A REPORT OF TWO MONTHS TRAINING

On

CAR PRICE PREDICTION

at

COURSERA

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE AWARD OF THE DEGREE OF

BACHELOR OF TECHNOLOGY

(Computer Science and Engineering)

SUBMITTED BY:

NAME: SACHIN KUMAR

UNIVERSITY ROLL NO.: 170280288

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

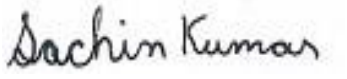
BABA BANDA SINGH BAHADUR ENGINEERING COLLEGE

FATEHGARH SAHIB

Table of Contents

CANDIDATE'S DECLARATION

I Sachin Kumar, hereby declare that I have undertaken two months training at Coursera during a period from June 1, 2020 to August 2, 2020. The project entitled Car Price Prediction submitted by Sachin Kumar, 170280288 in partial fulfillment of the requirement for the award of degree of the B. Tech. (Computer Science and Engineering) submitted in Baba Banda Singh Bahadur Engineering College, Fatehgarh Sahib is an authentic record of my own work carried out during eight weeks training, Training-III (BCSE1-734). The matter presented in this project has not formed the basis for the award of any other degree, diploma, fellowship or any other similar titles.

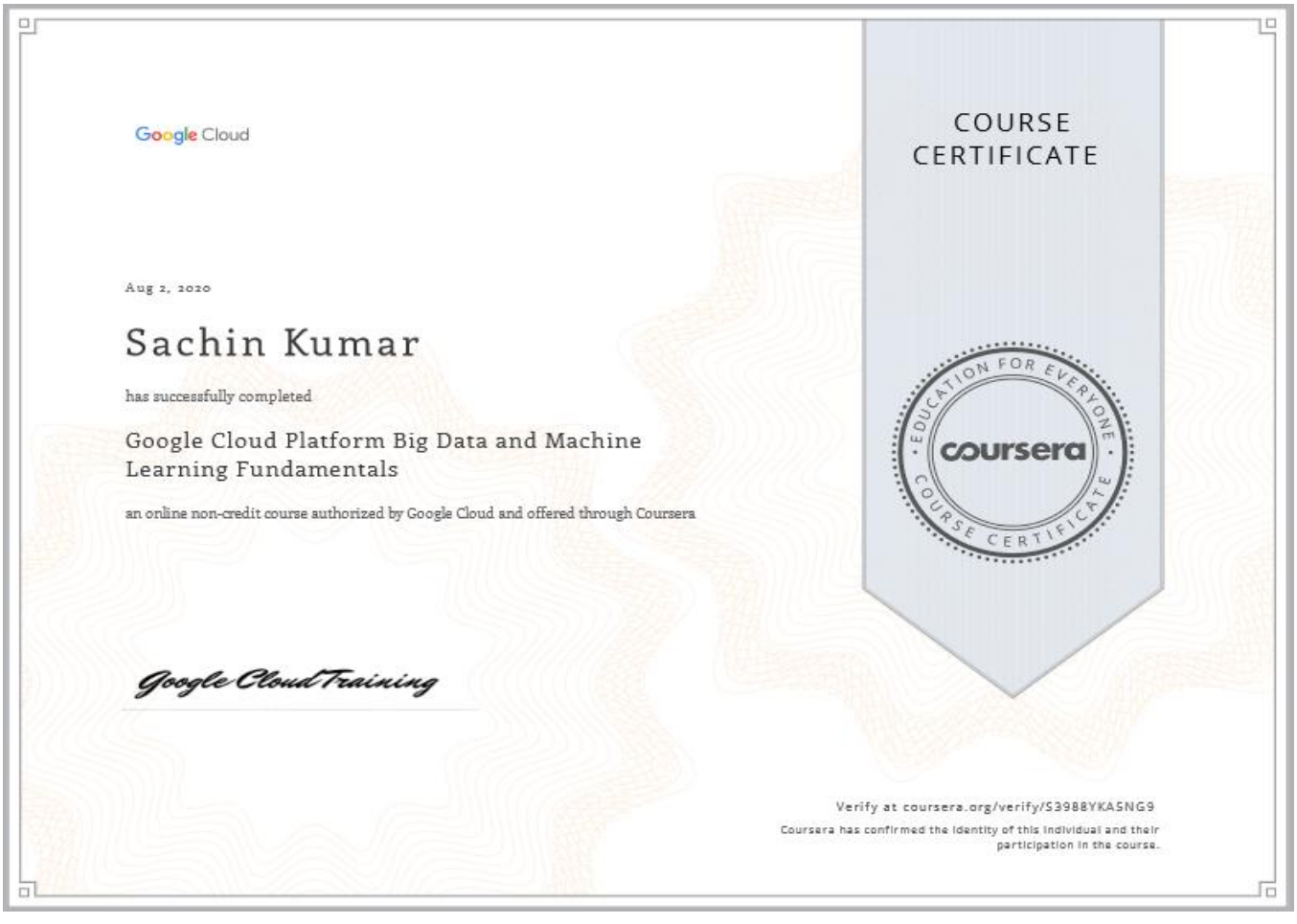


Signature of the Student

Place: Ludhiana

Date: October 6, 2020

CERTIFICATE



ACKNOWLEDGEMENT

I express my sincere gratitude to the Maharaja Ranjit Singh Punjab Technical University, Bathinda for giving me the opportunity to undergo eight weeks training, Training-III (BCSE1-734), after my 6th Semester of B.Tech. (CSE)

I would like to thank Dr. Gurcharan Singh Lamba, Principal and Dr. Kanwalvir Singh Dhindsa, Head of Department, CSE at Baba Banda Singh Bahadur Engineering College, Fatehgarh Sahib for their kind support.

I also owe my sincere gratitude towards Mr. Atul Kumar, Instructor for his valuable advice and healthy criticism throughout my training which helped me immensely to complete my work successfully.

I would also like to thank everyone who has knowingly and unknowingly helped me throughout my work.

Last but not least, a word of thanks for the authors of all those books and papers which I have consulted during my training as well as for preparing the report.

ABSTRACT

In this project, we 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, k-nearest neighbours, naive bayes and decision trees have been used to make the predictions. The predictions are then evaluated and compared in order to find those which provide the best performances. A seemingly easy problem turned out to be indeed very difficult to resolve with high accuracy. All the four methods provided comparable performance. In the future, we intend to use more sophisticated algorithms to make the predictions.

1. INTRODUCTION

Predicting the price of used cars in both an important and interesting problem. According to data obtained from the National Transport Authority, the number of cars registered between 2003 and 2013 has witnessed a spectacular increase of 234%. From 68, 524 cars registered in 2003, this number has now reached 160, 701. With difficult economic conditions, it is likely that sales of second-hand imported (reconditioned) cars and used cars will increase. It is reported in that the sales of new cars has registered a decrease of 8% in 2013.

In many developed countries, it is common to lease a car rather than buying it outright. A lease is a binding contract between a buyer and a seller (or a third party – usually a bank, insurance firm or other financial institutions) in which the buyer must pay fixed instalments for a pre-defined number of months/years to the seller/financer. After the lease period is over, the buyer has the possibility to buy the car at its residual value, i.e. its expected resale value. Thus, it is of commercial interest to seller/financers to be able to predict the salvage value (residual value) of cars with accuracy. If the residual value is under-estimated by the seller/financer at the beginning, the instalments will be higher for the clients who will certainly then opt for another seller/financer. If the residual value is over-estimated, the instalments will be lower for the clients but then the seller/financer may have much difficulty at selling these high-priced used cars at this over-estimated residual value. Thus, we can see that estimating the price of used cars is of very high commercial importance as well. Manufacturers' from Germany made a loss of 1 billion Euros in their USA market because of mis-calculating the residual value of leased cars. Most individuals in Mauritius who buy new cars are also very apprehensive about the resale value of their cars after certain number of years when they will possibly sell it in the used cars market.

Predicting the resale value of a car is not a simple task. It is trite knowledge that the value of used cars depends on a number of factors. The most important ones are usually the age of the car, its make (and model), the origin of the car (the original country of the manufacturer), its mileage (the number of kilometers it has run) and its horsepower. Due to rising fuel prices, fuel economy is also of prime importance. Unfortunately, in practice, most people do not know exactly how much fuel their car consumes for each km driven. Other factors such as the type of fuel it uses, the interior style, the braking system, acceleration, the volume of its cylinders (measured in cc), safety index, its size, number of doors, paint colour, weight of the car, consumer reviews, prestigious awards won by the car manufacturer, its physical state, whether it is a sports car, whether it has cruise

control, whether it is automatic or manual transmission, whether it belonged to an individual or a company and other options such as air conditioner, sound system, power steering, cosmic wheels, GPS navigator all may influence the price as well. Some special factors which buyers attach importance in Mauritius is the local of previous owners, whether the car had been involved in serious accidents and whether it is a lady-driven car. The look and feel of the car certainly contributes a lot to the price. As we can see, the price depends on a large number of factors. Unfortunately, information about all these factors are not always available and the buyer must make the decision to purchase at a certain price based on few factors only. In this work, we have considered only a small subset of the factors mentioned above. The following are the variables used:

- **Price:** The calculated retail price of GM cars. The cars which were selected for this data set were all less than a year old and were considered to be in good condition.
- **Mileage:** The total number of miles the car has been driven.
- **Make:** The manufacturer of the car.
- **Model:** The specific models for each car.
- **Trim:** The type of car model.
- **Type:** The car's body type.
- **Cylinder:** The number of cylinders present in the engine.
- **Liter:** The fuel capacity of the engine.
- **Doors:** The number of doors in the car.
- **Sound:** A categorical variable (binary), that represents whether upgraded speakers are present in the car (coded 1 if present).
- **Leather:** A categorical variable (binary), that represents whether the car has leather interiors (coded 1 if present).

Using these attributes, we will try to predict the price by using the Statistical Analysis System (SAS) for exploratory data analysis.

1.1 Hardware Specifications

- Version: 64-bit
- RAM: 4GB or more
- Hard disk: 2.5GB or more

- Screen Resolution: 1024x768

1.2 Software Specifications

- Operating System: Windows/MACOS/Linux
- Interpreter: Python 2.4 or higher, Jython, PyPy or IronPython
- IDE (Integrated Development Environment): PyCharm, Jupiter
- Browser: Chrome/Firefox

2. LITERATURE SURVEY

According to author Sameerchand, they have done the predictions of car price from the historical data that has been collected from daily newspapers. They have used the supervised machine learning techniques for predicting the price of vehicles. Many other algorithms such as multiple linear regression, k-nearest neighbor algorithms, naive based, and some decision tree algorithms also been used. All the four algorithms are compared and found the best algorithm for prediction. They have faced some difficulties in comparing the algorithms, somehow they have managed.

According to authors Pattabiraman, this paper is more concentrated on the relation between seller and buyer. In order to predict the price of four wheelers, more features are required such as already given price, mileage, make, model, trim, type, cylinder, liter, doors, cruise, sound, leather. Using these features the price of car has been predicted with the help of statistical analysis system for exploratory data analysis.

According to authors Enis Gegic et al, in this paper the mainly concentrate on collecting various data from web portal by using web scrap techniques. And those have been compared with the help of different machine learning algorithms to predict the vehicle price in easy manner. They classified the price according to different ranges of price that is already given. Artificial neural network, support vector machine, random forest algorithms were used on different datasets to build classifiers model.

Another approach was given by Richardsonin his thesis work. In his theory it states more durable cars will be produced by vehicle producer. He compared the hybrid cars and traditional cars in hoe it actually retains their value for longer time using multiple regression techniques. This improves the environmental conditions, and also it helps to provide huge efficiency of using fuels.

Wu et al, in this paper they have used neuro fuzzy knowledge based system to demonstrate car price prediction. By considering the following attributes such as brand, year of production and type of engine they predicted a model which has similar results as the simple regression model. Moreover, they made an expert system named ODAV (Optimal Distribution of Auction Vehicles) as there is a high demand for selling the by cars at the end of the leasing year by car dealers. This system gives insights into the best prices for cars, as well as the location where the best price can be gained. To predict a price of vehicles, the K – nearest neighbor machine learning algorithm has been used which is based on regression models. More number of vehicles has been exchanged through this system so this particular system is more successfully managed.

2.1 Study of Existing Systems

In the existing system, to predict the price of cars, a lot of data mining algorithms and machine learning algorithms were widely used. The major drawback of this existing system is they need more attributes in order to predict the car price. More comparison techniques must be used to get the result more efficiently. It is highly complicated to get sufficient data sets that were spread widely all over the world. The datasets can be collected only through online. But not on the offline mode. It is not possible for everyone to collect the data sets through online mode particularly in rural areas. The data sets will not have about the cars which were not used for long time and also the traditional model vehicles may or may not be included in the data sets. The major drawbacks of existing system is The system is very slow due to most of the works about the keyword query just analyze individual points, and they are inappropriate to many applications that call for analysis of groups of different car points. There are no fast query retrieval methods and is low due to lack of SVM under Constraints. Machine Learning systems includes:

- Justifications during data cleaning
- Exploratory Data Analysis (EDA)
- Performing machine learning models: Random Forest, Linear Regression, Ridge Regression, Lasso, KNN, XGBoost
- Comparison of the performance of the models
- Reporting the findings of the study in a professional manner

2.2 Problems in Existing System

The prices of new cars 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 to be worthy. But due to the increased price of new cars and the incapability of customers to buy new cars due to the lack of funds, used cars sales are on a global increase. There is a need for a used car price prediction system to effectively determine the worthiness of the car using a variety of features. Even though there are websites that offers this service, their prediction method may not be the best. Besides, different models and systems may contribute on predicting power for a used car's actual market value. It is important to know their actual market value while both buying and selling.

3. PURPOSED SYSTEM

Based on the varying features and factors, and also with the help of experts knowledge the vehicle price prediction has been done accurately. The most necessity ingredient for prediction is brand and model, period usage of vehicle, mileage of vehicle. The fuel type used in the vehicle as well as fuel consumption per mile highly affect price of a vehicle due to a frequent changes in the price of a fuel. Different features like exterior color, door number, type of transmission, dimensions, safety, air condition, interior, whether it has navigation or not will also influence the vehicle price. In this paper, we applied different methods and techniques in order to achieve higher precision of the used vehicle price prediction.

3.1 Problem Definition

The Client

To be able to predict used cars market value can help both buyers and sellers.

- **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.

The Data

The data used in this project was downloaded from Kaggle. It was uploaded on Kaggle by Austin Reese who Kaggle.com user. Austin Reese scraped this data from craigslist with non-profit purpose. It contains most all relevant information that Craigslist provides on car sales including columns like price, condition, manufacturer, latitude/longitude, and 22 other categories.

- **Data Wrangling:** In this section, some explorations and data visualizations were applied on data set. This gave some idea and guide about how to deal with missing values and extreme values. After data cleaning, data exploration was applied again in order to understand cleaned version of the data.
- **Data Cleaning:** First step for data cleaning was to remove unnecessary features. For this purpose, *'url','image_url', 'lat', 'long', 'city_url', 'desc', 'city', 'VIN'* features were dropped totally. As a next step, it was investigated number of null points and percentage of null data points for that feature (Table 1).

Table 1

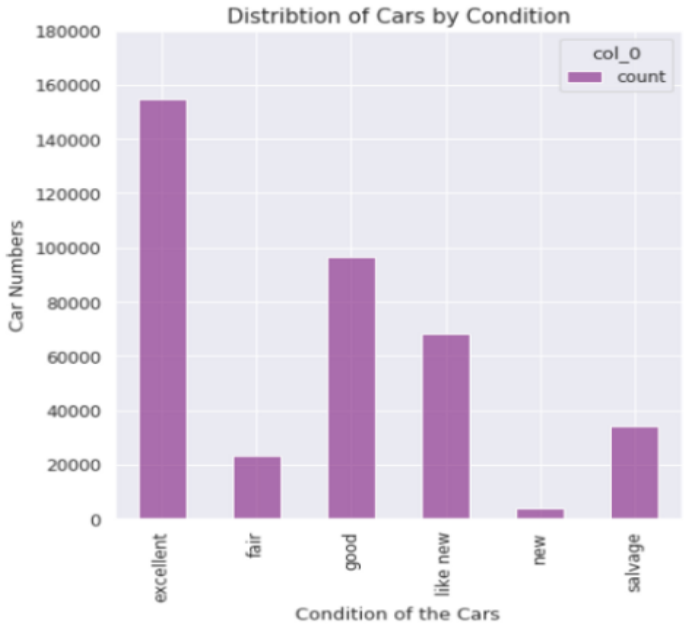
Missing Values in Car Dataset

Feature	Null Values	Percent
size	366256	67
condition	250074	45
cylinders	218997	40
paint_color	180021	33
drive	165838	30
type	159179	29
odometer	110800	20
manufacturer	26915	5
make	9677	2
fuel	4741	1
title_status	4024	1
transmission	4055	1
year	1487	0
price	0	0
Total	1502064	

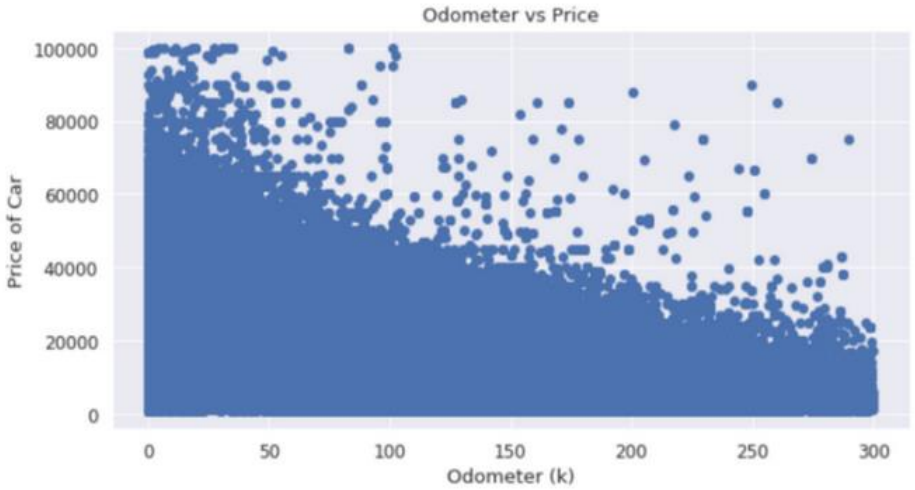
The Exploratory Data Analysis (EDA)

While exploring the data, we will look at the different combinations of features with the help of visuals. This will help us to understand our data better and give us some clue about pattern in data.

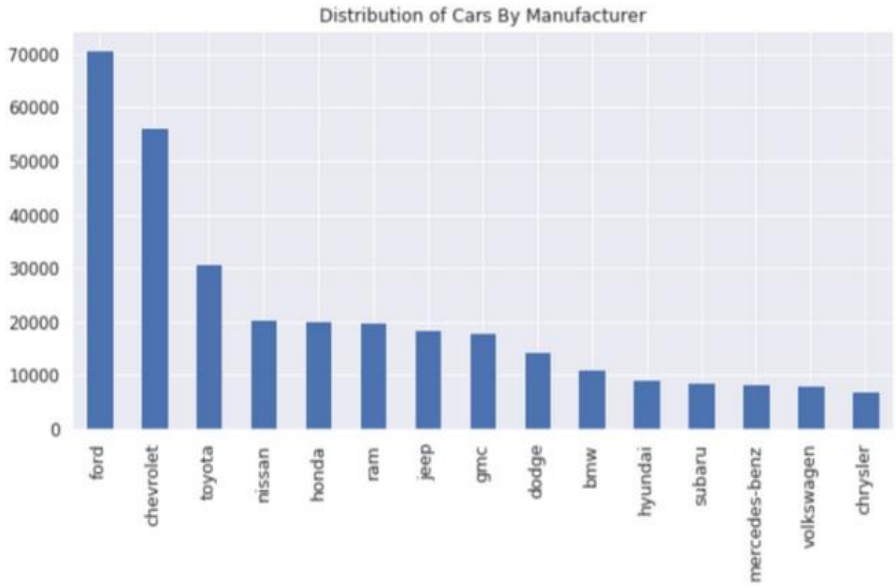
- **Distribution of Cars by Condition:** We can guess that all type of cars can be cheap or expensive. But, still excellent, like new, and good condition cars are the most popular cars in used car market. It is hard to make a strong estimate of a price of a car just by considering the type or condition of a car. But we can tell it certain condition cars are popular and higher chance to be sold.



- **Odometer vs Price:** When buying a used car, people pay serious attention to the odometer value on the car. We can see that odometer changes the price of a car significantly.



- **Distribution of Cars by Manufacturer:** Manufacturer of a car is another important variable on used car market. Ford and Chevrolet are one dominant manufacturer in North America. Toyota and Nissan follow the order as big manufacturers. It can be concluded that Japanese cars have a considerable share in used car market. However, American cars are still on demand and dominant.



3.2 Objectives of Car Price Prediction

The objective of our project is to be able to predict the price of a used car given various attributes (data) of that car. There is a saying that a car loses 10% of its value the moment you drive it off a lot. Given, that I would expect that one of the main predictors is the amount of miles driven in the car, since more driving wears down the car. Additionally, I would expect the brand (make) of the car to also be a factor in the price of a used car, since some brands of cars cost more and may be better made. I expect to encounter some issues with multicolinearity since some aspects of cars may be highly correlated. For example, larger cars will probably have larger engines and more doors. Larger engines are correlated with more cylinders.

4. DESIGN & IMPLEMENTATION

Design: Input Design plays a vital role in the life cycle of software development. The attention of developers is required to collect the information about vehicles. The most accurate data must be entered in the input design. The design of input is more important in minimizing the errors that has

been given by the user. By the rules of software engineering concepts, the validation control must be defined over the input limit in the input forms or screens. The validation control must take care of other input related errors. The input screens have been included in almost all the modules. The alert message will be displayed whenever user did any mistakes while giving input. And also some messages will be provided in order to guide the user in correct way. By this we can achieve to get only valid details.

The user created input has been converted in to computer related format. The input design is based on data entry logical. The main goal of input design is to make the form as free from errors. The input design will control the errors in the input form. The created application should be user-friendly manner. Wherever the cursor is placed in processing the input must be entered in that same place. By this way the form has been designed. There might be several options for a single input so that the user has to select suited input to get the best result. Each entered data must be validated accordingly. The error message must be displayed whenever the user enters any wrong data or irrelevant data as input. Even the user is in last page of input if he did not get the result properly then he can go to the first page and he can change the input given already.

The primary output form has been created in order to get communication between the administrator and the clients. The VPN system produces output in the form of managing clients by the project leaders, in a way such as creating new clients, allotting new projects to them, have a look over table in which to get the details about project status, and the same will be accessed by each clients. A new project will be assigned to every client when he completes his old project. At every initial stage of the new project, the user authentication should be maintained. A user registration can be done either by the administrator or the user can do by himself. But only the administrator must have the authority to assign the projects to each user.

When the application is executed it starts running. The used browser is internet explorer and the server will start its process. The project will run on the local area network so the server machine will serve as the administrator while the other connected systems can act as the clients.

Implementation

- **Seller:** In this module, the login has done by the seller by using valid user name and password. After login successful he can perform some operations such as View All Authorized Users, Add Category, Add Sub Category, Add Vehicles, View All Uploaded vehicles, View

Purchased Vehicles, View Search Transactions, View All Vehicle Reviews, View Vehicle Score Result, View keyword Score Result.

- **Viewing and Authorizing Users:** In this module, the Server can view all user details and authorization can be done them for login permission. User Details includes User Name, Address, Email Id and Mobile Number in the corresponding lignin page.

- **Adding Categories and vehicles:** In this module, the admin adds Categories and Vehicles with details such as Category Name, Vehicle Name, Description and Vehicle Image.

- **List all Vehicles with Ratings:** In this module, the admin can list his entire added Product with details along with rating. The ratings are calculated based on number of recommendations made on particular vehicle.

- **List all User's Vehicle Search History:** In this, the administrator can view all user's Vehicle search history details.

- **User:** In this module, there are n numbers of users can register. Once user registers, their details will be stored to the database.

  After registration successful, he has to login by using authorized user name and password that has been sent to user. Once Login is successful user can perform some operations like My Profile, Create Online Bank Account, Search Optimal Products, Search top Products, and View Search History.

- **Viewing Profile Details:** In this module, the user can see their own profile details, such as their address, email, mobile number, profile Image.

- **Search Vehicles:** In this, the user can Search vehicle based on brand and model, age, vehicle price, mileage. In My Profile, the user can Create Online Bank Account and also can View Search History.

5. METHODOLOGY

We utilized several classic and state-of-the-art methods, including ensemble learning techniques, with a 90% - 10% split for the training and test data. To reduce the time required for training, we

used 500 thousand examples from our dataset. Linear Regression, Random Forest and Gradient Boost were our baseline methods. For most of the model implementations, the open-source Scikit-Learn package was used.

I. **Linear Regression:** Linear Regression was chosen as the first model due to its simplicity and comparatively small training time. The features, without any feature mapping, were used directly as the feature vectors. No regularization was used since the results clearly showed low variance.

II. **Random Forest:** Random Forest is an ensemble learning based regression model. It uses a model called decision tree, specifically as the name suggests, multiple decision trees to generate the ensemble model which collectively produces a prediction. The benefit of this model is that the trees are produced in parallel and are relatively uncorrelated, thus producing good results as each tree is not prone to individual errors of other trees. This uncorrelated behavior is partly ensured by the use of Bootstrap Aggregation or bagging providing the randomness required to produce robust and uncorrelated trees. This model was hence chosen to account for the large number of features in the dataset and compare a bagging technique with the following gradient boosting methods.

III. **Gradient Boost:** Gradient Boosting is another decision tree based method that is generally described as "a method of transforming weak learners into strong learners". This means that like a typical boosting method, observations are assigned different weights and based on certain metrics, the weights of difficult to predict observations are increased and then fed into another tree to be trained. In this case the metric is the gradient of the loss function. This model was chosen to account for non-linear relationships between the features and predicted price, by splitting the data into 100 regions.

IV. **XGBoost:** Extreme Gradient Boosting or XGBoost is one of the most popular machine learning models in current times. XGBoost is quite similar at the core to the original gradient boosting algorithm but features many additive features that significantly improve its performance such as built in support for regularization, parallel processing as well as giving additional hyperparameters to tune such as tree pruning, sub sampling and number of decision trees. A maximum depth of 16 was used and the algorithm was run on all cores in parallel.

V. **LightGBM:** Light GBM is another gradient boosting based framework which is gaining popularity due it higher speed and accuracy compared to XGBoost or the original gradient boosting method. Similar to XGBoost, this LightGBM has a leaf-wise tree growth instead of a level-wise approach resulting in higher loss reduction. This framework can also handle categorical features, thus eliminating the need to one hot vectorize them and in turn, reducing memory usage. Make, Model and State and cities were declared as categorical features. The algorithm was run at tree depths in multiples of 12 and was run on all cores in parallel.

VI. **KMeans + Linear Regression:** In order to capitalize on the linear regression results and the apparent categorical linearity in the data as indicated, an ensemble method which used KMeans clustering of the features and linear regression on each cluster was used. Due to large training time, a three-cluster model was used. Then, the dataset was classified into these three clusters and passed through a linear regressor trained on each of the three training sets.

VII. **Deep Neural Network (MLP Regressor):** To introduce mode complexities in the model, the MLP regressor, which uses a deep neural net perceptron regressor model, was used. This model optimizes the squared-loss using LBFGS or stochastic gradient descent. Two hidden layers of width 200 and 20 were used. The learning rate was set at 0.001 and batch size at 200. ReLu was used as the activation function.

6. RESULTS

Learning Algorithm	R Score on Test 2 Data	R Score on Training 2 Data	Training Time
Linear Regression	0.87	0.87	15 minutes
Gradient Boost	0.64	0.64	130 minutes
Random Forest	0.88	0.98	75 minutes
Light GBM	0.81	0.82	104 seconds
XGBoost	0.78	0.81	180 minutes

KMeans + LinReg	0.88	0.89	70 minutes
Deep Neural Network	0.85	0.85	10 hours

The results of our tests were quantified in terms of the R score of our predictions. score is a statistical 2 R 2 measure of how close the data are to the fitted regression line.

Compared to Linear Regression, most Decision-Tree based methods did not perform comparably well. This can be attributed to the apparent linearity of the dataset. We believe that It can also be attributed to the difficulty in tuning the hyperparameters for most gradient boost methods. The exception to this is the Random Forest method which marginally outperforms Linear Regression. However Random Forests tend to overfit the dataset due to the tendency of growing longer trees. This was worked upon by restricting the depth of trees to different values and it was observed that beyond limiting depth to 36 resulted in negligible improvement in prediction performance but progressively increased overfitting. As expected lightGBM performed marginally better than XGBoost but had a significantly faster training time.

Building up from the relatively good performance of Linear Regression, the KMeans + Linear Regression Ensemble Learning Method (with K = 3) produced the best R score on test data without high variance as 2 it fits linear relationships categorically. The deep neural network was converging to local minima due to small batch-sizes.

7. CONCLUSION

Vehicle price prediction can be a challenging task due to the more number of attributes that should be considered for the accurate prediction. The collection and preprocessing of data is the major step in prediction. In this paper, to normalize, standardize and to clean the data, PHP scripts were built. This will used to avoid unnecessary noise for machine learning algorithms.

The prediction performance must be increased by using data cleaning processes. But in this paper, the insufficient set of complex data is the drawback here. We will get only 50 percent result on applying the single machine algorithm. Therefore, we proposed multiple groups of machine learning algorithm to gain more accuracy and it achieved 93 percent of efficiency. This comparison

of single and multiple groups of machine learning algorithm is significant. And also it overcome the drawback of single machine algorithm which is given in proposed system.

Although, this system has achieved valuable performance in vehicle price prediction, our aim for the future work is to test this system to work successfully with various data sets.

SCREENSHOTS

REFERENCES

1. https://github.com/krishnaik06/Car-Price-Prediction
2. https://www.kaggle.com/jpayne/852k-used-car-listings
3. https://youtu.be/ZhGDcTOzAFM
4. Listiani M. 2009. Support Vector Regression Analysis for Price Prediction in a Car Leasing Application. Master Thesis. Hamburg University of Technology.