, and **Posting_Date** features are dropped in the first place. Some of these features are links, the others are free text uncategorical features that seem useless to analyze the data. On the other hand, the **Size** feature is highly unusable due to an extremely low fill rate which is only 25 percent (mentioned in Table 2). Thus, this feature is also dropped.

Region and **State** information are required on the website and well-categorized features. The distribution of the data and price correlation is explained in detail in Section 2.3.1. Thus, this research does not require **lat** and **long** features which contains many default and erroneous data. These are dropped.

The **Year** feature is transformed into age feature which is explained in Section 2.2.4. After feature engineering, the year feature is removed, and the **Age** feature is added to the dataset. On the other hand, **Title Status** and **Model** features are dropped as well. The **Title Status** feature is optional user input. The data is significantly imbalanced, up to ninety percent of the listing data has "clean" status. The **Model** feature is also an optional free-text user input. Its data is very dirty.

The remaining features are **Manufacturer, Condition, Cylinders, Fuel, Odometer, Drive, Type,** and **Paint Color.** These features and preprocessing steps are explained in further sections.

After all preprocessing steps, the dataset is split into test and train sets by 0.2 test size ratio. This partition is recorded in different comma-separated text files for the purpose of reproducibility. All models are trained in the same train dataset and tested in the same test dataset partitions. By this approach, the comparison of the model results is meaningful.

The very last step of the data processing is the Standardization of the transformed features. Since the features have different ranges, The Min-Max Scaler from the scikit library of the python language is used to transform all predictor features (other features from Price) to the same scale.

2.2. Data Preprocessing

The raw data information is given below in Table 2. After dropped features, the remaining data information is also given in Table 3.

Table 2: Dataset General Information

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 458213 entries, 0 to 458212
Data columns (total 26 columns):
Unnamed: 0      458213 non-null int64
id              458213 non-null int64
url             458213 non-null object
region          458213 non-null object
region_url      458213 non-null object
price           458213 non-null int64
year            457163 non-null float64
manufacturer    439993 non-null object
model           453367 non-null object
condition       265273 non-null object
cylinders       287073 non-null object
fuel            454976 non-null object
odometer        402910 non-null float64
title_status    455636 non-null object
transmission    455771 non-null object
VIN             270664 non-null object
drive           324025 non-null object
size            136865 non-null object
type            345475 non-null object
paint_color     317370 non-null object
image_url       458185 non-null object
description     458143 non-null object
state           458213 non-null object
lat             450765 non-null float64
long            450765 non-null float64
posting_date    458185 non-null object
dtypes: float64(4), int64(3), object(19)
memory usage: 90.9+ MB
```

Table 3: Dataset After Dropped Features

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 366984 entries, 0 to 458212
Data columns (total 12 columns):
region          366984 non-null object
price           366984 non-null int64
year            366984 non-null float64
manufacturer    356990 non-null object
condition       221613 non-null object
cylinders       233649 non-null object
fuel            364705 non-null object
odometer        330345 non-null float64
drive           263031 non-null object
type            280111 non-null object
paint_color     262236 non-null object
state           366984 non-null object
dtypes: float64(2), int64(1), object(9)
memory usage: 46.4+ MB
```

2.2.1. Removing Outliers and Grouping Categories

The most important part of the data preprocessing is getting rid of the outliers and dealing with the imbalanced data distribution. In this dataset, the price and the year features have outliers which may result in an unsuccessful analysis.

The **Year** distribution of the data is given in Figure 1. The data distribution is imbalanced and significantly in low numbers before 2000. According to the interquartile range analysis, the data between 2000 and 2020 is used in the analysis. Other observations are stated as outliers and dropped from the dataset.

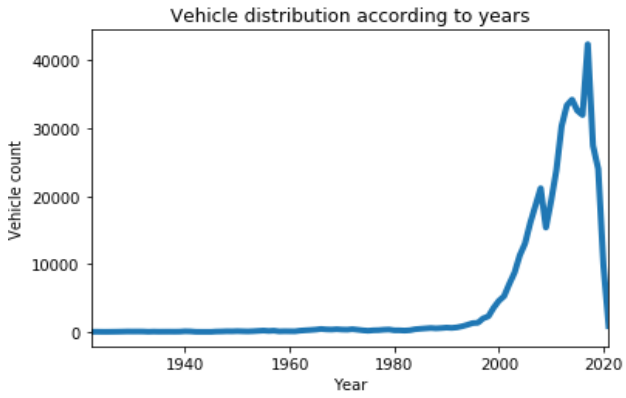

Figure 1: Year Distribution

The **Price** feature has also outlier records. The box plot in Figure 2 is used as an outlier analysis. After the detailed examination of this feature, observations between 100$ and 50000$ are selected, others are removed as outliers.

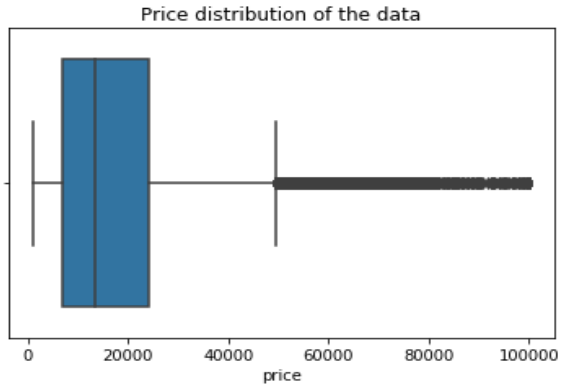

Figure 2: Price Distribution - Box Plot

In imbalanced datasets removing the very infrequent concurrences may be a good alternative to make better conclusions (Bo, et al., 2020). The distribution of the other categorical variables is examined in this perspective. The **Manufacturer** feature distribution of the data given in Figure 3. Alfa-Romeo, Harley-Davidson, Land Rover, Aston-Martin, Hennessey, and Datsun manufacturers are rare in the dataset and are removed.

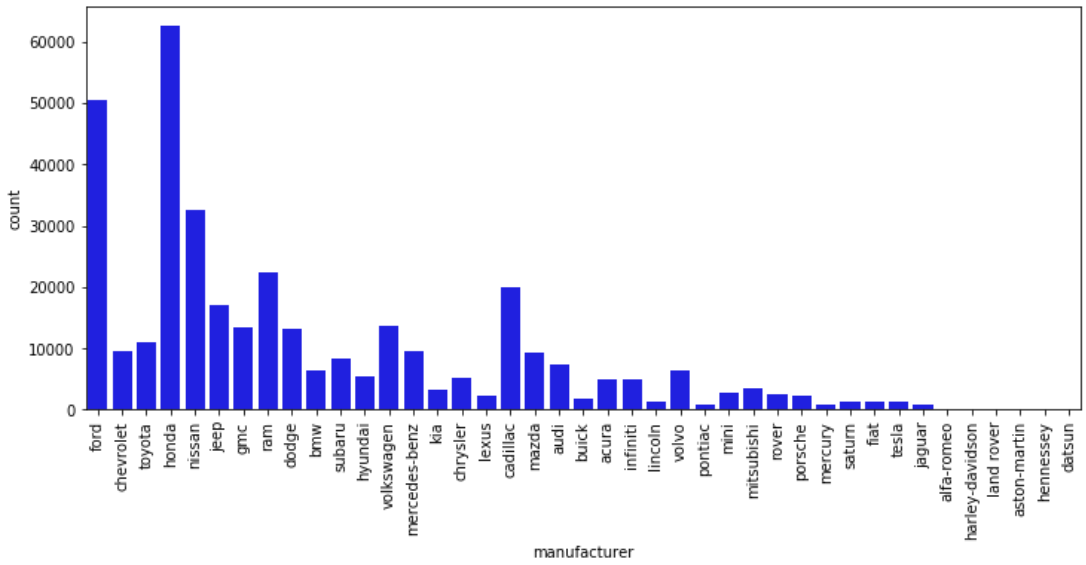

Figure 3: Manufacturer Distribution

The **Condition** distribution is given in Figure 4. “New” and “Salvage” occurrences are rare. The input of these categories is no longer available on the website. Thus, these observations are removed as outliers.

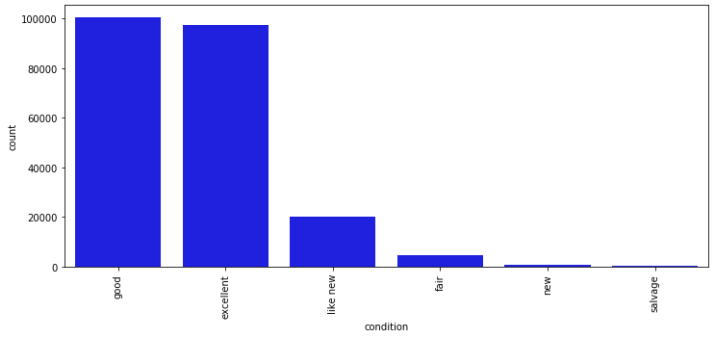

Figure 4: Condition Distribution

A different approach to the imbalanced datasets is not removing the rare occurrences but grouping them into a new category like ‘other’ or ‘unknown’. The different characteristic apart from the common categories may be useful to analyze the data (Abdellatif, Safa, et al.,2018)

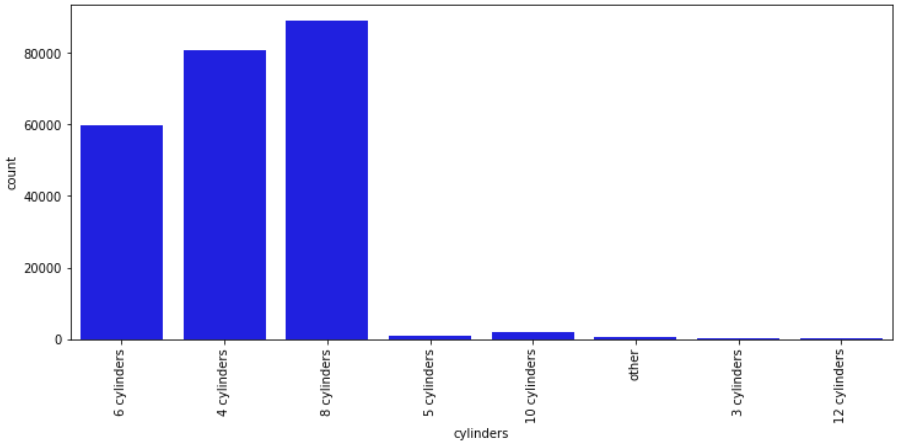

Figure 5: Cylinder Distribution

The distribution of the **Cylinders** feature is given in Figure 5. Above 5-cylinders observations are regrouped into the ‘other’ category. The **Fuel** feature shows the same characteristics. The interpretation will be the same. Thus, “hybrid”, “diesel” and “electric” categories are regrouped into the ‘other’ category as well. Figure 6 is the distribution of the fuel feature.

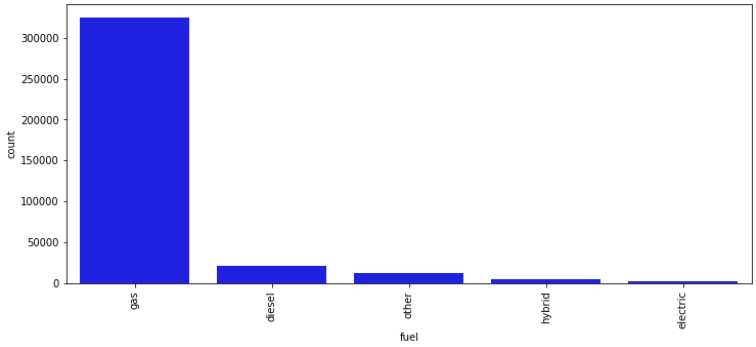

Figure 6: Fuel Distribution

The other categorical features, **Cylinders, Drive,** and **Paint Color** have good distributions and no outliers. Thus, no preprocessing activity is required. The last activity in this section is the interpretation of the **Odometer** feature. This feature is the mileage information of the data and a

numerical one. To help our models to work with clearer data, this feature will be rounded to base 500. After interpretation of the data, the dataset is seemed as in Table 4.

Table 4: Dataset After Pre-Processing

	region	price	year	manufacturer	condition	cylinders	fuel	odometer	drive	paint_color	state
0	auburn	35990	2010.0	chevrolet	good	8 cylinders	gas	32500.0	rwd	NaN	al
1	auburn	7500	2014.0	hyundai	excellent	4 cylinders	gas	93500.0	fwd	NaN	al
2	auburn	4900	2006.0	bmw	good	6 cylinders	gas	87000.0	NaN	blue	al
4	auburn	19500	2005.0	ford	excellent	8 cylinders	other	116000.0	4wd	blue	al
5	auburn	29590	2016.0	toyota	good	6 cylinders	gas	33000.0	NaN	red	al

2.2.2. Data Cleaning - Dealing with N/A Values

After outlier removing and regrouping some of the features, the first job is to handle N/A Values. **Manufacturer, Condition, Cylinders, Fuel, Drive,** and **Paint Color** features have null values. In the previous step, we have seen that each of the features has ‘unknown’ or ‘other’ category. So, the null values will be replaced with these values. After the null replacement, the dataset information is given in Table 5. In this research, models are built on this dataset, 330K observations and 11 features. (**Year** feature will be replaced soon in Section 2.2.4 by **Age** feature)

Table 5: Clean Dataset Information

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 330345 entries, 0 to 458212
Data columns (total 11 columns):
region          330345 non-null object
price           330345 non-null int64
year            330345 non-null float64
manufacturer    330345 non-null object
condition       330345 non-null object
cylinders       330345 non-null object
fuel            330345 non-null object
odometer        330345 non-null float64
drive           330345 non-null object
paint_color     330345 non-null object
state           330345 non-null object
dtypes: float64(2), int64(1), object(8)
memory usage: 30.2+ MB
```

2.2.3. Label Encoding - Categorical Features

The Region, manufacturer, condition, cylinders, fuel, drive, paint color, and state features are categorical variables. Categorical features may be encoded by different techniques. Since there are eight predictor categorical variables, choosing the right algorithm is important. One-Hot Encoding and Label-Encoding are two popular categorical feature transformation techniques.

One-hot encoding is a sparse way of representing data in a binary string in which only a single bit can be **one**, while all others are **zero**. Label Encoding is a popular encoding technique for handling categorical variables. In this technique, each label is assigned a unique integer based on alphabetical ordering. According to Stefanowski (2016), label-encoding can be used for this dataset. After encoding the dataset is shown in Table 6.

Since some of the categorical features are not ordinal, one-hot encoding may be more suitable for that features. But because of the large size of the dataset and due to restrictions on the computation power and memory size, label encoding is preferred. There is a growth area on the categorical encoding in this research. The methodology is not perfect, and there is a glitch in this research because of this.

Table 6: Dataset After Label Encoding

	region	price	manufacturer	condition	cylinders	fuel	odometer	drive	paint_color	state	age
0	16	35990	5	2	2	0	32500.0	2	10	1	10
1	16	7500	12	0	0	0	93500.0	1	10	1	6
2	16	4900	2	2	1	0	87000.0	3	1	1	14
4	16	19500	9	0	2	1	116000.0	0	1	1	15
5	16	29590	33	2	1	0	33000.0	3	8	1	4

2.2.4. Feature-Engineering

The year feature is the production year of the car. But the year since the production information may be more useful. To obtain this information **'age'** variable is added to the data. The definition of Age is 2020 minus the production year. For example, a vehicle that the manufactured in 2015, has the age of five.

Table 7: Age feature is Engineered by Year Feature

	region	price	manufacturer	condition	cylinders	fuel	odometer	drive	paint_color	state	age
0	auburn	35990	chevrolet	good	8 cylinders	gas	32500.0	rwd	unknown	al	10
1	auburn	7500	hyundai	excellent	4 cylinders	gas	93500.0	fwd	unknown	al	6
2	auburn	4900	bmw	good	6 cylinders	gas	87000.0	unknown	blue	al	14
4	auburn	19500	ford	excellent	8 cylinders	other	116000.0	4wd	blue	al	15
5	auburn	29590	toyota	good	6 cylinders	gas	33000.0	unknown	red	al	4

2.3. Exploratory Data Analysis

After data pre-processing steps, the dataset is ready to further analysis. The initial perspective is to find out which feature has a strong relationship with the price information. This part gives us a descriptive view. It is expected that the first insights which are discovered in this part should be parallel to the model results.

2.3.1. Region and State Distribution Analysis

Region and **State** distribution are quite good. Especially the region information has a uniform distribution. In Figures 7 and 8, although it seems that the region average price may differ, the listing count of the regions is very uniform. Both characteristics make the region is a piece of useful information in the price prediction process. **State** distribution is highly related the population information. Since CA (California) is one of the biggest states in the U.S of America, the listing count is maximum here. The population information is not in the dataset, this deduction is made by general knowledge. In Figures 9 and 10 it is shown both the counterplot and the average price bar chart of the state feature.

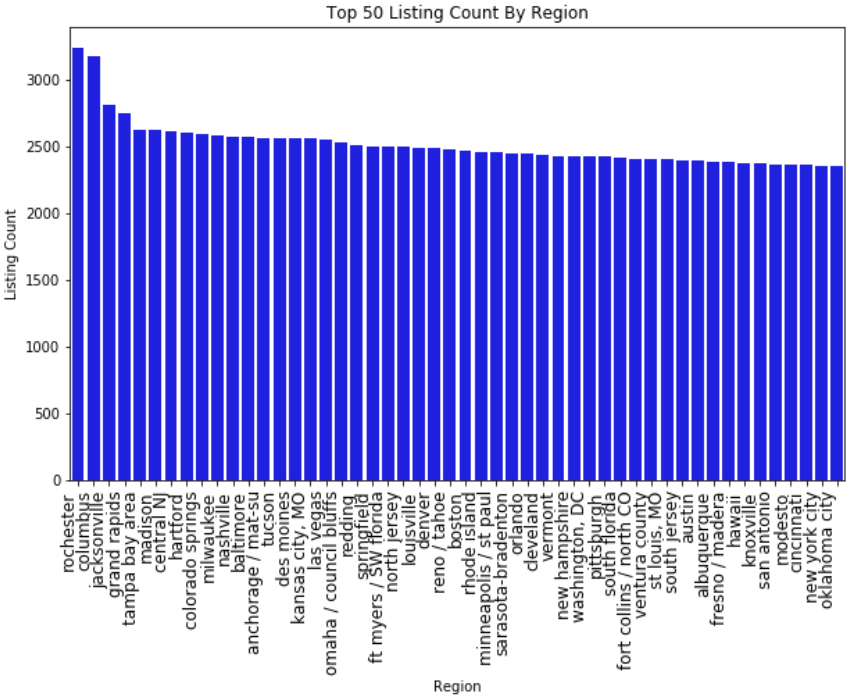

Figure 7: The Region Distribution

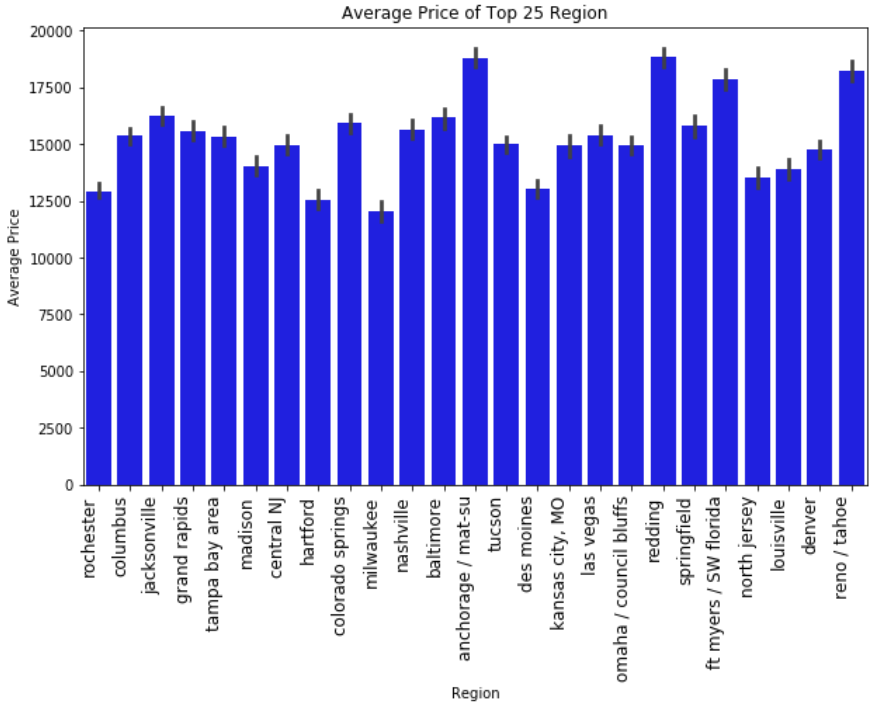

Figure 8: Average Price of Top 25 Region

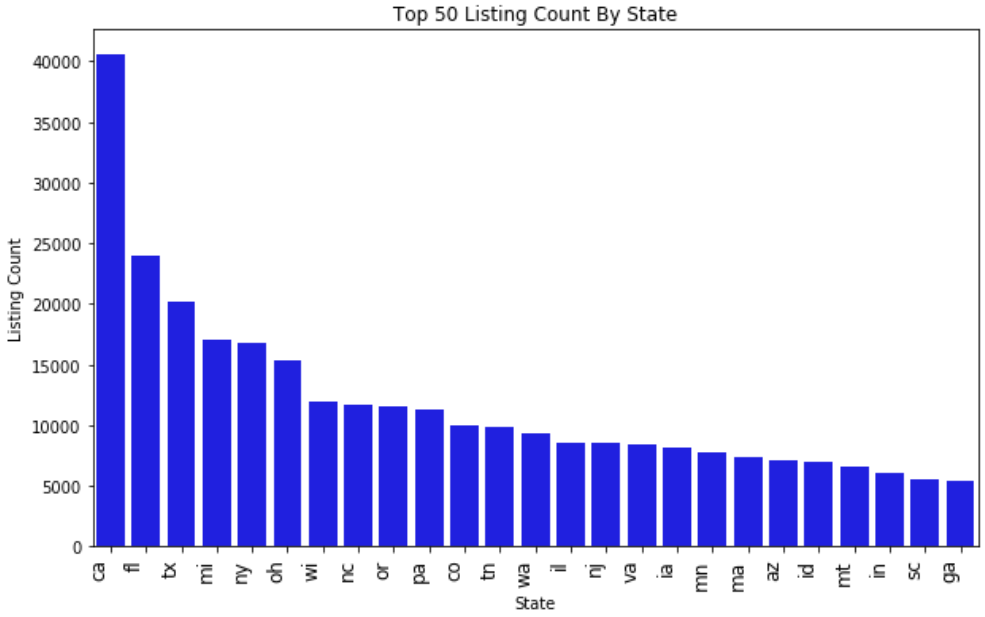

Figure 9: State Distribution

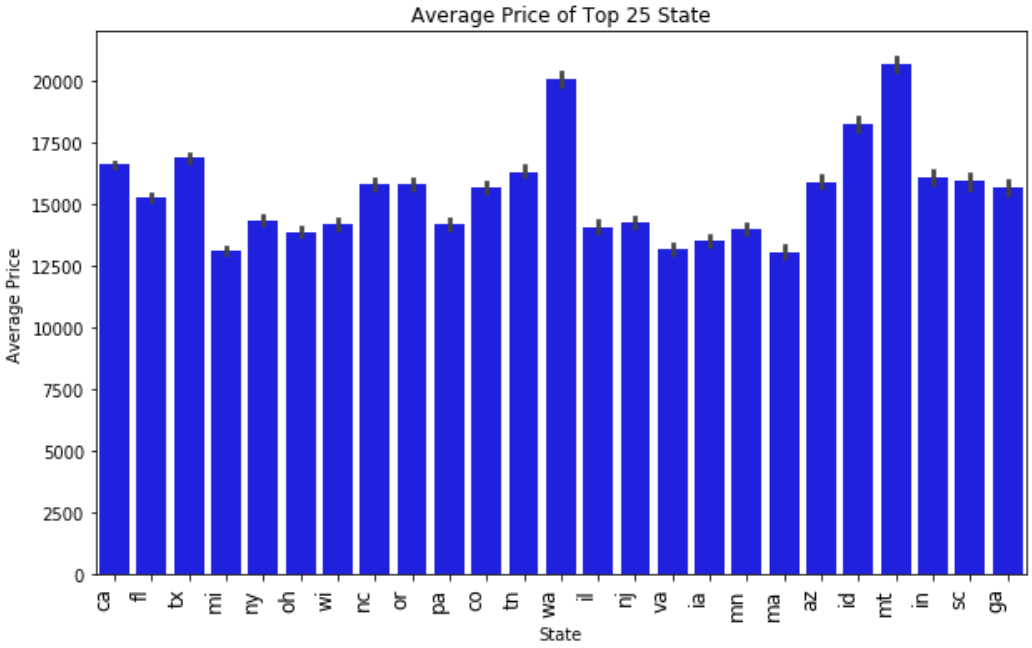

Figure 10: Average Price of State

2.3.2. Yearly Distribution Analysis

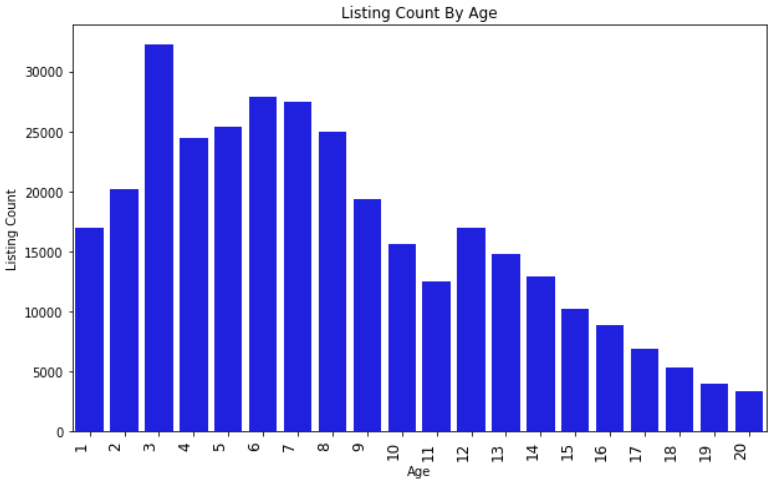

Figure 11: Vehicle Age Distribution

Since we have the **Age** feature engineered in Section 2.2.4 from the **Year**, this feature is used from here. Most of the data has vehicle age below nine. The distribution is left-skewed where the older vehicles are less listed on the website (Figure 11). In Figure 12, there is negative linearity between age and average price. This gives us insights into the high importance and linearity of the age feature.

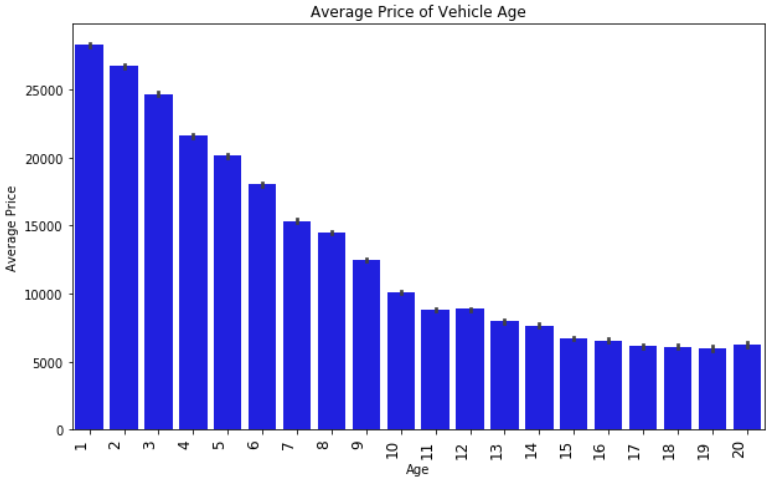

Figure 12: Average Price of Age

2.3.3. Manufacturer Feature Distribution Analysis

Ford and Chevrolet are common American brands. It is expected that the listing count of these brands is higher than the other brands. In Figure 13, this expectation is met. Also, the exclusive brands like Porsche and Jaguar, and foreign brands like Fiat and Mitsubishi have very few listing records.

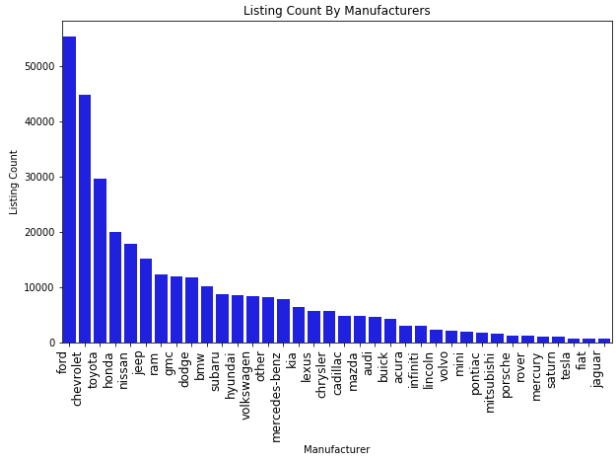

Figure 13: The Manufacturer Distribution

Although the domination in the number of American Brands, the average price of manufacturers implies this feature's importance. The average is highly parallel with the common knowledge of Car Brands as shown in Figure 14. The short line over the bars defines the confidence interval. This line is short; hence it can be said that the distribution of the data in a specific manufacturer is reasonable. If the line were long, we would say that the price of the record of a specific manufacturer differs too much.

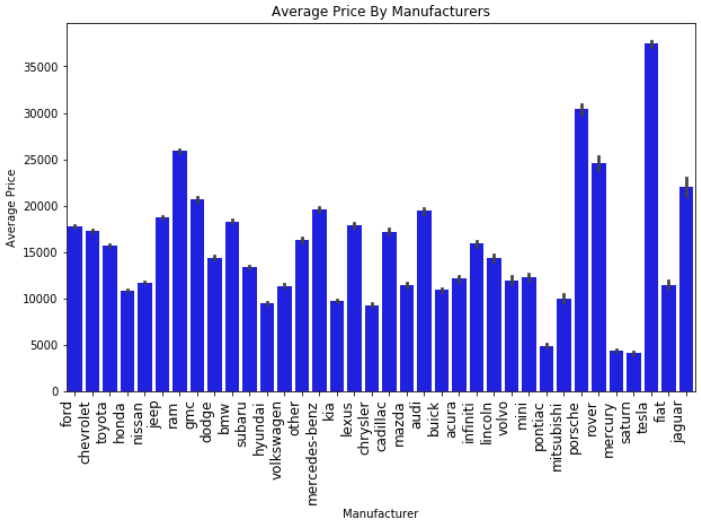

Figure 14: Average Price of Manufacturers

2.3.4. Cylinder Feature Distribution Analysis

The Cylinder is optional information on Craigslist. Therefore, there are unknown values that are labeled as 'other'. '8 Cylinders vehicles' are relatively rare (Figure 15) but have significantly high prices (Figure 16). That can be an important feature related to price.

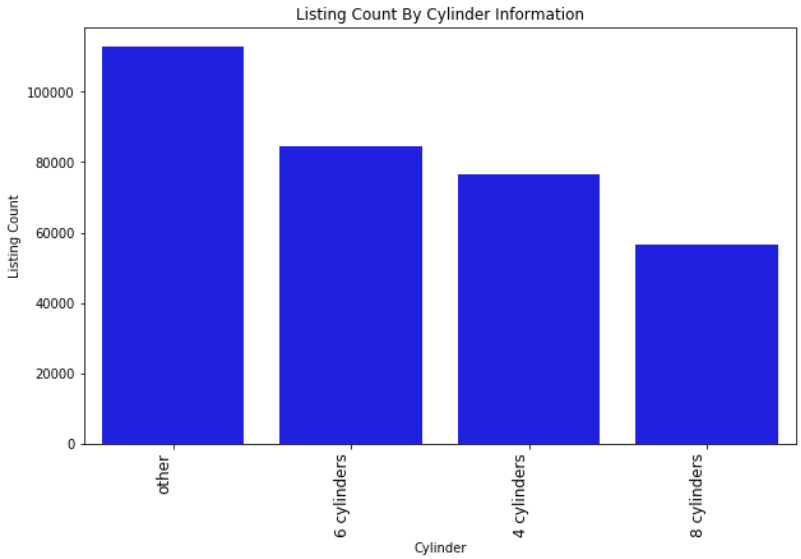

Figure 15: The Cylinders Distribution

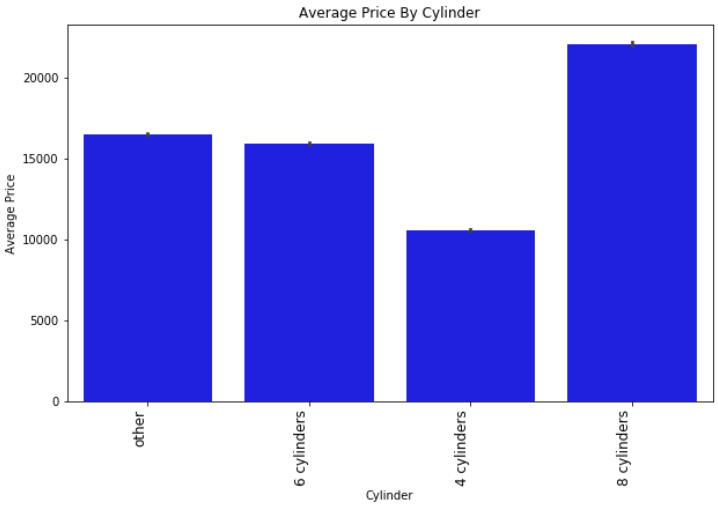

Figure 16: The Average Price of Cylinders

2.3.5. Condition Feature Distribution Analysis

Condition feature may be considered as an important feature. But this information is user input. Cross-validation through age is applied, and the results show us that this information is highly subjective. Although the average age of the excellent condition was eight, the average age of good condition is likely seven (Figure 17).

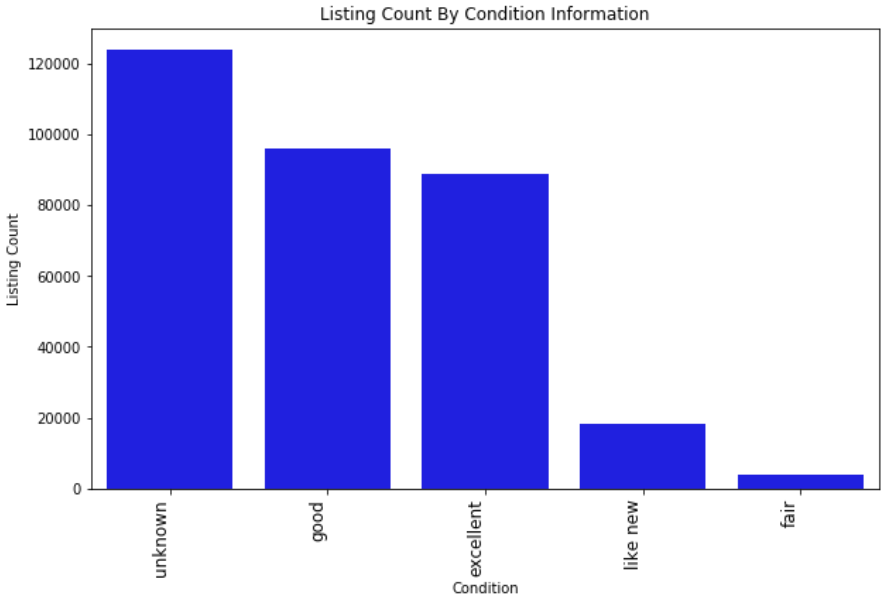

Figure 17: The Condition Distribution

If it is assumed that the condition feature is reliable, the average price of the fair condition is significantly lower than the other categories. Maybe there is no significant average price difference in good, excellent and like new categories, but the characteristics of the fair condition would enhance the models (Figure 18).

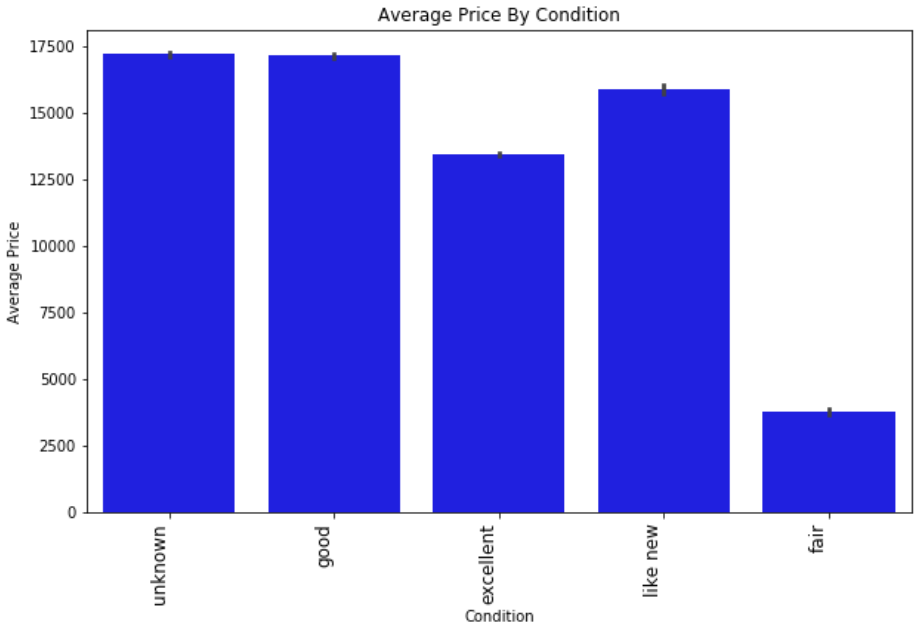

Figure 18: Average Price of Condition

3.PROJECT DEFINITION

3.1. Problem Statement

The vehicle listing information is obtained from one of the largest second-hand websites. Craigslist Used Car Dataset is a useful playground where supervised learning techniques can be used. The effects of the different features on the target price feature are analyzed. The insight that is obtained from Exploratory Data Analysis is enhanced by machine learning techniques.

In this research, the price prediction for the vehicle dataset is the focus. The key aspect is the comparison of models. The most common fifteen regression models are applied to the data. Results are compared in R-Square, Relative Error, and Root-Mean-Squared-Error terms. For boosting methods, a detailed hyper-parameter tuning process is applied to achieve better results. Both Regression algorithms and boosting techniques are used

After selecting the best-fitting models, an empirical analysis is going to be used. The best-fitting models predict the test dataset which is split in Data Preprocessing Section. After this prediction, actual and predicted prices are compared by a percentage error approach. The results are categorized as Normal, Cheap, Expensive, Unclassified, and Extreme. This part adds a Fraud Detection property to this research.

3.2. Methods, Tools, and Techniques

The raw data obtained from Kaggle.com references the CraigsList.org. Machine learning algorithms to predict prices such as Linear Regression, Support Vector Machine, Linear Support Vector Regressor, LGBM, Bagging Regressor, XGBBoosting, Gradient-Boosting Regressor, MLPRegressor, AdaBoostRegressor, Ridge Regressor, Voting Regressor, Stochastic Gradient Descent are used.

The raw data was relatively clean. After detailed data preprocessing, which is mentioned in 2.2. Data Preprocessing section, exploratory data analysis is completed afterward. In this section, the positively and negatively related features are identified. Especially the odometer and manufacturer features seemed to have a relation with the price information. The synthetic feature age which engineered from year feature is very dominant and negatively related to the price information.

Linear Regression is a linear approach to modeling the relationship between a scalar response. That algorithm was the first option to analyze the data. The mathematical background of the technique is given in Section 1.2. The feature importance of the linear regression model can be seen in Table 8. The Paint Color feature has an incredibly low significance on the price prediction. Therefore, this feature is dropped from our dataset. After dropping the feature, the re-evaluation of the model generates the same results. For the Linear Regression model Linear Regression package is used from the sklearn-linear_model library. To evaluate the method metrics package is used from the same library.

Table 8: Feature Importance of Linear Regression

Feature	region	manufacturer	condition	cylinders	fuel	odometer	drive	paint_color	state	age
Importance	-60	-185	106	2186	1941	-439	-1379	-2.7	-75	-6491

The importance coefficients tell us how one unit change in the feature will affect the target price value. For example, let us say the odometer feature is thousand-kilometer precision after normalization process. If we change the odometer value from 0.5 to 0.6 of an observation, then the price value will decrease 43.9. The -439 coefficient of the odometer parameter means that.

Support Vector Regressor and Linear SVR methods of sklearn-svm library are used to learn about empirical results. Since the coefficient determination (R-Square value) which is shown in Equation 8 is only 14 percent, and the RMSE value which is mathematically described in Equation 9 is significantly high, these two algorithms are classified as inappropriate for this dataset.

$$R^2 = 1 - \frac{\text{Unexplained Variation}}{\text{Total Variation}}$$

Equation 8: The coefficient of determination formula

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} \left(P_i - O_i \right)^2}{n}}$$

Equation 9: Root Mean Squared Error Formula

The **coefficient of determination (R-squared)** is a statistical measure that represents the proportion of the variance for a dependent variable that is explained by an independent variable or variables in a regression model. **Root Mean Square Error (RMSE)** is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. In this research RMSE values of the models are compared selecting the best-fitting model.

The Multi-Layer Perceptron is also a supervised learning algorithm. By mathematically it minimizes the loss function of each layer according to their weights. After the computation of the loss of a layer, a backward directional evaluation is placed. From the output layer to previous layers there is an update value meant to decrease the loss. The final loss function is given in Figure 31, where alpha is the hyperparameter that controls the magnitude of the penalty between layers. The node-wise layer representation is given in Equation 10.

$$Loss(\hat{y}, y, W) = \frac{1}{2} ||\hat{y} - y||_2^2 + \frac{\alpha}{2} ||W||_2^2$$

Equation 10: Loss Function of Multi-Layer Perceptron

The **MLPRegressor** package in sklearn-neural_network library uses the gradient descent methodology and the loss function in implementation is given in Equation 11. Detailed information about the gradient descent approach is given in Section 1.5.

$$W^{i+1} = W^i - \epsilon \nabla Loss_W^i$$

Equation 11: Gradient Descent of the Loss Function

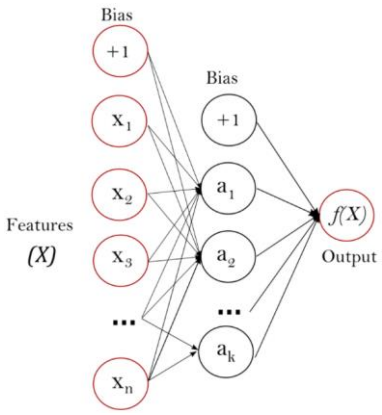

Figure 19: One Layer MLPFig

Note. Reprinted from *Figure 1: One Hidden Layer*, *retrieved* from https://scikit-learn.org/stable/modules/neural_networks_supervised.

The **GridSearchCV** package from "sklearn-model_selection" library is used in the MLPRegression implementation. The instance implements the usual estimator API: when "fitting" it on a dataset all the possible combinations of parameter values are evaluated, and the best combination is retained. It is a time-consuming effort, but to find the best-fitting model, hyper-parameter tuning of the models is especially important. Therefore, four different combinations of alpha and learning rate parameters are used by selecting the best-hidden layer sizes.

SGDRegressor package from the sklearn-linear_model library is also used to compare the models. The mathematical background, which objective function is in Equation 12 is like other gradient descent algorithms, but some implementation differences exist. This implementation is used as an empirical analysis to compare with other models. The "w" value is the parameter that minimizes Q(w).

$$Q(w) = \frac{1}{n} \sum_{i=1}^{n} Q_i(w),$$

Equation 12: Objective Function of Stochastic Gradient Descent

In the literature survey, there were significant signs that **XGBoosting** and **LightGBM** algorithms may perform very well with our data. Thus, these algorithms are used with hyperparameter tuning and cross-validation processes. The mathematical approach is explained in Section 1.4. The implementation is done by using the python packages xgboost and lgbm.

For **XGBoosting** algorithm 50 to 150 estimators with 0.01 to 0.1 learning rate intervals are used in GridSearchCV. The maximum depth variable was selected as four to seven, and the objective function was selected as Regression-Squared Error. With a fast result, the best parameters and score are found.

The **LightGBM** was used with the validation set. First, train and validation sets were split with 0.2 ratios. The training procedure was completed thirty thousand rounds with eight thousand early stopping numbers. In each iteration, better RMSE values are obtained and thirty thousand iterations were selected by this observation.

GradientBoostingRegressor algorithm was used after hyper-parameter optimization. According to cross-validation scores, the different estimators and maximum depth were tried with ten evaluation steps. After a huge time-consumption optimization process the best maximum depth and number of estimators were calculated 2 and 222, respectively. After calculation of the parameters algorithm was applied to the data and results were recorded. This is an alternative approach in implementation by using boosting techniques to the gradient descent methodology. There is no significant difference mathematically, the loss function interpretation is given in Equations 13, 14, and 15.

Given m number of items in our dataset, with x(predictor) and y(target) features, the objective is to solve Equation 13. The cost function derived from that objective is given in Equation 14 where x(i) and y(i) are the x, y values for that component. The goal of the gradient descent is to minimize Equation 14. Thus, each iteration of the gradient descent can be formulated like Equation 15.

$$h_\theta(x) = \theta_0 + \theta_1 x \, .$$

Equation 13: Regression Equation

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (\theta_0 + \theta_1 x^{(i)} - y^{(i)})^2$$

Equation 14: Cost Function of Gradient Descent

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1)$$

Equation 15: Each Iteration of the Gradient Descent

4.RESULTS

This work is completed on a personal computer whose operating system is Windows 10 with a 16GB ram capacity. The Processor is 2.20 GHz, 2208 MHz with six-core seventh-generation Intel processor. All code part is written in Anaconda Jupyter Notebook at 3.7.4 Python version. Because of the limitation of resources, optimization activities like Hyperparameter Tuning are done in low cycles. If a workstation or a distributed system like Azure is used, the tuning and optimization operations may be extended comprehensively.

In this research, we discovered many significant points. First, according to Table 9, the highest coefficient of determination factor is obtained in the LightGBM and XGBoosting models. The reason behind this is hyperparameter-tuning. Also, by RMSE values in Table 10, the same models have the best RMSE scores. According to the literature survey, the minimum RMSE values are meant to the most fitting algorithm for the test data.

The XGBoosting model is selected after many Grid Search optimization with the variation of estimators, max depth and, learning rate. The best score is obtained as **0.786** and the best parameters are shown in Figure 22. The LGBM model has different RMSE values on the train and test datasets. The values are lower than the other algorithms, so it can be said that the errors in prediction are relatively better. But the difference between the train and test datasets tells us that LGBM may be erroneous when predicting the test dataset.

```
Best parameters set: {'learning_rate': 0.1, 'max_depth': 7, 'n_estimators': 140, 'reg_lambda': 0.5}
```

Figure 20: Best Parameters of XGBoostingRegressor

Table 9: R-Squared Score on Test Data

	Model	RSquare Train	RSquare Test	Relative Train Error	Relative Test Error	RMSE Train	RMSE Test
13	LGBM	89.57	85.23	15.06	17.66	359,360.14	429,207.50
8	XGB	80.03	79.07	20.94	21.40	497,168.38	510,946.96
10	GradientBoostingRegressor	44.80	45.52	39.39	40.20	826,544.59	824,232.30
5	MLPRegressor - Alpha(0.0005) - Learning Rate (...	44.80	45.51	39.35	40.59	826,532.52	824,340.46
2	MLPRegressor - Alpha(0.0001) - Learning Rate (...	44.82	45.46	39.46	40.01	826,404.94	824,698.00
3	MLPRegressor - Alpha(0.0005) - Learning Rate (...	44.83	45.40	39.34	39.71	826,353.92	825,149.63
16	AdaBoostRegressor	52.29	42.60	35.74	43.10	768,406.23	846,060.64
0	Linear Regression	42.10	42.48	41.49	41.70	846,501.03	846,961.41
17	VotingRegressor	42.06	42.35	41.55	41.69	846,847.17	847,872.62
15	RidgeRegressor	42.06	42.31	41.48	41.59	846,835.55	848,180.31
11	BaggingRegressor	42.06	42.30	41.44	41.55	846,853.00	848,294.15
9	Stochastic Gradient Descent	97.05	30.42	6.38	58.41	191,218.86	931,515.86
1	Linear SVR	29.12	29.14	52.68	52.87	936,618.32	940,057.89
4	MLPRegressor - Alpha(0.0001) - Learning Rate (...	43.83	3.62	40.12	86.43	833,751.43	1,096,344.50

Table 10: RMSE Score on Test Data

	Model	RSquare Train	RSquare Test	Relative Train Error	Relative Test Error	RMSE Train	RMSE Test
13	LGBM	89.57	85.23	15.06	17.66	359,360.14	429,207.50
8	XGB	80.03	79.07	20.94	21.40	497,168.38	510,946.96
10	GradientBoostingRegressor	44.80	45.52	39.39	40.20	826,544.59	824,232.30
5	MLPRegressor - Alpha(0.0005) - Learning Rate (...	44.80	45.51	39.35	40.59	826,532.52	824,340.46
2	MLPRegressor - Alpha(0.0001) - Learning Rate (...	44.82	45.46	39.46	40.01	826,404.94	824,698.00
3	MLPRegressor - Alpha(0.0005) - Learning Rate (...	44.83	45.40	39.34	39.71	826,353.92	825,149.63
16	AdaBoostRegressor	52.29	42.60	35.74	43.10	768,406.23	846,060.64
0	Linear Regression	42.10	42.48	41.49	41.70	846,501.03	846,961.41
17	VotingRegressor	42.06	42.35	41.55	41.69	846,847.17	847,872.62
15	RidgeRegressor	42.06	42.31	41.48	41.59	846,835.55	848,180.31
11	BaggingRegressor	42.06	42.30	41.44	41.55	846,853.00	848,294.15
9	Stochastic Gradient Descent	97.05	30.42	6.38	58.41	191,218.86	931,515.86
1	Linear SVR	29.12	29.14	52.68	52.87	936,618.32	940,057.89
4	MLPRegressor - Alpha(0.0001) - Learning Rate (...	43.83	3.62	40.12	86.43	833,751.43	1,096,344.50

According to XGBoosting model, the most important features are age, cylinders, fuel, and drive ordinally. On the other hand, LGBM is built on the odometer, region, age, and state features (Figure 37). The dominance of the Age importance is expected according to the Literature Survey and Exploratory Data Analysis. Both algorithms recorded the low importance of the condition relatively.

Feature	region	manufacturer	condition	cylinders	fuel	odometer	drive	state	age
LGBM	248323	198382	105748	116026	48464	287472	112728	186681	196176
XGBossting	1.273908	4.202685	2.044225	19.264445	12.666367	6.13598	13.226637	1.752608	39.433149

Figure 21: Feature Importance of models

In this research, an empirical analysis is made different from the other related works. After model selection, all test dataset is again predicted by these models. Predicted values and observed values are compared by a percentage error approach.

First, data is obtained which is shown in Figure 38. Then prediction differences of each model are calculated in percentage. This error is formulated as Actual Price minus Predicted Price over Actual Price. This value is how much far the predicted value is from the actual value. Then the classification of each observation is made. The methodology is given below.

- If the difference in deviation of predicted values is above twenty percent these observations are classified as **'Unclassified'**. This means the two models we choose predicted quite different results.
- If the average deviation of predicted values in percentage is below 10, these observations are classified as '**Normal'**. It can be said that these observations are predicted highly likely to the actual price.
- If the average deviation of predicted values in percentage is below 50 and above 10, these observations are classified as **'Cheap'** or **'Expensive'** according to predicted values' difference from the actual ones.
- If the average deviation of predicted values in percentage is above 50 but with respect to the first bullet, two models calculated similar results to each other, these observations are classified as **'Extreme'**

	Actual Price	XGBoosting Prediction	LGBM Prediction
271457	25904	21,020.68	15,149.93
421228	14995	9,775.06	13,527.55
130709	15180	17,426.14	16,875.68
432101	7900	5,831.33	6,520.04
230239	11998	15,470.67	14,050.05
...	...	...	...
292909	2995	6,373.35	6,406.19
134337	13999	13,805.69	13,164.61
400995	12995	13,810.73	12,906.66
434364	9900	7,544.93	7,089.60
370286	10909	14,107.57	14,323.65

Figure 22: Actual Prices versus Predicted Values

After classification, the distribution graph (extreme and unclassified classes), average price, and the counterplot of the data are given in Figure 25, Figure 26, and Figure 27, respectively. In figure

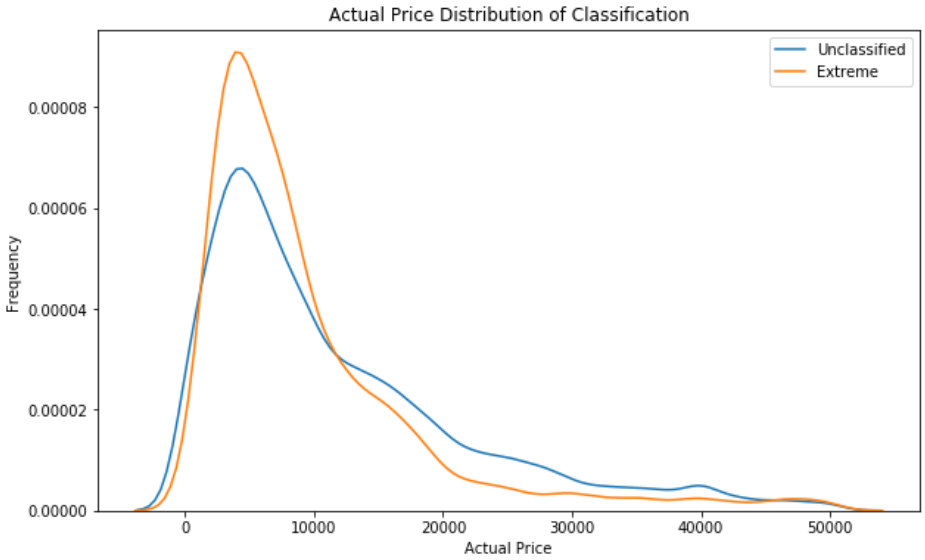

Figure 23: Classification Distribution Plot (Extreme and Unclassified)

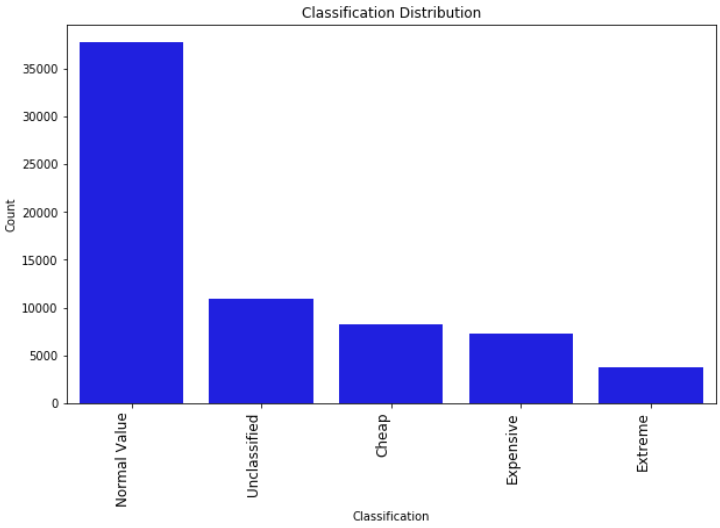

Figure 24: Classification Distribution in Numbers

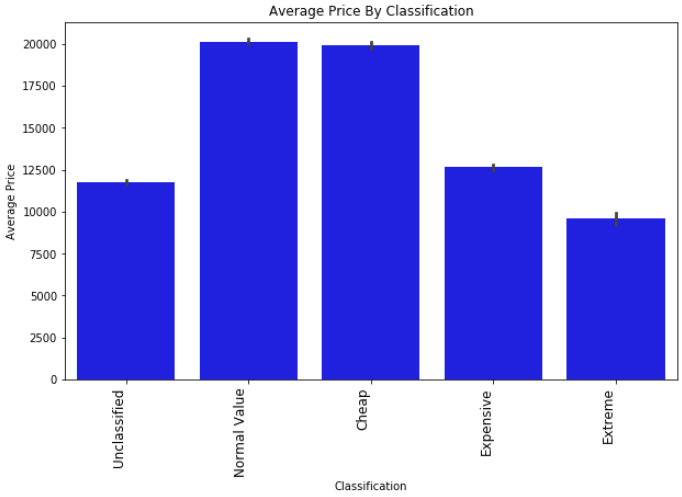

Figure 25: Average Price of Classification

It is observed that extreme and unclassified observations exist when the actual price is quite low. There are nearly 7500 Unclassified and 10000 extreme observations which are condensed in the left part of the distribution (Figure 28). The model accuracy to the observations which have low prices is not particularly good. After eliminating the observations whose actual price is over 5000, the percentage of Unclassified and Extreme classes are dropped sharply (Figure 28).

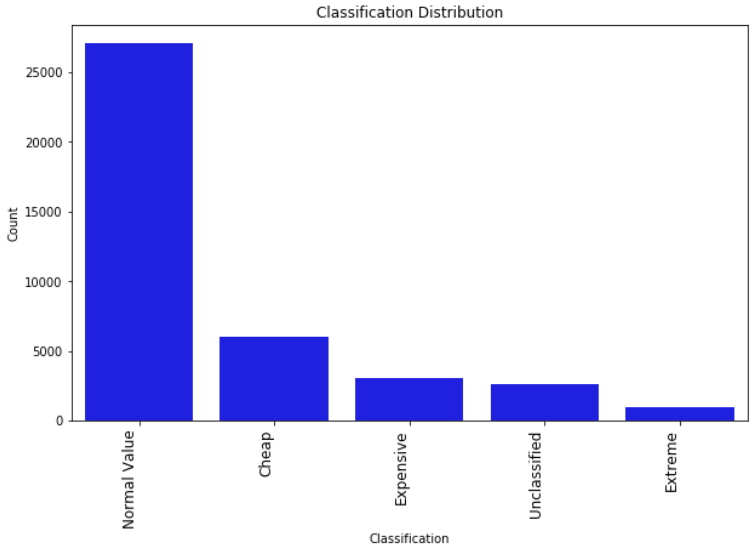

Figure 26: Classification Distribution in Numbers (Over 5000 Actual Price)

5.CONCLUSION

In this research, Craigslist used card data is examined. After pre-processing efforts, nine predictor variables are selected to predict the target price variable. The train and test partitions are selected with a 0.2 test ratio. Twelve different regression algorithms are implemented in the Python language. Both boosting and begging methods are used. All models are trained with the same train set and tested with the same test set. The complex models like XGBoosting, LGBM, Gradient Descent Regressor, and Ridge Regressor have resulted in better fit metrics. XGBoosting and LGBM models are selected after the coefficient of determination and RMSE comparison.

In these two models feature importance of the variables is expected to be similar. The age feature is dominant in both algorithms as expected. The odometer importance is relatively low in the XGB algorithm. This can be explained that the age feature contains similar information with the odometer feature in the dataset, therefore the XGB algorithm may transfer the odometer's importance to the age feature. The odometers' importance is the leading one in the LGBM, but in fifth position in the XGB. Although region and state information are important for the LGBM, drive and cylinders are more dominant for the XGB. Cylinders' importance is unexpected according to the literature survey but had moderate coefficients in both algorithms.

The simple regression models had difficulties explaining the data. The coefficients of determination are relatively low than the ensembled or tuned complex algorithms. The insight about the high success rate of complex models in related works is supported in this research. Hyper-Parameter Tuning was a crucial aspect of the research. In the absence of tuning, models generated significantly bad results with default parameters.

Until this part, some other related works have similar results and approaches. Many of them are used the comparison of the model methodology. In this research, an empirical effort is added. The test dataset is predicted by two algorithms, then the predicted values are compared with the observed ones. If two algorithms have different percentage deviation and this deviation is above twenty percent, it is stated that two models behaved differently, and the classification is set to **Unclassified**. If two algorithms are predicted values in the margin of ten percent, these observations are classified as **Normal Value.** If this margin is between ten and fifty, the observations are classified as **Cheap** or **Expensive** according to margin direction. If both algorithms generated similar results, but the margin is above fifty, these observations are classified

as **Extreme**. This approach is used with the aim of creating a Decision Support mechanism. The customers or firms may use this mechanism to identify buying opportunities. Also, individuals can control if the selling price of their car is reasonable.

If this work is improved in the future, there is a usage are about Fraud Detection. If the accuracy of models is increased, the **Extreme** class may be used to detect fraudulent listing information. If a vehicle is being sold much less than the predicted price, there can be an illegal activity and need to be investigated.

As a future work, the empirical approach to the predicted values may be replaced with an statistical approach. The cut-offs of the percentages are determined in an empirical way, but the best conclusion can be achieved by the prediction interval approach. The prediction interval approach predicts the distribution of future points. The classifications that mentioned in previous parts should be made by this predictive inference.

REFERENCES

[1] Gegic, Enis & Isakovic, Becir & Kečo, Dino & Mašetić, Zerina & Kevric, Jasmin. (2019). Car price prediction using machine learning techniques. TEM Journal. 8. 113-118. 10.18421/TEM81-16.

[2] Shonda Kuiper (2008) Introduction to Multiple Regression: How Much Is Your Car Worth? Journal of Statistics Education, 16:3, DOI: 10.1080/10691898.2008.11889579

[3] Pudaruth, Sameerchand. "Predicting the price of used cars using machine learning techniques." Int. J. Inf. Comput. Technol 4.7 (2014): 753-764

[4] Noor, Kanwal, and Sadaqat Jan. "Vehicle Price Prediction System using Machine Learning Techniques." International Journal of Computer Applications 167.9 (2017).

[5] Xinyuan Zhang, Zhiye Zhang and Changtong Qiu, "Model of Predicting the Price Range of Used Car", 2017

[6] Richardson, M., 2009. Determinants of Used Car Resale Value. Thesis (BSc). The Colorado College.

[7] Shim, Joo Yong, and Chang Ha Hwang. "Support Vector Quantile Regression Using Asymmetric e-Insensitive Loss Function." Communications for Statistical Applications and Methods 18, no. 2 (2011): 165–170

[8] Truong, Quang & Dang, Hy & Mei, Bo. (2020). Housing Price Prediction via Improved Machine Learning Techniques. Procedia Computer Science. 174. 433-442. 10.1016/j.procs.2020.06.111.

[9] Mr. S. Siva Prakash, Ahubhakumar, S.Appash, J.Cibiragul, 2019, Credit Card Fraud Detection using Adaboost and Majority Voting, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) RTI ACT – 2019 (Volume 7 – Issue 01)

[10] Roger W. Hoerl. (2020) Ridge Regression: A Historical Context. Technometrics 62:4, pages 420-425.

[11] Stefanowski, J.: Dealing with data difficulty factors while learning from imbalanced data. In: Challenges in Computational Statistics and Data Mining, pp. 333–363 (2016)

[12] Abdellatif, Safa & Ben Hassine, Mohamed Ali & Ben Yahia, Sadok & Bouzeghoub, Amel. (2018). ARCID: A New Approach to Deal with Imbalanced Datasets Classification.

[13] Pal, N., Arora, P., Kohli, P., Sundararaman, D., & Palakurthy, S. S. (2018, April). How Much Is My Car Worth? A Methodology for Predicting Used Cars' Prices Using Random Forest. In Future of Information and Communication Conference (pp. 413–422). Springer, Cham.

[14] Hurwitz, E., & Marwala, T. (2012). Common mistakes when applying computational intelligence and machine learning to stock market modelling. arXiv preprint arXiv:1208.4429.

[15] Potdar, Kedar & Pardawala, Taher & Pai, Chinmay. (2017). A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers. International Journal of Computer Applications. 175. 7-9. 10.5120/ijca2017915495.

[16] Figueiredo, Dalson & Júnior, Silva, & Rocha, Enivaldo. (2011). What is R2 all about? Leviathan-Cadernos de Pesquisa Polútica. 3. 60-68. 10.11606/issn.2237-4485.lev.2011.132282.