**PROJECT REPORT**

**ON**

**Used Car Price Prediction**

Logo

Description automatically generated

Team 9 (Section-02)

Vishakha Vinayak - 001550683

Pranali Vedpathak - 001563326

**Abstract**

This report consists of the analyses and conclusions drawn from Supervised Machine Learning algorithms used for the prediction of the used car price based on Craigslist ads.

The dataset for the project has been taken from Kaggle. The data is then cleaned as per the requirement and then various models were built in Jupyter Notebook (Anaconda 3) to examine the models’ performance on certain parameters.

The study focuses on various algorithms like Linear Regression, Decision Trees, Random Forest, Gradient Boosting, XGBoost.

**Table of Contents**

Contents

[**1.** **Introduction** 4](#_Toc89547125)

[**1.1 Background** 4](#_Toc89547126)

[**1.2 Motivation** 4](#_Toc89547127)

[**1.3 Goal** 4](#_Toc89547128)

[**2.** **Methodology & Algorithm** 4](#_Toc89547129)

[**2.1 Data Review** 4](#_Toc89547130)

[**2.2 Software and Libraries used** 4](#_Toc89547131)

[**2.3 EXPLORATORY DATA ANALYSIS** 5](#_Toc89547132)

[2.3.1 Removing unwanted features 5](#_Toc89547133)

[2.3.2 Handling Missing Values & Normalizing 5](#_Toc89547134)

[**2.4** **Data Visualization & Cleaning** 6](#_Toc89547135)

[3.5 Data Preprocessing and Modelling 10](#_Toc89547136)

[3.5.1 Data Preparation 10](#_Toc89547137)

[3.5.2 Modelling 10](#_Toc89547138)

[**Conclusion** 11](#_Toc89547139)

[**References** 11](#_Toc89547140)

# **Introduction**

## **1.1 Background**

Used Car Price system aims to predict the price of the secondhand cars based on the ads posted on craigslist. This system has ‘Price’ as the target variable and its predictions are based on a various attributes such as model, manufacturer, age of the car, odometer, condition, type, fuel, transmission, car status.

## **1.2 Motivation**

In US, most of the places are not accessible through public transport, rental cars are not readily available which leaves you with having your own vehicle. Buying a brand-new car is not within the reach of everyone instead buying a used car is a good option for people with less purchasing power and serve the purpose.

## **1.3 Goal**

This project intends to illustrate the modeling of a data set using machine learning to better predict the used car prediction. It is helpful for both who are buying as well as selling the car to propose/ buy the car based on market value taking various factors into account like odometer, make, transmission, type, model, fuel.

# **Methodology & Algorithm**

## **2.1 Data Review**

The dataset is from Kaggle by Austin Reese. This data is scrapped from craigslist which has the world’s largest collection of used vehicles for sale. It contains most recent relevant information from Craigslist including columns like price, condition, manufacturer, type and 23 other features provided in the site. Data size is 1.45 GB consisting of more than 400,000 samples.

Data Source link: <https://www.kaggle.com/austinreese/craigslist-carstrucks-data>

## **2.2 Software and Libraries used**

**Software** – Jupyter Notebook

**Libraries**

Data Manipulation: Pandas, Numpy, Matplotlib, Seaborn

Preprocessing: LabelEncoder, OneHotEncoder

Models : LinearRegression, DescisionTreeRegressor, GradientBoostingRegressor

Evaluation: metrics

Data Cleaning: missingno

## **2.3 EXPLORATORY DATA ANALYSIS**

In this step, the dataset is analyzed to perform the following:

* Removing unwanted features
* Handling Missing value & Normalizing

## Removing unwanted features

Rows that have identical data are probably useless can be misleading during model evaluation. The dataset is checked for values that can be ignored like image\_url,region\_url, url, region, description, size, lat, long, VIN. These records and empty values like county are removed from the dataset.

Chart, bar chart

Description automatically generated

## Handling Missing Values & Normalizing

There is a possibility of having few missing values in one or more columns of the dataset. Before modelling, these missing values are handled by dropping the few missing values

Text, table

Description automatically generated

After dropping the missing values and duplicates and changing the datatype of year, dataframe looks as follows:

Graphical user interface

Description automatically generated with low confidence

# **Data Visualization & Cleaning**

Visualization provides a better overview of data and helps to understand the relation between various features using correlation matrix, scatter & box plots

Understanding Non-Categorical features – In our case, price and odometer are two non-categorical

Table

Description automatically generated

Price values varies from $0 – $3,736,928,711.00 and mean price is $44,802.26 while odometer ranges from 0 – 10,000,000.00. These values are little realistic so, we are narrowing down the price values between 2k to 50k dollars

Chart, histogram

Description automatically generated

Similarly, for odometer, its highly unlikely to run 10 million miles eventually engine will be useless. Analyzing the number entries having 0 and more than 200K odometer values.

Text

Description automatically generated Text

Description automatically generated

There are about 9k values more than 200K which is very less compared to the no of values we have so can be ignored. Based on the above stats, 100 – 200K can be a good dataset. We can verify that, by calculating the skewness for all the data and between 100 to 200K odometer values.

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Skewness for first figure is 48.58 and for second is 0.37, so our data is evenly distributed around the range we have selected.

If we boxplot the values of price vs year, we can get the min and max values of car price for that year.

Chart, bar chart

Description automatically generated

People don’t buy very old models, filtering the dataset further to year 2000-2022. After 2000, the price are in increasing fashion

Further, if we compare the age of the car with price, we can see that as the model becomes older, price decreases. Similarly for the cars that are driven more are less expensive.

Chart, scatter chart

Description automatically generated

It is evident that luxuary cars are expensive over cheap manufactured one’s

Chart, scatter chart

Description automatically generated

Most of the features in this dataset are categorical will use Label Encoding. We will be carrying out encoding on all categorical features expect for models since it has highest cardinality

Table

Description automatically generated with medium confidence

Using correlation matrix, we see that price is positively correlated with year and negatively with age, odometer. We don’t have any feature that is highly correlated, so won’t be dropping any feature at this stage.

Timeline

Description automatically generated

## 3.5 Data Preprocessing and Modelling

### 3.5.1 Data Preparation

1. X will be all the features except target value i.e price

2. Y will contain target value

3. Splitting the data into training(70%) and testing(30%) sets

### 3.5.2 Modelling

Our target data is non-categorical field, we are using regression models over classifiers. To evaluate the model, we are using following metrics:

1. Coefficient of Determination or R2 – variation in dependent variable(Y)
2. Root mean square error (RMSE) – root of absolute measure of error
3. Mean absolute error (MAE) – sum of absolute value of error
   * + 1. Linear Regression

Text

Description automatically generated

* + - 1. Decision Tree Regressor

Graphical user interface, text, application, email

Description automatically generated

* + - 1. Random Forest Regressor

Text

Description automatically generated

* + - 1. Gradient Boosting Regression

Graphical user interface, text, email

Description automatically generated

* + - 1. XGBoost Regression

Text

Description automatically generated

* + 1. Summary

Table

Description automatically generated

From the above results, Random Forest has the best testing accuracy compared to other models used.

Decision Tree is highly overfitting model. Gradient Boost is a well generalized model with low over fit and low bias.

Chart, bar chart

Description automatically generated

# **Conclusion**

With Accuracy Score of all the used Algorithms, the Random Forest Regressor produced a significant accuracy for prediction of the used car price

# **References**

* Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. [“A Countrywide Traffic Accident Dataset.”](https://arxiv.org/abs/1906.05409), arXiv preprint arXiv:1906.05409 (2019).