Ultimate Student's Hunt – A Machine Learning competition

Approach (Team_datadragos)

GBM Model

- First of all our approach was to try out the best algorithms possible for the data after making the data complete. The data was full of missing values which we analyzed by using mice package. Using the "aggr" function which showed some variables with <10% of data and some variables with 27% and 35% of missing values respectively. Thus we decided to replace the missing values with -999 and also splitted the date variable to "day", "month" and year in M.S Excel and removed the variables Location, Park ID, ID from the final complete dataset. Splitted the complete dataset in train and test data and applied Gradient Boost Model upon it giving an iteration of 5000 trees and training the model and saving the prediction of given test dataset in one file.

Xtreme Gradient Boost and Artificial Neural Network Model

- In this case we combined the given train and test data and did some feature engineering by creating some new variables by computing the difference of max and min values of each variables and also taking the average of max and min values of the old variables. Also did some feature engineering for direction of wind by converting it into categorical variable by binning the values of the variable and assigning some value to the missing data also. "ID", "Date" variable were removed after converting the date variable into month variable. Then dummy variables were created for the categorical variable such as park_id, location, direction_of_wind, month and the new data was saved as total_dummy data. It was split into train and test data and then train data was used to train the xgb model. First the data was converted into sparse matrix form after splitting the train_dummy data into train and test data and then Footfall variable was removed from either of the data after storing the predictor variable into the variable x and y. Finally after conversion of the both test and train data into sparse matrix using xgb.Dmatrix form, nested loops were executed to iterate three parameters of xgboost model ie.. "eta", "colsample_bytree" and "subsample" and list of parameters were passed through the xgb.train function to train the xgb model. Prediction were made for the test_dummy data and then it was saved in a file.csv format. For knowing feature importance values a feature importance plot was plotted after computing an importance matrix. Then again the predictions were made for the tuned parameters and a max_depth =10 for the tree also.

- ANN model was trained after imputing the missing values of the combined test and train data first using the mice package and then splitting the train data and then without using cv mehod, just by neuralnet function, the predictions were made and once also by using

cross validation method to tune parameters and used the nnet method and then making the predictions and comparing the rmse score.

- Randomforest model and CART Model were also used which gave a RMSE Score of approx. 126 and 132 respectively. But we discarded them from our main model due to XGB+ANN and GBM model giving a better rmse score than the former two models.

- Some change in the values were made in the predicted data by observing the training footfall for the past years after ensembling the xgb model with the artificial neural network model to achieve a low rmse score and better accuracy. The Rmse Score was also checked without combining neural net model and also by combining the two different neural net model with and without cross validation with the xgb model but the rmse score was not improved that much when xgb model was taken once as with the ensemble model. The predicted data footfall was replace by mean year wise and median values for the December month for year 2002 and 2003 and rounding off the footfall less than 685 and 721 to 700 and 900 respectively. These were some steps adopted to achieve some better prediction results and low rmse score.

Ensemble Model

- Then we ensembled the individual GBM model with the already ensembled ANN and XGBoost Model giving higher proportion to the model giving a better score by weighted average method. At a particular optimum point the lowest RMSE score was achieved and we got the best results.