Car-Valuate : Car Valuation Made Easy Using Machine Learning

Sumit Tiwari, Priyanshu, Dr. R. Mahendran

Department of Computer Science and Engineering
Galgotias University –203201, INDIA
Corresponding Author: sumit.22scse1010915@galgotiasuniversity.edu.in

ABSTRACT

The swift growth of online used-car market place has raised the pressure on proper, transparent, and automated systems of valuations od car prices. Pricing of used car is a complicated fact, because it is related to several variable which are interconnected including age of the vehicle, miles covered, type of fuel used, transmission, ownership record, and market demand. Traditional valuation techniques founded on subjective appraisal or rudimentary statistical model may be subjective as well as weak in predictive ability particularly when non-linear depreciation patterns are dealt with. In this paper, Car-Valuate is proposed, a car valuation system that works based on machine learning and uses historical data to predict real used car prices. First, a Linear Regression model is used as a baseline because they are not complex and easy to interpret but in the actual experiment, it is shown that it has significant mistakes in prediction because it assumes linear relations between variable. In order to address these shortcomings a Random Forest Regress or is used which uses ensemble learning to learn the intricate non-linear interactions of features. The systematic data pre-processing steps to be included in the proposed system are the processing of missing values, encoding categorical features, feature selection, and the logarithmic transformation of the price values. The trained model will be implemented with the help of the Flask-based backend and combined with a web-based frontend written in HTML, CSS, and Bootstrap, which will allow interacting with the user in real time. The evaluation of the results of the performance shows that the predictions made by the Random Forest model are in close correlation with the real prices in the market, and they are far more effective and robust as compared to the Linear Regression. Evidence proves that Car-Valuate is a powerful system.

Keywords: Car price prediction, Machine learning, Regression techniques, Linear regression, Random Forest Regressor.

I. INTRODUCTION

The drastic digitalization of the automotive sector has given rise to a great number of online websites that are used to purchase and sell used cars. Precise and reasonable car price valuation in such platforms is important in the transparency, trust, and efficiency of both sellers and buyers. A wrong pricing can lead to losses, time wastage in sales and lack of trust in online stores by the users. Through this, the automated and data based valuation systems are today as much of a necessity in contemporary automotive ecosystem.

The primary methods of the traditional used car valuation methods are based on either a manual or expert inspection process or prices based on rules. Although such techniques offer domain knowledge, they tend to be subjective, inconsistent and cannot scale easily. Besides, they are not responsive to changing market trends and regional price differences. Accurate prediction of depreciation is a complicated process since there are various interacting factors that affect the depreciation of a vehicle, such as the manufacturing year, the vehicle mileage, the type of fuel used, the transmission system, the history of ownership as well as the brand value.

The progress of machine learning has made it possible to create predictive models that can be trained to analysis massive historical data and extract complicated patterns and associations. Machine learning-based solutions are more accurate and more flexible than the classical statistical techniques. In the paper, Car-Valuate, a machine learning- based used car price prediction system, will be introduced, and its purpose is to offer quality and automated valuation. The analysis will first use Linear Regression as a reference model and then a Random Forest Regressor will be proposed which will be able to avoid the weakness of Linear modelling. The usefulness of ensemble learning in non-linear pricing behaviours is proved by comparative analysis to reveal the practical nature of proposed system.

II. LITERATURE REVIEW

There have been a few studies that discuss the application of machine learning algorithms in predicting the price of used cars. The first studies mainly focused on the classic statistical methods such as the Linear Regression as it is easy to understand

ISBN: 978-81-985770-0-9

and interpret. In reality, however, it is assumed that the relationships between the features of the input and the variables of the target are linear and this limits the models' performance in real-life pricing scenarios [1][2].

Recent research has shown that price depreciation of cars has non-linear trends that depend on several interacting characteristics such as the age of the vehicle, vehicle mileage, brand name, the type of fuel and the type of transmission system. As mentioned in the previous research, Linear Regression is sensitive to be poorly fitted model when using such a complex data [3].

It has been explained that Decision Tree-based models have gained popularity because they are able to model non-linear relationships without having to do much transformation of the data. Random Forest, as an ensemble learning method, is a technique developed by Breiman, which involves the combination of a number of decision trees to minimize variance and enhance predictive accuracy [4]. The effectiveness of Random Forest has been pointed out by many researchers in car price prediction in comparison to single regression models [5][6].

The model based on Gradient Boosting and XG Boost is also more accurate as it corrects the error of the previous model. These models however need hyper parameter tuning and are more expensive to compute, might restrict their application to real-time web-based systems [7][8].

There is feature engineering and data pre-processing, which are important in enhancing prediction performance. It is shown that it is necessary to encode categorical variables and also to deal with the outliers to improve the accuracy of the model [9]. Furthermore, as in a number of earlier studies, the price values have been subjected to log-transformation, which has been shown to provide better regression results and to stabilize the variance [10].

A number of vehicle valuation tools have been created on Python and Flask frameworks that allow them to be easily integrated with machine learning models into the real world. Such systems are successful tests of the suitability of implementing predictive models as REST full services that could be applied through the web interface [11][12].

From the analysis of the sources, it is concluded that Random Forest Regressor is one of the most accurate and stable models which can be used to predict the price of a used car owing to its accuracy, resistance and it can be deployed.

III. DESIGN FLOW

The Car-Valuate system is designed in a systematic and structured manner and can be used to develop a used car price prediction model based on machine learning. It starts with data collection, collecting historic car used data from reputable sources like online auto marketplaces, dealership databases, and public data sets. The information gleaned includes key car features like model year, mileage, engine details, transmission type, car history, and selling price. This dataset will be used as ground truth to train and evaluate the prediction models.

Following the data collection phase, data are pre-processed to improve data quality and boost the performance of the model. Data is checked for consistency and inconsistencies are identified and deleted. Label encoding is applied to categorical features (such as fuel type and transmission type) to map them into numbers, and numerical features are standardized if needed. In addition, the attribute of selling price is log transformed to help correct skewness and stabilize variance, which makes more effective the predictive models in this case.

After pre-processing, Exploratory Data Analysis (EDA) is performed to understand the distribution and behaviour of the data. Scatter plots and line plots are used to detect patterns, trends and relationships between variables. This analysis is helpful in understanding which factors have a significant influence on used car prices and can give helpful guidance on feature selection and model building.

Then, feature selection is conducted to determine the most salient feature, which can help in making accurate price prediction. Limited features that do not provide the same amount of predictability are eliminated to decrease the complexity of the model, avoid over fitting and increase the generalization capability of the system. This reduces the overall efficiency and interpretability of the prediction model since only the most influential variables will be kept.

A machine learning algorithm is implemented in the model development phase. Linear Regression is used as a baseline model as one of the simplest and interpretable models. However, used cars often have non-linear and complex relationships between the variables and a Random Forest Regressor is also created. The Random Forest model combines multiple decision trees and ensemble learning methods to detect complex patterns in data and offer enhanced and more accurate predictions.

ISBN: 978-81-985770-0-9

The models are evaluated by comparing the actual selling prices with the corresponding values predicted by the models, to assess their effectiveness. Performance metrics are used to assess the accuracy and reliability of predictions. These evaluation findings suggest that the Random Forest Regressor has a greater ability to predict market prices, as it consistently provides more accurate predictions than the Linear Regression model, especially when dealing with complex pricing patterns.

After the optimal model is found, the trained Random Forest model is serialized using the Pickle library and deployed using a Flask-based web app. The user interface is created with HTML, CSS and Bootstrap which facilitates users to enter vehicle information and get actual price forecasts. The machine learning model is integrated with the web application, making it accessible and providing practical usage for the end-users.

Lastly, the system computes the approximate market value of the chosen used car, given the input features. This automated valuation process streamlines the manual tasks, provides accurate pricing, and helps buyers and sellers make informed decisions. The whole design flow guarantees a scalable, reliable, and effective used car price prediction system which can be applied in actual car marketplaces.

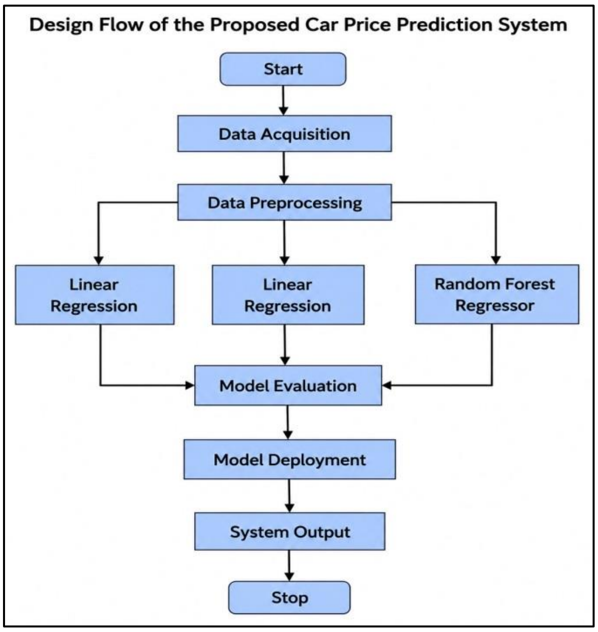

Fig. 1. Design Flow of proposed car price prediction system

IV. IMPLEMENTATION / METHODOLOGY

The Car Valuate system is developed by the proposed machine learning pipeline that accurately and efficiently predicts the price of a used car. The overall workflow comprises of data collection, data pre-processing, dataset preparation, model development, performance evaluation and deployment. The first step is to gather real-life data about used cars from trustworthy, online, automotive-related sources and publicly available datasets. The data gathered are significant vehicle characteristics such as model year, mileage, fuel type, transmission type, ownership information, and the selling price. The features are the main inputs to be used when creating predictive models.

The data is subjected to intensive data pre-processing prior to the model training for improving the quality and consistency of the data. The records with missing, inconsistent or invalid values are deleted and duplicate values are deduplicated to ensure data integrity. Categorical variables like fuel type and transmission type are mapped to numerical values by encoding the categories into numbers to be used by machine learning algorithms. In order to avoid unnecessary complexity and to increase the efficiency of the model, a number of feature selection methods are applied to select only the most important and predictive features. Moreover, the attribute selling price is log-transformed to minimize skewness and stabilize the variance for improved learning and prediction.

After pre-processing, the data is split into a training set and a test set. The training data is used to train the model and identify patterns and relationships between different vehicle attributes and selling prices, and the testing data is used to test how well the model can predict the selling prices of unseen vehicles. This separation will allow for an unbiased evaluation of the model performance and the ability to generalize this model.

For the development of models, two machine learning algorithms are developed and compared. Linear Regression is used as a base line regression method because of its simplicity and interpretability. In Linear Regression it is assumed that the target selling price is linearly related to the features of the input. The pricing model, however, has very limited predictive capacity due to the complex and non-linear interactions between various factors and the results of the pricing model are often quite different from the actual market price.

These are overcome by using a Regression Random Forest as the main Prediction Model. Random Forest is an ensemble learning method that multiple decision trees are used to improve the prediction accuracy and robustness. The model is also able to account for the non-linear interactions between the vehicle characteristics and selling prices, and avoids over

ISBN: 978-81-985770-0-9

fitting by means of random feature selection and data sampling. Based on the experimental observations, the predictions obtained by the Random Forest Regressor are much closer to the real car prices than the Linear Regression model.

Both models are assessed by graphically comparing the predicted prices with actual selling prices, and by standard regression evaluation methods. From the comparative results, it can be seen that the Random Forest Regressor model gained better accuracy, reliability and robustness in making prediction; hence it is the most preferred model for the proposed system than the Linear Regression model.

Once the best performing model is identified, the trained Random Forest Regressor is then serialized and stored, using the Pickle library, for use in future. The model is then deployed using Flask a backend framework that allows easy integration with a web application. A web-based user interface, which has been built with modern web technologies, enables users to input vehicle specifications and get instant market prices. The deployment strategy provides for accessibility, scalability and practical usability of the Car-Valuate system in the real world of used car valuations.

V. RESULTS AND DISCUSSIONS

The accuracy of the suggested car price prediction model is measured between the forecasted prices of the proposed models of Linear Regression and Random Forest Regressor and the real market prices. The analysis is based on ideal forecasting, the inclination of the models, and the strength of the models during the description of the actual car prices in the real world.

Result can be expressed as Fig 2, Fig 3, Fig 4.

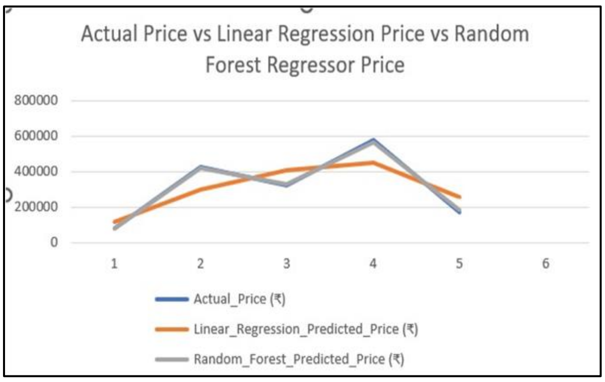

Fig 2. Comparison between the actual car prices**,** Linear Regression predicted prices**,** and Random Forest predicted prices

As can be seen on the figure, the predictions of the Random Forest Regressor are quite close to the actual price curve, but the predictions of the Linear Regression model have prominent deviations. Linear Regression curve does not reflect price movement, especially among more expensive vehicles, which means that the curve is under fitting as it includes the linear assumption.

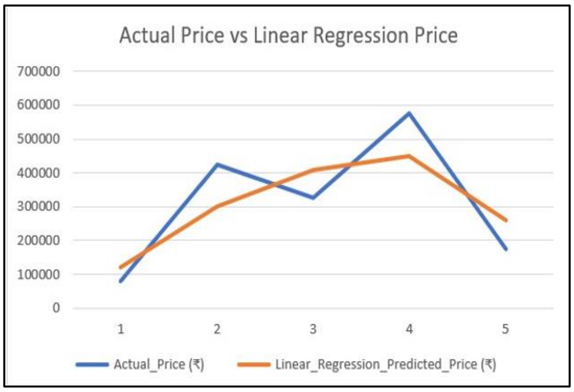

Fig 3. Comparison between actual price vs Linear regression predicted price

The graph is indicative of large deviations in real and forecasted values. The curve of predicted prices is not always trending in the direction of the real prices and has some underestimation and overestimation at several points on the data. This tendency proves that Linear Regression cannot be used to model the non-linear depreciation patterns and intricate interdependences among car attributes including the age, mileage, and fuel type and transmission amongst others.

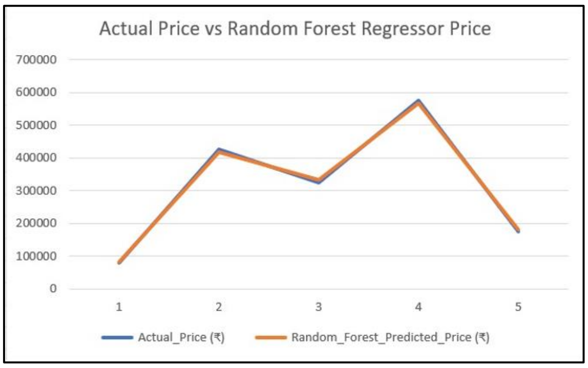

Fig 4. Comparison between actual price vs Random Forest Regressor Predicted price

The random forest model has a value that was predicted to be close to the actual prices in all the samples. The intersecting of the curves shows that the ensemble-based method works well in capturing non-linear correlations and diminishing prediction

ISBN: 978-81-985770-0-9

variance. This shows how strong and better generalization is of the Random Forest Regressor.

On the whole, the findings of the experiment prove that the Random Forest Regressor is much more accurate and reliable than the Linear Regression. Even though the existing system is promising in its output, there are more features that can be added to improve its performance, including brand recognition, accident history and geographical market demand. Further research on the topic of advanced ensemble methods and deep learning models to enhance prediction accuracy and scalability may also be considered in the future.

VI. CONCLUSION AND FUTURE SCOPE

The paper proposed a machine learning based car-to-X communication system value of a price prediction system to predict the fair value, historical data of used cars, to obtain acceptable values of vehicles. The subjects of Linear Regression and Random Forest Regressor are examined. Linear Regression and Random Forest Regressor are explored, models were comparatively analysed to assess their usefulness in forecasting the car prices in the real world. The experimental These data and graphical comparisons were found to be effective for Linear Regression, can be described as having severe forecasting errors due to on the assumption of the linear relationships, this results in under fitting and failing to get the same performance across various price ranges.

On the other hand, the Random Forest Regressor showed an improved predictive power with the ability to effectively represent non- correlations between car characteristics such as the vehicle's year, mileage and fuel, transmission and ownership record. In the real price series, the correlation between the two was higher and they were both statistically significant. the predicted prices as shown in the result graphs is an indicator that the ensemble learning method is sound and can be used to implement. to generalization. The ability to get the trained model to be able to be the same application, converted to become a web application using Flask. Indications that the proposed system would be applicable in, to be skilful in real-time car price estimation is a practice.

The future work will be based on enhancing the accuracy of predictions and add more features such as the brand reputation, accident history, records of service and trends in the local economy. Besides, it can be thought of that incorporated advanced ensemble methods and deep learning to improve performance of the performance. It will also be argued to train the models updating the data sets periodically so as to make them the capacity to be reliable and respond to new market conditions in the long-term.

REFERENCES

[1] N. R. Draper and H. Smith, *Applied Regression Analysis*, 3rd ed., Wiley, New York, 1998.

[2] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis*, 5th ed., Wiley, Hoboken, NJ, 2012.

[3] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., Springer, New York, 2009.

[4] L. Breiman, “Random Forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[5] J. H. Friedman, “Greedy Function Approximation: A Gradient Boosting Machine,” *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[6] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD)*, 2016, pp. 785–794.

[7] M. Kuhn and K. Johnson, *Applied Predictive Modeling*, Springer, New York, 2013.

[8] G. E. P. Box and D. R. Cox, “An Analysis of Transformations,” *Journal of the Royal Statistical Society, Series B*, vol. 26, no. 2, pp. 211–252, 1964.

[9] M. Grinberg, *Flask Web Development: Developing Web Applications with Python*, 2nd ed., O’Reilly Media, 2018.

[10] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[11] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning with Applications in R*, Springer, New York, 2013.

[12] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 2nd ed., O’Reilly Media, 2019.

[13] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, 2006.

[14] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*, MIT Press, Cambridge, MA, 2012.

[15] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, Cambridge, MA, 2016.

Cite this article as:

Sumit Tiwari, Priyanshu and et. al. " Car-Valuate : Car Valuation Made Easy Using Machine Learning ", Proceedings of 13th international conference on Microelectronics, Circuits and Systems, Micro2026

Displayed as online on 04th July 2026.

ISBN: 978-81-985770-0-9

Link: http://actsoft.org/science/act2026-pro/371-micro2026.pdf

ISBN: 978-81-985770-0-9