Logo, company name

Description automatically generated

**PROJECT REPORT ON:**

**“Car Price Prediction”**

**SUBMITTED BY**

**Ajit Madame**

**ACKNOWLEDGMENT**

I would like to express my special gratitude to “Flip Robo” team, who has given me this opportunity to deal with a beautiful dataset and it has helped me to improve my analyzation skills. And I want to express my huge gratitude to Ms.Gulshana Chaudhary (SME Flip Robo), she is the person who has helped me to get out of all the difficulties I faced while doing the project.

A huge thanks to my academic team “Data trained” who are the reason behind what I am today. Last but not least my parents who have been my backbone in every step of my life. And also thank you for many other persons who has helped me directly or indirectly to complete the project

**Contents:**

1. **Introduction**
   * Business Problem Framing
   * Conceptual Background of the Domain Problem
   * Review of literature
   * Motivation for the Problem undertaken

**2. Analytical Problem Framing**

* + Mathematical/ Analytical Modelling of the Problem
  + Data Sources and their formats
  + Data Pre-processing Done
  + Data Input – Logic – Output Relationships
  + Hardware, Software and Tools Used

**3. Data Analysis and Visualization**

* + Univariate Visualization
  + Bivariate Visualizations
  + Multivariate Visualization

**4. Model Developments and Evaluation**

* + The model algorithms used
  + Interpretation of the result
  + Hyperparameter tuning

**5. Conclusions**

* + Key Finding and conclusions
  + Limitation of this works and scope for future works

**1.INTRODUCTION**

1.1Business Problem Framing:

With the covid 19 impact in the market, we have seen lot of changes in the car market. Now some cars are in demand hence making them costly and some are not in demand hence cheaper. With the change in market due to covid 19 impact. For this I am collecting data from cars24.com and make data frame and used this data for building the model.

The project Car Price Prediction deals with providing the solution to this problem. Through this project, we will get to know which of the factors are significant and tell us how they affected the cars market.

* 1. Conceptual Background of the Domain Problem

A good knowledge of after sales market of cars is necessary. What makes a car valuable will be key. As the mobile internet improves by leaps and bounds, the model traditional offline used car trading has gradually lost the ability to lives up to the needs of customers, and online used car trading platforms have emerged as the times require. Second-hand car price prediction is the premise of second-hand car trading, and reasonable price can reflect the objective, fair and true nature of the second-hand car market.

* 1. Review of Literature

The first paper is Predicting the price of Used Car Using Machine Learning Techniques. In this paper, they investigate the application of supervised machine learning techniques to predict the price of used cars in Mauritius. The predictions are based on historical data collected from daily newspapers. Different techniques like multiple linear regression analysis, Random forest regressor, Gradient Boosting Regression, XGBoost regressor havebeen used to make the predictions.The Second paper is Car Price Prediction Using MachineLearning Techniques. Considerable number of distinctattributes are examined for the reliable and accurateprediction.

* 1. Motivation for the Problem Undertaken

The goal of this project is to create machine learning models that can properly forecast the price of a used car based on its attributes so that buyers can make educated decisions. On a dataset containing the sale prices of various brands and models, we build and analyses several learning approaches. Due to covid-19 the car market has changed a lot, some cars have shot up in popularity and some gone down in price.

**2.Analytical Problem Framing**

* 1. Mathematical/ Analytical Modelling of the Problem

Inbuilt function such as standardising and log will be used in tackling this problem.

R-square is a comparison of residual sum of squares *(SSres)* with total sum of squares*(SStot)*. Total sum of squares is calculated by summation of squares of perpendicular distance between data points and the average line.

https://media.geeksforgeeks.org/wp-content/uploads/20190415232942/CodeCogsEqn-10.gif

Where *SSres* is the residual sum of squares and *SStot* is the total sum of squares.

R-square is the main metric which I will use in this regression analysis.

Concordance index was also used. The concordance index or c-index is a metric to evaluate the predictions made by an algorithm. It is defined as the proportion of concordant pairs divided by the total number of possible evaluation pairs.

2.2 Data Sources and their formats

The data was scraped from cars24 website; data was scraped for different cities where prices differ.

Table

Description automatically generated

2.3 Data Pre-processing Done

* First step I have imported required libraries and I have imported the dataset which was in csv format.
* Then I did all the statistical analysis like checking shape, nunique, value counts, info etc.
* I found that, few features are containing null value so fill it with mean value.
* Numerical variables were converted to integer type (form string) so I could perform deeper analysis on them.
* Then doing some EDA and Building Models.

2.4 Data Inputs - Logic - Output Relationships

The main assumption is that there is no selection bias in the data which we have.

This is because we have cars from varying years and varying city; each city doesn’t have equal amount of data. Here we can see the count of data per city.

2.5 Hardware, Software and Tool Used

**Hardware Used:**

Processor – Intel core i3

Physical Memory – 8 GB

**Software Used:**

* Windows 10 Operating System
* Anaconda Package and Environment Manager
* Jupyter Notebook
* Python Libraries used: In Which Pandas, Seaborn, Matplotlib, Numpy and Scipy
* sklearn for Modelling Machine learning algorithms, Data Encoding, Evaluation metrics, Data Transformation, Data Scaling, Component analysis, Feature selection etc.

**3.Data Analysis and Visualization**

3.1 Univariate Visualization

A picture containing chart

Description automatically generated

* We can see selling price is look like that data has normally distributed.
* We can see, maximum selling price lies in the range of 4 to 6 Lakh.

Chart, histogram

Description automatically generated

* We can see, km driven has not a normally distributed but maximum data lies in the range of 0 to 100000 km.

Chart, bar chart

Description automatically generated

* We can see, Fuel as petrol has maximum count followed by Diesel.
* Petrol + CNG has lowest count than others.
* It means that maximum people are used petrol vehicle than others. But petrol vehicles are more expensive than others.

Chart, bar chart

Description automatically generated

* We can see, 1st Owner owned vehicle are highest for selling followed by 2nd Owner.
* It simple because 1st owner vehicles are having more selling than others.

3.2 Bivariate Visualization

Chart, scatter chart

Description automatically generated

* We can see, there are no direct relationship of both feature to each other’s.
* But the vehicle km driven are in 0 to 80000km those are having good price. And we can see maximum vehicle km driven lies in this range.

Chart, bar chart

Description automatically generated

* We can see, Diesel vehicles are having maximum selling price than others.
* Whereas petrol vehicles are having lowest selling price than others. But all are having very little difference in their selling price.

Chart, bar chart

Description automatically generatedChart

Description automatically generated

* We can see, Diesel cars are having maximum km\_driven. It means that, diesel cars runing km is high they used more. Petrol cars are having lowest km\_driven. But here is little contrast that, generally those cars driven less km they are having high price but here is opposite.
* We can see, manual and automatic both type of transmission is having almost same km\_driven.
* But Automatic transmission little lower than Manual transmission cars.

Chart, bar chart

Description automatically generatedChart, box and whisker chart

Description automatically generated

* We can see, Automatic transmission cars are having higher selling price than Manual transmission cars. But there is no major difference in selling price of both transmission cars. We can see, up to the range 0 to 450000 both are having same price.
* We can see, both transmissions are having almost same km driven.

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

* We can see, 3rd owner cars are having maximum km\_driven. It obviously because these are cars owned by multiple owners and they drive these cars more. 1st owner cars are having less km driven.
* We can see, 1st owner cars are having maximum selling price than others. But all are having almost same range of selling price.

3.2 Multivariate Visualization

Chart, bar chart

Description automatically generated

* We can see, in all variants of fuel, 1st owner are having maximum cars selling price than others owner.
* But in Petrol fuel variant, all owners are having almost same selling price there is no major difference in selling price.

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

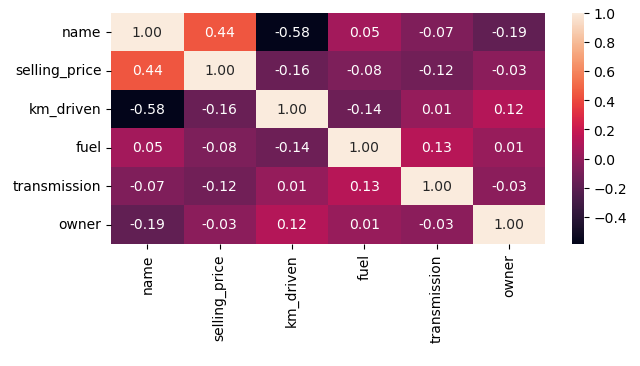
* 2nd owner is top in petrol + CNG and diesel variants of car. It means in these two variants of fuel in which 2nd owner are drive a cars maximum km than others. But in Petrol fuel cars, in which 3rd owner cars are having maximum km driven.
* We can see, in all fuel type, Manual transmission cars are having maximum km\_driven. But in Petrol+CNG variants cars, there is Automatic transmission cars.

Chart, bar chart

Description automatically generated

* We can see, In Petrol fuel type, In which automatic transmission cars are having maximum selling price than others.
* But petrol+CNG fuel type, there is no automatic transmission cars.

**Correlation of the features with target columns**



* name has 44% positive correlation with target column.
* km\_driven has 16% negative correlation with target column.
* fuel has 8% negative correlation with target column.
* transmission has 12% negative correlation with target column.
* owner has 3% negative correlation with target columns.

**Visualizing correlation of feature columns with label column.**

**Chart, waterfall chart

Description automatically generated**

* Name have the strongest positive correlation with Selling Price.
* While km\_dricven, transmission, fuel have the strongest negative correlation with Selling Price.

**Checking Outliers**

Chart, box and whisker chart

Description automatically generated

* We can see, selling price and km\_driven are having some outliers.

**Removing Outliers**

**Text

Description automatically generated**

**Percentage of data loss**

**Graphical user interface, text, application, chat or text message

Description automatically generated**

4. Models Development and Evaluation

4.1 The model algorithms used

**Checking Multicollinearity**

**Graphical user interface, text, application, email

Description automatically generated**

**Graphical user interface, application

Description automatically generated**

**Linear Regression:**

**Finding Best Random State**

**Text

Description automatically generated**

**Train and Test**

**Text

Description automatically generated**

**Graphical user interface, text, application

Description automatically generated**

**Cross Validation of the Model**

**Graphical user interface, text, application

Description automatically generatedText

Description automatically generated**

**Graphical user interface, text

Description automatically generated**

**Random Forest Regressor**

**Finding Best Random State**

**Graphical user interface, text, application

Description automatically generated**

**Train and Test**

**Graphical user interface, text

Description automatically generated**

**Graphical user interface, text, application

Description automatically generated**

**Cross Validation of the Model**

**Text

Description automatically generatedGraphical user interface, text

Description automatically generated**

**Graphical user interface, text, application

Description automatically generated**

**Gradient Boost Regressor**

**Finding Best Random State**

**Text

Description automatically generated**

**Train and Test**

**Graphical user interface, text

Description automatically generated**

**Graphical user interface, text, application

Description automatically generated**

**Cross Validation of the Model**

**Text

Description automatically generated**

**Text

Description automatically generatedText

Description automatically generated**

**XGBRegressor**

**Finding Best Random State**

**Graphical user interface, text, application

Description automatically generated**

**Train and Test**

****

**Graphical user interface, text, application, email

Description automatically generated**

**Cross Validation of the Model**

**Text

Description automatically generatedText

Description automatically generated**

**Graphical user interface, text, application

Description automatically generated**

**Regularization**

**Text

Description automatically generatedGraphical user interface, text, application

Description automatically generated**

4.3 Interpretation of the results

Based on comparing Accuracy Score results with Cross Validation results, it is determined XGboost Regressor is the best model. It has least difference between accuracy score and cross validation score.

4.4 Hyperparameter Tuning

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Based on the input parameter values and after fitting the train datasets The XGBoost Regressor model was further tuned based on the parameter values yielded from GridSearchCV. The XGBoost Regressor model displayed an accuracy of 68.62%.

**Scatter plot Between Actual and Predicted Selling Price of Car**

**Chart, scatter chart

Description automatically generated**

**The Model Saving and Testing**

**Graphical user interface, text, application, email

Description automatically generatedText

Description automatically generated**

**Graphical user interface, application

Description automatically generated**

**5. Conclusions**

5.1 Key Finding and Conclusions

The main component on which the price of a car depends is the engine size, the year which car was bought, the mileage on the car etc.

The price also depends on which city the car was registered, as some cities have different tax rates and restrictions.

XGBRegressor works best for this particular data set, hyper parameter tuning was performed, and optimal parameters were found.

EDA is very powerful in understanding the data and pre-processing it before feeding it to the algorithm. Statistical methods work the best.

5.2 Limitation of this works and scope for future works

Post covid-19 car market is still evolving, and it will keep evolving for the foreseeable future. The algorithms will need to keep changing to keep up with the evolution.