
# Car Price Prediction

Submitted by:

Laxmikant Deepak

# ACKNOWLEDGMENT

I have referred below resources that helped and guided me in completion of this project as below :-

www.w3resource.com

www.towardsdatascience.com

www.stackoverflow.com

# INTRODUCTION

- Business Problem Framing

Price prediction of second-hand cars depends on numerous factors. The most important ones are manufacturing year, make, model, mileage, horsepower and country of origin. Some other factors are type and amount of fuel per usage, the type of braking system, its acceleration, the interior style, its physical state, volume of cylinders (measured in cubic centimeters), size of the car, number of doors, weight of the car, consumer reviews, paint color and type, transmission type, whether it is a sports car, sound system, cosmic wheels, power steering, air conditioner, GPS navigator, safety index etc. In the Mauritian context, there are some special factors that are also usually considered such as who were the previous owners and whether the car has had any serious accidents.

- Conceptual Background of the Domain Problem

  Domain model is a structured visual representation of interconnected concepts or real-world objects that incorporates vocabulary, key concepts, behavior, and relationships of all of its entities.

- Review of Literature

  They are one of the biggest target group that can be interested in results of this study. If used car sellers better understand what makes a car desirable, what the important features are for a used car, then they may consider this knowledge and offer a better service.

- Motivation for the Problem Undertaken

**Used car sellers (dealers):** They are one of the biggest target group that can be interested in results of this study. If used car sellers better understand what makes a car desirable, what the important features are for a used car, then they may consider this knowledge and offer a better service.

**Online pricing services:** There are websites that offers an estimate value of a car. They may have a good prediction model. However, having a second model may help them to give a better prediction to their users. Therefore, the model developed in this study may help online web services that tells a used car's market value.

**Individuals:** There are lots of individuals who are interested in the used car market at some points in their life because they wanted to sell their car or buy a used car. In this process, it's a big corner to pay too much or sell less then it's market value.

# Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

  Describe the mathematical, statistical and analytics modelling done during this project along with the proper justification.

- Data Sources and their formats

  I am use for data web scrapping for extract dataset  save in excel format.

- Data Preprocessing Done

  First I am use for data web scrapping for extract dataset  save in excel format and the data analysis with machine learning .

- Data Inputs- Logic- Output Relationships

  Input data for feature list and target is in numeric format and hence classification model  best suits for this dataset

- State the set of assumptions (if any) related to the problem under consideration

  Here, you can describe any presumptions taken by you.

- Hardware and Software Requirements and Tools Used

  - **Hardware Requirement-**Laptop with below configurations-
  - Windows Edition-Windows 10 Pro
  - Processor-Intel i3
  - Memory-12 GB
  - System Type-64 bit OS

- **Software Requirement-** Anaconda 3.7 & above , Jupiter Notebook 6.

## Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)
- **Analytical Approach –**Based on type of data by performing EDA I have decided which model to be used for this data.
- **Statistical Approach –** Data should be in scaled manner,it should not be distorted,for that I have replace all null values using mean methgod due to continuous data numbers.
- Testing of Identified Approaches (Algorithms)
  - Below are classification algorithms used for the training and testing this dataset.

- Run and Evaluate selected models

### Training and predicting

I'll create a **Linear Regression** model and a **Random Forest** model to train on the data and use it for future predictions.

```
In [24]: linearRegression = LinearRegression()
         linearRegression.fit(X_train, y_train)
         y_pred = linearRegression.predict(X_test)
         r2_score(y_test, y_pred)

Out[24]: 0.7008908549416721

In [25]: rf = RandomForestRegressor(n_estimators = 100)
         rf.fit(X_train, y_train)
         y_pred = rf.predict(X_test)
         r2_score(y_test, y_pred)

Out[25]: 0.8860868487769373
```

The **Random Forest** model performed the best with a R2 score of **0.88**.

- Key Metrics for success in solving problem under consideration

  Used cross validation matrix to overcome under-fitting /over-fitting this model by deciding number of folds

- Visualizations

  Mention all the plots made along with their pictures and what were the inferences and observations obtained from those. Describe them in detail.

  If different platforms were used, mention that as well.

- Interpretation of the Results

  Visualisation shows outliers which need to be removed / corrected.

  Data Pre-processing done by performing EDA (Exploratory Data Analysis), checking for best accuracy score.

  Modelling done based on type of data as this is categorical data, we have to go with multiple classification models & finalise the best score giving model.

## CONCLUSION

- Key Findings and Conclusions of the Study

  By performing different models, it was aimed to get different perspectives and eventually compared their performance. With this study, it purpose was to predict prices of used cars by using a dataset that has

observations. With the help of the data visualizations and exploratory data analysis, the dataset was uncovered and features were explored deeply.

- Learning Outcomes of the Study in respect of Data Science

This dataset is categorical in nature ,we can verify data by using read method & get stats related information for each column using describe method.

As its categorical data, classification model best suits for this.

- Limitations of this work and Scope for Future Work

As suggestion for further studies, while pre-processing data, instead of using label encoder, one hot encoder method can be used. Thus, all non-numeric features can be converted to nominal data instead of ordinal data (Raschka & Mirjalili, 2017). This may cause a serious change in performance of predictive models. Also, after training the data, instead of min-max scaler, standard scaler can be performned and results can be compared. Different scalers can be checked whether there is an improvement in prediction power of models or not..