Prediction of Used Car Prices Using Random Forest Algorithm

K Shruthi[1], B Priya[2], S Meeha[3]
[1, 2, 3] Dept of CSE
[1, 2, 3] Vins Christian College Of Engineering

Abstract- *With the extensive growth in usage of cars, the newly produced cars are unable to reach the customers for various reasons like high prices, less availability, financial incapability, and so on. Hence the used car market is escalated across the globe but in India, the used car market is in a very nascent stage and mostly dominated by the unorganized sector. This gives chance for fraud while buying a used car. Hence a high precision model is required which will estimate the price of an used car with none bias towards customer or merchandiser. This project presents a working model for used car price prediction with a low error value. A considerable number of distinct attributes are examined for reliable and accurate predictions. The results obtained agree with theoretical predictions and have shown improvement over models which use simple linear models. Random Forest Algorithm built to predict the results. These algorithms are tested with the car dataset and predict the results.*

Keywords- car market, car price prediction, Random Forest Algorithm

I. INTRODUCTION

Car price prediction is anyhow interesting and popular problem. Accurate car price prediction involves expert knowledge, because price usually depends on many unique features and factors. Generally, most important ones are brand name and model, years, KMs driven and mileage. The fuel type used in the car as well as fuel consumption per mile highly affected price of a car due to often changes in the price of a fuel. Distinct features like exterior color, door number, type of transmission, dimensions, safety, air condition, interior, whether it has navigation or not will also results in the car price. Existing System includes a process where a seller decides a price randomly and buyer has no idea about the car and it's value in the present day scenario. In fact, seller also has no idea about the car's existing value or the price he should be selling the car at. To overcome this problem applied dissimilar and unique methodologies in order to achieve higher precision of car price like Car Model Name, Year, Selling, Price, Present Price, KMs Driven, Fuel Type, Seller Type, Transmission Type, Owner Type. Random Forest Algorithm and models in order to accomplish higher precision of the used car price prediction. A model to predict the price of a used car should be developed in order to assess its value based on a variety of characteristics. Several factors affect the price of a used car, such as company, model, year, transmission, distance driven, fuel type, seller type, and owner type. As a result, it is crucial to know the car's actual market value before purchasing or selling it. presents a working model for used car price prediction with a low error value. They provide us with continuous value as an output and not a categorized value. Because of which it will be possible to predict the actual price a car rather than the price range of a car.

Machine learning is a subfield of artificial intelligence (AI). The goal of machine learning generally is to understand the structure of data and fit that data into models that can be understood and utilized by people. Although machine learning is a field within computer science, it differs from traditional computational approaches. In traditional computing, algorithms are sets of explicitly programmed instructions used by computers to calculate or problem solve. Machine learning algorithms instead allow for computers to train on data inputs and use statistical analysis in order to output values that fall within a specific range. Because of this, machine learning facilitates computers in building models from sample data in order to automate decision-making processes based on data inputs.

Machine learning is a subfield of artificial intelligence (AI). The goal of machine learning generally is to understand the structure of data and fit that data into models that can be understood and utilized by people.

Although machine learning is a field within computer science, it differs from traditional computational approaches. In traditional computing, algorithms are sets of explicitly programmed instructions used by computers to calculate or problem solve. Machine learning algorithms instead allow for computers to train on data inputs and use statistical analysis in order to output values that fall within a specific range. Because of this, machine learning facilitates computers in building models from sample data in order to automate decision-making processes based on data inputs.

In supervised learning, the computer is provided with example inputs that are labeled with their desired outputs. The purpose of this method is for the algorithm to be able to "learn" by comparing its actual output with the "taught" outputs to find errors, and modify the model accordingly. Supervised learning therefore uses patterns to predict label values on additional unlabeled data.

For example, with supervised learning, an algorithm may be fed data with images of sharks labeled as fish and images of oceans labeled as water. By being trained on this data, the supervised learning algorithm should be able to later identify unlabeled shark images as fish and unlabeled ocean images as water.

A common use case of supervised learning is to use historical data to predict statistically likely future events. It may use historical stock market information to anticipate upcoming fluctuations, or be employed to filter out spam emails. In supervised learning, tagged photos of dogs can be used as input data to classify untagged photos of dogs.

In unsupervised learning, data is unlabeled, so the learning algorithm is left to find commonalities among its input data. As unlabeled data are more abundant than labeled data, machine learning methods that facilitate unsupervised learning are particularly valuable.

The goal of unsupervised learning may be as straightforward as discovering hidden patterns within a dataset, but it may also have a goal of feature learning, which allows the computational machine to automatically discover the representations that are needed to classify raw data.

Unsupervised learning is commonly used for transactional data. You may have a large dataset of customers and their purchases, but as a human you will likely not be able to make sense of what similar attributes can be drawn from customer profiles and their types of purchases.

With this data fed into an unsupervised learning algorithm, it may be determined that women of a certain age range who buy unscented soaps are likely to be pregnant, and therefore a marketing campaign related to pregnancy and baby products can be targeted to this audience in order to increase their number of purchases.

Without being told a "correct" answer, unsupervised learning methods can look at complex data that is more expansive and seemingly unrelated in order to organize it in potentially meaningful ways. Unsupervised learning is often used for anomaly detection including for fraudulent credit card purchases, and recommender systems that recommend what products to buy next. In unsupervised learning, untagged photos of dogs can be used as input data for the algorithm to find likenesses and classify dog photos together.

II. LITERATURE SURVEY

The number of cars on Mauritian roads has been rising consistently by 5% during the last decade. In 2014, 173 954 cars were registered at the National Transport Authority. Thus, one Mauritian in every six owns a car, most of which are second hand reconditioned cars and used cars. The aim of this study is to assess whether it is possible to predict the price of second-hand cars using artificial neural networks. Thus, data for 200 cars from different sources was gathered and fed to four different machine learning algorithms. Found that support vector machine regression produced slightly better results than using a neural network or linear regression. However, some of the predicted values are quite far away from the actual prices, especially for higher priced cars. Thus, more investigations with a larger data set are required and more experimentation with different network type and structures is still required in order to obtain better predictions.

When buying and selling cars, it can be a challenge to assign the correct price. Artificial neural networks, a branch of artificial intelligence, are frequently used for such calculations. Designed two different artificial neural networks for car price forecasting and tested them using data from a car sales website. For data, a software was developed using the C# programming language and the MSSQL Server database management system, also HTMLAgilityPack library was used to read the data on the website. A procedure was written with T SQL language to digitize the data. As a result of the study, using data from approximately 1000 cars, a success rate of 91.38% was obtained. More data is needed for better results.

In the real world, entities have two or more representations in databases. Duplicate records do not share a common key and/or they contain errors that make duplicate matching a difficult task. Errors are introduced as the result of transcription errors, incomplete information, lack of standard formats, or any combination of these factors. Present a thorough analysis of the literature on duplicate record detection. Cover similarity metrics that are commonly used to detect similar field entries, and we present an extensive set of duplicate detection algorithms that can detect approximately duplicate records in a database. Also cover multiple techniques for improving the efficiency and scalability of approximate duplicate detection algorithms. Conclude with coverage of existing tools and with a brief discussion of the big open problems in the area.

Conventional artificial intelligence techniques and their hybrid models are incapable of handling several hypotheses at a time. The limitation in the performance of certain techniques has made the ensemble learning paradigm a desirable alternative for better predictions. The petroleum industry stands to gain immensely from this learning methodology due to the persistent quest for better prediction accuracies of reservoir properties for improved hydrocarbon exploration, production, and management activities. Artificial Neural Networks (ANN) has been applied in petroleum engineering but widely reported to be lacking in global optima caused mainly by the great challenge involved in the determination of optimal number of hidden neurons. This paper presents a novel ensemble model of ANN that uses a randomized algorithm to generate the number of hidden neurons in the prediction of petroleum reservoir properties. Ten base learners of the ANN model were created with each using a randomly generated number of hidden neurons. Each learner contributed in solving the problem and a single ensemble solution was evolved. The performance of the ensemble model was evaluated using standard evaluation criteria. The results showed that the performance of the proposed ensemble model is better than the average performance of the individual base learners. This study is a successful proof of concept of randomization of the number of hidden neurons and demonstrated the great potential for the application of this learning paradigm in petroleum reservoir characterization.

Glycosylation is one of the most complex post-translational modifications (PTMs) of proteins in eukaryotic cells. Glycosylation plays an important role in biological processes ranging from protein folding and sub cellular localization, to ligand recognition and cell-cell interactions. Experimental identification of glycosylation sites is expensive and laborious. Hence, there is significant interest in the development of computational methods for reliable prediction of glycosylation sites from amino acid sequences. explore machine learning methods for training classifiers to predict the amino acid residues that are likely to be glycosylated using information derived from the target amino acid residue and its sequence neighbors. Compare the performance of Support Vector Machine classifiers and ensembles of Support Vector Machine classifiers trained on a dataset of experimentally determined N-linked, O-linked, and C-linked glycosylation sites extracted fromO-GlycBase version 6.00, a database of 242 proteins from several different species. The results of our experiments show that the ensembles of Support Vector Machine classifiers outperform single Support Vector Machine classifiers on the problem of predicting glycosylation sites in terms of arrange of standard measures for comparing the performance of classifiers. The resulting methods have been implemented in EnsembleGly, a web server for glycosylation site prediction.

The need to make judicious use of organizational resources has put a lot of pressure on production managers and demand planners; thus, it is necessary to accurately predict what resources will yield what output. Well planned activities result in improved performance of organizational goals among which are productivity, price recovery and profitability. This work uses artificial neural network, Back Propagation Artificial Neural Network (BP-ANN), as an alternative predictive tool to multi-linear regression, for establishing the interrelationships among productivity, price recovery and profitability as performance measures. A 2-20-20-1 back propagation artificial neural network was proposed. Productivity and price recovery served as independent variables while profitability was used as the dependent variable in the BP ANN architecture. It was observed that BA-ANN model has Mean Square Error (MSE) of 0.02 while Multiple Linear Regression (MLR) has MSE of 0.036. This study concluded that artificial neural network is a more efficient tool for modeling interrelationships among productivity, price recovery and profitability. This approach can be applied in predicting performance measures of firms.

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, naïve 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.

III. PROPOSED SYSTEM

In proposed system using Random Forest algorithm. In this algorithm provided good accuracy in comparison to prior work using these data sets. Based on the varying features and factors, and also with the help of experts knowledge the vehicle price prediction has been done accurately and achieve higher precision of the used vehicle price prediction.

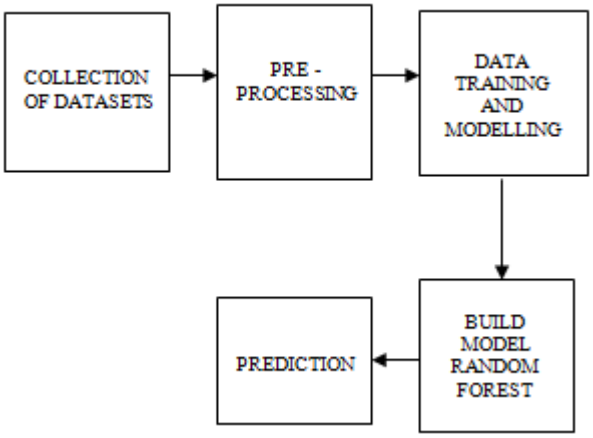
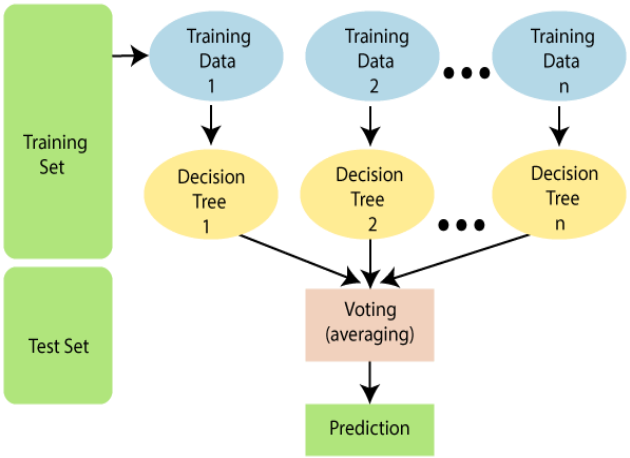


Fig 1 System Architecture

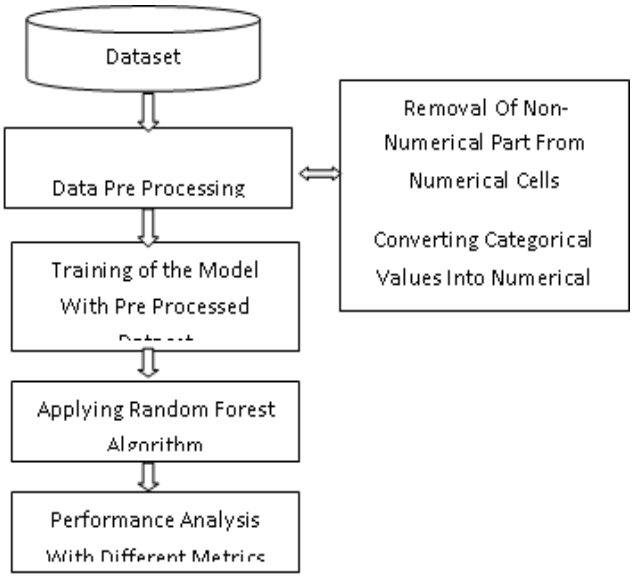
Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of over fitting.



Random Forest algorithm

Since the random forest combines multiple trees to predict the class of the dataset, it is possible that some decision trees may predict the correct output, while others may not. But together, all the trees predict the correct output. Therefore, below are two assumptions for a better Random forest classifier: There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result. The predictions from each tree must have very low correlations.

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.



Flow Diagram

IV. MODULES DESCRIPTION

Collection of Datasets

Data set downloaded from Kaggle. It was uploaded from Cardekho.com . The dataset consists of 301 rows and 9 columns with no null values. Column data consist of independent Features. The independent features contain both categorical and numeric values. Selling price the price at which car is being sold this will be target label for further prediction of price.km_driven number of kilometer car has been driven. Fuel this feature the fuel type of car (CNG, petrol, diesel etc).seller type tells whether the seller is individual or a dealer. transmission gives information about

the whether the car is automatic and manual. Owner number of previous owner of the car.Present_price what is the current showroom price of the car.

Pre-processing

This step is one of the important steps in supervised machine learning. It includes the following.

i) Removal of Non-numerical part from numerical features

This step removes the non-numerical words from the features like Mileage, Engine, and power for data processing.

Step1: Converting the data frame into the list.
Step2: Splits the list based on a delimiter.
Step3: Store the required data back to the data frame.

Converting Categorical values into numerical

Here, the categorical values like Name, Location, Fuel Type, Transmission, Owner Type are converted to numerical because machine learning deal with numerical values easily because of the machine-readable form. This is done by using Label Encoder which is a python package.

Step1: We have to select categorical values based on its data type.
Step2: Converting the categorical values into numerical values by using Label encoder API in python.

Separate the target variable

Here, we have to separate the target feature which is we are going to predict. In this case, price is the target variable.

Step1: The target variable price is assigned to the variable ‘y’.
Step2: The preprocessed data set except the target variable is assigned to the variable ‘X’.

Data Training and Modelling

We must first define the dependent and independent variables in order to train and construct a model. To find these variables, first used to find the correlation between the output variables, and then separated my variables into two axes, which we call x and y, with the x-axis containing all the independent variables and the y-axis containing the dependent variable, which in our model is the Used Cars selling price. This dataset is further dispersed in the train-test dataset using RandomizedSearch tuning to discover the optimum hyper parameters for our model prediction using the sklearn.model selection package and its train test split function

Prediction

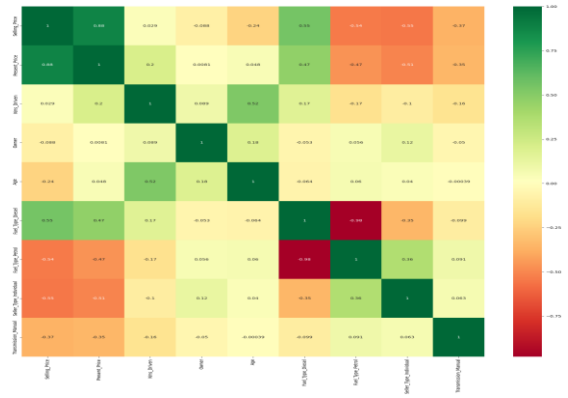
Random Forest is a classifier containing various decision trees on different subsets of the given dataset and takes the average to increase the accuracy of the dataset. The data from these trees are then combines to guarantee the most accurate predictions.

V. SCREEN SHOTS

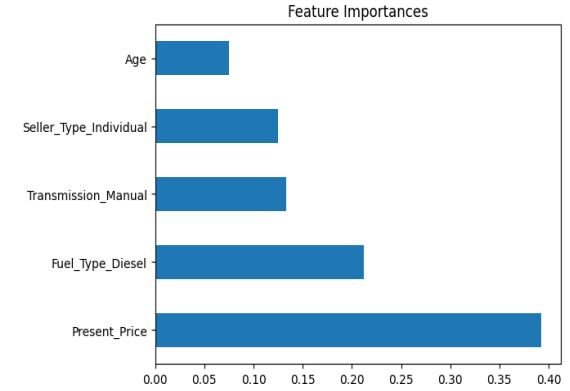
	Car_Name	Year	Selling_Price	Present_Price	Kms_Driven	Fuel_Type	Seller_Type	Transmission	Owner
0	ritz	2014	3.35	5.59	27000	Petrol	Dealer	Manual	0
1	sx4	2013	4.75	9.54	43000	Diesel	Dealer	Manual	0
2	ciaz	2017	7.25	9.85	6900	Petrol	Dealer	Manual	0
3	wagon r	2011	2.85	4.15	5200	Petrol	Dealer	Manual	0
4	swift	2014	4.60	6.87	42450	Diesel	Dealer	Manual	0

Loading data set

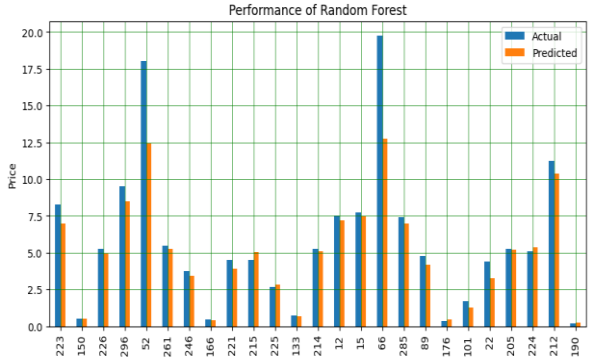
Heatmap Correlation



Heat map between every pair of features



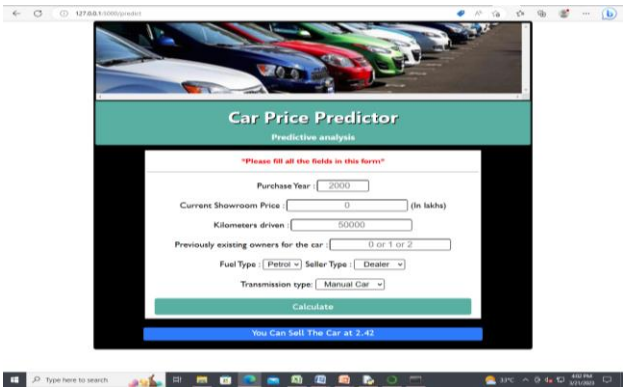
Feature importance



Algorithm performance

	precision	recall	f1-score	support
0.0	1.00	0.93	0.97	61
1.0	0.00	0.00	0.00	0
accuracy			0.93	61
macro avg	0.50	0.47	0.48	61
weighted avg	1.00	0.93	0.97	61

Accuracy Table



Car price predictor

VI. CONCLUSION

Car price prediction will be a challenging task because of the high number of attributes that should be considered for the accurate prediction. The most important step within the prediction process is collection and preprocessing of the information. During the research, Car data collected from kaggle.com is converted into CSV form and used for building the machine learning algorithms.. The Final result was predicted consistent with the algorithm which achieves higher accuracy. It can be concluded by saying that increased prices of new cars and the monetary lack of ability of clients to get them Used Car market is expanding globally. Therefore, there is an urgent need for a Used Car Price Prediction system that viably determines price of the car using a variety of features. The process of predicting used cars price involves high caution and great knowledge in the field of cars and their models. Among all the proposed models, Random Forest determines the price of a used car with minimum possible error.

REFERENCES

[1] Anyaeche, C. O. (2013). “Predicting Performance Measures using Linear Regression and Neural Network: A Comparison”. African Journal of Engineering Research, Vol. 1, No. 3, pp. 84-89.

[2] C. V. Narayana, C. L. Likhitha, S. Bademiya and K. Kusumanjali, "Machine Learning Techniques To Predict The Price Of Used Cars: Predictive Analytics in Retail Business," 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC), 2021, pp. 1680-1687,

[3] Cheng, B. and Titterington, D. M. (1994). “Neural Networks: A Review from a Statistical Perspective”. Statistical Science, Vol. 9, pp. 2-54.

[4] Ganesh, Mukkesh & Venkatasubbu, Pattabiraman. (2019). Used Cars Price Prediction using Supervised Learning Techniques. International Journal of Engineering and Advanced Technology.9. 216-223.

[5] He, Q. (1999) “Neural Network and its Application in IR”. Thesis (BSc). University of Illinois.

[6] LE DEFI MEDIA GROUP. 2014. [Online] Available at: http://www.defimedia.info/ [Accessed 23 September 2014].

[7] LEXPRESS.MU ONLINE. 2014. [Online] Available at: http://www.lexpress.mu/ [Accessed 23 September 2014].

[8] N. Monburinon, P. Chertchom, T. Kaewkiriya, S. Rungpheung, S. Buya and P. Boonpou, "Prediction of prices for used car by using regression models," 2018 pp. 115-119.

[9] N. Sun, H. Bai, Y. Geng and H. Shi, "Price evaluation model in second hand car system based on BP neural network theory," 2017 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence,

[10] Pandey, Abhishek and Rastogi, Vanshika and Singh, Sanika, Car's Selling Price Prediction using Random Forest Machine Learning Algorithm 2020.
[11] Peerun, Saamiyah & Chummun, Nushrah & Pudaruth, Sameerchand. (2015). Predicting the Price of Second-hand Cars using Artificial Neural Networks.
[12] Published in International Journal of Trend inScientificResearch and Development (ijtsrd), ISSN: 2456-6470, Volume-5 | Issue-4June 2021,
[13] Rose, D. (2003) "Predicting Car Production using a Neural Network Technical Paper- Vetronics (In house)". Thesis, U.S. Army Tank Automotive Research, Development and Engineering Center (TARDEC).
[14] Vehicle Price Prediction using SVM Techniques S.E.Viswapriya, Durbaka Sai Sandeep Sharma, Gandavarapu Sathya kiran International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-9 Issue-8, June 2020.
[15] Networking and Parallel/Distributed Computing (SNPD), 2017, pp. 431-436.