LITERATURE SURVEY

Price Prediction for Pre-Owned Cars Using Ensemble Machine Learning Techniques

LEARNING - Chetna Longania, Sai Prasad Potharaju, Sandhya Deore, Dept of Computer Engineering, Sanjivani College of Engineering, Kopargaon, Maharashtra, India.

The Pre-owned cars or so-called used cars have capacious markets across the globe. 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. This paper presents a system that has been implemented to predict a fair price for any pre-owned car. The system works well to anticipate the price of used cars for the Mumbai region. Ensemble techniques in machine learning namely Random Forest Algorithm, extreme Gradient Boost are deployed to develop models that can predict an appropriate price for the used cars. The techniques are compared so as to determine an optimal one. Both the methods provided comparable performance wherein extreme Boost outperformed the random forest algorithm. The researchers applied three machine learning technologies namely, Artificial Intelligence (AI), Support Vector Machine (SVM) and Random Forest (RF) separately. Authors collected data through different web portals.

METHODOLOGY:

We used ensemble machine learning techniques to implement the system. Ensemble techniques build multiple models, and then blend them. Thereby produces upgraded results than a single model would. We can train an ensemble and further use it to make predictions. Hence, an ensemble is a supervised learning algorithm. Using different ensemble methods, we can combine various models, thereby, moving on the path of achieving better accuracy. Suppose that you have designed an android application, before making it public you wish to know its ratings. What you can do is either ask your friends or family or colleagues to rate your app. This process would give you limited feedback. How about cumulating the reviews from fifty or more people who could be your family, friends or even strangers? Now the response that you will seek will be more assorted as the people in the closure possess different skills. This task of accumulating feedback will give you honest and accurate ratings. We can here conclude that a group of miscellaneous people make better decisions as compared to an individual. And the same is true if rather than using a single model, we use a group of diverse models. To achieve diversification in machine learning, we have ensemble techniques. Bagging, boosting, stacking and voting are famous ensemble methods. Bagging configures one ensemble model by deploying Bootstrapping as well as Aggregation, where bagging replaces observations from original datasets and creates multiple subsets holding observations. Further, a base model is fabricated for each of these subsets, which is run in parallel for each of these subsets, independent of each other. At the end, the results from these models are aggregated to give a

final prediction. Unlike bagging, the boosting process works sequentially on models. Here, different models are erected; each of the subsequent models corrects errors of the previous model. The weighted mean of all the models produces the final outcome, thereby combining weak learners to form a strong learner. Each of these models contributes to boost the performance of the ensemble. Stacking is an ensemble technique that uses predictions outputted from multiple models to construct a new model, which is further used to make predictions on the test set. Random Forest is a bagging algorithm whereas XGBoost is a boosting algorithm; these algorithms are used to implement the proposed system.

CONCLUSION:

The proposed system works well to predict a fair price for the pre-owned cars. The system skillfully projects prices for used cars in Mumbai region. The user of the system either the seller or the buyer, will get the honest price for the used car. Two popular ensemble machine learning algorithms namely Random Forest and XGBoost are deployed in order to implement a regression system for predicting used car prices. Both the techniques are comparable and offer high accuracy. Random Forest prevents overfitting by making use of more trees. With an ability to handle missing values, prevent overfitting, XGBoost is a widely used algorithm. As displayed by the results, XGBoost outperforms the Random Forest Algorithm. XGBoost is fast to execute and gives appreciable accuracy. The system proposed here is implemented for the Mumbai region only.

Price Prediction of Used Cars Using Machine Learning

LEARNING - Mr. Ram Prashath R, Nithish C N, Ajith Kumar J. Ijraset Journal For Research in Applied Science and Engineering Technology.

The goal of this study is to develop a model that can anticipate fair used car pricing based on a variety of factors such as vehicle model, year of manufacture, fuel type, Price, Kms, Driven. In the used car market, this strategy can benefit vendors, purchasers, and car manufacturers. It can then produce a reasonably accurate price estimate based on the data that users provide. Machine learning and data science are used in the model-building process. The data was taken from classified ads for second hand autos. To attain the maximum accuracy, the researchers used a variety of regression approaches, including linear regression, polynomial regression, support vector regression, decision tree regression, and random forest regression. This project visualized the data to better comprehend the dataset before starting the model-building process. To assure the regression's performance, the dataset was partitioned and changed to fit the regression. R-square was used to evaluate the performance of each regression. The final model contains more elements of used autos than earlier research while also having a higher forecast accuracy. Price prediction of used car using machine learning techniques is the first paper. They look at how supervised machine learning techniques can be used to estimate the price of second hand cars in Mauritius in this study. The forecasts are based on historical data taken from daily publications. To make the predictions, various techniques such as multiple linear regression analysis, were employed. According to author Sameer Chand, car price estimates on historical data gathered from daily newspapers. For estimating the price of cars, they employed supervised machine learning algorithms. Other methods that have been employed include multiple linear regression, k-nearest neighbor algorithms, nave based, and various decision tree algorithms. The best algorithm for prediction was identified after comparing all four algorithms. They had some issues comparing the algorithms, but they succeeded to do so.

METHODOLOGY:

In this section, we'll go over the many algorithms and datasets that were used to create this module. The model will be trained using a dataset with 92386 records. The value of an automobile is determined by factors such as kilometers travelled, year of registration, fuel type, car model, financial power, car brand, and gear type. We implemented five algorithms because this is a regression problem: Lasso Regression, Ridge Regression, Linear Regression.

A. *Lasso Regression*

The lasso regression allows you to shrink or regularize these coefficients to avoid overfitting and make them work better on different datasets. This type of regression is used when the dataset shows high multi collinearity or when you want to automate variable elimination and feature selection.

B. *Ridge Regression*

Ridge regression is a method of estimating the coefficients of multiple-regression models in scenarios where linearly independent variables are highly correlated. It has been used in many fields including econometrics, chemistry, and engineering. Ridge regression is a sort of linear regression that introduces a little degree of bias in order to improve long-term predictions.

C. *Linear Regression*

Quick to train and test as a baseline algorithm

CONCLUSION:

Because of the large number of characteristics that must be examined for an effective prediction, car price prediction will be a difficult assignment. The collecting and preparation of data is the most crucial step in the prediction process. Car data collected from kaggle.com is transformed into CSV format and used to create machine learning algorithms during the research. In this study, three algorithms were used: Linear, Lasso, and Ridge Regression. SVM classifier separated the data into two portions for training and testing purposes (Support Vector Machine). i.e., 75% of the data was used for machine learning training and 25% of the data was used for machine learning testing. The three machine learning models accuracy was tested and compared against one another. This is an important comparison between single and multiple groups of machine learning algorithms. As a result, this model will assist in predicting the car's actual price.

PREDICTING THE PRICE OF USED CARS USING MACHINE LEARNING-SAMEERCHAND PUDARUTH, Computer Science and Engineering Department, University of Mauritius, Reduit, Mauritius.

Predicting the price of used car is both an interesting and important problem. The purpose of this paper is to 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 newspaper on daily basis. The predictions are then evaluated and compared to find those which provide the best performance. Predicting the resale value of car is not a simple task because the value of used cars depend 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. 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.

METHODOLOGY:

There are four different methodologies used to predict the value of resale car. The first method is the multiple linear regression analysis, the Pearson correlation coefficient (r) was computed between different pairs of features. The correlation between mileage and volume of cylinder, manufactured year and its mileage, mileage and price, cylinder volume and price are plotted in graph to obtain a linear graph. Also, the linear regression, logarithmic regression and regression coefficient are noted. Next K-Nearest Neighbours (kNN), is a machine learning technique in which the new data is compared to all the existing records in order to locate the best match. Only three attributes were considered namely the make, year and cylinder volume. The data for year and cylinder had to be normalise to prevent large values from over-shadowing smaller values. Performance starts to degrade for higher values of k. The next method is Decision Trees, here the prices were grouped into six nominal categories as most of the popular decision tree algorithms cannot handle numeric outputs. There are many gaps in the ranges that have been defined because there were no cars within these ranges although it is certainly possible to get new data which fits within these zones. The Random Forest algorithm is very good at classifying the data based on the whole training set only however when the data is split between a training set and a testing set. The last method is the Naïve Bayes, it is one of the most useful machine learning technique. It is very easy to implement in software and the

accuracy is usually as good as more complex algorithms.

CONCLUSION:

Four different machine learning techniques have been used to forecast the price of used cars in Mauritius.The main weakness of decision trees and naïve bayes is their inability to handle output classes with numeric values.The main limitation of this study is the low number of records that have been used. As future work, we intend to collect more data and to use more advanced techniques like artificial neural networks, fuzzy logic and genetic algorithms to predict car prices.

USED CAR PRICE PREDICTION --Praful Rane, Deep Pandya, Dhawal Kotak, Information Technology Engineering, Padmabhushan Vasantdada Patil Pratisthan College of Engineering, Maharashtra India.

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.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.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.User Interface has also been developed which acquires input from any user and displays the Price of a car according to user's inputs.

METHODOLOGY:

There are two primary phases in the system: 1. Training phase: The system is trained by using the data in the data set and fits a model (line/curve) based on the algorithm chosen accordingly. 2. Testing phase: the system is provided with the inputs and is tested for its working. The accuracy is checked. Therefore, the data that is used to train the model or test it, has to be appropriate. The system is designed to detect and predict price of used car and hence appropriate algorithms must be used to do the two different tasks. Before the algorithms are selected for further use, different algorithms were compared for its accuracy. The well-suited one for the task was chosen.

Linear Regression:

Linear Regression attempt to model the relationship between two variables by fitting a linear equation to observed data. The other is considered to be dependent variable.

Ridge Regression:

A Ridge regressor is basically a regularized version of Linear Regressor. The regularized term has the parameter 'alpha' which controls the regularization of the model i.e helps in reducing the variance of the estimates.

Lasso Regression :

The "LASSO" stands for Least Absolute Shrinkage and Selection Operator. Lasso regression is a regularization technique. It is used over regression methods for a more accurate prediction. This model uses shrinkage. Shrinkage is where data values are shrunk towards a central point as the mean. The lasso procedure encourages simple, sparse models (i.e. models

with fewer parameters).

CONCLUSION:

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. For better performance, it is planned to judiciously design deep learning network structures, use adaptive learning rates and train on clusters of data rather than the whole dataset.

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. They have implemented and evaluated various learning methods on a dataset consisting of the sale prices of different makes and models across cities in the United States.Deciding whether a used car is worth the posted price when you see listings online can be difficult. Several factors, including mileage, make, model, year, etc. can influence the actual worth of a car. From the perspective of a seller, it is also a dilemma to price a used car appropriately. Based on existing data, the aim is to use machine learning algorithms to develop models for predicting used car prices.

METHODOLOGY:

They have 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, they used 500 thousand examples from dataset.

1. 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.

2. 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.

3.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.

4. 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.

5. 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.

6. KMeans + Linear Regression:

In order to capitalize on the linear regression results and the apparent categorical linearity in the data, 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.

7. 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.

CONCLUSION:

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. 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.As expected lightGBM performed marginally better than XGBoost but had a significantly faster training time. To correct for overfitting in Random Forest, different selections of features and number of trees will be tested to check for change in performance.