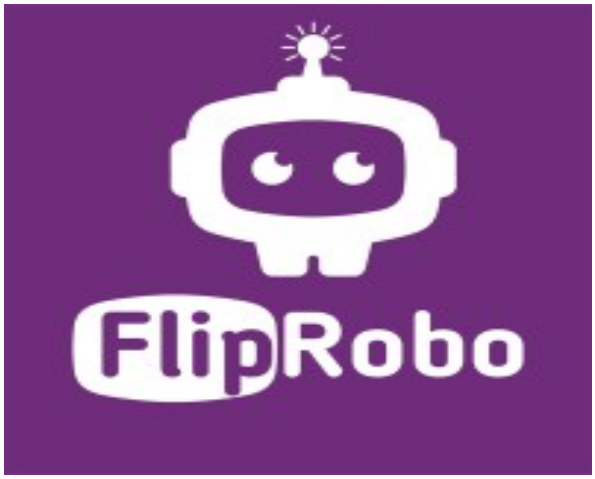
NAME OF THE PROJECT

CAR PRICE PREDICTION



Submitted by:

Pratyush Raj

ACKNOWLEDGMENT

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

A huge thanks to “Data trained” who are the reason behind my Internship at Fliprobo. Last but not least my parents have been my backbone in every step of my life.

References used in this project:

1. SCIKIT Learn Library Documentation
2. Blogs from towardsdatascience, Analytics Vidya, Medium
3. Andrew Ng Notes on Machine Learning (GitHub)
4. Data Science Projects with Python Second Edition by Packt
5. Hands-on Machine learning with scikit learn and tensor flow by Aurelien Geron
6. Pudaruth, S. (2014). Predicting the price of used cars using machine learning techniques. Int. J. Inf. Comput. Technol, 4(7), 753-764 – [1]
7. Agencija za statistiku BiH. (n.d.), retrieved from: http://www.bhas.ba . [accessed July 18, 2018.] – [2]

Chap 1. Introduction

1.1 Business Problem Framing

The price of a new car in the industry is fixed by the manufacturer with some additional costs incurred by the Government in the form of taxes. So, customers buying a new car can be assured of the money they invest is worthy. But, due to the increased prices of new cars and the financial incapability of the customers to buy them, Used Car sales are on a global increase. 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.

Therefore, there is an urgent need for a Used Car Price Prediction system that effectively determines the worthiness of the car using a variety of features. The existing System includes a process where a seller decides a price randomly and the buyer has no idea about the car and its value in the present-day scenario. In fact, the seller also has no idea about the car's existing value or the price he should be selling the car. To overcome this problem there is a need to develop a model which will be highly effective to predict the actual price of a car rather than the price range of a car.

1.2 Conceptual Background of the Domain Problem

Determining the listed price of a used car is a challenging task, due to the many factors that drive a used vehicle's price on the market. The focus of this project is developing machine learning models that can accurately predict the price of a used car based on its features, in order to make informed purchases.

Accurate car price prediction involves expert knowledge because price usually depends on many distinctive features and factors. Typically, the most significant ones are brand and model, age, horsepower, and mileage. The fuel type used in the car as well as fuel consumption per mile highly affects the price of a car due to frequent changes in the price of fuel. Different features like exterior color, door number, type of transmission, dimensions, safety, air condition, and interior, and whether it has navigation or not will also influence the car price. In this project, we applied different methods and techniques in order to achieve higher precision in the used car price prediction.

Regression Algorithms are used because they provide us with continuous value as output and not a categorized value because of which it will be possible to predict the actual price of a car rather than the price range of a car.

The data associated with the investigation was very large because there are thousands of used cars and each car's data comprises values of many features. Both data gathering and analysis are complex. Used cars data scrape from www. cardheko.com which is a well-known online platform for reselling used and new cars in India. Features like car's model, make, seating capacity, color, mileage, engine capacity, brakes, Torque, Transmission type, Maximum power, Gearbox type, power steering, engine type, turbocharger, supercharger, and price were included.

Review of Literature

Pudaruth applied various machine learning algorithms, namely: k-nearest neighbors, multiple linear regression analysis, decision trees, and naïve Bayes for car price prediction in Mauritius. The dataset used to create a prediction model was collected manually from local newspapers in a period of less than one month, as time can have a noticeable impact on the price of the car. He studied the following attributes: brand, model, cubic capacity, mileage in kilometers, production year, exterior color, transmission type, and price. However, the author found out that Naive Bayes and Decision Tree were unable to predict and classify numeric values. Additionally, a limited number of dataset instances could not give high classification performances, i.e., accuracies less than 70%.

As per the information that was gotten from the Agency for Statistics of BiH, 921.456 vehicles were registered in 2014 of which 84% of them are cars for personal usage [2]. This number is increased by 2.7% since 2013 and it is likely that this trend will continue, and the number of cars will increase in the future. This adds additional significance to the problem of car price prediction.

Motivation for the Problem Undertaken

The project was the first provided to me by Flip Robo Technologies as a part of the internship program. The exposure to real-world data and the opportunity to deploy my skillset in solving a real-time problem has been my primary motivation.

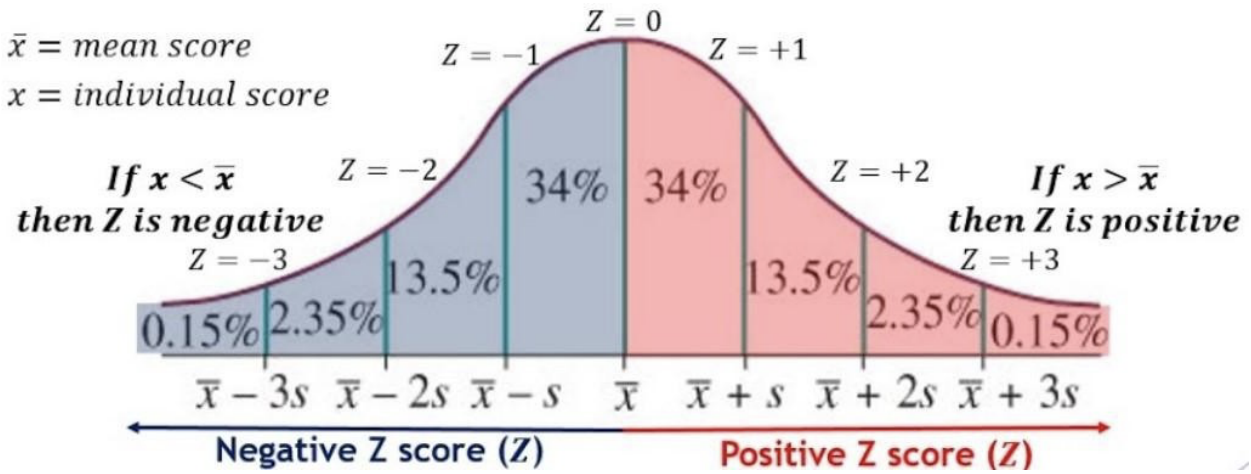
Data needed for this project is required to scrap from the internet and work over it. Deciding whether a used car is worth the posted price when you see listings online can be difficult. The focus of this project is developing machine learning models that can accurately predict the price of a used car based on its features, in order to make informed purchases. The model

developed in this study may help online web services that tells a used car's market value.

Analytical Problem Framing

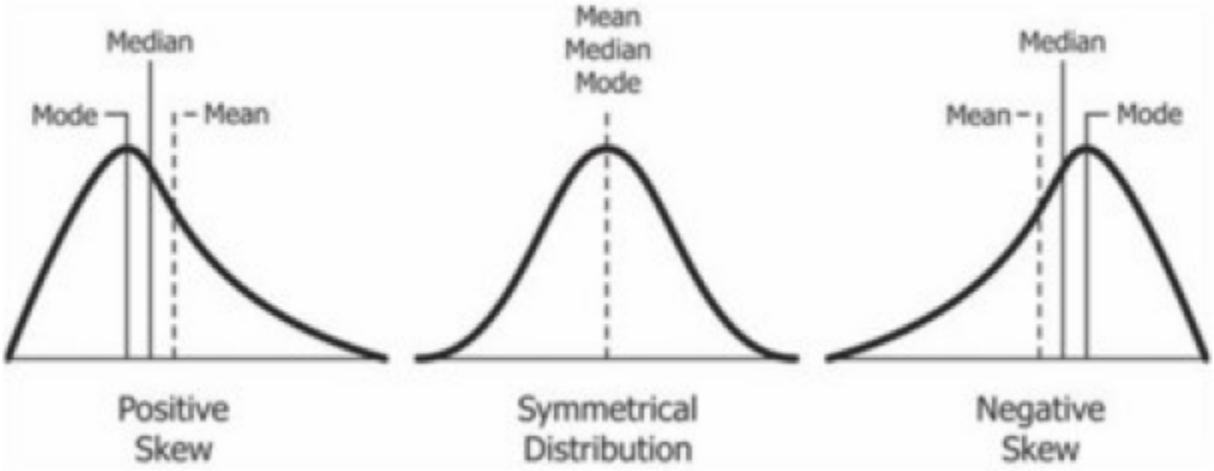
● Mathematical/ Analytical Modeling of the Problem Various mathematical tools are required to build the model such as

Zscore-



Through this method, we can remove outliers present in data so that it should be normally distributed which is essential for model building.

2. SKewness-



● Data Sources and their formats

The data source is taken from cardekho.com where data is scrapped extensively through web scraping using the selenium tool and further it is converted into a CSV file for further prediction.

● Data Preprocessing Done

Various steps involved in data preprocessing

- Acquire the dataset. ...
- Import all the crucial libraries. ...
- Import the dataset. ...
- Identifying and handling the missing values. ...
- Encoding the categorical data. ...
- Splitting the dataset. ...
- Feature scaling.

● Data Inputs- Logic- Output Relationships

Various independent Features like a brand, and model which show the change in car price which changes of the input features

● Hardware and Software Requirements and Tools Used

Hardware required:

• Processor: core i5 or above • RAM: 8 GB or above

• ROM/SSD: 250 GB or above

Software Required:

Anaconda

Python Programming Language Selenium

Chrome

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

Not every problem which has numbers involved in it is a machine-learning problem. There's a great saying, if the only tool you have is a hammer, you tend to see every problem as a nail.

Machine Learning can only be used in the following problems:

1. **Learning from the data is required.**
2. **Prediction of an outcome is asked for.**
3. **Automation is involved.**
4. **Understanding the pattern is required like that in the case of user sentiments.**
5. **Same as point d for building recommendation systems.**
6. **Identification/Detection of an entity/object is required. There are many other bullets to it too but the fundamentals are the ones mentioned above. A use case may have more than one bullet. There may be things where one might simply not need to have machine learning practice for the same in such a case he should go with one because simplicity is what is valued everywhere.**

Now coming up with how to solve a machine learning problem. A following stepwise approach would help you solve almost any machine learning problem.

Step 1(a). How to solve a Machine Learning problem?

Stepwise approach

1. **Read the data (from CSV, JSON, etc)**
2. **Identify the dependent and independent variables.**
3. **Check if the data has missing values or if the data is categorical or not.**
4. **If yes, apply basic data preprocessing operations to bring the data in a go-to-go format.**
5. **Now split the data into the groups of training and testing for the respective purpose.**

After splitting the data, fit it into the most suitable model. (How to find a suitable model is answered below)
Validate the model. If satisfactory, then go with it, else tune the parameters and keep testing. In a few cases, you can also try different algorithms for the same problem to understand the difference between the accuracies.
From step 7 one can also learn about the accuracy paradox.

Visualize the data.

Visualizing the data is important because we need to understand where our data is heading and also it looks more representative while storytelling about the data.

- **Testing of Identified Approaches (Algorithms)**

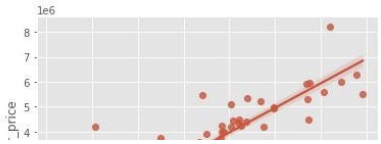
  The List of algorithms used are

  Random forest Regressor with R2_Score is 90% XGBoost Regressor with R2_Score is 93%
  KNN Regressor with R2_score is 88 %

- **Run and Evaluate selected models**

XGBoost

```
In [161…
xgb=XGBRegressor()
xgb.fit(x_train,y_train)

Out[161…
XGBRegressor(base_score=0.5, booster='gbtree', callbacks=None,
             colsample_bylevel=1, colsample_bynode=1, colsample_bytree=1,
             early_stopping_rounds=None, enable_categorical=False,
             eval_metric=None, gamma=0, gpu_id=-1, grow_policy='depthwise',
             importance_type=None, interaction_constraints='',
             learning_rate=0.300000012, max_bin=256, max_cat_to_onehot=4,
             max_delta_step=0, max_depth=6, max_leaves=0, min_child_weight=1,
             missing=nan, monotone_constraints='()', n_estimators=100, n_jobs=0,
             num_parallel_tree=1, predictor='auto', random_state=0, reg_alpha=0,
             reg_lambda=1, ...)

In [162…
model(xgb,x_train,x_test,y_train,y_test,train = True)

Traning r2_score:  99.68163402561375

In [163…
model(xgb,x_train,x_test,y_train,y_test,train = False)

The testing Score- 93.27566100609464
MSE:  53288845125.46947
MAE:  101866.94486768244
RMSE  230843.76778563778
```



Key Metrics for success in solving Problems under consideration

Mean Squared Error- It means the sum of all predicted value

The difference with an actual value should be lesser to get an accurate model.

Mean Absolute error-

- Visualizations

Univariate Analysis:-

Uni means one and variate means variable, so in univariate analysis, there is only one dependable variable. The objective of univariate analysis is to derive the data, define and summarize it, and analyze the pattern present in it. In a dataset, it explores each variable separately. It is possible for two kinds of variables- Categorical and Numerical.

Bivariate Analysis:-

Bi means two and variate means variable, so here there are two variables. The analysis is related to the cause and the relationship between the two variables. There are three types of bivariate analysis.

Learning Outcomes of the Study in respect of Data Science

- ➢ Scraping data for the project from www.cardheko.com. This first such kind of project for me. Web scraping such a huge amount of data challenges my scraping skill.
- ➢ Data cleaning or data pre-processing aspect of the project is good hands-on for me in this area. There were a lot of discrepancies in data scrap with different units and different names for the same sub-categories. Data cleaning was a big part of this project.

3. Limitations of this work and Scope for Future Work

- ➢ Around data for more than 10000 car scrap from cardheko.com
- ➢ We can scrap more data from different online platforms like Olx, car24. More data obviously means more accurate predictions.
- ➢ Here we Scrap almost 24 features. But there are many different kinds of safety, comfort, and entertainment features that buyers weigh while buying a car. We can also include much more features in the future.