Mini Hackathon 2021

Preliminary Round

Report Submission

Team

Data Geeks

Team Leader: K. Sutharsan

Team Members:

1. I. Jayatharan
2. S. Piramilam
3. Y. Dilaxiha

Kaggle Username: teamdatageeks

Kaggle Display Name: Data Geeks

1. Introduction

In this competition we are dealing with a classification problem. Here we got a dataset related with the applications received to a nursery. The dataset consists of several characteristics from the applications as predictors and we have a variable called “app_status” as our response variable. It indicates the application status, that means whether the applicant got selected or not. Our goal of this competition is to successfully build a machine learning model to predict the application status of a given applicant based on the characteristics of the applicant.

2.Methodology

Initially we were given two datasets, training data and test data. Our goal is to train a machine learning model and predict the response of the test data. Since this is a classification problem, we have tried several classification machine learning algorithms to predict the response. Therefore, to compare the performances of the different models first, we divided the training data into train data and validation data. Then we trained the models with the train data and then, we used the validation data to compare the performances of the models. As our final model we choose XGBoost algorithm to predict the response since it gave a better performance amongst the other models.

XGBoost stands for “eXtreme Gradient Boosting”. XGBoost is a decision-tree-based ensemble machine learning algorithm which uses a gradient boosting framework. To put it another way, the method creates a strong classifier by chaining together numerous weak classifiers. It learns from previous models and adds new models to correct the errors that past models have made. Models are added in a simple manner until there are no more improvements to be made.

3.Results

We evaluated our model based on some evaluation metrices that we obtained based on the validation set. First, we obtained the confusion matrix for the validation set from our model and then we obtained ROC curve for our model as well. Following figures show the confusion matrix and ROC curve of our model,

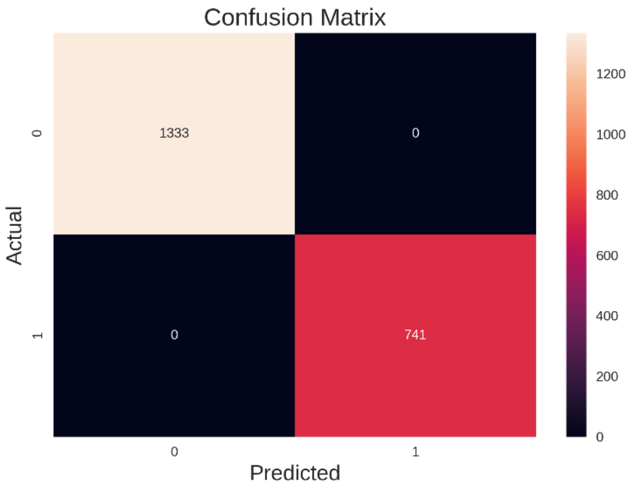

Figure 1: Confusion Matrix

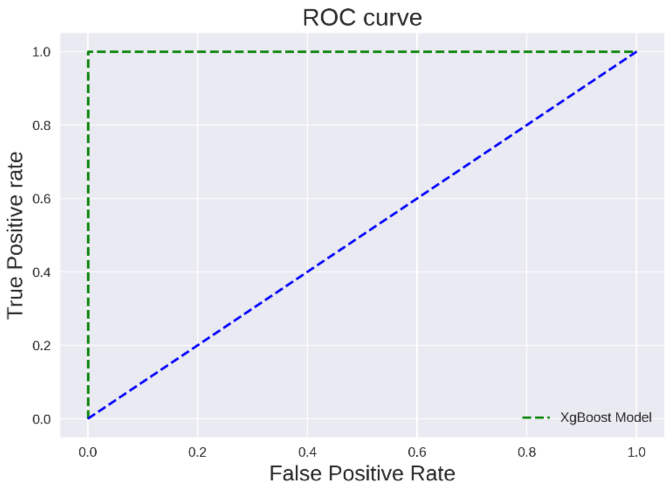

Figure 2: ROC Curve

From the confusion matrix we can see that, we didn’t get any False positive or False Negative predictions from our model. That is the class wise error rate of our model is 0 and the accuracy score of our model is **1**. We also obtained the ROC curve for our model and from the figure 2 we can see that we obtained the ideal line from our model. That is we got the AUC score of our model as **1**. Further, we also calculated the F1-score of our model and obtained **1** as F1-score from our model. These evaluation metrices shows that our model performed well and predicted the responses correctly all the time.

4.Conclusion

For the give classification problem, we used XGBoost model for predict the application status of the nursery applicants. Based on the evaluation metrices the results showed that our model performed really well. This model can be used to predict the application status of the nursery applicants in the future with high accuracy.