Second-hand Car Price Prediction

N Anil Kumar[0000−0002−4874−0767]

Department of Information Technology,
Vasavi College of Engineering,
Hyderabad
anil230@gmail.com

Abstract. Predicting the price of Second-hand or used cars is an important as well as an interesting problem. The efforts required in achieving the desired price of the used car gives a rough sketch about the amount at which the car can be sold at the best price. The challenging part is to find the best price of the used cars based upon the actual features of the car. The highest co-related features are considered and are used to build the model by using Random Forest Regression technique.
The features which were used to build the model, should be given as input to the system for predicting the price. Random Forest is the best technique as it is a classifier that contains a number of decision trees on various subsets of the given data set and takes the average to improve upon the predictive accuracy of that data set. This ensures that the predicted price is worthy

Keywords: Random forest · Used car price · Price prediction.

1 Introduction

1.1 Motivation

Estimating whether the price of a used car given to a dealer or individual is worth or not. Many factors such as year, kilometer driven, etc. They can affect the real value of a car. From in the seller's opinion, it is also a problem to properly price a used car. Based on existing data, the purpose of which is to use machine learning skills to improve the model of pricing used cars

1.2 Problem addressed

The price of a used car is especially predicted when the car is used and not recently made.This prediction is a very tough and important . With the increase in the number of used cars it is increasing and the cost is increasing. So consumers are finding other ways to buy new cars. There is a need for accurate predictor equipment for used vehicles.

1.3 Objectives

- The aim of this paper is to predict the best reasonable price of the used car.
- To determine on what attributes of used cars the price is depend.

1.4 Solution/Novelty

Firstly, we read the data set consists of features of the car and analyze the dependent and independent variables, then data is processed in order to handle the missing values .Later, with the help of Random forest Regression machine learning algorithm we build the model which can best predict the price of the used car. Once the model is built then we measure its accuracy and try to improve the model's accuracy using various approaches.

2 Related works

Enis Gegic et. al. [1] build a model for predicting the price of used cars in Bosnia and Herzegovina, has applied three machine learning techniques namely Artificial Neural Network, Support Vector Machine and Random Forest and has given a comparative analysis of these models.
Richardson [2] in his thesis he proved that Re- Manufacturing the old cars will give high fuel efficiency and he used multiple regression analysis to calculate the used cars price before it goes to the re-manufacturing. He also showed that hybrid cars are more able to keep there values than traditional vehicles.
By utilizing neuro-fuzzy knowledge-based system, Wu et al. [3] has worked on car price prediction study. Their data set consisted of attributes of the car like brand of the car, year of manufacturing and the type of engine it is made of. Their model predicted simple regression model's results. They have also made an system called Optimal Distribution of Auction Vehicles (ODAV) to meet the high demand of the cars which are for lease every year.

Pudaruth [4] has worked with various algorithms of machine learning , like naïve bayes , decision trees analysis, multiple linear regression and k-nearest neighbors for car price prediction in Mauritius.They collected their data set from local newspapers within a month or less. Their data set consisted of following attributes like brand, model,transmission type and price ,mileage in kilometers, cubic capacity, production year,exterior color. Finally he concluded that Decision Tree and Naive Bayes were not up to the expectations.

3 System Analysis

In this paper, we aim to find the best price of the used car by using Random Forest Regression model. All the car features remain the dependent features i.e, they will become the inputs to the model and the predicted price or estimated selling price will be the independent feature, which would be generated from the model. The model is entirely built on Random Forest Regression technique from sci-py library. To build the model a large data set is been used, which consist of different cars along with their own features.

The main aim of this project is to predict the worth full price of any used car which can be fair enough to both seller and buyer.

The entire code is hosted on flask server. Currently we would be using the local server for hosting the entire code and later on we can host it on any public server such that any person could access it and make use of it.

3.1 Data Set

The Data set comprises of the attributes of the car [5] as the columns and they are car name, year, present price, selling price, fuel type, transmission type, number of the owner and seller type .

During further processing the attributes like number of years was created and the one with multiple valued attribute was converted to single value attribute and all the attributed of the car except the selling price was used as the dependent values i.e, the input values and the selling price would be the independent value i.e, the final output

3.2 Algorithm-Random Forest Regression

Random Forest [6, 7] algorithm is the supervised learning technique which is very popular in machine learning algorithms. Both Classification and Regression problems in ML can be solved by Random Forest algorithm. In particular, we are using Random Forest Regression for building our model. To enhance the overall performance of the version and to resolve a complicated problem , Random Forest regression makes use of the idea of ensemble learning, this is a method of mixing a couple of classifiers.

Random Forest Regression is a classifier which includes greater wide variety of selection tree on diverse subsets of the given information set and to enhance the predictive accuracy it takes the average.The random forest regression considers the prediction from every tree rather than counting on one selected tree, primarily depending totally on bulk counting of predictions, and at last it finally predicts the majority counting.

Random Forest Regression, as its call suggests, includes a many wide variety of selection tree that paintings as an ensemble. Each person tree within side the random forest regression spits out a category prediction and the magnificence with the maximum votes will become our version's prediction. The essential idea at the back of random forest regression is a easy however effective one – the information of crowds. In information science, the motive that random woodland version works so nicely is: A massive wide variety of extraordinarily uncorrelated models (trees) working as a committee will outperform any of the person constituent models.

In the Figure 1 we could see that the initially entire data set is split into many N number of training data sets. Each training data set, leads to its own respective decision tree. These both phases i.e, training data set and decision tree falls under training data set section. Averages of all the results of the decision tree are considered and the majority is considered as the predicted final output. This entire process of averaging and prediction falls under test data set.

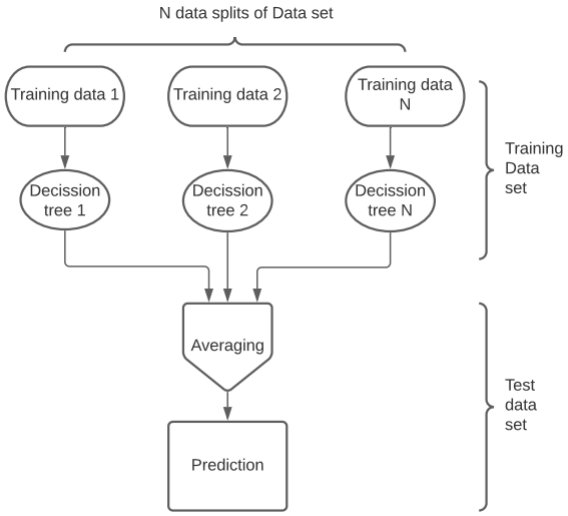


Fig. 1. Working of Random forest algorithm

3.3 Assumptions for Random Forest Regression

It is a conjunction of multiple trees to predict the class of the data set, there may be some possibility that decision trees may predict the correct output, where as others may not predict the correct output instead predict the wrong output . But altogether by averaging , all the decision trees predict the correct output in the last. Thus, below are two assumptions that should be followed for the best version of Random forest classifier to predict:

- For the Random forest classifier to predict the output accurately and correctly instead of the guessed result. it is expected to have some real values in the data set .
- Every selection decision tree is expected to have prediction with very less correlation

4 Implementation

4.1 Pre-Processing

Initially we upload the Data set and check for all the null values. After clearing all null values we go with the creation of the data set into proper format as some attributes (columns) may be in the textual format with more than one unique type of value. So, we need to make each column for each unique value while Pre-Processing the data. After all the things we finally add the new columns based on the requirement and go with the future model building.

4.2 Visualizing

The actual data visualization and Pre-processing are done before and now we just visualize the main features by using ExtraTreeRegressor and we extract the feature importance from the data set such that we can know about each feature's importance in calculating the selling price of the car.

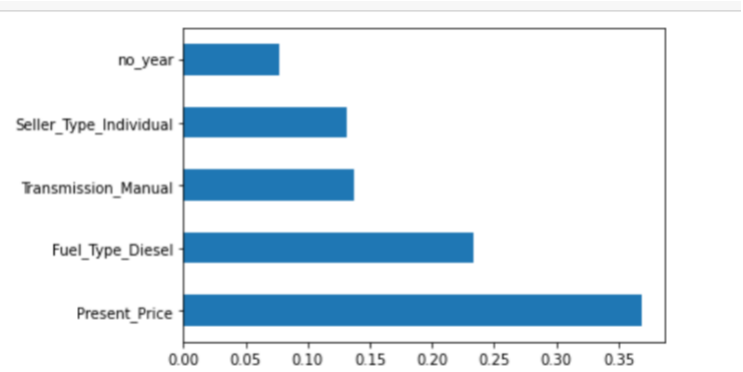


Fig. 2. Dist plot for our model vs testing data

4.3 Model Creation

We are using Random forest regression with the decision trees constructed using the Randomeized searchCV [8] attributes to the Model and coming to the Scoring Factor] [9] we use "neg_mean_squared_errror".

After the model is been built based on the given attributes provided to the Randomforest Regressor function we now fit the actual values to model by using the training variables of the dataset.

The entire Data set is distributed into 2 types the Independent values and Dependent values and each type is again distributed into 2 categories one for training the model and the other for testing. The Independent value is what we predict and the Dependent value is what we give to the model.

4.4 Input

The values to Model can be provided via basic python file by loading the model and giving values to the variable and by using those variable we need to create the array to give to the model.

4.5 Output

After the prediction is started with the loaded model and the given input the output could be displayed.

5 Analysis

In this paper while pre-processing we divide the entire data set into two splits one for training and other for testing and the entire project based on Random Forest Regression the model was generated based on the X train, and Y train the the prediction was done by using the X test and Y test.

X comprises of the entire dependable data(input) and Y comprises of the independent data(output).
After the creation of the model we then use a variable prediction to store the values of the all prediction generated with the X test. After the array is been generated the array values are given to test with the Y test via to plotting techniques.

Dist Plot The plotting of X axis is with the selling price and Y axis from prediction array.
The actual plotting shows the normalized curve for the model by using difference between the Y test and prediction.

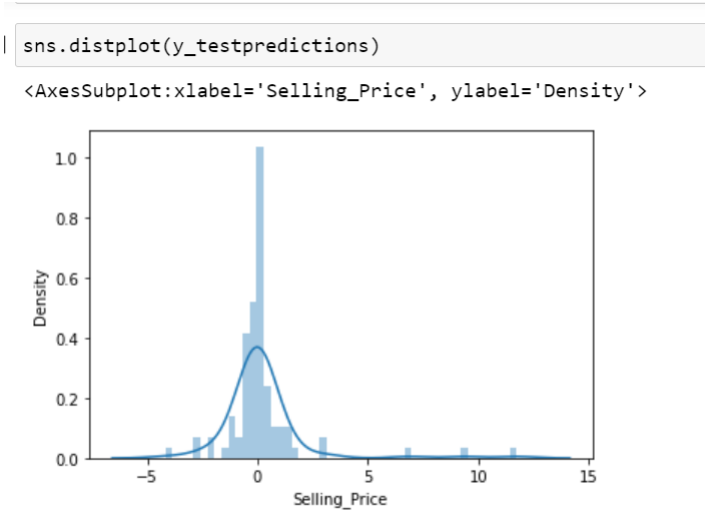
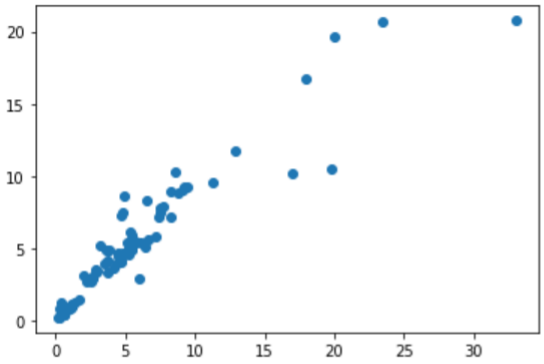


Fig. 3. Dist plot for our model vs testing data

Scatter plot The Scatter plot is also predicted for selling price vs prediction Here in the scatter plot we could see that the plot is exactly having an linear line between the attributes and the best regression model will have the linear exhibitions scattered points on the graph.

```
plt.scatter(y_test,predictions)
```

```
<matplotlib.collections.PathCollection at 0x2af5e26d248>
```




```
from sklearn import metrics
print('MAE:', metrics.mean_absolute_error(y_test, predictions))
print('MSE:', metrics.mean_squared_error(y_test, predictions))
print('RMSE:', np.sqrt(metrics.mean_squared_error(y_test, predictions)))
```

```
MAE: 0.8796358241758259
MSE: 3.930771725852751
RMSE: 1.9826173927040867
```

Fig. 4. Scatter plot for our model vs testing data

In the Figure 4 we could also see the final metrics for an regression model and the model with the these values near to zero can be considered with best accurate model. Even the values are near to zero and that indicates that the prediction made using This model we be accurate.

6 Conclusions and Future Scope

6.1 Conclusion

Finally we can say that the methods used in this paper are apt and the accuracy of this project is determined using a dist plot which gave a normalized curve indicating that the accuracy is very high , and even a scatter plot was used to display the accuracy which showed that the points form a straight line indicating in high accuracy even the scoring factor for our project nearly close to the 0 so we could say our project has the best model to predict the input values. Random Forest Regression also gave us the good accuracy value.

6.2 Future scope

For better higher performance, we are able to layout deep gaining knowledge of model, use adaptive studying quotes and educate on clusters of information in

preference to the complete dataset. To accurate for overfitting in Random Forest, extraordinary choices of capabilities and range of trees might be examined to test for alternate in performance.

References

1. Gegic, E., Isakovic, B., Keco, D., Masetic, Z., & Kevric, J. (2019). Car price prediction using machine learning techniques. TEM Journal, 8(1), 113.
2. Michael S Richardson. "Determinants of used car resale value". PhD thesis. Colorado College., 2009.
3. 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.
4. Pudaruth, S. (2014). Predicting the price of used cars using machine learning techniques. Int. J. Inf. Comput. Technol, 4(7), 753-764.
5. Vehicle dataset - Used cars data from websites `https://www.kaggle.com/nehalbirla/vehicle-dataset-from-cardekho`
6. Random Forest Regressor URL: `https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html`.
7. VanderPlas, J. (2016). Python data science handbook: Essential tools for working with data. O'Reilly Media, Inc.
8. Randomized search CV. URL: `https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html`
9. Scoring factors. URL: `https://scikit-learn.org/stable/modules/model_evaluation.html`