International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 05 | May - 2022 Impact Factor: 7.185 ISSN: 2582-3930

Car Price Prediction Using Machine Learning approach

Dr. Ajay B. Gadicha[1], Atharva Jahagirdar[2], Hrushikesh Pawar[3] , Siddhant Sontakke[4] , Ram Chaudhari[5], Lalit Thokal[6]

[1]Professor, Computer Science and Engineering, P. R.Pote College of Engg. and Mang. , Amravati
[2]Student, Computer Science and Engineering, P. R.Pote College of Engg. and Mang. , Amravati
[3]Student, Computer Science and Engineering, P. R.Pote College of Engg. and Mang. , Amravati
[4]Student, Computer Science and Engineering, P. R.Pote College of Engg. and Mang. , Amravati
[5]Student, Computer Science and Engineering, P. R.Pote College of Engg. and Mang. , Amravati
[6]Student, Computer Science and Engineering, P. R.Pote College of Engg. and Mang. , Amravati

---***---

Abstract - In this swiftly-moving world, managing our skills additionally as personal lives have become quite agitated and if we have a tendency to not have our own personal vehicle for transportation, life may be a heap of additional agitation. To get on the safe facet, one ought to have a additional reliable and straightforward mode for transportation and a private vehicle is usually the simplest
option. Having an automobile is extremely necessary for individuals recently because it offers an explicit social status and additionally offers an explicit extent of private management to individuals owning it. In some areas with low population, having an automobile becomes essential because it provides the only possibility for covering long distances just in case of an associated absence of transport. Old aged individuals, United Nations agencies have difficulties in walking or athletics to places, have driving the sole possibility for moving while not being dependent. And for those who don't have enough resources to buy a new automobile, shopping for an associated previous vehicle becomes a necessity which too at an affordable price. The automobile production has been increasing fleetly over the years throughout the past decade, with concerning ninety two million cars that were factory-made in 2019. This provides an enormous boost for the market of previous and used cars that is currently bobbing up as a more and more growing business. The recent entries of various websites and web-portals have consummated the wants of consumers up to some extent as they currently understand these trends and state of affairs to urge the value of unspecified vehicle gifts within the market. Machine Learning contains a heap of applications in the real world state of affairs however one amongst the foremost better-known applications is the use of Machine Learning in resolution of prediction issues. The project being mentioned here is very much based mostly upon one of such applications. Using varied Machine Learning Algorithms, we'll try and build an applied math model primarily based upon given information and choices set to estimate the prices of used cars.

***Key Words*:** Cars, Price, Analysis, Prediction, Features, Python, Algorithm, Regression.

1. INTRODUCTION

When we search via automobile websites for the purpose of purchasing or selling a used car, the price we receive is not precise enough. The purchase price is sometimes too high, and the selling price is sometimes too low. This causes us to be perplexed. whether to purchase or sell the vehicle at that price The used car sector operates with the goal of profiting from customers. and buyers. It covers their commission as well as any additional income they earn from customers. Whether or not a used car is worth the money It can be difficult to determine the advertised price when viewing ads online. KM driven, fuel type, number of owners, and year are all aspects to consider. etc. can have an impact on a car's true value. It might be difficult to price a secondhand car correctly from the standpoint of a seller.

The goal is to construct models for predicting used automobile prices using machine learning algorithms based on existing data. There are thousands of used cars on the globe of the automobile industry. We can create realistic used car values using their data on our own platform. Both buyers and sellers benefit from this arrangement. According to our research into the Indian automobile market, 65 percent of the industry is made up of secondhand cars. India's used car industry was worth 24.24 billion dollars in 2019. The value is also predicted to expand at a 15 percent compound annual growth rate.

2. TECHNOLOGIES USED

Python is mostly utilised in this project to implement machine learning techniques because it has a lot of built-in methods in the form of packaged libraries and modules. The following libraries were utilized during the project's implementation:

Pandas: Pandas is a popular Python library for data scientists. It offers a variety of structures and data analysis tools, all of which are simple to use and deliver excellent performance.

NumPy is a Python open-source package that performs fast mathematical calculations on matrices and arrays. 'Numeric Python' or 'Numerical Python' is what NumPy stands for.

NumPy provides a comprehensive Python Machine Learning Ecosystem when combined with other Machine Learning Modules such as Scikit-learn, Pandas, Matplotlib, and others.

Matplotlib: Matplotlib is mostly used to plot bars, pies, lines, scatter plots, and other data visualization elements. It is a graphics tool for data visualization in Python that is nicely integrated with libraries such as NumPy and Pandas. The pyplot package closely resembles the MATLAB charting commands.

Scikit-learn: The Scikit-learn module in Python provides a homogeneous interface for a number of supervised and unsupervised learning techniques. SciPy, or Scientific Python, must first be installed before using the scikit-learn package, as SciPy is the foundation upon which Scikit-learn is based. The goal of this library is to provide a level of reliability and support for use in production systems.

Pickle: The pickle module is used to serialize and deserialize a Python object structure using binary protocols that it implements. 'Pickling' is the process of converting a Python object hierarchy into a byte stream, whereas 'Unpickling' is the reverse of cheval cheval. Serialization, marshaling, and flattening are other terms for pickling.

3. IMPLEMENTATION

1. **Data Preprocessing:** "Any model utilizing any algorithm Data Preprocessing is the most important phase and will be the primary step before training." "10" is one of the checkpoints (steps) in the data preprocessing.

1. Import Libraries: I utilized Pandas for data manipulation and analysis, Numpy for numerical analysis, Matplotlib, and Seaborn for better data visualization and graphical statistics.

2. Import the Dataset: This obtained the dataset from Kaggle and then used the pandas library to import it.

3. Step 3: Addressing Missing Data in the Dataset: After evaluating this dataset, I discovered no missing values.

4. Step 4: Encoding categorical data: This dataset contains categorical values such as fuel type, owner type, and seller type, so we need to encode these categorical values into an encoded format to better train our model. To do this, I used the pandas get Dummies() method, which converted the entire dataset's categorical values into binary values.

5. Splitting the Dataset into the Training and Testing Sets: To train our machine learning model, I utilized the capable machine learning library of python, scikit-learn or sklearn to partition this dataset into Test and Train datasets. Supervised Learning, or using its model selection method to construct testing data by selecting random values from the given dataset for model prediction.

6. Feature Scaling: Because all of the data is in a standard format, I don't utilise any feature scaling approaches here.

2. **Data Training and Modeling**: The dependent and independent variables are required to train and construct a model. To find these variables, I first used to find the correlation between the output variables, and then I separated my variables into two axes, which we call x and y, with the x-axis containing all the independent variables and the y-axis containing the dependent variable, which in our model is the Used Cars selling price. This dataset is further dispersed in the train-test dataset using RandomizedSearch and the sklearn.model selection library's train test split function. CV The optimal hyperparameters for our model prediction are found by modifying this model.

3. **Algorithm**

The Random Forest technique is an Ensemble-Bagging approach that works by generating several decision trees during the training phase. The random forest chooses the ultimate option based on the majority of the outputs (trees). Random Forest has the advantage of combining both forms of supervised learning problems, namely regression and classification. • For Remote Sensing, such as ETM devices used to acquire images of the earth's surface, Random Forest is the first choice because it provides higher accuracy in a shorter training time. • For Multiclass Object Detection, Random Forest is used because it provides better detection in complex environments.This method is used by several gaming consoles to track body movement and replicate it in the game. The Random Forest algorithm is taught to recognise body parts and learns from it. It then recognises the users' body parts, such as their hands, feet, face, eyes, and nose. "A random forest is made up of a large number of individual decision trees that work together as an ensemble, with each tree producing a category prediction, and the category with the most votes becoming our model's forecast." Another advantage of the Random Forest Algorithm is that it can measure the relative relevance of each feature on the forecast. Another advantage of the random forest algorithm is that it is simple to implement.

4. **Cross-validation and model prediction:** "A statistical analysis technique for testing how effectively the results of a statistical analysis generalize to an independent data set is cross-validation." Cross-validation is mostly used to verify the correctness of the forecast and the model's performance." The following result is achieved after doing cross-validation and analyzing all other metrics of model performance and visualizations. 1. The Heat Map in Figure 1 depicts the relationship between all of the parameters. The dark blue color indicates a positive correlation between the qualities on the x and y axes, whereas the white color indicates substantially negatively linked variables. We can deduce from this map that the "Selling Price" is ``Price" is positively correlated, and it can be a key factor in predicting the current selling price of used cars. This map also shows a negative relationship between "Present Price" and "Fuel Type" and "Seller Type" and "Present Price."

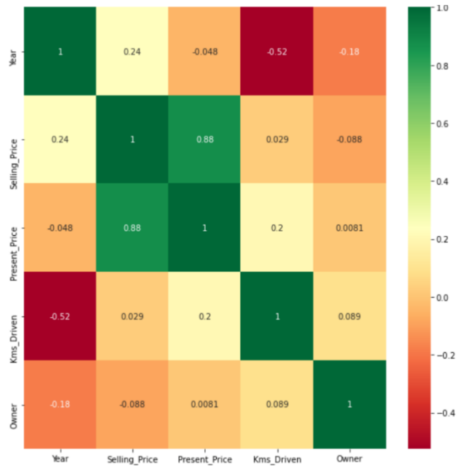

Fig -1: Heatmap

2. The distplot in figure 2 below displays the model's normal distribution with the test dataset, demonstrating its accuracy. As a result, we can conclude that this model's forecast is extremely accurate.

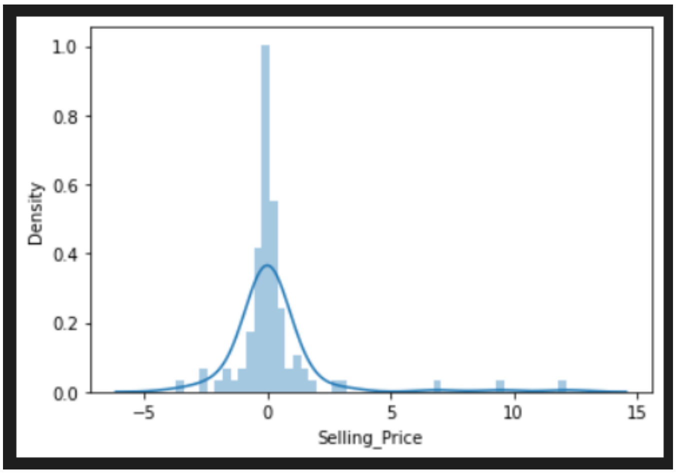

Fig -2: Distribution plot

3. The scatterplot in Figure 3 reveals a linear distribution, indicating that this model is accurate, therefore we can conclude that the selling price forecast using the provided dataset is accurate.

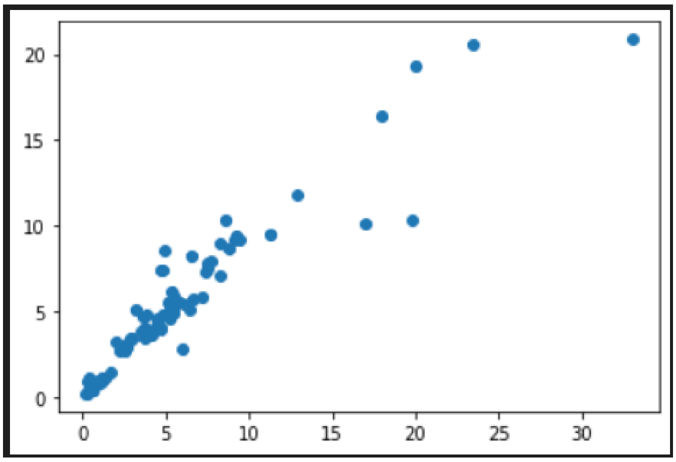

Fig -3: Scatter Plot

4. Finally, we can deploy this model as a web application using the Heroku platform. Figure 4 shows how I deployed this application on the same platform using Heroku.

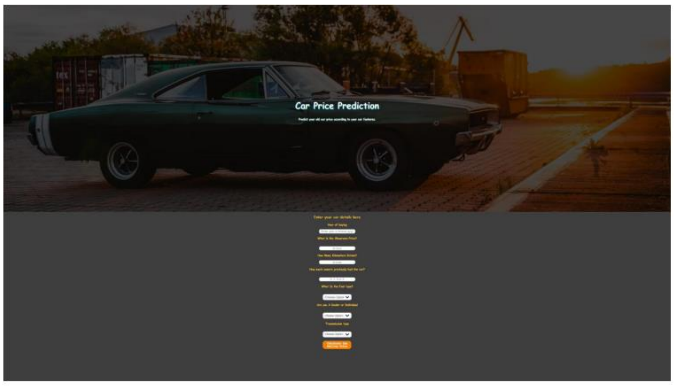

Fig -4: Web app

4. CONCLUSION

This model is built on machine learning algorithms, and we were attempting to forecast the selling price of used cars using the Kaggle dataset. We utilized two machine learning techniques to predict this dataset: Random Forest and Extra Tree Regressor. This model's prediction is then compared to a test dataset constructed by randomly selecting values from the original dataset, and the prediction is then evaluated using several approaches. Following a thorough examination of the prediction model, we can conclude that it is extremely accurate, and Random Forest and Extra Tree Regression are two of the best methods for regression problems. In terms of prediction, these two algorithms are extremely accurate and quick.

REFERENCES

[1] SAS Academy website - "https://www.sas.com/en_in/insights/analytics/machine-learning.html"

[2] Dr M. J. Garbade – "Clearing the Confusion: AI vs Machine Learning vs Deep Learning Differences" Available: https://towardsdatascience.com/clearing-the-confusion-ai-vs-machine-learning-vs-deep-learningdifferences-fce69b21d5eb

[3] A. Wilson – "A Brief Introduction to Supervised Learning" Available: https://towardsdatascience.com/a-brief-introduction-to-supervised-learning-54a3e3932590

[4] A. Wilson – "A Brief Introduction to Supervised Learning" Available: https://towardsdatascience.com/a-brief-introduction-to-supervised-learning-54a3e3932590

[5] J. Rocca – "Ensemble methods: bagging, boosting and stacking" Available: https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205

[6] J. Brownlee – "Bagging and Random Forest Ensemble Algorithms for Machine Learning" Available: https://machinelearningmastery.com/bagging-and-random-forest-ensemble-algorithms-for-machinelearning/

[7] T. Yiu –" Understanding Random Forest How the Algorithm Works and Why It Is So Effective" Available: https://towardsdatascience.com/understanding-random-forest-58381e0602d2

[8] N. Donges –" A COMPLETE GUIDE TO THE RANDOM FOREST ALGORITHM"Available: https://builtin.com/data-science/random-forest-algorithm

[9] N. Donges –" A COMPLETE GUIDE TO THE RANDOM FOREST ALGORITHM" Available: https://builtin.com/data-science/random-forest-algorithm

[10] A. Dey – "Data Pre-processing for Machine Learning" Available: https://medium.com/datadriveninvestor/data-preprocessing-for-machine-learning-188e9eef1d2c

[11] "Cross-Validation" – Available: https://www.techopedia.com/definition/32064/cross-validation

[12] Tomar, Ravi, Hanumat G. Sastry, and Manish Prateek. 2020. "A Novel Protocol for Information Dissemination in Vehicular Networks." Pp. 1–14 in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Vol. 11894 LNCS. Springer, Cham.

[13] Bansal, Parnika, Bhawna Aggarwal, and Ravi Tomar. 2019. "Low-Voltage Multi-Input High TransConductance Amplifier Using Flipped Voltage Follower and Its Application in High Pass Filter." Pp. 525–29 in 2019 International Conference on Automation, Computational and Technology Management, ICACTM 2019. IEEE.

[14] Tomar, Ravi, Rahul Tiwari, and Sarishma. 2019. "Information Delivery System for Early Forest Fire Detection Using Internet of Things." Pp. 477–86 in Communications in Computer and Information Science. Vol. 1045. Springer, Singapore.

[15] Tomar, Ravi, Manish Prateek, and Hanumat G. Sastry. 2017. "A Novel Approach to Multicast in VANET Using MQTT." Pp. 231–35 in Ada User Journal. Vol. 38. Ada-Europe.

[16] Tomar, Ravi, Hanumat Sastry, and Manish Prateek. 2020. "Establishing Parameters for Comparative Analysis of V2V Communication in VANET." Journal of Scientific and Industrial Research (JSIR) 79(01):26– 29.

[17] Tomar, Ravi and Sarishma. 2019. "Maintaining Trust in VANETs Using Blockchain." Ada User Journal 40(4):236–41.

[18] Kumar, Shiwanshu and Ravi Tomar. 2018. "The Role of Artificial Intelligence In Space Exploration." Pp. 499–503 in 2018 International Conference on Communication, Computing and Internet of Things (IC3IoT). IEEE.

[19] Sharma, S., Aggarwal, A., & Choudhury, T. (2018). Breast Cancer Detection Using Machine Learning Algorithms. 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), 114–118.

[20] Rohit, Sabitha, S., & Choudhury, T. (2018). Proposed approach for book recommendation based on user kNN. In

Advances in Intelligent Systems and Computing (Vol. 554). https://doi.org/10.1007/978-981-10-3773- 3_53

[21] Mehta, I. S., Chakraborty, A., Choudhury, T., & Sharma, M. (2018). Efficient approach towards bitcoin security algorithm. 2017 International Conference on Infocom Technologies and Unmanned Systems: Trends and Future Directions, ICTUS 2017, 2018-Janua. https://doi.org/10.1109/ICTUS.2017.8286117

[22] Chhabra, A. S., Choudhury, T., Srivastava, A. V., & Aggarwal, A. (2018). Prediction for big data and IoT in 2017. 2017 International Conference on Infocom Technologies and Unmanned Systems: Trends and Future Directions, ICTUS 2017, 2018-Janua. https://doi.org/10.1109/ICTUS.2017.8286001

[23] Kashyap, N., Choudhury, T., Chaudhary, D. K., & Lal, R. (2016). Mood based classification of music by analyzing lyrical data using text mining. Proceedings - 2016 International Conference on Micro-Electronics and Telecommunication Engineering, ICMETE 2016. https://doi.org/10.1109/ICMETE.2016.65

[24] Choudhury, T., Kaur, A., & Verma, U. S. (2017). Agricultural aid to seed cultivation: An Agribot. Proceeding - IEEE International Conference on Computing, Communication and Automation, ICCCA 2016. https://doi.org/10.1109/CCAA.2016.7813860

[25] Khunger, M., Choudhury, T., Satapathy, S. C., & Ting, K.-C. (2019). Automated detection of glaucoma using image processing techniques. In Advances in Intelligent Systems and Computing (Vol. 814). https://doi.org/10.1007/978-981-13-1501-5_28