



NAME OF THE PROJECT
“CAR PRICE PREDICTION”

Submitted by
VARSHA VASANT SHINDE

ACKNOWLEDGMENT

I would like to extend my thanks and appreciation to my SME MR.MOHD KASHIF SIR, my mentor, for his continuous support and guidance during the capstone project, we can also never forget the efforts of all the professors that taught me during the Data Scientist program and guided me through the world of data, which was a new realm for me. Moreover, I would like to extend my gratitude to Mr. Mohd Kashif sir for his guidance and patience with us during the capstone project.

INTRODUCTION

The used car market is an ever-rising industry, which has almost doubled its market value in the last few years. The emergence of online portals such as CarDheko, Quirk, Carpale, Cars24, and many others has facilitated the need for both the customer and the seller to be better informed about the trends and patterns that determine the value of the used car in the market.

Machine Learning algorithms can be used to predict the retail value of a car, based on a certain set of features. Different websites have different algorithms to generate the retail price of the used cars, and hence there isn't a unified Algorithm for determining the price.

By training statistical models for

Predicting the prices, one can easily get a rough estimate of the price without actually entering the details into the desired website. The main objective of this paper is to use three different prediction models to predict the retail price of a used car and compare their levels of accuracy.




Analytical Problem Framing

Pre-processing is a Data Mining technique that involves converting raw data into a comprehensible format. There is often a lack of specific activity or trend data, and many inaccurate facts are included in real-world data.




Consequently, this may result in poor-quality data collection, and, in turn, poor-quality models constructed from the data. Such problems can be resolved by pre-processing the data.

Pre-processing in Machine Learning is the process of modifying, or encoding, data so that the machine can parse it more easily. Thus, the algorithm can now properly interpret the data.

➤ **Hardware requirements**

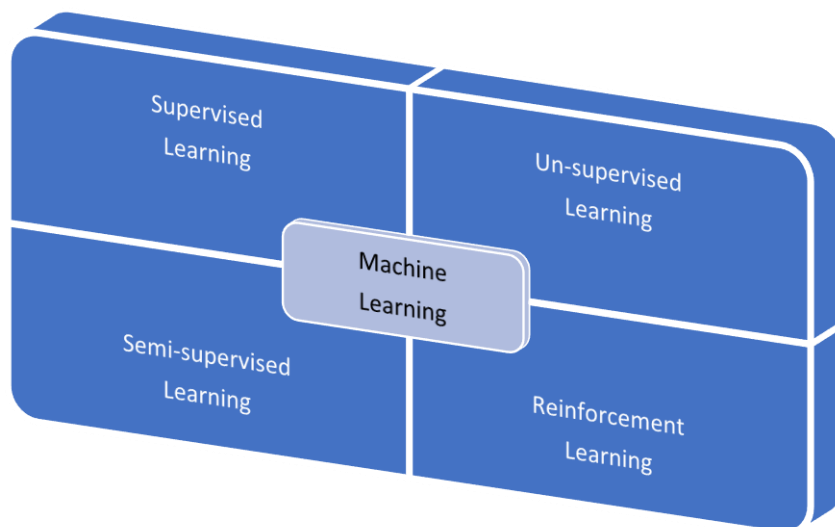
-  Operating system-Windows 7,8,10
-  Processor- dual core 2.4 GHz (i5 or i7 series Intel processor
-  Or equivalent AMD) RAM-4GB

➤ **Software Requirements**

-  Python
-  Jupyter Notebook
-  Chrome

Model/s Development and Evaluation

The goal of machine learning (ML) is to help a computer learn without being explicitly instructed to do so by means of mathematical models of data. Artificial intelligence (AI) is a subset of machine learning. Data is analysed using algorithms to identify patterns, which are then used to create predictive models. Like humans, machine learning becomes more accurate with more data and experience. With machine learning, you can adapt to situations where data is constantly changing, the nature of the request or task is shifting, or coding a solution isn't feasible. Three important categories of machine learning are:



Supervised and un-supervised learning are commonly used types while reinforcement is sequential decision maker technique.

Main categories of supervised and un-supervised machine learning are:

Supervised Learning:-

Working and details of some famous supervised machine learning algorithms which are used in this project are: Logistic Regression: Whenever the dependent variable is non-numerical (categorical) and the class should be predicted, not classified, the logistic regression algorithm needs to be abandoned.

Machine learning technique Logistic Regression is commonly used to classify binary data. The function of Logistic Regression is to optimize results based on various datasets.

To predict results, the default label class is always employed, but the results and probability are always calculated after all categorical values have been converted into numerical values and all data has been normalized.

Random Forest Regressor:

Random Forest is already revealing that it creates forest and then somehow randomizes it. It builds the forest through the ensemble of Decision Trees and most of the time trains it using a method called the Bagging Method.

Since it uses the ensemble method, the result is improved. Decision tree and bagging classifier hyperparameters are the same. Each feature in the tree can be made random simply by adding thresholds.

The following steps enable us to understand how Random Forest operates:

1. First, choose random samples from the given dataset.
2. Following that, each sample will be given a choice tree. Based on those choices, it will get a prediction end result.
3. For each anticipated outcome, voting may occur in this step.
4. In the end, choose the prediction outcome with the most votes because it is the very last prediction outcome.

Un-supervised learning:-

An unsupervised learning algorithm is trained on information that has neither been classified nor labelled, allowing it to act unsupervised on that information.

Using this algorithm, unsorted data is grouped using patterns, resemblances, and differences without any prior training.

Unlike supervised learning, the algorithm does not receive any instruction from a trainer. The algorithm, therefore, focuses on discovering the hidden pattern in unlabelled facts by ourselves.

The regression model can be evaluated on following parameters:

1. Mean Square Error (MSE):

MSE is the single value that provides information about goodness of regression line. Smaller the MSE value, better the fit because smaller value implies smaller magnitude of errors.

2. Root Mean Square Error (RMSE):

RMSE is the quadratic scoring rule that also measures the average magnitude of the error. It is the square root of average squared difference between prediction and actual observation.

3. Mean Absolute Error (MAE):

This measure represents the average absolute difference between the actual and predicted values in the dataset. It represents the average residual from the dataset.

Random Forest Regressor

Normally, random forests or random decision forests are used for classification, regression, and other tasks where they construct a multitude of decision trees at training time and output the class that is the mode of the classes (classification) or mean/average prediction (regression) of the individual trees.

A random decision forest corrects decision trees' habit of over fitting their training set. While random forests are generally better than decision trees, they are not as accurate as gradient boosted trees. They are, however, affected by data characteristics.

In this project, the random forest Regressor was trained with the intercept property. Score, MSE, RMSE and MAE errors are used to evaluate the model. Following results are achieved from there.

CONCLUSION

- The prediction error rate of all the models was well under the Accepted 5% of error. But, on further analysis, the mean error of the regression tree model was found to be more than the mean error rate of the multiple regressions and lasso regression models.
- Even though for some seeds the regression tree has better accuracy, its error rates are higher for the rest. This has been confirmed by performing an ANOVA.
- Significantly different from each other. To get even more accurate models, we can also choose more advanced machine learning algorithms such as random forests, an ensemble learning algorithm which creates multiple decision/regression trees, which brings down over fitting massively or Boosting,
- Which tries to bias the overall model by weighing in the favour of good performers? More data from newer websites and different countries can also be scraped and this data can be used to retrain these models to check for reproducibility.
- Using data mining and machine learning approaches, this project proposed a scalable framework for Mumbai based used cars price prediction.
- Cars24 website was scraped using the Parse Hub scraping tool to collect the benchmark data.
- An efficient machine learning model is built by training, testing, and evaluating three machine learning repressors named Random Forest Repressor, Linear Regression.
- As a result of pre-processing and transformation, Random Forest Repressor came out on top with 95% accuracy followed by Bagging Repressor with 88%.