XGBoost Machine Learning Model For Car Condition Prediction

Mohit Verma,
Galgotias college of Engineering And Technology Greater Noida, India
mohitvermaca37@gmail.com

Suraj Kushwaha,
Galgotias college of Engineering and Technology Greater Noida, India
srjofficx@gmail.com

***Abstract*—The worth of latest cars within the business is fastened by the manufacturer with some extra prices incurred by the govt within the style of taxes. So, customers shopping for a replacement automobile are often assured of the money they invest to be worthy. however, because of the increased worth of latest cars and therefore the incapability of consumers to shop for new cars because of the shortage of funds, used automobile sales area unit on a world increase there's a desire for a second-hand automobile condition prediction system to effectively verify the goodness of the automobile employing a sort of options. even supposing their area unit websites that supply this service, their prediction technique might not be the most effective. Besides, completely different models and systems could contribute to predicting power for a second-hand car's actual value. This paper encompasses a technique of XGBoost. The advantage of applying the XGBoost model is that the procedure method doesn't need an extended time, and possesses satisfactory accuracy in various regression and classification cases. supported the results, it is often over that XGBoost produces higher accuracy than alternative machine learning strategies like provision Regression, KNN Classifier, Random Forests Classifier.**

***Keywords*—** *Machine Learning*, Logistic Regression, KNN Classifier, *Random Forest, XGBoost.*

I. Introduction

Car condition prediction is somehow an interesting and standard drawback. As per info that was gotten from the Agency for Statistics of BiH, 921.456 vehicles were registered in 2014 from that 84% of them area unit cars for personal usage [1]. This range has inflated by 2.7% since 2013 and it's seemingly that this trend can continue, and also the range of cars can increase in future. Predicting the condition of used cars is a vital and interesting problem. it's vital to know their actual value whereas both buying and marketing. There are websites that offer an estimated condition of a car. They may have a good prediction model. However, having a second model may help them to give a better prediction to their users. Therefore, the model developed in this study may help online web services that tell a used car's market value. There are lots of individuals who are interested in the used car market at some points in their life because they wanted to sell their car or buy a used car. In this process, it's a big corner to pay too much or sell less than its market value. Dealers are one of the biggest target groups that can be interested in the results of this study. If used car sellers better understand what makes a car desirable, what the important features are for a used car, and then they may consider this knowledge and offer a better service.

This Car Evaluation Database was derived from a simple hierarchical decision model originally developed for the demonstration of DEX, M. Bohanec, V. Rajkovic. The data has six attributes (*buying, maintenance, doors, lug_boot, persons, and safety*) which help to predict the condition of a used vehicle.

II. Literature Review

In this Methodology most of the machine learning approaches were compared but only the xgboost technique was giving the best accuracy and best car condition prediction compared to other classifiers, XGBoost is well known to provide better solutions than other machine learning algorithms. XGBoost (Extreme Gradient Boosting) belongs to a family of boosting algorithms and uses the gradient boosting framework at its core. It is an optimized distributed gradient boosting library. Boosting is a sequential technique which works on the principle of an ensemble. It combines a set of weak learners and delivers improved prediction accuracy.

Predicting the price of a used car has been studied extensively in various researches. Listian discussed, in her paper written for Master thesis [2].

Another approach was given by Richardson in his thesis work [3]. His theory was that car producers produce more durable cars. Richardson applied multiple regression analysis and demonstrated that hybrid cars retain their value for longer time than traditional cars. This has roots in environmental concerns about the climate and it gives higher fuel efficiency.

Wu et al. [4] conducted a car price prediction study, by using a neuro-fuzzy knowledge-based system. They took into consideration the following attributes: brand, year of production and type of engine. Their prediction model produced similar results as the simple regression model. Moreover, they made an expert system named ODAV (Optimal Distribution of Auction Vehicles) as there is a high demand for selling the cars at the end of the leasing year by car dealers. This system gives insights into the best prices for vehicles, as well as the location where the best price can be gained.

Gonggie [5] proposed a model that is built using ANN (Artificial Neural Networks) for the price prediction of a used car. He considered several attributes: miles passed, estimated car life and brand. The proposed model was built so it could deal with nonlinear relations in data which was not the case with previous models that were utilizing the simple linear regression techniques

Furthermore, Pudaruth [6] applied various machine learning algorithms, namely: k-nearest neighbors, multiple linear regression analysis, decision trees and naïve bayes for car price prediction in Mauritius. The dataset used to create a prediction model was collected manually from local newspapers in less than one month, as time can have a noticeable impact on price of the car. He studied the following attributes: brand, model, cubic capacity, mileage in kilometers, production year, exterior

color, transmission type and price. However, the author found out that Naive Bayes and Decision Tree were unable to predict and classify numeric values. Additionally, a limited number of dataset instances could not give high classification performances, i.e. accuracies less than 70%.

Noor and Jan [7] build a model for car price prediction by using multiple linear regression. The dataset was created during the two-months period and included the following features: price, cubic capacity, exterior color, date when the ad was posted, number of ad views, power steering, mileage in kilometer, rims type, type of transmission, engine type, city, registered city, model, version, make and model year. After applying feature selection, the authors considered only engine type, price, model year and model as input features. With the given setup authors were able to achieve prediction accuracy of 98%.

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

III. METHODOLOGY

In our analysis, we have a tendency to found That Random Forest works fine for automobile condition Prediction with nice accuracy however it may be any raised by another fresh arrived algorithmic program XGBoost, during this analysis we have a tendency to proposing a technique to create a predictive model with larger accuracy than Random Forest.

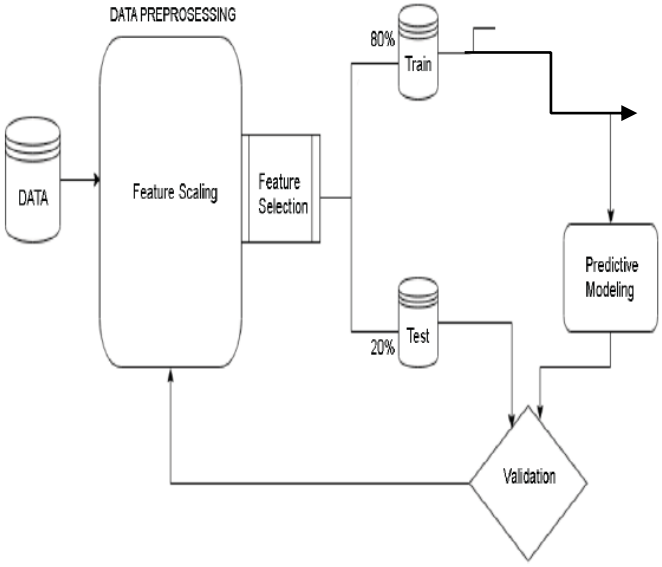

Fig.1 Flowchart Representing Execution Process

A. Data Preprocessing

Data Preprocessing is a crucial step in development of an effective prediction model. If we have any missing values in the dataset then that can interrupt our model to the prediction, data preprocessing helps to found any missing values in the dataset.

a. Label Encoding - In the dataset, there are 6 attributes. 2 of them are numerical variables while the rest of them are categorical. In order to apply machine learning models, we need numeric representation of the features. Therefore, all non-numeric features were transformed into numerical form.

B. Car Evaluation Dataset

In order to find the effectiveness of our approach, we choose a standard car evaluation dataset of UCI machine learning Repository This repository is publicly available at https://archive.ics.uci.edu/ml/datasets/car+evaluation dataset contain six attributes respectively.

In fig. 2 we can effortlessly observe the condition of vehicles that are given within the data set. there are approx 1200 cars which can be unacceptable, approx four hundred cars are in acceptable condition beneath two hundred cars are in vgood condition and also approx a hundred and fifty cars are in good condition.

Fig. 2 Distribution of Cars by Condition

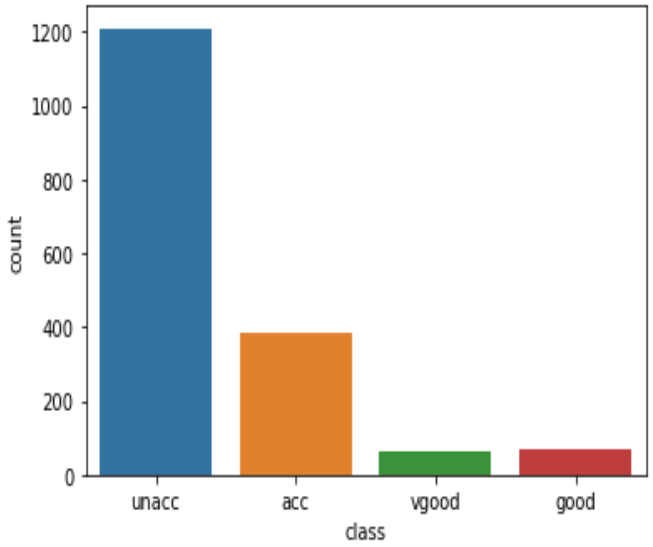

Fig. 3 Dataset insights

	buying	maint	doors	persons	lug_boot	safety
0	vhigh	vhigh	2	2	small	low
1	vhigh	vhigh	2	2	small	med
2	vhigh	vhigh	2	2	small	high
3	vhigh	vhigh	2	2	med	low
4	vhigh	vhigh	2	2	med	med

C. Data Split in Training Data & Testing Data

We cannot train our model by all points of the dataset. If we tend to train our model, for all data points, it'll be the difficulty of overfitting and in such a situation; our model is presumably the incorrect prediction of a new statement. to check the effectiveness and responsibleness of our model we've determined to split our dataset into parts within the quantitative relation of 80:20 severally. In 80% of the data set is employed in model coaching whereas the opposite 20% is employed to evaluate the model performance with expected and actual price comparisons.

IV. RESULTS

In this experiment, we tested 4 algorithms which are logistic regression, knn classifier, Random Forest and XGBoost. The results achieved for models are:

Logistic Regression: Logistic regression could be a technique borrowed by machine learning from the sector of statistics. logistic Regression could be a supervised machine learning classifier. it's a special case of regression, however, regression predicts during a continuous manner whereas logistic regression predicts in binary. within the next section, you'll see that supplying regression did not perform well, it provides solely 65% accuracy and 59% F1 score therefore we tend to affect on toward tree-based classifiers.
The results achieved using Logistic Regression is shown in Table:

Performance Measures of Logistic Regression

Performance Parameter	Results
Accuracy	65%
Precision	55%
Recall	66%
F1 Score	59%
Cross validation score	65%

Random Forest: Random forest (RF) also known as random decision forest belongs to the category of ensemble methods. RF can be used for classification and regression problems. The algorithm was developed by Ho as an improvement for overfitting of the decision tree algorithms [8]. Random Forest algorithm was applied on the whole dataset, to test how accurately the classifier can categorize samples into unacceptable, acceptable, good, vgood car classes. RF is a meta estimator that fits a number of decision tree classifiers on various subsamples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting [9]. From the self-experiment, we found that Random Forest performed outstanding in car condition prediction with a great Accuracy of 96%.
The results achieved using Random Forest is shown in Table:

\

Performance Measures of Random Forest

Performance Parameter	Results
Accuracy	96%
Precision	81%
Recall	66%
F1 Score	71%
Cross validation score	85%

KNN Classifier: KNN-classifier can be used when your data set is small enough, so that KNN-Classifier completes running in a shorter time. The KNN algorithm can compete with the most accurate models because it makes highly accurate predictions. Therefore, we can use the KNN algorithm for applications that require a good prediction but do not require a human-readable model. The quality of the predictions depends on the distance measure. Therefore, the KNN algorithm is suitable for applications for which sufficient domain knowledge is available. KNN classifier offers Accuracy of 89%.

Performance Measures of KNN Classifier

Performance Parameter	Results
Accuracy	89%
Precision	81%
Recall	66%
F1 Score	71%
Cross validation score	74%

XGBoost: XGBoost may be a powerful machine learning algorithmic program. it had been recently introduced. It comes below supervised Learning. it's essentially supported the thought of Gradient Boosting. XG Boost works on parallel tree boosting that predicts targets by combined results of multiple weak models it offers nice speed and accuracy. XGBoost Performed very well on our model and got an accuracy of 98% without tuning.

The results achieved using XGBoost are shown in Table:

Performance Measures of XGBoost

Performance Parameter	Results
Accuracy	98%
Precision	91%
Recall	94%
F1 Score	91%
Cross validation score	88%

V. CONCLUSION

The major step within the prediction method is that the assortment and preprocessing of the information. during this analysis, we tend to pre-process the data. we tend to summary the data then used the machine learning models — logistic regression, Random forest, KNN classifier, and XGBoost then we tend to see XGBoost performs very over totally different models while not tuning the model it provides the simplest accuracy of 98%.

Then we use our model for prediction. We see our model can predict nicely and offer excellent results.

REFERENCES

[1] Agencija za statistiku BiH. (n.d.), retrieved from: http://www.bhas.ba . [accessed July 18, 2018.]

[2] Listiani, M. (2009). Support vector regression analysis for price prediction in a car leasing application (Doctoral dissertation, Master thesis, TU Hamburg-Harburg).

[3] Richardson, M. S. (2009). Determinants of used car resale value. Retrieved from: https://digitalcc.coloradocollege.edu/islandora/object /coccc%3A1346 [accessed: August 1, 2018.]

[4] Wu, J. D., Hsu, C. C., & Chen, H. C. (2009). An expert system of price forecasting for used cars using adaptive neuro-fuzzy inference. Expert Systems with Applications, 36(4), 7809-7817.

[5] Gongqi, S., Yansong, W., & Qiang, Z. (2011, January). New Model for Residual Value Prediction of the Used Car Based on BP Neural Network and Nonlinear Curve Fit. In Measuring Technology and Mechatronics Automation (ICMTMA), 2011 Third International Conference on (Vol. 2, pp. 682-685). IEEE.

[6] Pudaruth, S. (2014). Predicting the price of used cars using machine learning techniques. Int. J. Inf. Comput. Technol, 4(7), 753-764

[7] Noor, K., & Jan, S. (2017). Vehicle Price Prediction System using Machine Learning Techniques. International Journal of Computer Applications, 167(9), 27-31.

[8] Ho, T. K. (1995, August). Random decision forests. In Document analysis and recognition, 1995., proceedings of the third international conference on (Vol. 1, pp. 278-282). IEEE.

[9] 3.2.4.3.1. sklearn.ensemble.RandomForestClassifier — scikit-learn 0.19.2 documentation. (n.d.). Retrieved from: http://scikitlearn.org/stable/modules/generated/sklearn.ensemb le .RandomForestClassifier.html [accessed: August 30, 2018].