ML Developer Coding Problem Report

By- Yash Mittal

In this question I have initially fetched data from marketing-data.csv file. It has 45211 rows and 17 columns. We have to predict the column is_success in this data file using all other column. Age is not a great feature as we have seen through boxplot. I have separated categorical and quantitative variables in two parts. As we can see we have is_success data in yes/no form but we need to have this in numerical form to put this data in our algorithms. So I have used dictionary method to convert that data in form of 0 and 1(0 for no, 1 for yes). Next, I have splitted data into training data and testing data. After breaking the data it is also required to scale the data otherwise it may create anomalies in prediction. Splitted data and complete data both are scaled using min max scaler.

Now we have data and we have to apply different algorithms on it. First I have applied knn algorithm and I got accuracy of around 90% on complete data set. I have also used cross validation technique for different values of k.

Next I have used logistic regression in that I got accuracy of around 88.9% on complete data set. I have also checked my result through cross validation which gave me a fine result.

Next, I have used SVM algorithm for prediction which gave me 88.3% accuracy which is definitely not a good choice.

Then, I used Decision Tree Classifier which gave me 99.9% accuracy on training data set and 96.2% on complete data set. The accuracy of testing data is reduced and it is 84.8%. This is definitely a case of overfitting. Then I set max_depth=10 to reduce the problem of overfitting and it gave me an accuracy of 91% on complete data set.

At last I have Random Forest Classifier in which I got 92.5% accuracy on training data set and 92% approx. on complete data set.

Result-

After using all algorithms on data set I found that Random forest classifier will be the best for this data with an accuracy of more than 92%