Car Resale Value Prediction

Mrs. Rekha M[1], Gowtham S[2], Selvaganapathy A[3] ,Naveen M[4], Revanth Reddy Bapathu [5]

[1]Assistant Professor, Department of Information Technology, R.M.K. Engineering College
[2-5]Final Year B.Tech Information Technology, R.M.K Engineering College

Literature Survey

Introduction:

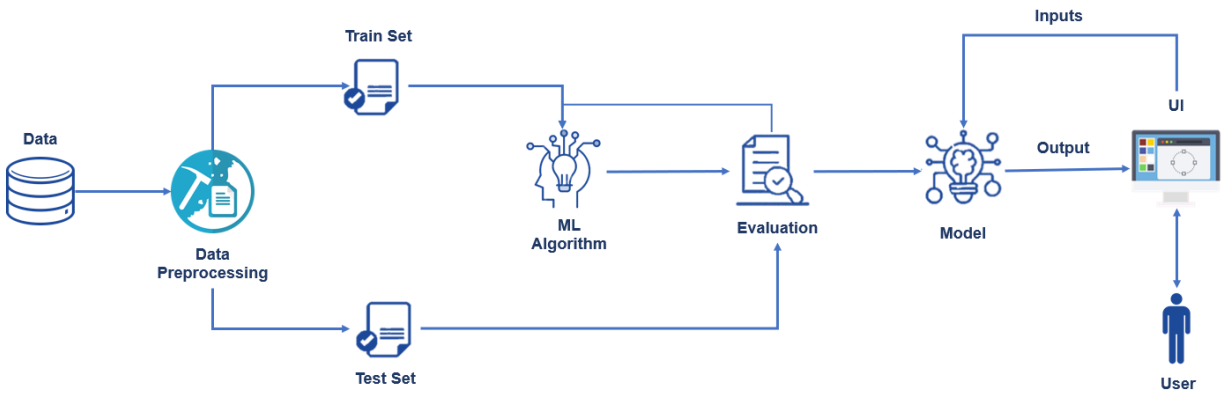
Car resale value prediction is the system to predict the amount of resale value based on the parameters provided by the user. User enters the details of the car into the form given and accordingly the car resale value is predicted. In this project we have used different algorithms with different techniques for developing Car resale value prediction systems considering different features of the car. In a nutshell, car resale value prediction helps the user to predict the resale value of the car depending upon various features like kilometers driven, fuel type, etc.The main idea of making a car resale value prediction system is to get hands-on practice for python using Data Science. Car resale value prediction is the system to predict the amount of resale value based on the parameters provided by the user. User enters the details of the car into the form given and accordingly the car resale value is predicted. The system is defined in the python language that predicts the amount of resale value based on the given information. The system works on the trained dataset of the machine learning program that evaluates the precise value of the car. User can enter details only of fields like purchase price of car, kilometers driven, fuel of car, year of purchase.

Approach:

With difficult economic conditions, it is likely that sales of second-hand imported (reconditioned) cars and used cars will increase. In many developed countries, it is common to lease a car rather than buying it outright. 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 sellers/financers to be able to predict the salvage value (residual value) of cars with accuracy.

In order to predict the resale value of the car, we proposed an intelligent, flexible, and effective system that is based on using regression algorithms. Considering the main factors which would affect the resale value of a vehicle a regression model is to be built that would give the nearest resale value of the vehicle. We will be using various regression algorithms and algorithm with the best accuracy will be taken as a solution, then it will be integrated to the web-based application where the user is notified with the status of his product.

Architecture:



- **Used Cars Price Prediction and Valuation**
    - Abdulla AlShared

Methodology:

The project's methodology is after data collection the dataset was pre-processed to remove samples that have missing value, and remove non-numerical part from numerical attributes, converting categorical values into numerical (if needed), fix any discrepancies in the units, as well as removing attributes that doesn't affect the price evaluations if needed to reduce the complexity of the model.

Data Understanding and preparation is an essential part of building a model as it gives the insight into the data and what corrections or modifications shall be done before designing and executing the model, preliminary analysis of the data must be done to have deeper understanding into the quality of the data, in terms of outliers and the skewedness of the figures, descriptive Statistics of categorical and numerical variables was done for that to be achieved. That was done through a correlation matrix for every attribute to understand the relations between the different factors. Afterwards when the data is organized and transformed into a form that could be processed by the data mining technique. Different data mining models were designed to predict prices and values of used cars. In this study three models are proposed to be built using Logistic Regression model technique, Random Forest Regressor and Bagging Regressor. Firstly, the data was portioned into section for training and the other part for testing, portioning percentage can be tested with different ratios to analyse different results. All three models were evaluated on four evaluation matrices known as model score, Mean Square Error (MSE), Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). From all, the Random Forest Regressor outperformed.

Technology:

Using data mining and machine learning approaches, this project proposed a scalable framework for Dubai based used cars price prediction. Buyanycar.com 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 regressors named Random Forest Regressor, Linear Regression, and Bagging Regressor. As a result of pre-processing and transformation, Random Forest Regressor came out on top with 95% accuracy followed by Bagging Regressor with 88%. Each experiment was performed in real-time within the Google Colab environment. In comparison to the system's integrated Jupyter notebook and Anaconda's platform, algorithms took less training time in Google Colab.In the future, more data will be collected using different web-scraping techniques, and deep learning classifiers will be tested. Algorithms like Quantile Regression, ANN and SVM will be tested.

Afterwards, the intelligent model will be integrated with web and mobile-based applications for public use. Moreover, after the data collection phase Semiconductor shortages have incurred after the pandemic which led to an increase in car prices, and greatly affected the secondhand market. Hence having a regular Data collection and analysis is required periodically, ideally, we would be having a real time processing program.

- **Used Cars Price Prediction using Supervised Learning Techniques**
    - Mukkesh Ganesh

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, Quikr, Carwale, 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.

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 regression 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. Also, the post-hoc test revealed that the error rates in multiple regression models and lasso regression models aren't 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 overfitting massively or Boosting, which tries to bias the overall model by weighing in the favor 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.

- **VEHICLE RESALE PRICE PREDICTION USING MACHINE LEARNING**
    - B.Lavanya, Sk.Reshma, N.Nikitha, M.Namitha

In this project, we mainly focus on the analysis of the Vehicle Resale Predict and then predict the results through them using training data. The trade- in vehicle market is an always rising industry, which has nearly multiplied its fairly estimated worth over the most recent couple of years. The rise of online entrances like CarDheko, Quikr, Carwale, Cars24, and numerous others has worked with the requirement for both the client and the merchant to afterward Linear Regression calculation.These predictions are done using the previous results of Previous Data Set.

In this paper, four distinctive AI procedures have been utilized to figure the cost of pre-owned vehicles in Mauritius. The mean blunder with direct relapse was about Rs 51,000 while for kNN it was about Rs 27,000 for Nissan vehicles and about Rs 45,000 for Toyota vehicles. J48 and NaiveBayes exactness hung between 60-70% for various blends of boundaries. The primary shortcoming of choice trees and credulous bayes is their powerlessness to deal with yield classes with numeric qualities. Consequently, the value quality must be ordered into classes which contained a scope of costs yet this clearly presented further justification for errors. The primary limit of this examination is the low number of records that have been utilized. As future work, we plan to gather more information and to utilizes further developed methods like counterfeit neural organizations, fluffy logic and hereditary calculations to foresee vehicle costs.

Hence, this investigation utilized various models to foresee utilized vehicle costs. Nonetheless, there was a generally little dataset for making a solid induction as a result of the quantity of perceptions. Assembling more information can yield more powerful expectations. Furthermore, there could be more highlights that can be acceptable indicators. For instance, here are a few factors that may work on the

model: number of entryways, gas/mile (per gallon), shading, mechanical and restorative reconditioning time, used-to-new proportion, examination to-exchange proportion. Another point that has space to improve is that the information cleaning cycle should be possible all the more enthusiastically with the assistance of more specialized data. For instance, rather than utilizing the 'fill' technique, there may be pointers that assistance to fill missing qualities all the more genuinely.

- **Predicting Used Car Prices**
    - Kshitij Kumbar, Pranav Gadre and Varun Nayak

Determining whether 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. We implement and evaluate various learning methods on a dataset consisting of the sale prices of different makes and models across cities in the United States. Our results show that Random Forest model and K-Means clustering with linear regression yield the best results, but are compute heavy. Conventional linear regression also yielded satisfactory results, with the advantage of a significantly lower training time in comparison to the aforementioned methods.

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. Linear Regression, Random Forest, Gradient Boost, XGBoost, LightGBM, KMeans + Linear Regression, Deep Neural Network (MLP Regressor).For better performance, we plan to judiciously design deep learning network structures, use adaptive learning rates and train on clusters of data rather than the whole dataset. To correct for overfitting in Random Forest, different

selections of features and number of trees will be tested to check for change in performance.

- **Price Prediction for Used Cars**
    - Marcus Collard

This work will use a quantitative method to achieve the scientific goals. The evaluation of models will be done by collecting and comparing various performance metrics for each of the machine learning algorithms to be tested in this work.Machine learning models need a large amount of data to train on. The first step in performing this study is to source a sufficiently large and reliable dataset. There are several criteria for such a dataset. It must be large enough, include sufficiently many relevant features, have very few null values for those features, have reliable values, and must be distributed over several years.
To ensure the highest possible accuracy for the various models, a result- driven iterative process including data cleaning, model training, and model testing will be used to refine the models.

This work will focus on answering the research questions. They all entail a comparison of different ML algorithms for price prediction. This will be accomplished by sourcing and preparing a dataset on which all the algorithms can be trained on and compared fairly. The algorithms selected must therefore be similar enough for the same dataset to be used for all of them. This also means that no large optimization efforts on the dataset will be made to boost the performance, if these changes do not benefit the other models. Maximizing price prediction performance of any one algorithm in ways that do not offer better comparisons is outside the scope of this work.

Possible future research that can expand upon the knowledge gained through research on Applying the Method to Other ML Models, Adding Additional Features Related to the Year

- **USED CAR PRICE PREDICTION**
    - Praful Rane, Deep Pandya, Dhawal Kotak

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 to be 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. 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. 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 we have developed a model which will be highly effective. Regression Algorithms are used because 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. User Interface has also been developed which acquires input from any user and displays the Price of a car according to user's inputs.

The process starts by collecting the dataset. The next step is to do Data Preprocessing which includes Data cleaning, Data reduction, Data Transformation. Then, using various machine learning algorithms we will predict the price. The algorithms involve Linear Regression, Ridge Regression and Lasso Regression.

In future this machine learning model may bind with various website which can provide real time data for price prediction. Also we may add large historical data of car price which can help to improve accuracy of the machine learning model. We can build an android app as user interface for interacting with user. For better performance, we plan to judiciously design deep learning network structures, use adaptive learning rates and train on clusters of data rather than the whole dataset.

- **PREDICTIVE ANALYSIS OF USED CAR PRICES USING MACHINE LEARNING**
    - Ashutosh Datt Sharma, Vibhor Sharma, Sahil Mittal, Gautam Jain, Sudha Narang

A prediction model like this would not only help the buyers but sellers can also consider it to get an estimate of the value of vehicle they are looking forward to sell. Additionally, various online websites and portals can employ this model to improve prediction power and accuracy of their own system.

Lasso Regression: It is a type of linear regression itself which uses shrinkage which means that the data values are shrunk towards a data point in the center or in simple term, mean of the data. Lasso procedure supports simple and sparse models that have a lesser number of parameters. When any model has a high level of multicollinearity then this regression is best suited for that particular model. This model can also be employed in case certain parts of model selection are needed to be automated such as variable selection or parameter elimination. 'LASSO' is an acronym for Least Absolute Shrinkage and Selection Operator.

Ridge Regression: It is a regression method used for tuning of a model and analyzing a data that has multicollinearity. L2 regularization are performed under this method. The multicollinearity of data results in unbiased least-squares, large variance and thus the predicted values are quite far from the actual values.

Bayesian Ridge Regression: This regression is used to estimate any probabilistic model of any regression problem allowing a natural mechanism that survives data insufficiency or poor data distribution by linear regression formulation with the use of probability distributors avoiding any point estimates.

Random Forest Regression: Random-forest uses ensemble learning method for classification and regression and thus is a Supervised Learning Algorithm. Random forests have trees that run parallel to each other and have no interaction while they are being built. Random forest is a meta-estimator that assembles the results of multiple predictions. It also aggregates multiple decision trees with the help of some modifications.

Decision Tree Regression: This algorithm is used to build regression and classification models in the form of a tree structure. A dataset is broken into smaller subsets and simultaneously an associated decision tree is also created in an incremental manner. The final tree consists of decision nodes or leaf nodes as the

results. The algorithm used to construct a decision tree employs a top-down greedy search throughout the tree and possible branches in it without any backtracking.

XGBoost Regression: For building supervised regression models XGBoost is a very powerful algorithm to approach. XGBoost is one of the ensemble learning methods which involves training of individual models and then combining these individual models (base learners) to generate a single prediction.

Gradient Boosting Regression: It is a technique in machine learning for regression and classification problems to generate a prediction model. The prediction model produce is an ensemble of weak prediction models which typically are the decision trees. This technique generally outperforms the random forest method.

Python is majorly used for implementing machine learning concepts during this project as there are a number of inbuilt methods in the form of packaged libraries and modules present in python. The libraries used during the project implementation are the following:

Pandas: Pandas is one of the most used python libraries in data science. It supports various structures and data analysis tools using which is quite easy and they provide a high level of performance.

NumPy: NumPy is an open-source module in Python that provides very quick mathematical calculations on matrices and arrays. NumPy stands for ‘Numeric Python’ or ‘Numerical Python’. NumPy in combination with some other Machine Learning Modules like: Scikit-learn, Pandas, Matplotlib etc. provides a complete Python Machine Learning Ecosystem.

Limitations of these Studies:

In the past year the world of automobiles has seen a drastic change with the semiconductor shortages after the pandemic, which led to spike in used car prices. Hence, there was fast change in car prices during this study which will affect the actual car pricing prediction future. As the current dataset will undervalue the cars in the market. Therefore, a model that is built on real time data can be best integrated into a mobile app for public use would be the idea solution.

Future Development:

A potential improvement to the predictive power of all ML models, if they are able to take advantage of the information, is to add more correlated features. There are some features which are not related to the attributes of the car, such as the price of fuel. A car that uses more fuel will be worth less when fuel costs more. Other such features could include the economic conditions, or changes in the climate.

Conclusion:

Car price prediction can be a challenging task due to the high number of attributes that should be considered for the accurate prediction. The major step in the prediction process is collection and preprocessing of the data. In this research, PHP scripts were built to normalize, standardize and clean data to avoid unnecessary noise for machine learning algorithms.Data cleaning is one of the processes that increases prediction performance, yet insufficient for the cases of complex data sets as the one in this research. Although, this system has achieved astonishing performance in car price prediction problem our aim for the future research is to test this system to work successfully with various data sets.

References

1. Used Cars Price Prediction using Supervised Learning Techniques
   Source: ResearchGate.
   Authors: Mukkesh Ganesh
   https://www.researchgate.net/publication/343878698_Used_Cars_Price_Prediction_using_Supervised_Learning_Techniques

2. VEHICLE RESALE PRICE PREDICTION USING MACHINE LEARNING
   Author:B.Lavanya, Sk.Reshma, N.Nikitha, M.Namitha
   http://junikhyatjournal.in/no_1_Online_21/68.pdf

3. Predicting Used Car Prices
   Source : Stanford
   Kshitij Kumbar, Pranav Gadre and Varun Nayak
   http://cs229.stanford.edu/proj2019aut/data/assignment_308832_raw/26612934.pdf

4. Used Cars Price Prediction and Valuation
   Source : Rochester Institute of Technology
   Author: Abdulla AlShared
   https://scholarworks.rit.edu/cgi/viewcontent.cgi?article=12220&context=theses

5. Price Prediction for Used Cars
   Author : Marcus Collard
   https://www.diva-portal.org/smash/get/diva2:1674070/FULLTEXT01.pdf

6. USED CAR PRICE PREDICTION
   Authors : Praful Rane, Deep Pandya, Dhawal Kotak
   https://www.irjet.net/archives/V8/i4/IRJET-V8I4278.pdf

7. PREDICTIVE ANALYSIS OF USED CAR PRICES USING MACHINE LEARNING
   Authors: Ashutosh Datt Sharma, Vibhor Sharma, Sahil Mittal, Gautam Jain, Sudha Narang
   https://www.irjmets.com/uploadedfiles/paper/volume3/issue_6_june_2021/12071/1628083486.pdf