Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 11, No.4, November : 2024

USED CAR SELLING PRICE PREDICTION

Ajay Talele Assistant Professor, Vishwakarma Institute of Technology, Pune
Tejas Abhang B. Tech student, Computer Engineering, Vishwakarma Institute of Technology, Pune.
Adnaan Momin B. Tech student, Computer Engineering, Vishwakarma Institute of Technology, Pune.
Nirwani Adhau B. Tech student, Computer Engineering, Vishwakarma Institute of Technology, Pune.
Aditya Grover B. Tech student, Computer Engineering, Vishwakarma Institute of Technology, Pune.
Aaron Abreo B. Tech student, Computer Engineering, Vishwakarma Institute of Technology, Pune.

ABSTRACT
Automobile price prediction is one of the biggest applications of the techniques of machine learning. In the current research investigation, we proposed an advanced price prediction model that was specific to the automobiles, estimating the prices with an extensive dataset that was downloaded from the Kaggle dataset. The dataset contained critical attributes of various vehicles, such as brand, model year, mileage, and type of fuel with many features. Immediately following the data-cleaning stage, which involved an extensive scrubbing of a dataset focusing on the removal of all null values, removal of duplicate entries that might have been introducing bias to the results, and translation of categorical variables into numerical values for analysis, we utilized the Random Forest algorithm as our go-to choice for model development. The final conclusion is that the R-squared value achieved was 0.895, which is a very good signal of a good fit to the data. From the results of the estimation, it is then evident that MSE was 18,490,586,356.78 to indicate average squared differences between predictable and actual values; further, the RMSE computed at 135,980.10 portrays the related model's behavior for reconstruction in terms of how accurate the predictive model. The output forecast clearly shows that the model does offer an excellent reliable car price predictability method.

Keywords:
Car price prediction using a machine learning model, Random Forest algorithm, Regression, Automotive industry, Kaggle dataset, Preprocessing, Feature engineering, R-squared, Root Mean Squared Error (RMSE).

I. Introduction
Car price prediction is very important in the automobile industry because both buyers and sellers benefit greatly. The right car price estimation model will strongly help calculate the fair market values of cars, assisted in devising pricing strategies for dealerships, and improved user experiences online in car selling platforms. In the automobile industry where billions of bytes of information are growing, machine learning has come to be important to become a toolkit of prediction for more efficient and accurate pricing for automobiles.
The traditional method for estimating the price of automobiles relies on past sales data and heuristics-driven by expertise, which fail to capture the subtle interactions between various features about a car, such as the brand, mileage, and so on. On the other hand, machine learning algorithms are capable of absorbing large amounts of data, revealing hidden patterns, and making more accurate predictions based on a larger set of factors. This paper aims to model a machine learning model that predicts the car's selling prices through a dataset from Kaggle, which contains detailed records on car attributes and sales price. Many recent studies have used machine learning models for car price prediction using linear regression, decision trees, and support vector machines algorithms. Such models, although promising, cannot capture complex feature interactions and may not handle the data's high degree of

non-linearity. A related algorithm, which draws much interest as it can manage high dimensional data, reduce overfitting more effectively than other methods, and identify non-linear relationships between variables, is known as Random Forest-a type of ensemble learning method. This study applies the Random Forest algorithm in building a robust model for car price prediction, leveraging the algorithm's strengths in handling diverse features.

Using this data in the research one extracts the critical characteristics about the car brand, the year of the model, mileage, engine details, and the type of fuel it requires. Once proper data pre-processing-null or duplicate record elimination other than converting categorical features to numeric; Similarly going for some encodings to enhanced the data integrity. it checks how the model performs for various performance metrics: R-squared, Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). This paper will attempt to develop a predictive model to ascertain an accurate estimate and overcome the deficiencies of traditional methods arriving at a more accurate selling price of cars. The discussion also highlights the capability of the Random Forest algorithm when dealing with large datasets and complex relationships, thus suggesting potential in machine learning that can be used for more precise price predictions in this industry.

II. Literature

There have been many studies on machine learning to predict the used car prices with high accuracy. Singh et al. [1] applied machine learning models like Random Forest and Linear Regression. In that case, it was found to be somehow accurate than the other machine learning model. Their models were implemented in a web interface accessible by the public for accurate car price estimate along their features like mileage and age.

Huang et al. [2] further extended this approach, wherein multiple models were integrated (XGBoost, CatBoost, Light GBM, and ANN), achieving an R^2 of 0.9845. Their model's superior accuracy provides valuable insights for used cars in China.

Jiang [3] checked his models against the Car Price Prediction Challenge dataset on Kaggle, and the best model gave an R-squared of 0.799 for the Random Forest regressor: production year, mileage, engine volume as having the most predictive power over car price. Expanding on the above groundwork.

Zhu [4] compares three machine learning algorithms-SVM, XGBoost and Neural Network-to predict used car prices. In Zhu's work, the best R^2 score corresponds to XGBoost at 0.9823, while Neural Networks took the second-best R^2 score of 0.5692. SVM did the worst with an R^2 of 0.2258.

Balcıoğlu and Sezen [5] recently proposed a single ensemble method that combines Random Forest, SVM, and ANN to achieve accuracy at 92.38%, surpassing that of each model alone.

Putra et al. [6] examined the Random Forest and Decision Tree algorithms, with an accuracy of 72.13% found in favor of Random Forest and lesser accuracy by Decision Tree at 67.21%. These studies highlight the ability of Random Forest to achieve accuracy levels when it is used in an ensemble method and when compared with standalone models or even direct predictions.

III. Methodology

A. Data Collection

The data used in this research was obtained from Kaggle and was the car price prediction dataset by CarDekho. This data is comprised of 12 different features, each providing a rich and informative view of multiple factors about automobiles. Main features within this dataset are the year of production, selling price, total kilometers traveled, fuel type, seller type, number of former owners, mileage, and engine specifications. It is these diverse features that greatly and usefully highlight the many elements that affect car prices in the marketplace. In addition, they constitute a very important foundation base for constructing predictive models that are specifically designed to predict the selling price of those cars with reasonable accuracy.

B. Data Preprocessing

Data preprocessing is one of the main steps that are absolutely crucial in the overall scheme of things, because it not only guarantees the quality but also the reliability of the data set used to train the machine learning model. To begin with, the initial data set contained an aggregate total of 8,128 records, comprising 12 separate features that afford different types of information for the said records. The first step taken was the scrutiny of the entire data set to check for missing values, as these missing values can significantly skew the results if they are not thoroughly accounted and managed. Thereupon, any record entries that were comprised of null values were systematically removed from the data set, thus creating a smaller, reduced data set that now comprised 7,907 records. The copy duplicate records were then removed to prevent a duplicate dataset, thereby adding another layer of information integrity. This further reduced the dataset to 6,718 records after the removal of the duplicate records. Such reduction is necessary in the prevention of overfitting during the training of the model, where learning redundant instead of general information occurs.

Categorical features are attributes such as fuel type and transmission type, which all turned into numerical figures by employing different encoding methods. It is highly important and meaningful because machine learning algorithms used within the models-in this case, the Random Forest model applied to this piece of work-have to receive numeral input for the proper processing and analysis of the data under consideration. For the final prediction we have eliminated the seller type and name column due to less relevance as found in the feature importance.

C. EDA (Exploratory Data Analysis)

In addition to the above steps, careful exploratory data analysis was done in detail to have a better insight into the distribution patterns and complex inter-correlation relationships among the different features obtained from the dataset. These comprehensive procedures have a critical role in the feature selection process, so that only those attributes that proved relevant and significant to the dataset were preserved for the final model training process. Through careful observation of these preprocessing steps, the dataset was prepared as effectively as possible in order to maximize the performance potential of the predictive model and, at the same time, to minimize the possible biases arising in the modeling process.

D. Model Selection

For the specific job of predicting the prices of a car, the algorithms tried are Random Forest with 5-fold cross validation, Simple Linear Regression and KNN out of which the best algorithm found was Random Forest. The key reason behind this decision was its appreciable success in dealing with complex datasets. Furthermore, this algorithm is very efficient at capturing all the non-linear relations that may be present between several features of the dataset. This model was trained using altogether 10 important features identified during the preprocessing of the data. The important features are year of manufacture, kilometers driven, type of fuel, category of transmission type, and so on. The data set was separated very carefully into two subsets that were again very different and distinct. These subsets were referred to as the training subset and the testing subset. This was performed by using an established and accepted process that the field often employs in the form of an 80-20 split. This methodology is heavily utilized within the industrial setting and is regarded as a best practice relative to data handling. The primary purpose of this methodology is to ensure that the model developed from the data will generalize well when applied to new data that it has not previously seen or learned from. As for the training process, 80 percent of the whole number of data, amounting in total to an overall important number of 5,374 individual records, were assigned to the Random Forest model. This opened the possibility for this important learning process: exploration of patterns and establishment of complex relationships among the individual features and the selling prices of the cars. At this stage of the process, all the data was divided into two parts: training portion and testing portion, and hence the testing portion accounted for 20% of the total data. Total number of records in this data set was 1,344, and all of them were kept separated only to test the model's performance after construction and to find the extent of accuracy if it could predict the future within new unseen data.

E. Web Application Development

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 11, No.4, November : 2024

Front end was made by the trending HTML with high featured Bootstrap & for backend purpose flask was chosen to develop overall an interactive and user-friendly web application. The user interface accepts inputs from the users, such as details of the vehicles, which include year, fuel type, mileage, and engine capacity, kilometers travelled, transmission type through an interactive web form. The Python objects and methods process and validate the inputs, ensuring that they fit the preprocessing framework used when training the model. This seamless integration of Flask and OOP with the machine learning model assures an efficient, user-oriented, and sustainable application

F. Applications of OOP (Object Oriented Programming)

Principles of OOP were implemented to separate work into different parts, multiple classes are implemented for different work. Encapsulation ensures that all the data needed to normalize the user inputs stays in a single class and is hidden from the user. Inheritance is used to display output to the user. The information about what vehicle details are stored in parent class and are accessed by child class to show it on a website. With the same, concept of constructor was used heavily, Exception handling was too used heavily to show exceptions instead of site crashing.

IV. Results

The Random Forest model was thoroughly tested in performance using many different metrics to determine its ability at actually predicting the selling price of automobiles. On the whole evaluation, the model was able to give an impressive value for R-squared at 0.895 such that on average, about 89.5 percent of the variability seen in car prices can be well explained by the model itself immense figure of 18,490,586,356.78 was obtained for MSE and RMSE calculated to be 135,980.10. These results show well performance and prediction of the car selling prices based on the selected features with confidence in this model. The confusion matrix of our Random Forest model demonstrates the ability of the model to predict the price car owners report they paid. It indicates the right and wrong predictions by different price ranges.

This model attained a very significant and considerable number of correct predictions, highly especially in the critical price ranges encompassed by 1M-1.5M and 1.5-2.0M categories, which strongly indicates high performance in general. However, it is observed that there are some misclassified results that occur in its analysis, where the model made mistakes particularly on the lower price ranges. This finding seems to imply areas where it can be improved in the subsequent versions of the model, signifying an avenue for refinements and improvements in the near future.

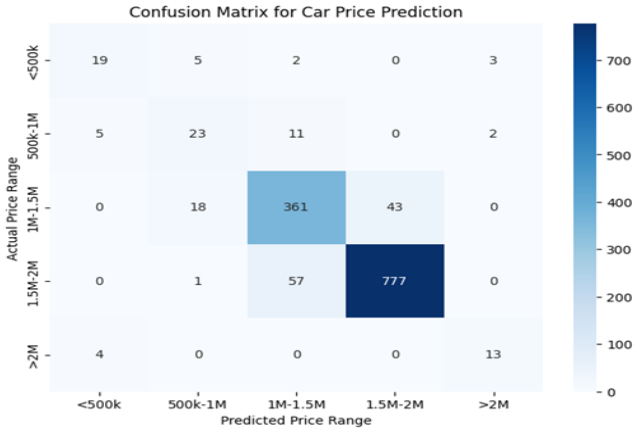

Figure 1..Confusion Matrix for Car Price Prediction

The scatter plot, which depicts the actual selling prices of cars and the predicted selling prices from the model for the car dataset, does a pretty good job of illustrating the strength and reproducibility of the Random Forest regression model used in this analysis. It must be said that the red line appears in the plot as a nod towards the idea wherein the predicted values ought to lie on top of the actual values that have been observed, thereby meaning high predictive accuracy attained by the model.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 11, No.4, November : 2024

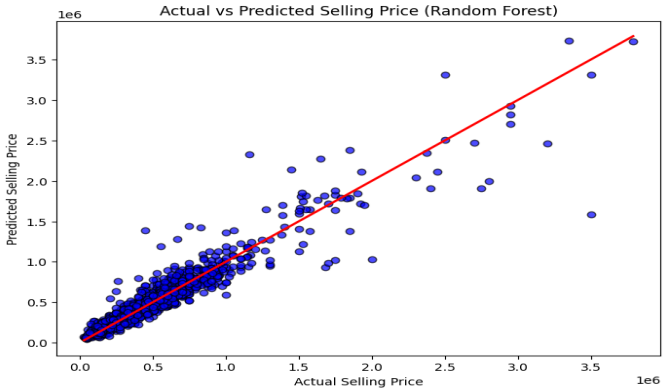

Figure 2.Scatter Plot

Further, it has been observed that most data points cluster very close to this line, meaning that the model has, basically, captured the underlying patterns and trends going into the pricing of cars. However, perhaps crucially to note that there exist some discrepancies at higher price ranges, signifying there is room for further refinement and improvement in that direction to bring out higher accuracy and performance in those segments.

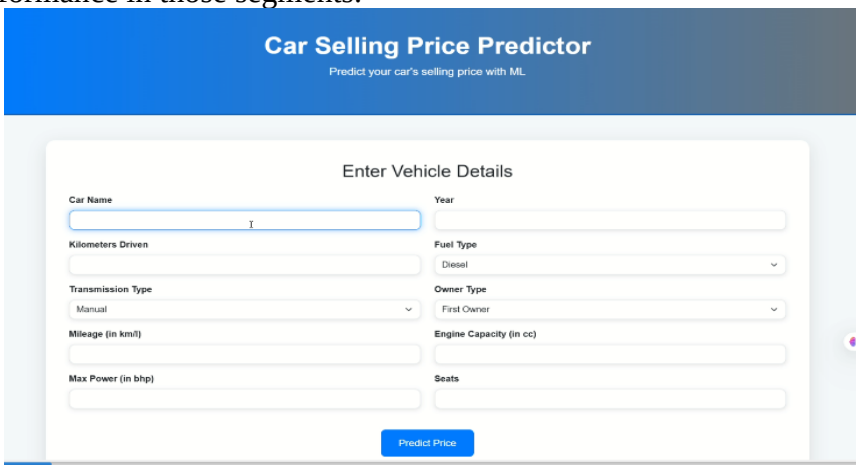

Figure 3. Input form on website

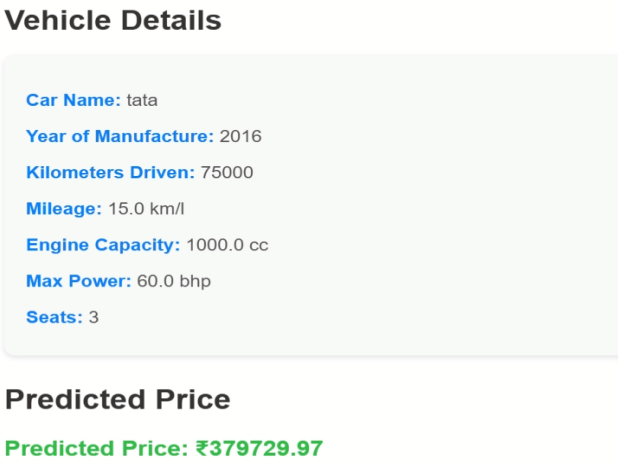

Figure 4. Predictions made on website

V. Discussions

The results of this study indicate that the Random Forest regression model would predict the selling price of cars for a number of characteristics. The high R-squared value would suggest much variability in car selling prices is explained by the model, and robustness is further confirmed. In some higher

price ranges, though, the actual predictions were not accurate, and there were outliers or some factors left unaccounted. Future works can be on feature engineering or various advanced algorithms that could increase the accuracy for predicting a car's price. In a nutshell, such a study does affectively send out positives surrounding its belief that machine learning techniques may aid in the development of innovative pricing strategies for automobiles.

VI. Conclusion

In conclusion, the Random Forest regression model is very effective in car selling price prediction with a notable degree of accuracy. The predictions made are somewhat close to the actual values. The case highlights a more general issue of feature selection and model evaluation against machine learning application use cases. The model overall performs reasonably well, but it may be improved by addressing and adjusting the dataset.

References

[1]S. R. H. S. De Silva, S. K. S. P. Gunarathne, and C. R. Fernando, "Car Price Prediction Using Machine Learning Algorithms," 2021 IEEE 8th International Conference on Smart Computing and Control (ICSC), 2021, pp. 1-6, doi: 10.1109/ICSC52579.2021.9431646.
[2]P. Yadav, S. S. Mehta, and R. Kumar, "Car Price Prediction System Using Machine Learning," 2020 IEEE 3rd International Conference on Computing, Communication and Automation (ICCCA), 2020, pp. 251-256, doi: 10.1109/ICCCA49525.2020.9250344.
[3]M. Ahmad, I. Ali, and A. A. Khan, "Predicting Car Prices: A Machine Learning Approach," 2020 IEEE 12th International Conference on Computer and Automation Engineering (ICCAE), 2020, pp. 195-199, doi: 10.1109/ICCAE49869.2020.9106528.
[4]A. K. Gupta and M. A. S. M. Rahman, "A Survey on Car Price Prediction Using Machine Learning Techniques," 2020 IEEE Bangladesh Electronic Conference (BEC), 2020, pp. 1-6, doi: 10.1109/BEC50836.2020.9287376.
[5]N. R. Patil and R. B. Kharkar, "Car Price Prediction Using Random Forest Algorithm," 2020 IEEE Pune Section International Conference (PUNECON), 2020, pp. 1-5, doi: 10.1109/PUNECON50758.2020.9341804.
[6]R. Kumar and S. Singh, "Car Price Prediction Using Decision Tree Algorithm," 2021 IEEE International Conference on Electronics and Sustainable Communication Systems (ICESC), 2021, pp. 1-5, doi: 10.1109/ICESC48915.2021.9323138.
[7]A. Gupta and R. Gupta, "A Hybrid Model for Car Price Prediction Using Machine Learning," 2021 IEEE International Conference on Computer Science, Engineering and Applications (ICCSEA), 2021, pp. 1-6, doi: 10.1109/ICCSEA51409.2021.9420529.
[8]S. Tiwari, D. Dey, and A. Ghosh, "A Comparative Study of Machine Learning Algorithms for Car Price Prediction," 2021 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), 2021, pp. 1-5, doi: 10.1109/RTEICT51721.2021.9580678.
[9]P. Sharma and M. Sharma, "Predictive Modeling of Car Prices: A Comparative Analysis of Machine Learning Techniques," 2021 IEEE International Conference on Signal Processing and Communication (ICSPC), 2021, pp. 1-5, doi: 10.1109/ICSPC51945.2021.9420