NITTE EDUCATION TRUST | NITTE MEENAKSHI INSTITUTE OF TECHNOLOGY

AN AUTONOMOUS INSTITUTION AFFILIATED TO VTU, BELGAUM, ACCREDITED by NAAC (‘A+’ Grade)
YELAHANKA, BANGALORE-560064

A Project Report on Pre-owned Car Price Prediction Using Machine Learning

Submitted in partial fulfillment of the requirement for the award of the degree of

BACHELOR OF ENGINEERING IN INFORMATION SCIENCE AND ENGINEERING

By

ANNAPURNA B NEELAKARI	1NT20IS401
NIKHIL P G	1NT20IS406
REUBEN V KUDNAVAR	1NT20IS410
YASHODHA R NAIK	1NT20IS416

Under the Guidance of

Ms.Ullal Akshatha Nayak
Assistant Professor
Department of Information Science and Engineering
Nitte Meenakshi Institute of Technology, Bengaluru - 560064

DEPARTMENT OF INFORMATION SCIENCE AND ENGINEERING

(Accredited by NBA Tier-1) 2020-23

NITTE MEENAKSHI INSTITUTE OF TECHNOLOGY
(AN AUTONOMOUS INSTITUTION, AFFILIATED TO VTU, BELGAUM)

DEPARTMENT OF INFORMATION SCIENCE AND ENGINEERING

Accredited by NBA Tier-1

CERTIFICATE

Certified that the project work entitled " **Pre-owned Car Price Prediction Using Machine Learning**" carried out by Mr./Ms. **Annapurna B Neelakari-1NT20IS401, Nikhil P G - 1NT20IS406, Reuben V Kudnavar-1NT20IS410, Yashodha R Naik-1NT20IS416,** a bonafide student of **Department of Information and Engineering and College Nitte Meenakshi Institute of Technology** in partial fulfillment for the award of Bachelor of Engineering in **Information Science and Engineering** of the Visvesraya Technological University, Belgaum during the year **2021 - 22**. It is certified that all corrections/suggestions indicated for Internal Assessment have been incorporated in the Report deposited in the departmental library. The project report has been approved as it satisfies the academic requirements in respect of Project work prescribed for the said Degree.

Name and Signature of the Guide

Name and Signature of the HOD

Signature of the Principal

External Viva

Name of the examiners

Signature with date

1.

2.

NITTE MEENAKSHI INSTITUTE OF TECHNOLOGY
(AN AUTONOMOUS INSTITUTION, AFFILIATED TO VTU, BELGAUM)

DEPARTMENT OF INFORMATION SCIENCE AND ENGINEERING

Accredited by NBA Tier-1

DECLARATION

I, Annapurna B Neelakari - 1NT20IS401, Nikhil P G-1NT20IS406, Reuben V Kudnavar-1NT20IS410, Yashodha R Naik -1NT20IS416, bonafide students of Nitte Meenakshi Institute of Technology, hereby declare that the project entitled "**Pre-owned Car Price Prediction Using Machine Learning**" submitted in partial fulfillment for the award of Bachelor Of Engineering in **Information Science & Engineering** of the Visvesvaraya Technological University, Belgaum during the year 2020-2023 is my original work and the project has not formed the basis for the award of any other degree, fellowship or any other similar titles.

Signature of the Student with Date

Place: Bangalore   Date:

ABSTRACT

It is generally known that, taking wise and challenging decisions is really a crucial task in every business and customers. The major objective of our project is to build a prediction model i.e a fair price mechanism to predict the pre-owned cars selling price based on their features.

A preowned car price prediction has been a high interest research area, as it requires noticeable effort and knowledge of the field expert for reliable and accurate prediction. With increased demand in the second-hand car market, the business for both buyers and sellers has increased. And because of the affordability of used cars in developing countries, people tend more purchase used cars.

Before acquiring a used car, the buyer should be able to decide whether the price affixed for the car is genuine. Several facets including mileage, year, model, make, run and many more are needed to be considered before getting a hold of any pre-owned car. Both the seller and the buyer should have a fair deal. Using a history of previously used cars selling data and using machine learning techniques.

In this project, we are going to propose a prediction system using a machine learning model combining random forest regression and a k-mean clustering algorithm to analyze the price of used cars. The predictions are then analysed and compared to determine which ones provide the best results.

ACKNOWLEDGMENT

We have immense pleasure in expressing our sincerest and deepest sense of gratitude towards our guide Ms.Ullal Akshatha Nayak for the assistance, valuable guidance and cooperation in carrying out this Project successfully. We have developed this project with the help of Faculty members of our institute and we are extremely grateful to all of them. We are greatly thankful to our parents, friends and faculty members for their motivation, guidance and help whenever needed.

Contents

Chapter 1

Introduction

With the rapid growth in the use of cars and the people showing interest in cars. Cars have been an important part of daily life. Several people are not capable of buying a new car because of the lack of funds. The global market for used cars valued USD 1332.2 billion in the year 2020. Market is looking forward to a stretch with a CAGR (Compound annual growth rate) of 5.5% during the next decade. Like the global market, Indian used car market is also forecasted to register a CAGR of 15% by next five years. Therefore, there is a need for predicting the price of pre-owned cars. Preowned cars are vehicles with one or more previous retail owners. Whenever a car is sold to dealers, the price should be calculated. The price of a pre-owned car depends on various factors. Customers who plan to purchase a pre-owned car often struggle to find an appropriate car within a budget. Even if a customer knows the type of car they want to purchase, it becomes challenging for them to estimate the price of the car. Car price prediction is a somehow interesting and popular problem. The market of the pre-owned car has reached twice the market for new car sales.

Here, during the purchase of pre-owned car. It has become a challenge to estimate the price of car. It is interest knowledge that the value of pre-owned cars depends on a number of factors. The most important ones are usually the age of the car, model, the original country of the manufacturer, mileage and its horsepower. Due to rising fuel prices, fuel economy is also of prime importance. 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, acceleration, safety index, its size, number of doors, colour, weight of the car, consumer reviews. The look and feel of the car also play important role in the price of the car.

With increase in demand for used cars, more and more vehicle buyers are finding alternatives of buying new cars outright. A number of online sites have assembled the automotive industry at one place, so that the end user can buy or sell with a click. These sites use different algorithms for generating the price for the used cars, hence may place incompatible results to the users. More upon, these systems provide the sell and purchase mostly in urban areas but we found out that there is less accuracy in the result. Our aim is to develop methodology and framework which will predict the car price, which will reduce human efforts in finding best deal for pre-owned cars. This will aid a buyer to make a more informed decision while buying a pre-owned car.

Chapter 2

Literature Review

In the literature, few researchers applied various machine learning techniques to predict the used car cost. Several studies and related works have been done previously to predict used car prices around the world using different methodologies and approaches, with varying results of accuracy from 50% to 90%.

Richardson [1] worked on the theory that car manufacturers are more likely to produce cars that do not depreciate rapidly in another university study. Richardson [1] applied multiple regression analysis and demonstrated that hybrid cars (cars that use two separate power sources to drive the vehicle, i.e. they have both an internal combustion engine and an electric motor) are more able to keep their value than conventional vehicles by using a multiple regression study. This has environmental concerns about the climate and it gives higher fuel efficiency. conducted car price prediction study, by using neuro-fuzzy knowledge-based system. They took into consideration the following attributes: brand, year of production and type of engine. Their prediction model produced 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 cars at the end of the leasing year by car dealers. This system gives insights into the best prices for vehicles, as well as the location where the best price can be gained. Regression model based on k-nearest neighbour machine learning algorithm was used to predict the price of a car. This system has a tendency to be exceptionally successful since more than two million vehicles were exchanged through it.

Noor & Jan [2] were able to achieve high level of accuracy using Multiple linear regression models to predict the price of cars collected from used cars website in Pakistan called Pak Wheels that totalled to 1699 records after pre-processing, and where able to achieve accuracy. Noor and Jan build a model for car price prediction by using multiple linear regression. The dataset was created during the two-months period and included the following features: price, cubic capacity, exterior color, date when the ad was posted, number of ad views, power steering, mileage in kilometer, rims type, type of transmission, engine type, city, registered city, model, version, make and model year. After applying feature selection, the authors considered only engine type, price, model year and model as input features. With the given setup authors were able to achieved pretty good accuracy.

Pudaruth [3] applied various machine learning algorithms, namely: k-nearest neighbours, 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 period less than one month, as time can have a noticeable impact on 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, limited number of dataset instances could not give high classification performances, i.e. accuracies less than 70%. He again researched to assess the neural network's success in predicting used car prices particularly on higher-priced vehicles, the predicted value is not very similar to the actual price. In predicting the price of a used car, they found that support vector machine regression outperformed neural networks and linear regression.

Listiani [4] used Support Vector Machines to evaluate leased cars prices, results have shown that SVM is far more accurate in large dataset with high dimensional data than Multiple linear regression. Whereas the computation Multiple linear regression can take several minutes and the SVM would take up to a day to compute the results. Multiple linear regression may be simple, but SVM is far more accurate. Moreover, the study includes Samples with up to 178 attributes which is far more than the proposed variable in our study, It was found that SVM also handles high dimensional data better and avoids both the underfitting and over-fitting issues. Genetic algorithm is used by Listiani [4] to find important features for SVM. However, the technique does not show in terms of variance and mean standard deviation why SVM is better than simple multiple regression.

Gonggie [5] He proposed a model that is built using ANN (Artificial Neural Networks) for the price prediction of a used car. He considered several attributes: miles passed, estimated car life and brand. The proposed model was built so it could deal with nonlinear relations in data which was not the case with previous models that were utilizing the simple linear regression techniques. The non-linear model was able to predict prices of cars with better precision (quality, condition, fact, exact, accurate) than other linear models. This model was found to be fairly effective at estimating the residual value of used vehicles.

In the related work shown above, authors proposed prediction model based on the single machine learning algorithm. However, it is noticeable that single machine learning algorithm approach did not give remarkable prediction results and could be enhanced by assembling.

Listian proposed model the only drawback of this study is that the improvement of SVM regression over simple regression was not expressed in simple measures like mean deviation or variance

In Pudaruth’s research it was found that decision tree algorithm and naïve bayes method were unable to classify and predict numeric values. Pudaruth’s research also concluded that limited number of instances in data set do not give high prediction accuracies

In the related work shown above, authors proposed prediction model based on the single machine learning algorithm. However, it is noticeable that single machine learning algorithm approach did not give remarkable prediction results and could be enhanced by assembling various machine learning methods in an ensemble

Chapter 3

Problem Statement

For the purposes of car valuation, it is common not to use machine learning. Instead, they source data from local sales and average the prices of many similar cars. This method works well if you have a common car with a common set of features. The condition of the car is judged very roughly, typically on a scale of one to three. Cars that are unusual are therefore hard to evaluate. Effectively, no inferences are drawn from similar cars but from a different make and model, whereas with machine learning, the entirety of the dataset and its features are used to train the model predictions. Using machine learning is a solution to the problem of utilization of all the data and will assist in utilizing all the features of a car to make valuations.

New cars of a particular make, model, location, and feature selection are identical in condition, function, and price. When new cars are sold for the first time they are then classified or called as used cars. As the car ages, its price changes because it declines in efficiency in the current and in all future periods. It reflects the change in net present value over time.

Generally, the pricing is summarized in the condition of the car. The value of repairs or custom modifications to the car are recognized only if they noticeably improve the overall condition of the car.

Using machine learning to better utilize data on all the less common features of a car can more accurately predict the value of a vehicle. This is a clear benefit to consumers, especially those who themselves cannot ascertain the value of the vehicle that they are buying or selling and must rely on a tool. A tool that can provide a more accurate price for used car and make the market fairer for all participants.

There are several machine learning models that can be applied to price prediction. This work will investigate which one offers the best performance according to several criteria. The nature of machine learning is to train on past data to predict unseen data. Applied to price prediction of cars, the data is sourced from past sales while the predictions are for the present value of cars. Hence, we can are going to use machine learning algorithm for better accuracy.

Chapter 4

Objective

This project aims to deliver price prediction models to the public, to help guide the individuals looking to buy or sell cars and to give them a better insight into the automotive sector. Buying a used car from a dealer can be a frustrating and an unsatisfying experience as some dealers are known to deploy deceitful sale tactics to close a deal. Therefore, to help consumers avoid falling victims to such tactics, this model hopes to equip consumers with right tools to guide them in buying used cars.

Another goal of the project is to explore new methods to evaluate used cars prices and to compare their accuracies. Considering this is an interesting topic in the research community, and in continuing their footsteps, we hope to build a predicting model to achieve significant results using more advanced methods. As our objective are as follows

- To predict pre-owned car price using machine learning based on suitable parameters like **Transmission of the vehicle, Fuel type, Condition of the vehicle, Price of the showroom**
- To compare machine learning for **accuracy and errors**
- Our aim is to provide **best valuation of the cars to the customers**
- To build a **web-based application**

Chapter 5

Framework and System Design

K-means clustering is performed first, and then based on the results of the clustering; a random forest model of different types is established.
We are going to use K-means clustering combined with random forests to form a group to predict pre-owned car price more accurately. Data of used cars is collected first.
The modelling process is as follows

- Use K-means clustering to divide the data into several categories.
- The measurement method usually used to compare the results of different k values is the average distance between a data point and its cluster centroid.
- Use k sets of data and random forest algorithm to train k models. After determining the category of the new data, the corresponding model can be used to calculate the price of car.
- When predicting new data, first we have to determine the category of the new data by calculating the Euclidean distance between the sample data and the centroids of multiple classes of data.
- The new data belongs to the category corresponding to the centroid with the smallest Euclidean distance.
- After category is determined. The value of used is obtained.

RANDOM FOREST:

Random Forest algorithm is a popular supervised machine learning algorithm that relies on the concept of ensemble learning and can be deployed for both classification and regression problems in machine learning. It operates by fabricating a number of decision trees during the training phase, further outputs the class. The class outputted is the mode of classes if the problem under consideration belongs to classification while in case of regression outputs the mean prediction of individual trees.

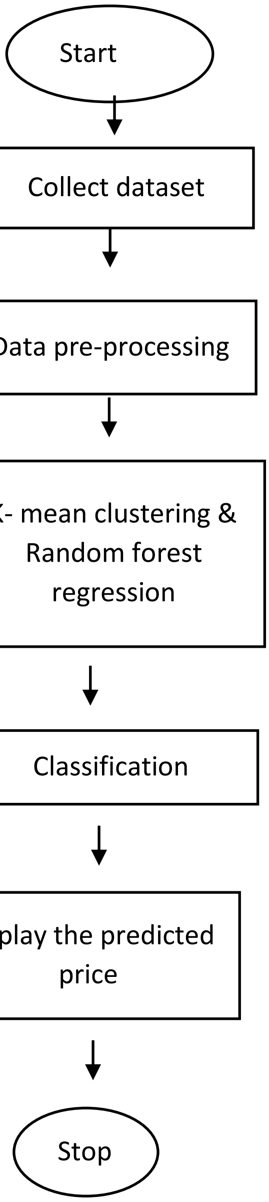
Being based on bagging ensemble learning, it deploys multiple uncorrelated decision trees on various subsets of a given dataset. Each of these decision trees outputs a certain prediction. Based on the average of predictions, final output is predicted.

The proposed system is a regression-based task, where we have to predict output that should be continuous numeric values. Hence, the average of previously observed labels will give us a final prediction. The greater number of trees in the forest makes a robust forest that leads to accurate and stable prediction.

Algorithm for Random Forest

i. Select random samples from a given dataset and build multiple subsets.
ii. Build a decision tree associated for every subset.
iii. Every decision tree will output a prediction.
iv. Take the average of these predicted values.
v. The average value will be the final prediction.

PROPOSED MODEL SYSTEM:


Start
Collect dataset
Data pre-processing
K- mean clustering & Random forest regression
Classification
Display the predicted price
Stop

Chapter 7

Conclusion and Future Scope

The increased prices of new cars and the financial incapability of the customers to buy them, Used Car sales are on a global increase. Therefore, there is an urgent need for a Used Car Price Prediction system which effectively determines the worthiness of the car using a variety of features. The proposed system will help to determine the accurate price of used car price prediction.

References

[1] Determinants of used car resale value Michael S. Richardson https://digitalccbeta.coloradocollege.edu/pid/coccc:1346/datastream/OBJ

[2] Noor, K., & Jan, S. Vehicle Price Prediction System using Machine Learning Techniques. International Journal of Computer Applications, 27-31. https://www.ijcaonline.org/archives/volume167/number9/noor-2017-ijca-914373.pdf

[3] S. Peerun, N. H. Chummun and a. S. Pudaruth, "Predicting the Price of Second-hand Cars using Artificial Neural Networks," The Second International Conference on Data Mining, Internet Computing, and Big Data, pp.

[4] Listiani, M. Support Vector Regression Analysis for Price Prediction in a Car Leasing Application. Unpublished. https://www.ifis.uni-luebeck.de/~moeller/publist-sts-pw-andm/source/papers/2009/list09.pdf

[5] GONGGI, S. New model for residual value prediction of used cars based on BP neural network and non-linear curve fit. In: Proceedings of the 3rd IEEE International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), Vol 2. pp. 682-685, IEEE Computer Society, Washington DC, ISA https://dl.acm.org/doi/10.1145/3449301.3449321

[6] Pudaruth, S. "Predicting the Price of Used Cars using Machine Learning Techniques". International Journal of Information & Computation Technology, Vol. 4, No. 7, pp.753-764. http://ripublication.com/irph/ijict_spl/ijictv4n7spl_17.pdf