**PRICE PREDICTION OF USED CARS**

**Milestone: Project Report**

Group 12

Student 1: Amar Sai Kiran Poosarla

Student 2: Venkata Sai Krishna Paruchuri

857-230-1191

857-334-6587

[poosarla.a@northeastern.edu](mailto:poosarla.a@northeastern.edu)

[paruchuri.v@northeastern.edu](mailto:paruchuri.v@northeastern.edu)

**Percentage of Effort contributed by Student 1:** 50%

**Percentage of Effort contributed by Student 2:** 50%

**Signature of Student 1:** Amar Sai Kiran Poosarla

**Signature of Student 2:** Venkata Sai Krishna Paruchuri

**Submission Date:** 04/25/2022

**Table of contents**

[**I. Problem Setting**](#one)  **3**

[**II. Problem Definition**](#two)  **3**

[**III. Data Source**](#three)  **3**

[**IV. Data Description**](#four)  **4** Table 1:- Description of variables  **4**

[**V. Data Collection**](#five) **5**Fig 1:- Dataset Information **5**

[**VI. Data Mining Tasks (Processing)**](#six)

Fig 2:- Cleaned Dataset preview **6** Fig 3:- Statistics of the dataset **7** Fig 4:- Pair Plot **7**

Fig 5: - Correlation **8** Fig 6:- Box Plot between Transmission and Fuel type **8**

[**VII. Data Exploration and Visualization**](#seven)  **9**

Fig 7 & 8: - Visualized with and without outliers **9** Fig 9: - Visualization using Heat map **10** Fig 10: - Visualization using Tree Map **11**

[**VIII. Model selection and exploration**](#eight) **14**

[**IX. Implementation of Selected models**](#nine) **15**

Fig 15 – Fig 24: - All models regression summary **15**

[**X. Project Result / Visualization**](#ten)  **21**

Fig 25 – Fig 27: - Visualization Insights **22**

Fig 28 :- RMSE Bar chart **23**

[**XI. Impact of the project outcomes**](#eleven) **23**

**DATA EXPLORATION AND VISUALIZATION**

**IE 7275: DATA MINING IN ENGINEERING**

**I. Problem Setting:**

According to a recent survey, the automotive industry has seen a decline in the production of cars due to chip shortages compared to previous years. This had a massive impact on car manufacturing and sales. But the demand from the customers has not decreased. In fact, the demand for the used cars has been inflated due to the global pandemic, also for safe and secured travel. This has raised the question of giving a reasonable car price to the customers.

**II. Problem Definition:**

The Dataset considered is about the reasonable prices of the past 10 years of various car brands, including luxury brands like AUDI and BMW and economy cars like Hyundai and Toyota with their models.

The purpose of this analysis is to predict the price of the used car based upon various specifications of the car like the model, year of manufacturing etc. The analysis we use will help customers to get the price predicted before they get a quote from the used cars companies.

**III. Data Source:**

The Dataset has been taken from Kaggle.com, an online platform for enthusiastic data scientists and machine learning professionals.

Kaggle.com is an online storehouse for a wide variety of datasets.

Topic of the Dataset: - **100,000 UK Used car dataset**

Dataset source link is <https://www.kaggle.com/adityadesai13/used-car-dataset-ford-and-mercedes>

**IV. Data Description:**

The Dataset consist of a total 108,540 instances and have 9 attributes with one target attribute “PRICE”

|  |  |  |
| --- | --- | --- |
| SNO | Variable | Description |
| 1 | Model | Car model |
| 2 | Year | Year of manufacturing |
| 3 | Price | Price of Used cars |
| 4 | Transmission | Type of transmission (Manual/automatic/ semi-auto) |
| 5 | Mileage | Total distance travelled |
| 6 | Fuel type | Type of fuel (Petrol/Diesel) |
| 7 | Tax | Sales tax |
| 8 | MPG (Miles per gallon) | Performance in terms of miles per gallon |
| 9 | Engine size | Capacity of the Engine |

**Table-1 Description of variables**

**V. Data Collection:**

Data collected in initial phase is from different sources (car brands) which is later merged. All the data required for exploration, visualization and prediction has been gathered, collected, and cleaned.

Data collection is primarily required for measuring information available on variables used for prediction. Variables used for price prediction are analyzed in systematic way, enabling one to test hypotheses and evaluate outcomes.

**Dataset Information:**

Text, table

Description automatically generated

FIG-1

**VI. Data Mining Tasks (Processing):**

Initially dataset consisted of 108,540 rows and 11 columns.

The first step towards processing of data is removing the duplicate values which decreases redundancy in dataset. There are around 2284 duplicate records which would have disturbed model’s accuracy if not removed.

The tax (£) columns which contained more than 90% of missing values cannot be fitted by mean values hence entire column has been dropped.

All the missing values in individual columns have been replaced by the mean of their category.

After these two steps of cleaning, we still ended up with 4784 and 18 missing values in tax and mileage columns respectively.

These missing values are due to non-availability of non-zero values in entire group by category.

The category is defined as data grouped by ‘model’, ’year’, ’transmission’,

’Fuel-type’, ’mpg’, ’mileage’.

To remove inconsistency in dataset all the remaining null value records have been removed from the dataset.

The final cleaned and preprocessed dataset consists of data cleaning we finally ended up with 101,475 and 10 columns.

Table

Description automatically generated

FIG-2

The columns transmission and fuel type which are categorical variables are encoded numerically for statistical interpretation purpose.

Statistics (mean, median etc.) have been calculated after following numerical transformation.

Table

Description automatically generated

FIG-3

Pairplot has been plotted between the variables to observe the correlation.

Chart

Description automatically generated with medium confidence

FIG-4

In support of above pairplot we calculated correlation between the numerical variable’s mileage, price, fuel type, tax and found that no correlation exceeded 0.5 value.

Table

Description automatically generated

FIG-5

Boxplot for price with respect to transmission and differentiation based on fuel type have given insight regarding the quantity of outliers present in data.

Chart, box and whisker chart

Description automatically generated

FIG-6

Since dimension reduction cannot be possible based on following two interpretation further data exploration is needed.

**VII. Data Exploration and Visualization:**

From the above figure **FIG-6** (previous page)

* We found that there are many outliers in our processed data set with the help of boxplot and decided to remove the outliers.
* Box plot has been plotted for price with respect to transmission and only for tax variable.
* After calculating interquartile region for tax distribution, it is observed that there are more than **6000 records** of outliers which is in turn significant percentage of dataset.
* Therefore, if the test score is not as expected for models, outliers will be substituted by their respective mean in every variable and re-test using new dataset.
* The best test score dataset and model will be finalized.

Chart, box and whisker chart

Description automatically generated

Fig – 7

In fig 7 it clearly shows that there are many outliers only with respect to TAX column

**Visualized with and without outliers:**

Insight: in the fig-8, bottom right boxplot, outliers are removed and that creates a huge data loss

Graphical user interface, chart, box and whisker chart

Description automatically generated

Fig-8

**Visualization using heatmap:**

Heatmap has been plotted with using the correlation values of four numerical response variables (transmission, fuel Type, mileage, tax).

Insight: It can be observed that there is no high correlation between the response variables

Hence dimension reduction is possible.

Chart, bar chart

Description automatically generated

Fig -9

**Visualization using Tree map:**

A tree map has been plotted for price with respect to brands, models.

Insight: It has been observed that the price for C Class brand and C Class model is highest.

Respective model price to brand is visualized using.

Chart, treemap chart

Description automatically generated

Fig-10

**Visualized variation of price (response variable) with respect to each predictor:**

Insight for Price vs Transmission: Total price is highest for transmission ‘Manual’

Note: - ‘0’=Manual, ‘1’= Automatic, ‘2’= Semi-auto, ‘3’= Others

Chart, bar chart

Description automatically generated

Fig-11

Insight for Price vs Brand: Total price is highest for Brand ‘merc’

Chart, bar chart

Description automatically generated

Fig – 12

Insight for Price vs fuel Type: Total price is highest for Petrol

Note: - ‘0’=Petrol, ‘1’= Diesel, ‘2’= Hybrid, ‘3’= Electric, ‘4’= Others

Chart, bar chart, histogram

Description automatically generated

Fig-13

Insight for Price vs year: Total price is highest for year 2019 as the count for cars is high in 2019

In the

Fig 14-a shows the Sum of total price for all brands vs Year

Fig 14-b shows the Count of highest price in all brands vs Year

Chart, histogram

Description automatically generated

Fig = 14-a Fig = 14-b

Since our goal is price prediction, insights have been concentrated on price with respect to predictor variables rather than total count

**VIII. Model selection and exploration:**

Based on the visualization, correlation and heatmaps it has been decided not to eliminate the outliers. There are more than 4 numerical and categorical variables for the model to be built.

Model building have few steps:

1. Training
2. Validation
3. Testing

The given dataset has been split into 75% for training and 25% for validation. Using 75% of training dataset model has been built. The built model has been validated using rest 25% of dataset.

Six models are chosen for training and validation.

1)Linear Regression

2)K-NN Algorithm

3) Lasso

4) Bayesian

5) Decision tree

6) Random forest

**Processes and Exploration involved:**

1)Scaling is performed to maintain uniformity of various predictors and price to get accurate result

2)Uniformity in the data type of variables is necessary. Therefore, the transmission object type data type is converted into float to maintain necessary uniformity

3)It is observed that there is slight change due to the “other” in transmission therefore 7 records containing “other” are eliminate

**Steps to build and execute for all models:**

1)Predictors (X) and response (y) data frames are sliced from our dataset to find train and validating predictors and response variables

2)Our dataset consists of 101475 records after cleaning and preprocessing which is split into 75% and 25% using test split function

3)The train and valid X, y is scaled, fit, and transformed using standard scaler function imported from sklearn library

4)Scaled X and y are used to train and validate the data. The regression summary in linear regression module is used to find various statistics which determine the performance of module.

Example: RMSE, MSE

5)Where as KNN model performance is based on the k value (nearest neighbor value). The accuracy of k for a particular range of k determines the performance of the model

6)Necessary modules are imported from sklearn for respective models and the train and test datasets are fitted accordingly. Fitted predictors and outcomes are used to predict the outcomes which in turn used to calculate residuals and regression summaries

**IX. Implementation of Selected models**

**Multiple Linear Regression:**

It is a statistical method which is based on several predictors and an extension of linear (OLS) regression.

Training Result

Text, table

Description automatically generated

FIG - 15

Validation Result

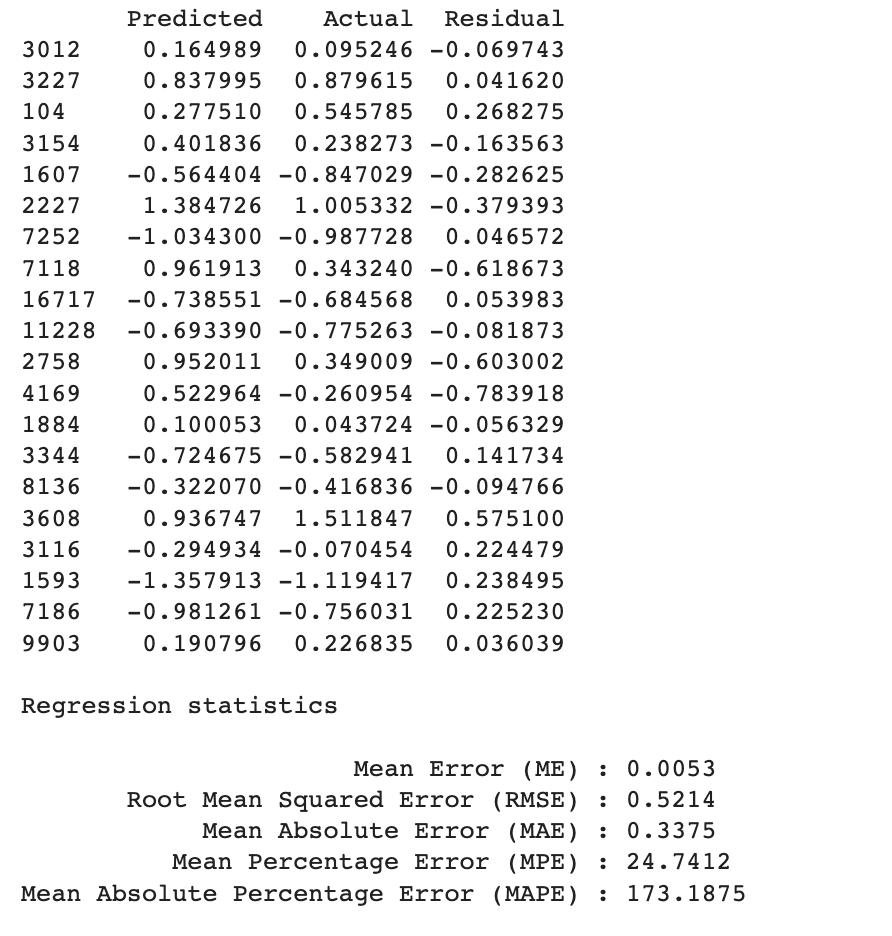


Fig - 16

Chart, histogram

Description automatically generated

FIG – 17

**K-NN:**

It is a supervised algorithm which is used for both classification and prediction. The model based on the similarity of the nearest neighbors (predictor observations) of response variables

KNN Result

Text

Description automatically generated

FIG – 18

Regression statistics for KNN

Text

Description automatically generated

Fig 19

**Lasso Result**

Text

Description automatically generated

Fig - 20

**Bayesian Model**

***A picture containing table

Description automatically generated***

Fig - 21

**Decision Tree:**

Decision tree is a supervised learning model which can be implemented in both classification and regression problems.

It uses tree representation to separate differently classified outcomes in classification problems and range of numerical outcomes using visual color representation

**A screenshot of a computer

Description automatically generated with medium confidence**

Fig - 22

**Random Forest:**

A random forest may search for the best predictor to split among a random subset of predictors.

Random forest is an ensemble of decision trees.

Each decision tree is trained on a bootstrap sample.

Final predictions is made considering predictions of all individual trees.

**Text

Description automatically generated**

Fig - 23

**R2 Summary for all the models**

**Text

Description automatically generated**

Fig - 24

**X. Visualization for all the models:**

Actual vs predicted

**Chart, scatter chart

Description automatically generated**

**A picture containing background pattern

Description automatically generated**

Fig - 25

**Visualization Insight:**

We can find that the plots are clumsy due to large number of records and similar residual values for various predictions however the model is accurate if the best fit line is identical to identity line.

The best fit line is identical to identity line for Random Forest model

Distribution

Graphical user interface

Description automatically generated with low confidence

Fig - 26

**Visualization Insight:**

The residual error distribution is spread around zero more for random forest model than any other model

**Visual representation in tabular form: -**

**Table

Description automatically generated**

Fig - 27

**RMSE Bar Chart**

**Chart, bar chart

Description automatically generated**

Fig - 28

**Result:**

All the regression summary statistics are minimum for Random Forest regression model

Thus, making it the most suitable for regression model for this dataset.

**XI. Impact of the Project Outcomes:**

The demand has been shifted to Used Cars in recent times due to shortage of vehicles as there is a dip in global semi-conductor manufacturing. Therefore, there is a need to calculate the price of used car accurately which would be beneficial for clients as well as the dealer.

Our model contains all the important factors like transmission, year of manufacturing, mileage, fuel type etc. However, there are few constraints we would face in real world i.e the demand surge and this prediction is limited to gas vehicles and doesn’t include electrical vehicles.

This model impact might become obsolete in upcoming years and needs to be updated regularly