PREDICTIVE ANALYTICS : ANALYSIS OF THE FACTORS FOR USED CAR PRICE PREDICTION IN THE UNITED STATES OF AMERICA

Abhinay Tanguturi
MSc. in Computing (Data Analytics)
20211181
abhinay.tanguturi2@mail.dcu.ie

Sahithi Rayidi
MSc. in Computing(Data Analytics)
20211181
sahithi.rayidi2@mail.dcu.ie

Nandhini Reddy Aileni
MSc. in computing(Data Analytics)
20211181
nandhini.aileni2@mail.dcu.ie

Krithika Sharon Komalapati
MSc. in Computing(Artificial Intelligence)
20211239
krithika.komalapati2@mail.dcu.ie

The Git Hub link : Group_18_Git_hub

***Abstract*—An exact used car value assessment is an impetus for the sound advancement of the trade-in car market. In many articles read, a used car price prediction was done using a data mining approach. This paper is about the comparison of different algorithms to predict the price of used cars that are listed throughout The United States of America by performing detailed analysis on comparison of 4 regressor algorithms: linear, lasso, AdaBoost and Random Forest. These 4 algorithms are employed to predict used car price based on one category i.e 'price' by correlating it with other listed features of the cars.Results prove that AdaBoost outperformed the other three algorithms, when handling data with a large number of variables and samples.**

***Index Terms*—Data Mining, linear, lasso, AdaBoost, Random Forest, correlation.**

I. Introduction

Predictive analysis is a term, which is derived from statistics,machine learning, database and optimization techniques. Future events and behavior of variables can be forecasted using the predictive analytics by analysing current and historic data [**article1**]. Resale is the common term that is often used these days in different industries. The price of new cars in the industry is fixed by the maker for certain extra charges, in the form of taxes. So, the clients can be guaranteed of the money they invest. However, because of the high cost of new vehicles and the inability of the buyers to invest in a new cars has increased the demand of used cars. So, there is a dire need for an automated system to forecast the price of used cars in the market. Despite the fact that there are sites that offer this assistance, their prediction technique may not be precise. Various models and frameworks contribute on foreseeing a trade-in vehicles real market value. It's essential to realize their genuine market value with both purchasing and selling. Used car value prediction framework helps both buyers and sellers. This paper is aligned as following. Section 2 introduces related work about predictive analytics and its usage in machine learning techniques . Section 3 gives the description of the techniques and their approach used in our system. Section 4 presents the results and their evaluation metrics. We make concluding remarks in Section 5.

II. Literature Review

In [11], the author gave a brief description of predictive analytics approach in general by identifying and capturing important trends in data chosen. It involves the following steps to forecast the future outcomes. Requirements are collected from the client by the analyst for predicting what the client needs [10]. Based on the client requirement data is collected by the analyst from databases and sensors. Later data analysis and massaging, where the gathered data is converted into structured data, data reduction, data transformation as well as features are extracted. Machine learning and statistics are applied on this analysed data. Later a model is developed for predictions. Finally analyst's make the predictions and monitor the models.

In [4] the authors defined and explained how predictive analysis in combination with machine learning algorithms helps in making effective future predictions. Predictive models are trained, such that they respond to new values and deliver the business needs, where, they are classified into two types : (1) classification models that help in predicting the class membership and (2) regression models that help in predicting the number [1]. On a large dataset the relationship is detected between a single dependent and one or more independent variables by identifying the key-patterns between them and knowing how they are related to each other is estimated by performing regression analysis.

[7] proposed a supervised learning method named Random Forest for used car price prediction, where impact of each feature on price prediction is determined. A Random Forest with about five hundred decision trees was created for training the data [3]. These trees are individually trained on parts of the dataset and further helped in learning highly unpredictable patterns by growing very deep. Accuracy of 95% was achieved on a trained dataset.

In [9], the authors discussed how the statistical and analytical data mining techniques are used in analysing big data and developed novel strategies for the future possibilities of prediction on different medical datasets to attain accuracy. The authors in [2] explained how predictive analysis is performed on big data that deal with extracting information from data and predict the trends and behaviour patterns.

In [5][8], they performed a comparative study on performance of supervised machine learning models like random forest regression and linear regression. Mean absolute error(MSE) is considered as an evaluation metric where random forest regression outperformed multiple linear regression algorithm.The authors in [6] performed predictions using supervised machine learning technique where price is a dependent variable which is being predicted. And this price is derived from factors like vehicle model, city, color and mileage. Multiple Linear Regression algorithm offered 98% prediction precision.

III. Methodology

A. Data Set Description

The data utilised to perform predictive analysis and also comparative analysis is taken from Kaggle, which a user scrapped from craigslist. This data contains about 458,213 listings along with all suitable information that Craigslist provides on car sales including columns like condition, manufacturer, price, and 23 other categories in four quarters of 2020. Before, proceeding with data cleaning few explorations and visualisations are performed on the raw data taken from kaggle to know about the features given, like missing values and extreme values are explored.

B. Data Cleaning

The initial step was to pull out the unwanted attributes from the given 26. As, a result characteristics like 'id', 'url','image_url', 'lat', 'long', 'city_url', 'desc', 'city', 'VIN' etc are eliminated, Leaving only 14 features to perform the further analysis. Further there data was scrutinised to identify if any extreme and missing values are there among these 14 features.

For our analysis, the data points below can be considered as outliers and are removed since they hinder the prediction power of the model. Firstly, cars that are listed in 'price' feature with more than $100,000 were eliminated as this contributed only the small percentage of the buyers/sellers which were only 580 out of 550313 records. And, also 61726 listings of the cars that had a price value less than $750 were dropped as these were considered as noise for the data. Further, odometer records that are greater and lower than 300.000 and 10 miles respectively were dropped. At last, car listings earlier than 1985 were eliminated because the car manufacturing year plays a predominant role in prediction.

In the next step, the null points are identified in a few features and are filled with relevant values. For the null points present in the 'condition' feature, were filled by paying attention to the different no. of categories in the 'condition' feature. Where, the average of the 'odometer' of all the sub-categories in 'condition' were calculated and then, null values were filled in view of these sub-category average odometer values. Further, where the feature 'model' values are higher than 2019 were filled as 'new' and cars between 2017–2019 were filled as 'like new'. By the end of this step, all null points in the 'condition' attribute were filled. Further, other categorical variables features with null points were filled by using the 'ffill' method, which means they are filled with the most repeating values among them so that no outliers exist and no new variable is needed to reconstruct the pattern. After this step, the data cleaning is done and we are left with only 380,962 tuples of data that will be further used for studying and analyzing.

C. Feature Engineering

This univariate and bivariate analysis is performed on the cleaned data to look and understand the features, their combinations and pattern trends in a better way.

By looking at the price feature, it clearly explains the car price values have a median around $15k-$17k which can be affordable by maximum all groups of people whereas, on the other hand cars with higher prices are also available with min. value of $50K which is for people with a high budget.

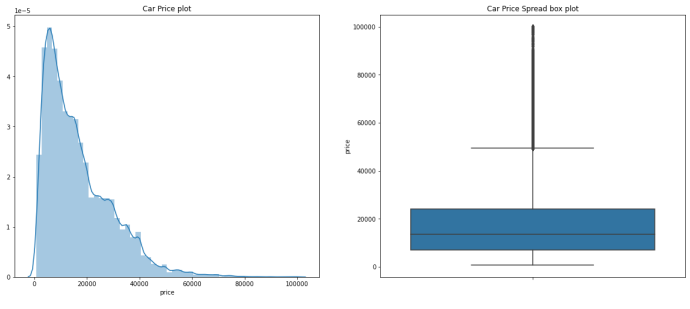

Figure 1. Uni-variate and Bi-variate Analysis of Price feature

By visualising the 'manufacturer' feature it is concluded that the "Ford" cars constitute 68,000 of car listings, which

indicates their demand is high in the market compared to other manufacturer types.

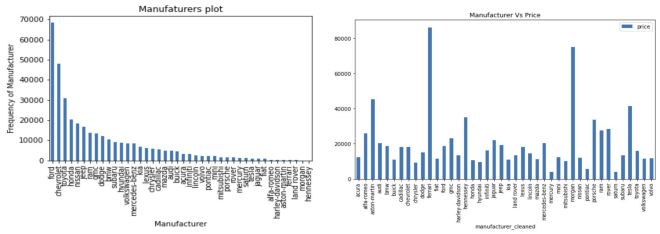

Figure 2. Uni-variate and Bi-variate Analysis of Manufacturer feature

By performing visualisation on condition attributes of the used cars it is clearly seen that the cars with excellent condition are in more demand followed by good condition. Similarly, when both the condition and odometer features are compared in a graph it distinctly shows that the cars with fair condition has a high average odometer value about 180k miles.

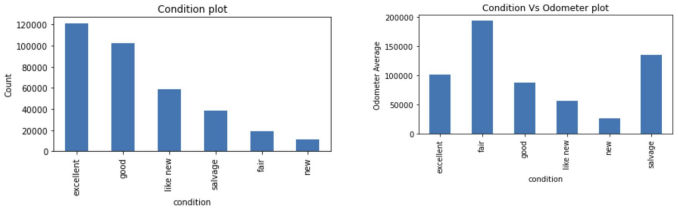

Figure 3. Uni-variate and Bi-variate Analysis of Condition feature

When the fuel is compared gas cars were more in count in the market whereas electric cars are quite low but the price is comparatively higher for electric cars as those are this era cars i.e., after 2016. Diesel cars have a better count when compared to other types except gas with a better price in the market.

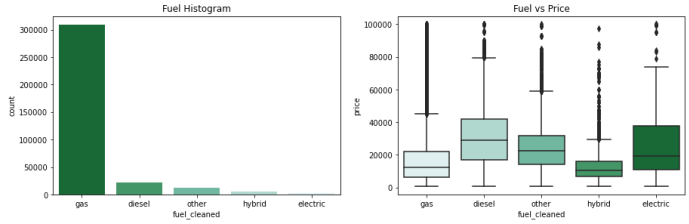

Figure 4. Uni-variate and Bi-variate Analysis of Fuel feature

Similar analytics is made to the other attributes of size, cylinders and drive type are analysed similarly to the fuel attribute. Analysing the cylinders variable along with price it is identified that 6 and 4 cylinder cars are relatively more for sale at a lower price rate. 12 cylinder cars are less in count but have a higher rate. Transmission in automatic cars is more, where the other(not specified) type has a higher price and the automatic cars are having relatively better price. The drive variable shows that 4wd types of cars are more in number but the price of 4wd and rwd are comparatively the same when compared.

When the odometer variable is visualised the peak values in the odometer are observed near 40K 120K miles, where 120k miles is the average value of the excellent condition value which again describes the importance of odometer and condition values importance in the car price prediction. Similarly when the odometer and price feature are analysed combined it shows that as the odometer value increases the price of the car decreases and are inversely proportional to each other.

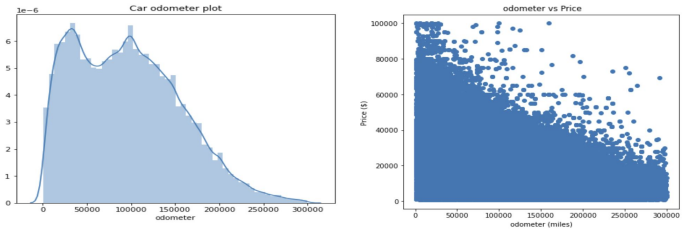

Figure 5. Uni-variate and Bi-variate Analysis of of Odometer feature

D. Hypothesis and Testing

Paired T-test is used to evaluate the attributes after cleaning for the algorithmic efficiency. The T-test is performed between price and odometer attribute as
Null Hypothesis: There is no significant relation between price and Odometer
Alternate Hypothesis: There is significant relation between Price and Odometer.
both the tests are validated with pvalue 0.05. Similar tests were performed with price and other attributes and are validated

E. Algorithms

The data set consists of 14 attribute after the feature engineering. Three are numerical variables and other are categorical variables which are transformed to numerical variables. To train 80% of the data is used and the other 20% data is tested for the efficiency.

Linear Regression algorithm is applied for the informational index which doesn't give any momentous outcomes where the Mean Square Error (MSE) is 73632452.73 and the Root Mean Square Error is which are moderately high. Test and Predicted values are looked at and negative qualities found in the predicted value which demonstrate the failure of the calculation. The Attribute significance for the calculation is visualized.

Lasso Regression is applied to check whether it changes the negative outcomes happened from the Linear Regression. Although the data-set isn't such huge for the lasso Regression yet applying it will get another viewpoint for the study. this lasso regression results were moderately same with the linear regression, where the variable significance marginally contrasts which isn't honorable.

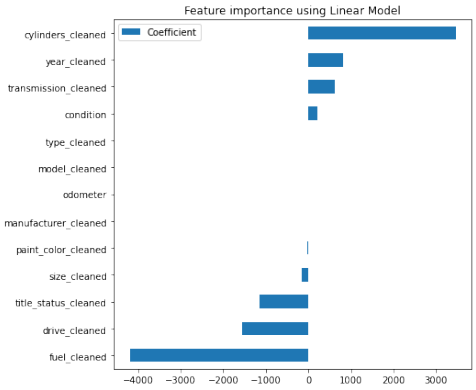

Figure 6. Variable Importance in Linear and Lasso Regression's Algorithm

AdaBoost is a particular execution of the inclination boosting technique which utilizes more exact approximations to track down the best tree model. It utilizes various techniques that make it astoundingly effective, especially with organized information. AdaBoost has extra benefits where preparing is extremely quick and can be parallelized across clusters. Hence, AdaBoost was another model utilized in the study. The visualization for the attribute importance for the prediction in AdaBoost algorithm. The algorithm performed really well and the efficiency is high. The Mean Square Logarithmic Error (MSLE) and Root Mean Square Logarithmic Error (RMSLE) are very low.

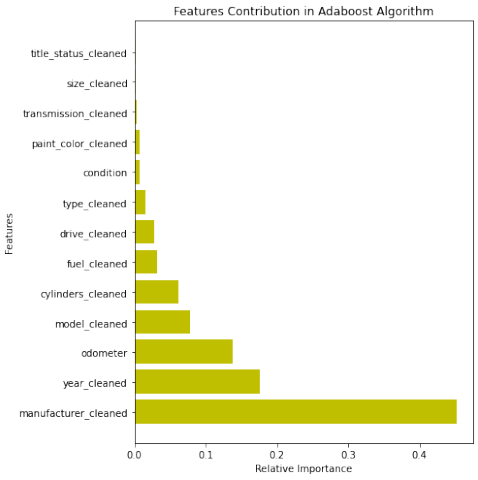

Figure 7. Variable Importance in AdaBoost Algorithm

Unlike the linear and lasso regression the variable importance is completely where in linear cylinders has more importance but in AdaBoost manufacturer has became most important variable. The year, model and odometer takes top other places in contributing to the algorithm.

Random Forest is a bunch of numerous decision trees. Profound Decision trees may experience the adverse effects of overfitting, however random forest prevents overfitting, by creating trees on the random subsets. That is the reason it's a decent model in the examination. In this investigation, 200 trees were made for better effectiveness. The greater the number of trees,the better the outcome. This is shown empirically through low RMSE and MSE values. The variable importance a major role in the random forest regression model. In the random forest model, the "year"

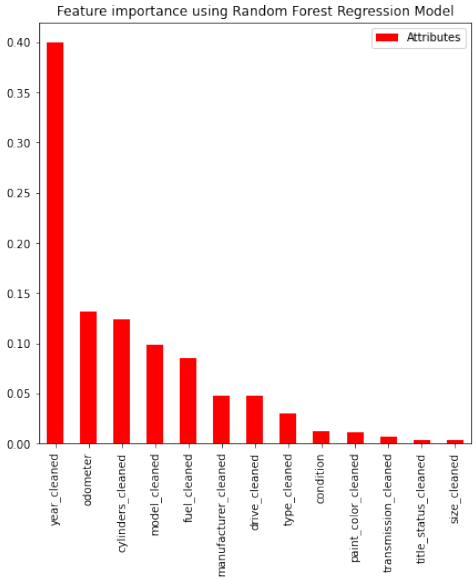

Figure 8. Variable Importance in Random Forest regressor Algorithm

attribute is a major contributor. As the efficiency difference between AdaBoost and random forest algorithms is negligible but the variable contribution is pretty much different. When the best attributes from the random forest are extracted and applied in the advanced random forest regressor model, a slight variation in the prediction efficiency is observed which is reduced by 1%. This also proves the prediction efficiency is achieved by most dependent variables of the data-set.

IV. Results Evaluation

By comparing all 4 models from the table 1 given below it can be deduced that the AdaBoost regressor has generated best values for MSLE, RMSLE, R2 Score and accuracy parameter when compared with Linear, Lasso, Random Forest and Advanced Random Forest Regressor. The most important features identified by AdaBoost regressor are here: Manufacturer, year, model and odometer.

Table I
MODEL RESULTS

Metrics	Linear	Lasso	AdaBoost	RF	ARF
MSLE	3.13211	3.11583	0.091891	0.084713	0.08972
RMSLE	1.76977	1.76517	0.30313	0.29105	0.299541
R2 Score	0.56383	0.56370	0.93149	0.93031	0.928306
Accuracy	56.38	56.37	93.14	93.03	92.83

RF - Random Forest Regresssor,
ARF - Adavanced RAndom Forest Regressor

V. CONCLUSION

Predictive models are used to forecast used car prices and their performance is compared using different machine learning techniques in this paper. A data-set with 13 predictors and 380962 observations is used for prediction.Feature exploration is performed by employing data visualisations and exploratory data analysis. Relationship between the features were identified. Finally 4 regressor models are applied for predicting the used car price. However by adding few new variables, which are not in the current data-set, like number of doors, gas/mile (per gallon), mechanical and cosmetic reconditioning time,and used-to-new ratio might improve the predictive models to forecast the used car prices. These predictive models can also be used in financial, retail, health insurance and public sectors.

REFERENCES

[1] Adem Akdogan. "Predicting Car Prices Using Machine Learning Models-Python". In: *Analytics Vidhya* (2020). DOI: Medium.com.

[2] R. Banjade and S. Maharjan. "Product recommendations using linear predictive modeling". In: *2011 Second Asian Himalayas International Conference on Internet (AH-ICI)*. 2011, pp. 1–4. DOI: 10.1109/AHICI.2011.6113930.

[3] Chuancan Chen, Lulu Hao, and Cong Xu. "Comparative analysis of used car price evaluation models". In: *AIP Conference Proceedings* 1839.1 (2017), p. 020165. DOI: 10.1063/1.4982530. eprint: https://aip.scitation.org/doi/pdf/10.1063/1.4982530.

[4] Wakefield Katrina. ""Predictive analytics and machine learning." SAS". In: (). DOI: https://www.sas.com/en_gb/insights/articles/analytics/a-guide-to-predictive-analytics-and-machine-learning.html#:~:text=Predictive20analytics\%20is\%20driven\%20by\%20predictive\%20modelling.&text=Predictive\%20analytics\%20and\%20machine\%20learning\%20go\%20hand\%2.

[5] N. Monburinon et al. "Prediction of prices for used car by using regression models". In: (2018), pp. 115–119. DOI: 10.1109/ICBIR.2018.8391177.

[6] Kanwal Noor and Sadaqat Jan. "Vehicle Price Prediction System using Machine Learning Techniques". In: *International Journal of Computer Applications* 167.9 (June 2017), pp. 27–31. ISSN: 0975-8887. DOI: 10.5120/ijca2017914373. URL: http://www.ijcaonline.org/archives/volume167/number9/27802-2017914373.

[7] Nabarun Pal et al. "How much is my car worth? A methodology for predicting used cars' prices using random forest". In: *Future of Information and Communication Conference*. Springer. 2018, pp. 413–422.

[8] N. H. Chummun S. Peerun and S. Pudaruth. "Predicting the Price of Second-hand Cars using Artificial Neural Networks". In: *The Second International Conference on Data Mining Internet Computing and Big Data* (2015). DOI: pp.17-21.

[9] Poornima Selvaraj and Pushpalatha Marudappa. "A survey of predictive analytics using big data with data mining". In: *International Journal of Bioinformatics Research and Applications* 14 (Jan. 2018), p. 269. DOI: 10.1504/IJBRA.2018.092697.

[10] Nilay D. Shah, Ewout W. Steyerberg, and David M. Kent. "Big Data and Predictive Analytics: Recalibrating Expectations". In: *JAMA* 320.1 (July 2018), pp. 27–28. DOI: 10.1001/jama.2018.5602.

[11] Shubham Vartak. "An Overview of Predictive Analysis: Techniques and Applications". In: *International Journal for Research in Applied Science and Engineering Technology* 8 (Nov. 2020), pp. 652–662. DOI: 10.22214/ijraset.2020.32250.