Machine Learning Assignment Part-4

Report on the Investigation Words: 1282

Introduction :
In this report I am going to explain about the opening of a new hotel in the area and to predict whether it will be a profitable or not. This I am going to do by using the Machine Learning techniques and the result of it will be the deciding factor to it. As a manager of a large hotel chain contact me to do this work for her. In Machine Learning we have many models which help us to predict the results on the basics of data which we get from the hotel company. As I have implemented the Machine Learning technique and after doing all the required things for this project I camme to know that the opening of hotel in this particular area will be profitable for the company.

Task Definition:
The problem which here I am going to address is that the opening of a new hotel and to predict and to tell this to the manager whether opening of the hotel here will be going to be good or not. If it's good then what are the chances that we are going to make profit from it or not. This is an interesting and important problem for me because while working on this project I realise that in today's world the technology has grown up to this much extends that now we have things to check the outcome of all the business before investing in it. As now by using machine learning we can find that whether this business will be profitable or not or should we have to change our business and to try on some other which surely gives us the profit.

Algorithms used:
At first here I am going to tell you about how we are going to fill the missing value in the first part of the project. When we get the data from the company it has so many missing values. So, to fill those values we used mean and median to fill all the missing values in columns, deleting the missing values column and replacing the missing values with zero. After filling in all these values we check the datatypes is same for the entire column or not. The best result I got from Imputation is by replacing the missing values with the mean values. Then here comes our important task to do in the first part is to use the classifiers.

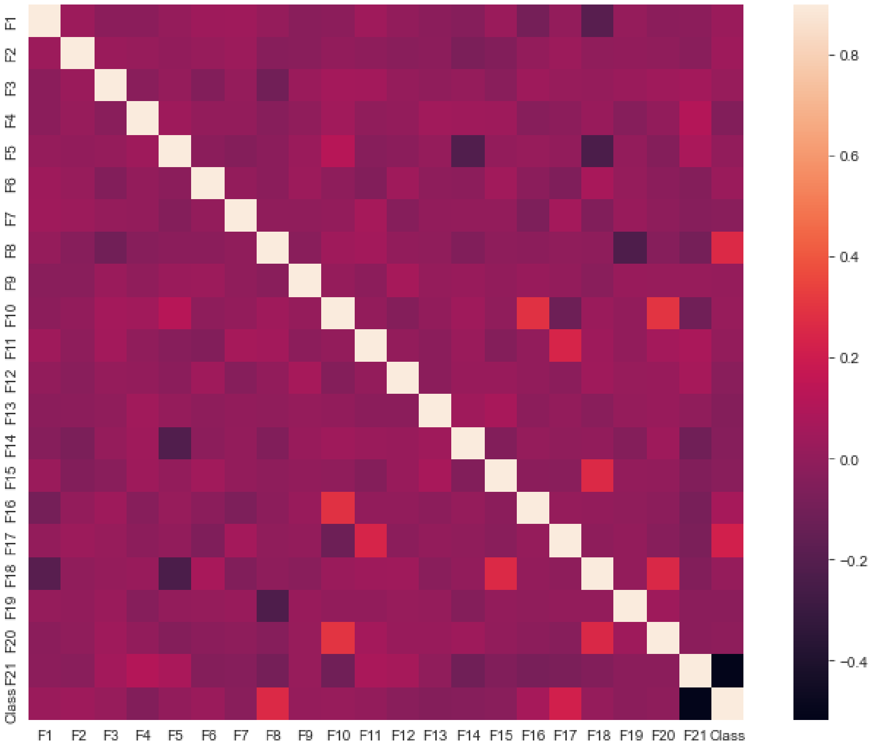

Figure for subplots of heatmap for Classifier

In data science, a classifier is a form of machine learning algorithm used to give a class label to data input. Classifier algorithms are trained using labeled data; in this prediction example, for instance, the classifier receives training data that labels different factors. After sufficient training, the classifier then can receive unlabelled data as inputs and will output classification labels for each data. In classifiers, we have so many models but from them some of the good predictive models I implemented here. The best accuracy rate I got out of all models I used is Random Forest Classifier.

Random forest classifier is a type of supervised machine learning model used for the predictive task. The "forest" it creates, looks like what we've got in the decision tree. It typically gets trained with the learning models which helps in increasing the final results.

Another quality of the random forest classifier model is that it is very easy to measure the relative importance of each label on the prediction. Scikit learn provides very good tool for this model that helps to measures a feature's importance by looking at how much the tree nodes can use that feature to reduce impurity across all trees in the model. It computes this score automatically for each label after training and measure the results so that the sum of all required label is equal to one.

The important hyperparameters we can use in the random forest to increase the accuracy are n_estimator, max_features, and min_sample_leaf out of these I used n_estimators = 100 to get the better predictive value. It is just the number of trees the algorithm builds before taking the maximum voting or taking the averages of predictions. In general, more the number of trees nodes more will be the chances of getting better output.

Now I am going to tell you about all that is done in the second part of this project which is to check whether it is going to make a profit or not. When we uploaded the data we came to know that we have different types of datatypes in the given data by the company. So our first task is to make all the datatypes of the same type. To do this we used the cleanup inbuilt function of the sklearn which help us out to make all the data of the same datatypes. As in this part of project, we are going to predict the profit so to get it we are going to use the regression technique for it.

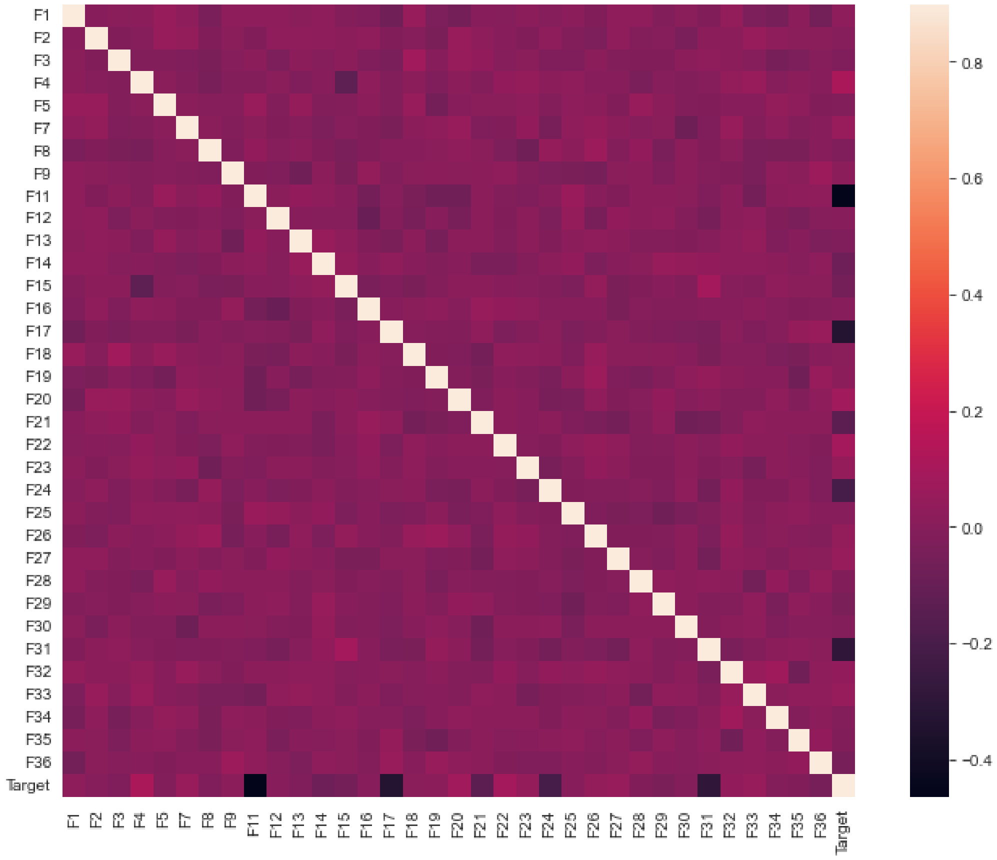

Figure for subplots of heatmap for Regression

Regression in machine learning is based on mathematical methods that allow scientist to find out a continuous results based on the value of one or more predictor variables. Linear regression is probably the most popular form of regression analysis because of its ease of use in predicting the output with good accuracy. And when I used different types of regression for prediction I also got the best results from Linear regression only.

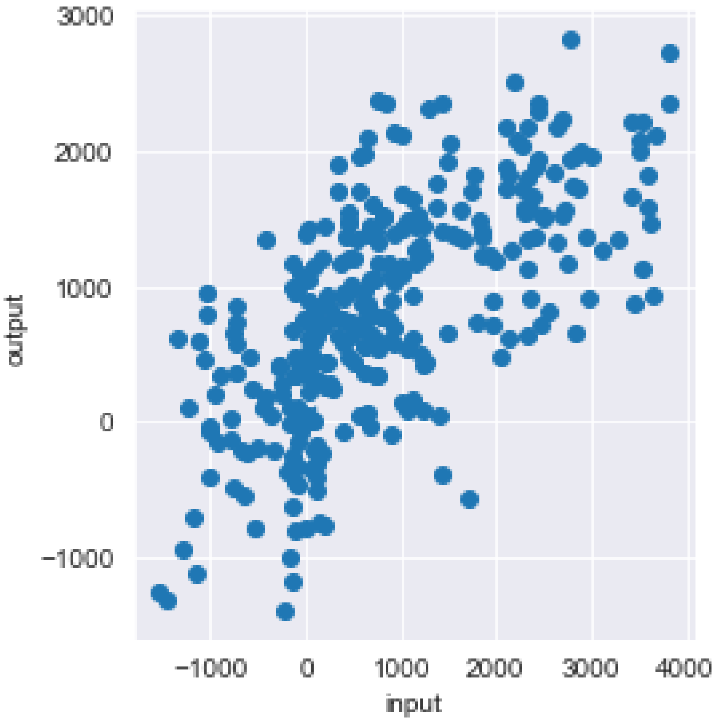

Figure of Linear Regression.

Linear regression finds the linear relationship between the dependent variable and independent variables by using a best-fit straight line technique. Generally, a linear regression model makes a prediction by simply computing a weighted sum of the input features, plus a constant called the bias term. In this technique, the dependent variable is continuous, the independent variables can be continuous or discrete, and the nature of the regression line is linear. The prediction done is then shown us the type of metrics it is fitting it. Here we have three types of fitting that is underfitting, ideal and overfitting.

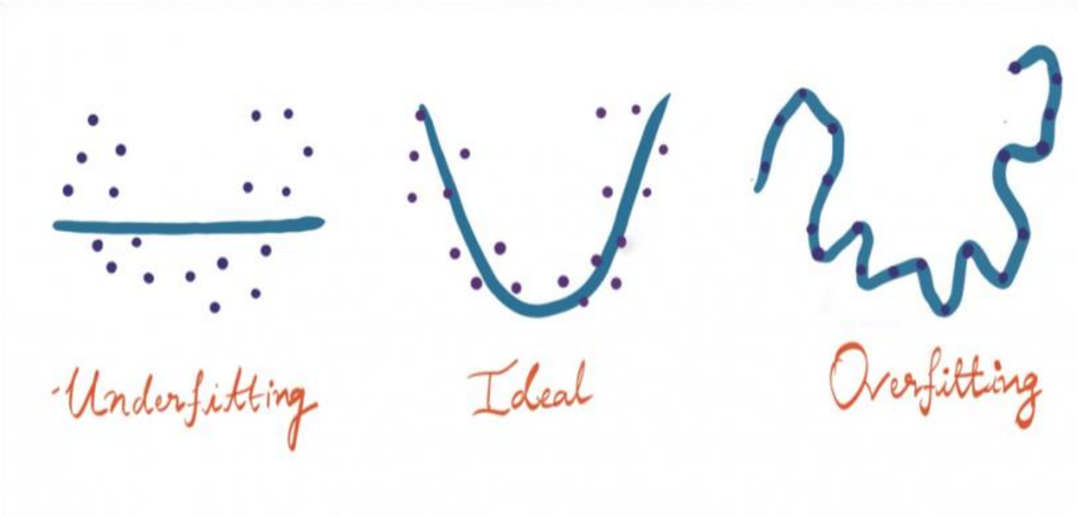

Figure of Types of fitting

Conclusion :
Here I want to conclude on the note that for this project I have implemented the classifiers model to help me out to say whether we can open the hotel at a particular place or not. I implemented a Decision tree, Random Forest, Support Vector Machine, and k-NN classifiers from all these I got the best accuracy score of 87.50% by using Random Forest Classifier, and in this part, we have also used the mean method for filling the missing values. After getting to know by the accuracy results that yes we can now open the hotel at this place, the next task that comes up is to check whether the hotel will make a profit at last or not. For this, we have Regression techniques in Machine Learning and here I implemented Linear, Logistic, and Lasso Regression but the best results I got is from Linear Regression with an accuracy of 47.24% with RMSE value 853.02 and MSE value is 727643. 22 and that helps me to tell the manager that surely you are going to make the profit from this business. The proper implementation of Machine Learning techniques is the most important task to do and to get the best out of it. So, according to me, this is the main point of this project. If we talk about the improvement of the results I can say that the data provided by the company is more accurate and full then we can assure that the results will be much better. As some of the data are missing and to fill those missing data we try other methods which help to some extent to give good accuracy but not the best.

References:

1. https://builtin.com/data-science/regression-machine-learning
2. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
3. https://www.kaggle.com/niklasdonges/end-to-end-project-with-python
4. https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html
5. https://www.analyticsvidhya.com/blog/2020/12/understand-machine-learning-and-its-end-to-end-process/
6. https://towardsdatascience.com/selecting-the-correct-predictive-modeling-technique-ba459c370d59
7. https://moodle.essex.ac.uk/course/view.php?id=3510
8.https://moodle.essex.ac.uk/pluginfile.php/1431352/mod_resource/content/2/CE802_Assignment_2021.pdf