"Car's Selling Price Prediction using Random Forest Machine Learning Algorithm."

Abhishek Pandey[1], Vanshika Rastogi[2], Sanika Singh[3]

[1] Student, BTech (IT)KCC Institute of Technology and Management, Greater Noida, E-mail: abhishekpandeyite@gmail.com

[2]Assistant Professor, ABES Engineering College, Ghaziabad, E-Mail: rastogi.vanshika21@gmail.com

[3]Assistant Professor, ABES Engineering College, Ghaziabad, E-Mail:sanikasingh39@gmail.com

Abstract:

India has one of the biggest automobile markets all over the globe every day many buyers usually sell their cars after using for the time to another buyer, we call them as 2nd /3rd owner etc. Many platforms such as cars24.com, cardekho.com and OLX.com provides these buyers with a platform where they can sell their used cars, but what should be the price of the car, this is the toughest question ever. Machine Learning algorithms can bring a solution to this problem. Using a history of previously used cars selling data and using machine learning techniques such as Supervised Learning can predict a fair price of the car, here I also used machine learning algorithms such as Random Forest and Extra Tree Regression along with powerful python library Scikit-Learn to predict the selling price of the used car. The result has shown that these both algorithms are highly accurate in prediction even the dataset is large or small, irrespective of the size of the dataset they give a precise result.

Keywords:

Machine Learning, Supervised Learning, Random Forest, Extra Tress Regression, RandomSearchCV, Algorithm, Kaggle, Car dataset, cardekho.

Introduction:

1. **Machine Learning:**

Machine Learning is a science of training computers to act without giving any command to it. Machine Learning is a subset of the most popular term of the 21st Century i.e. Artificial Intelligence. In Artificial Intelligence we make computers artificially intelligence to do work on their own. These systems are highly accurate and fast in doing their task. While in Machine learning we create and then train a model using different machine Learning Techniques such as Supervised Learning, Un-supervised Learning and Reinforcement Learning.

The pattern recognition can be considered as the origin of Machine Learning. Machine learning works on different regression and classification algorithms to train the models so that they can learn on their own. In the present scenario the term Data is not just a Data, we call it Big Data. "While many machine learning algorithms have been around for a long time, the ability to automatically apply complex mathematical calculations to big data – over and over, faster and faster – is a recent development." [1] Therefore the iterative aspect of machine learning is getting more important, as the models are exposed to new data which they are adapting by their own ability to learn from previous computations which are also responsible to produce reliable, repeatable decisions and results. The three machine learning techniques can be described as:

1. **Supervised Learning:** Models are presented with both input data and the desired results. Then the model will attempt to learn rules that map the input data to the desired results.
2. **Unsupervised Learning:** Models are presented with datasets that have no labels/ predefined patterns, and the model will attempt to infer the underlying structures from the dataset.
3. **Reinforcement learning:** The model or agent will interact with a dynamic world to achieve a certain goal. The dynamic world will reward or punish the agent supported its actions.. Over time, the agent will learn to navigate the dynamic world and accomplish its goal(s) based on the rewards and punishments that it has received.

Majority of the peoples are getting confused between Artificial Intelligence, Machine Learning and Deep Learning, but the fact is they all are different domain but works together to make a self-aware model which can learn by their own, and can complete tasks without any master to command them. Here is the short difference between these three domains.

"Whenever a machine completes tasks supported a group of stipulated rules that solve problems (algorithms), such an "intelligent" behaviour is we call it as **"Artificial Intelligence"**. Whereas ML is a subset of AI. It enables the machines to learn by themselves using the provided data and make accurate predictions. When we talk about deep learning, then it is a subset of Machine Learning. DL algorithms are roughly inspired by the information processing patterns found in the human brain known as Neural

Networks. DL is a collection of Neural Networks. Just like we use our brains to spot patterns and classify various sorts of information, deep learning algorithms are often taught to accomplish an equivalent tasks for machines. Examples such as RNN (Recurrent Neural Network), CNN (Convolutional Neural Networks) etc."[2]

2. **Supervised Learning:**

Supervised Learning is a process of algorithm learning from the training dataset. It maps an input to an output supported example input-output pairs. The test data is already provided for the model thus the model predicts the result based on the randomly picked test data values available within the original dataset. A supervised learning algorithm are often written simply as:

$$Y=f(x)$$

Where Y = The predicted output
f(x) = A mapping function that assigns a class to an input value *x*. This determines the predicted output Y. This function is used to connect input features to a predicted output Y and it is created by the machine learning model during training.[3]

The Supervised learning technique is one of the most famous techniques and can be used to solve two major types of real-world problems:

a. **Classification Problem:** This type of problems categorize all those variables which form the output. Examples include demographic data such as marital status, sex, or age. A most common model used to solve this type of problems is Support Vector Machine (SVM). There are numerous amount of algorithms available to solve classification problems, depends on their requirement. Some of the algorithms are ***Linear Classifiers, Support Vector Machines, Decision Trees, K-Nearest Neighbour, Random Forest.***

b. **Regression Problems:** This can be classified as problem types where the output variables are set as a real number. Often it follows a linear format. The equation for basic rectilinear regression are often written as:

$$\hat{y}= w[0] * x[0] +w[1]*x[1]+\ldots+w[i]*x[i]+b$$

Where $x[i]$ = Features for the data
$w[i]$ and b = Parameters, developed during training.
For models using two features, the plane will be used. Finally, for a model using more than two features, a hyperplane will be used. Some of the Regression Algorithms are: ***Linear Regression, Logistic Regression, Polynomial Regression, ExtraTree Regression, Random Forest.***[4]

3. **Ensemble and Bagging:**

In simplest words, The term Ensemble refers to the word "Assemble" which means to join multiple different parts and make it one working model/object. "Ensemble technique also works within the same way where multiple models (often called "weak learners") are trained to unravel an equivalent problem and combined to urge better results, the main hypothesis is that when weak models are correctly combined we will obtain more accurate and/or robust models."[5]

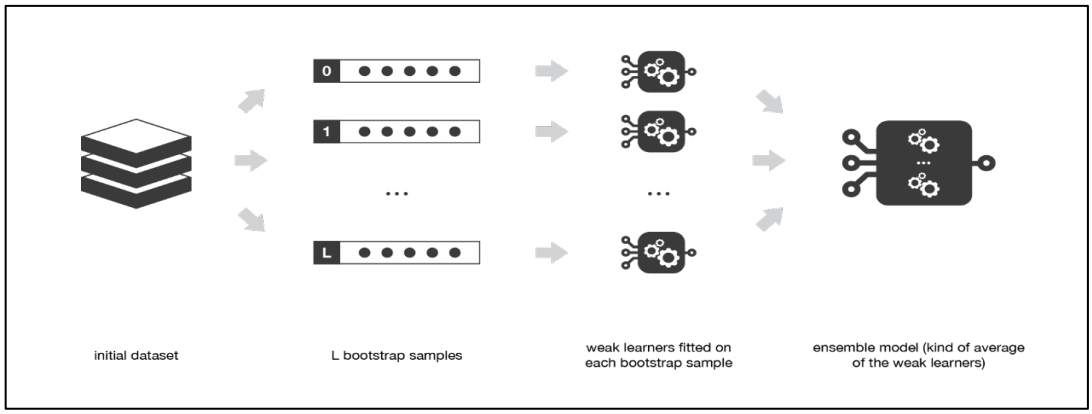

Figure 1: Bagging (Bootstrapping Aggregation) Technique

Whereas Bagging is a type of Ensemble Technique where the result is calculated on the voting-win method i.e. the output received by different models after testing on the test dataset are aggregated together then voting of the output decides which output will be considered as the final output. The output in the majority will be the final output of the model. Bagging consists of fitting several base models on different bootstrap samples and build an ensemble model that "average" the results of these weak learners.

"Bootstrap Aggregation or Bagging is usually wont to reduce the variance of the algorithms with high variances, like decision trees, classification and regression trees (CART). The only parameters when bagging decision trees is the number of samples and this can be selected by running the algorithm with different sample size until there isn't any improvement in the accuracy of the output. One of the best advantages of this technique is that a large number of the dataset may take a long time to run but it will never overfit the training data."[6]

4. **Random Forest Algorithm:**

Random Forest algorithm is Ensemble-Bagging method which operates by constructing multiple decision trees during the training phase. The Decision of the majority of the outputs(trees) is chosen by the random forest as the final decision. The main advantage of using Random Forest is that it is a mixture of both types of supervised learning problems i.e. Regression and Classification. The Random Forest algorithms are used in many machine learning applications such as :

- For Remote Sensing such as ETM devices used to acquire images of earth's surface, Random Forest is the first choice as it provides Higher Accuracy in a less training time.
- For Multiclass Object Detection, Random Forest is used as it provides better detection in complicated environments.
- Some gaming consoles use this algorithm as it is used to track body movement and recreates it in the game. Random Forest algorithm is trained to identify the body parts and algorithm learns from it. Then it identifies the body parts of the users such as hands, feet, face, eyes, nose etc.

"Random forest consists of an outsized number of individual decision trees that operate as an ensemble where each tree within the random forest spits out a category prediction then the category with the foremost votes becomes our model's prediction."[7]

Measurement of the relative importance of each feature on the prediction is another advantage to the Random Forest Algorithm. Another excellence of the random forest algorithm is that it is effortless too. In Random Forest each tree is picked from a random subset of features. This high level of variation results in lower correlation among trees and introduces more diversification.

Hyper Parameters in Random Forest:

The hyperparameters help to increase the accuracy and speed of the prediction model. The parameters provided in Sklearn library are:

1. **n_estimators**: "To select the number of trees the algorithm builds before taking the maximum voting or taking the averages of predictions. In general, a better number of trees increases the performance and makes the predictions more stable, but it also slows down the computation."[8]
2. **max_features:** The maximum number of features random forest considers to split a node.
3. **min_sample_leaf:** This determines the minimum number of leaf required to split an internal node.
4. **n_jobs:** This explains the number of processors can be used to run the model.
5. **random_state: "**This makes the model's output replicable. The model will always produce the same results when it has a definite value of random_state and if it has been given the same hyperparameters and the same training data."[9]
6. **oob_score**: It is used as a cross-validation method in the Random Forest Algorithm.

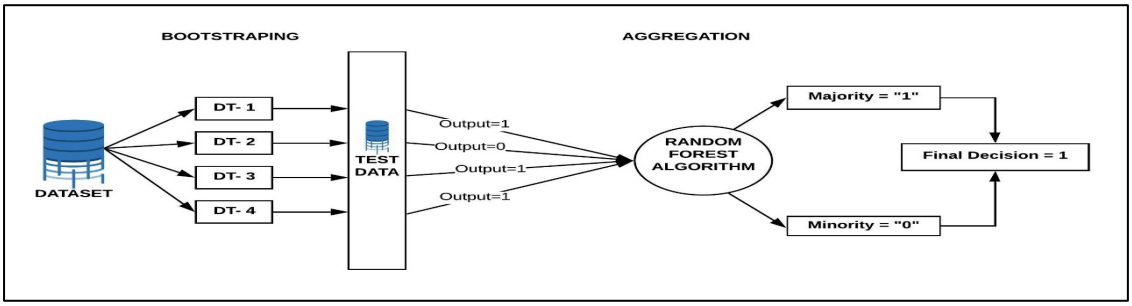

Figure 2: Implementation of the Random Forest Algorithm

Advantages of the Random Forest Algorithm:

1. No Overfitting and takes less training time.
2. High accuracy and runs efficiently on large databases.
3. Estimate missing values and still maintains high accuracy even having an extensive amount of missing data.

5. **RandomisedSearchCV:**
It is a hyper-parameter tuning method of python's scikit-learn library. It is used to implements a "fit" and a "score" method and to predict the best model.
6. **ExtraTrees Regression Algorithm:**
ExtraTrees Classifier is an ensemble method which is technically three times faster than the RandomForest method and equally accurate. But the difference between both the algorithms is:
    1. ExtraTree seems to keep a higher performance in the presence of a noisy feature in comparison to RandomForest.
    2. When all the variables are relevant, both methods achieve nearly the same performance.

Methodologies:

1. **Data Preprocessing:**
"Before Training, any model using any algorithm Data Preprocessing is that the most significant step and will be the primary step. the data Preprocessing contains several checkpoints (steps) such as: [10]
    1. **Step 1: Import Libraries:** The essential Libraries for Data preprocessing I used are *Pandas* for data manipulation and analysis, *Numpy* for numerical analysis, *Matplotlib* and *Seaborn* for better visuals and graphical stats of the data.
    2. **Step 2: Import the Dataset:** This downloaded this dataset from Kaggle, and then downloaded the dataset using *the pandas* library.
    3. **Step 3: Taking care of Missing Data in Dataset:** After evaluation of this dataset, I found no missing values in the dataset.
    4. **Step 4: Encoding categorical data:** This dataset contains some Categorical values such as fuel type, owner type, seller type, so we need to encode these categorical data into an encoded format to better train our model, to do this I used *get_Dummies()* method of pandas and this converted the whole Categorical values in the dataset into binary values.
    5. **Step 5: Splitting the Dataset into the Training set and Test Set:** To split this dataset into Test and Train dataset to train our machine learning model I used the capable machine learning library of python, scikit-learn or sklearn. Using its model selection method to create testing data by picking random values from the available dataset for model prediction, or we can say Supervised Learning.
    6. **Step 6: Feature Scaling:** Since all the data, available in a standard format, so here I do not use any feature scaling techniques.

2. **Data Training and Modelling:**
To train and develop a model, first of all, we need to the dependent and independent variables. To find these variables, first I used to find the correlation between the variables of the output and then separates my variables into two different axes we call it x and y where the x-axis contains all the independent variable and y-axis having the dependent variable, in our model its selling price of the Used Cars.
Using ***sklearn.model_selection*** library and its ***train_test_split*** function, further this dataset is distributed in the train-test dataset using ***RandomizedSearchCV*** tuning of this model is done to find the best hyperparameters for our model prediction.

3. **Proposed Model:**

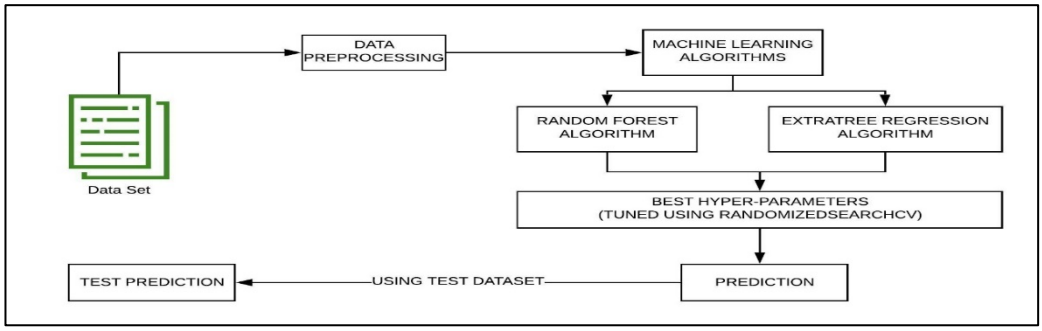

Figure 3: Flow Chart of Proposed Model

The proposed model is an application of the two machine learning algorithms i.e. Random Forest Algorithm and Extra Tree Regression algorithm. In this model first, the dataset is loaded for further exploration. In this specific model, I used a Dataset available at Kaggle. After performing the Data preprocessing steps on this dataset such as handling missing values, Hot encoding of Categorical Values, we start training the model for distributed dataset into two 1. Training Dataset and 2. Test Dataset. This test data is picked randomly from the original dataset. Applied the two machine Learning algorithms i.e. Random Forest Algorithm and ExtraTree Regression Algorithm and done tuning of the Hyperparameters using RandomizedSearchCV to get the best Hyper-Parameters for result prediction. Once the model predicts a result, I'll test the prediction using test dataset created using the scikit-Learn library and calculate its accuracy.

4. Model Prediction and cross-validation.

"Cross-validation is an analysis technique used for the assessment of the results of statistical analysis that how it generalizes to an independent data set. **Cross-validation** is mainly used for checking the accuracy of the prediction and to check the performance of the model."[11]

After performing cross-validation and evaluating all other metrics of the model performance and visualisations the following result are obtained.[12-25]

1. The Heat Map in figure 4 explains how all the attributes are correlated with each other. The dark blue colour shows a positive correlation among the attributes on x and y-axis and in this order, the sidebar shows the white colour as the highly negatively correlated variables.

   According to this map, we can say that the "Selling Price" and "Present Price" are positively correlated and they can be an important factor in predicting the current selling price after cars being used. This Map also indicates that there is a negative correlation between "Present Price" with "Fuel Type" and "Seller Type"

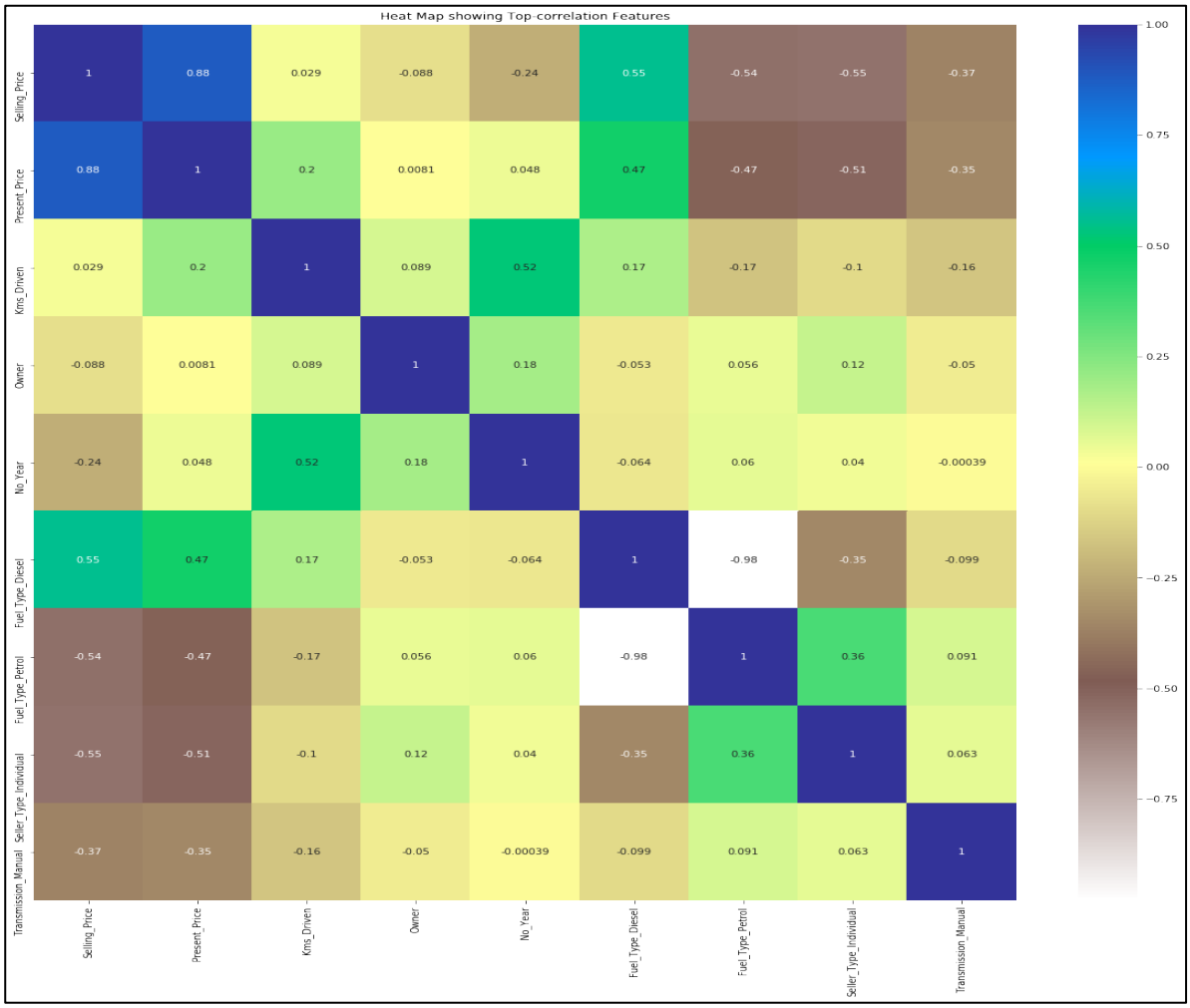

Figure 4: Heatmap showing Top correlation Features

2. The distplot in figure 5 below shows a normal distribution of the model with test dataset, this proves the accuracy of this model. Hence we can say that the prediction of this model is highly accurate.

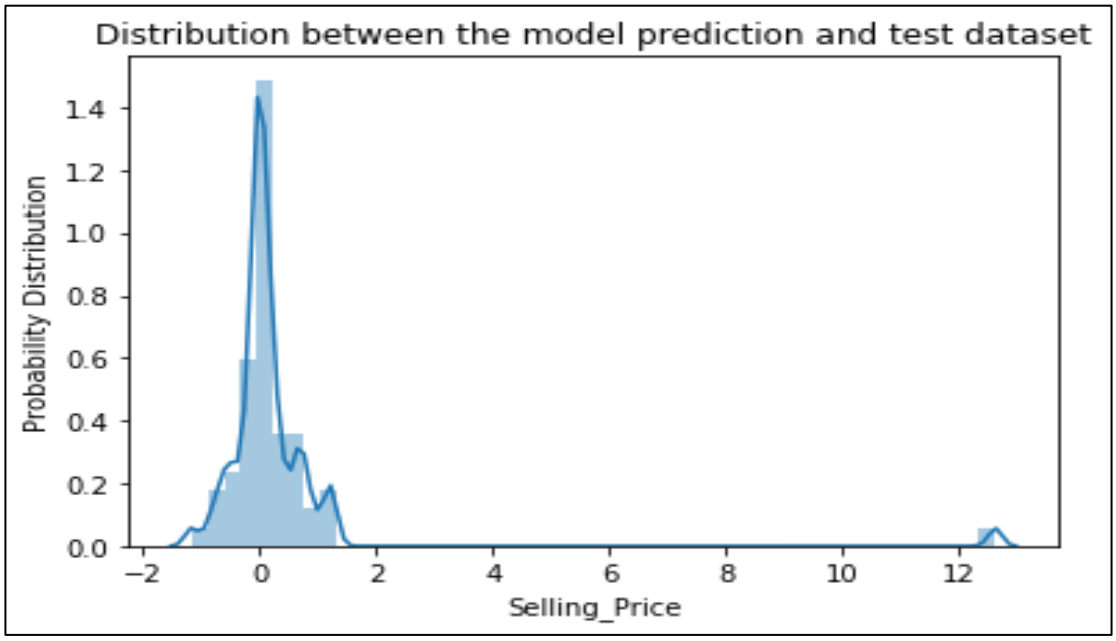

Figure 5: Distplot showing the distribution

3. The scatterplot in Figure 6 shows a linear distribution which ensures the accuracy of this model so we can finally say that prediction of the selling price using available dataset is accurate.

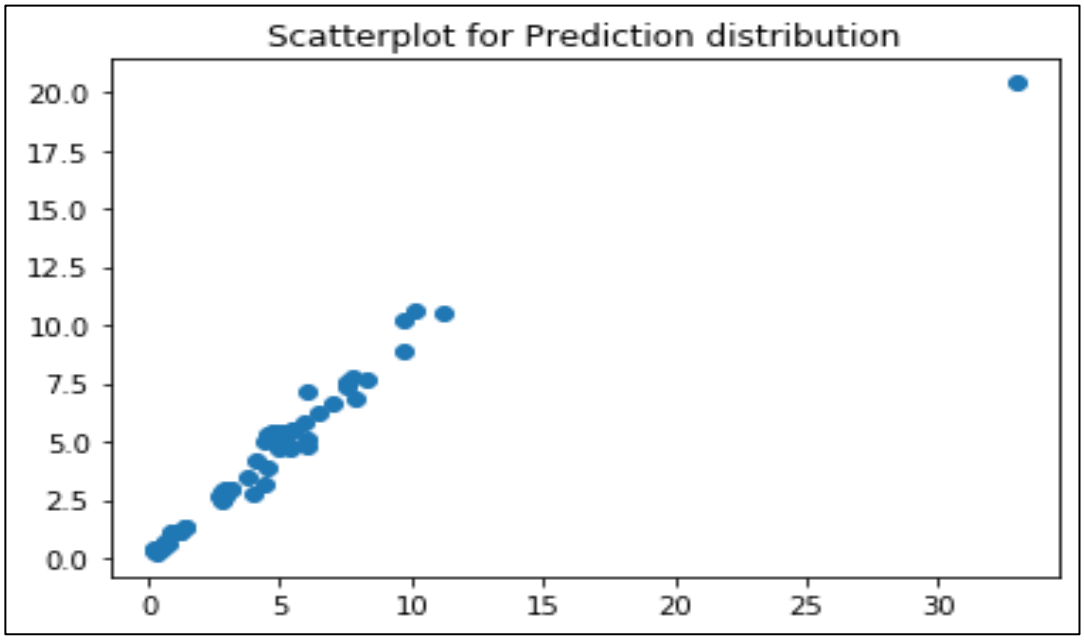

Figure 6: Scatterplot showing the distribution

4. Finally using the Heroku platform we can deploy this model as a web-based application. I also deployed this application using the same platform using Heroku as shown in figure 7.

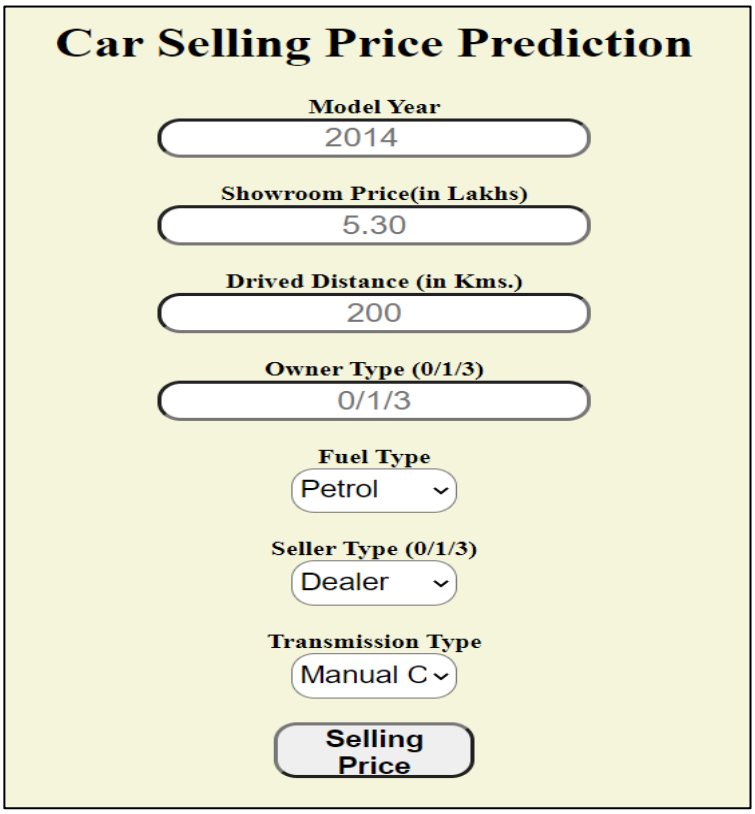

Figure 7: Deployment of the model using Heroku

Conclusion:

This model is based on the machine learning algorithms and we were trying to predict the selling price of the used cars based on the dataset provided at Kaggle. To predict this dataset we used two machine learning algorithms i.e. Random Forest and Extra Tress Regressor. The prediction of this model is further compared with the test dataset created by picking random values from the original dataset and the evaluation of the prediction is further evaluated using different methods. After a complete evaluation of the predictive model, we can conclude that the accuracy of this model is very and Random Forest and Extra Tree Regression is one of the best algorithms for regression problems. These two algorithms are highly accurate and fast in prediction irrespective of the size of the dataset.

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-learning-differences-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-machine-learning/
[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 Trans-Conductance 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 k-NN. 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