Audify Data Science Assignment

Problem Statement - You are given a sample dataset from 'Football World' of Audify which consists of a player's data of on a particular day, using which you have to predict whether the player will play the game next day or not.

Data Exploration

Size of data : 117997 rows * 143 columns
Target variable: next_day_label
Days_from_first_open_max,days_from_first_open_max has 117997 number of zeros which means that these two features do not impact the next_day_labe so we can drop these columns for building models.

Missing_value_Analysis

Data does not contains any missing values and perfectly collected

Feature Engineering

User_pseudo_id is not useful for model building as we have data for unique user.
Event_date_ can be used to extract date like features such as weekday,dayoftheyear etc which could tell us about the playing behaviour of the user.
Other variables which are of numerical data type except opp_kicked_status_list are used to get a mean and median to form new features which would be helpful for the model to find patterns for prediction.

Object data type columns

There are 18 columns which has object data type which has been feature engineered to columns details as follows:
User_pseudo_id_ - not relevant for classification
Event_date_ - Weekday,DayOfYear,month
Defend_ball_final_x_list - defend_ball_final_x_list_median,defend_ball_final_x_list_mean
Defend_ball_final_y_list - defend_ball_final_y_list_median,defend_ball_final_y_list_mean
Kicked_ball_final_x_list - kicked_ball_final_x_list_median,kicked_ball_final_x_list_mean
Kicked_ball_final_y_list - kicked_ball_final_y_list_median,kicked_ball_final_y_list_mean
Kicked_speed_list - kicked_speed_list_median,kicked_speed_list_mean
Opp_kicked_ball_final_x_list - opp_kicked_ball_final_x_list_median,opp_kicked_ball_final_x_list_mean
Opp_kicked_ball_final_y_list - opp_kicked_ball_final_y_list_median,opp_kicked_ball_final_y_list_mean
Opp_kicked_angle_list - opp_kicked_angle_list_median,opp_kicked_angle_list_mean
Opp_kicked_speed_list - opp_kicked_speed_list_median,opp_kicked_speed_list_mean
Opp_kicked_status_list - opp_kicked_status_list_Goal,opp_kicked_status_list_Saved,opp_kicked_status_list_Timeout,opp_kicked_status_list_TimeOut
Striker_draw_angle_corrected_list - striker_draw_angle_corrected_list_median,striker_draw_angle_corrected_list_mean

Striker_draw_length_corrected_list - striker_draw_length_corrected_list_median,striker_draw_length_corrected_list_mean
Striker_draw_time_taken_corrected_list - striker_draw_time_taken_corrected_list_median,striker_draw_time_taken_corrected_list_mean
Striker_draw_speed - striker_draw_speed_median,striker_draw_speed_mean
Striker_delta_x - some data points where missing due to which not considered during model building
Striker_delta_y - some data points where missing due to which not considered during model building

Imbalance data

0 92826
1 25171
Name: next_day_label, dtype: int64
The given data is imbalanced meaning there is unequal number of 0's and 1's. So we will use algorithms that are less sensitive to class imbalance, such as decision trees, random forests, or gradient boosting, as they tend to perform well even with imbalanced data. We will also perform class weighting by assigning different weights to different classes during model training to account for the data imbalance. This can be done by using the class_weight parameter in various machine learning algorithms. Models will pay more attention to the minority class, leading to better generalisation.

Feature Selection

As there are more than 150 variable, we performed permutation importance and Stepwise Feature Selection and select 12 features for classification task.

'user_skill_level_max', 'opp_kicked_status_list_Goal',
'is_opponent_bot_mean', 'user_won_mean', 'opp_kicked_status_list_Saved',
'DayOfYear', 'score_diff_proxy_max', 'defend_status_ct_Saved_max',
'score_diff_mean', 'next_day_label', 'game_number_count',
'kicked_status_mr_Saved_max', 'kicked_status_mr_TimeOut_min'

Out which next_day_label is our target variable.

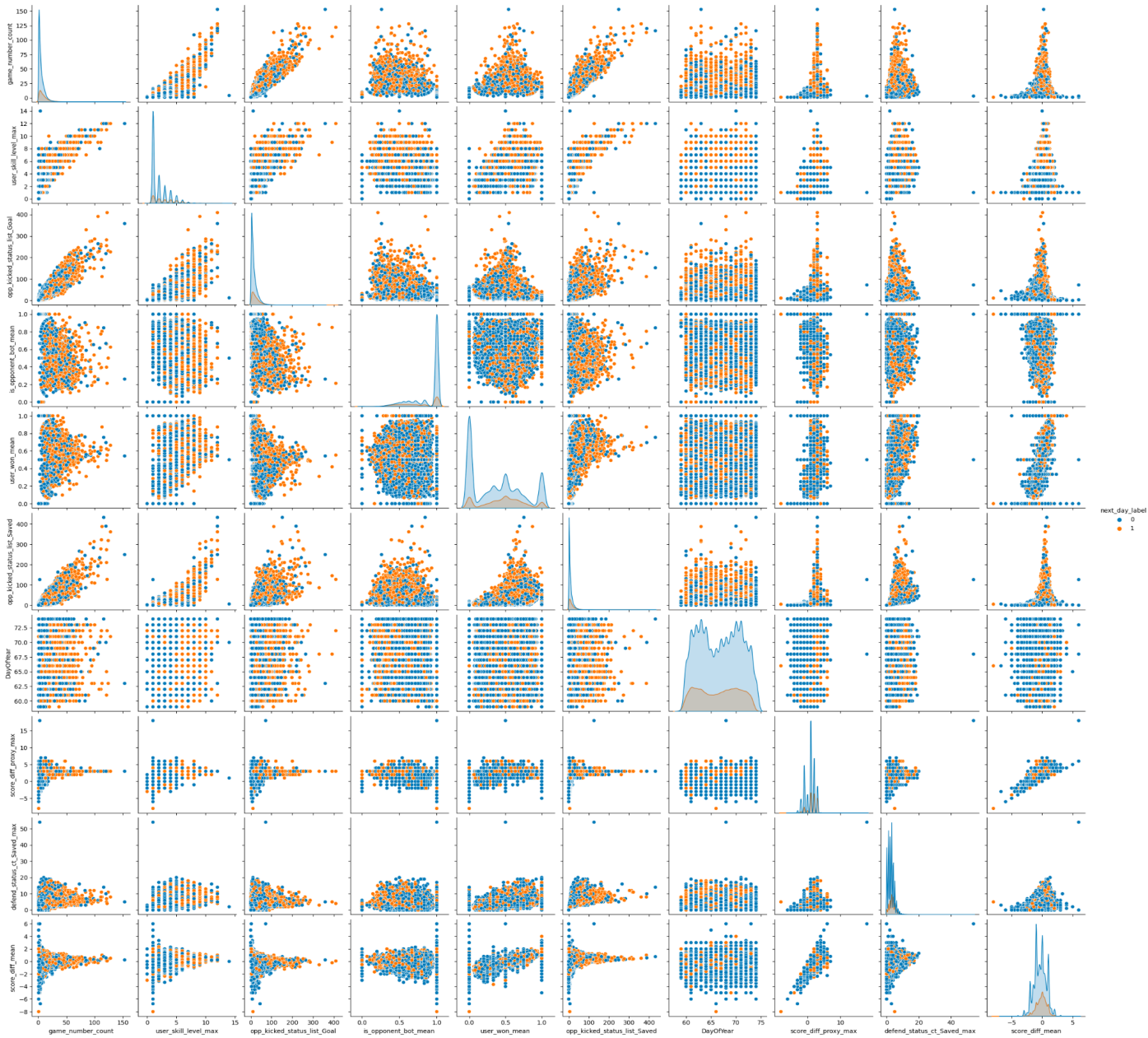
Model Building and Selection:
We tried three model out of which default random forest outperformed other models:
Model Accuracy and other parameters are as follows:
Decision Tree:
Cross-Validation Accuracy: 0.7875673990066984
Test Accuracy: 0.7908898305084746
Test Precision: 0.7478048132797959
Test Recall: 0.7908898305084746
Test F1 Score: 0.7367935639735835

Random Forest:
Cross-Validation Accuracy: 0.7907242757399711
Test Accuracy: 0.7919491525423729
Test Precision: 0.7579470183129493

Test Recall: 0.7919491525423729
Test F1 Score: 0.7178210910607353

XGBoost:
Cross-Validation Accuracy: 0.790628936195809
Test Accuracy: 0.7934322033898306
Test Precision: 0.6032608695652174
Test Recall: 0.08832305550029838
Test F1 Score: 0.1540864133263925

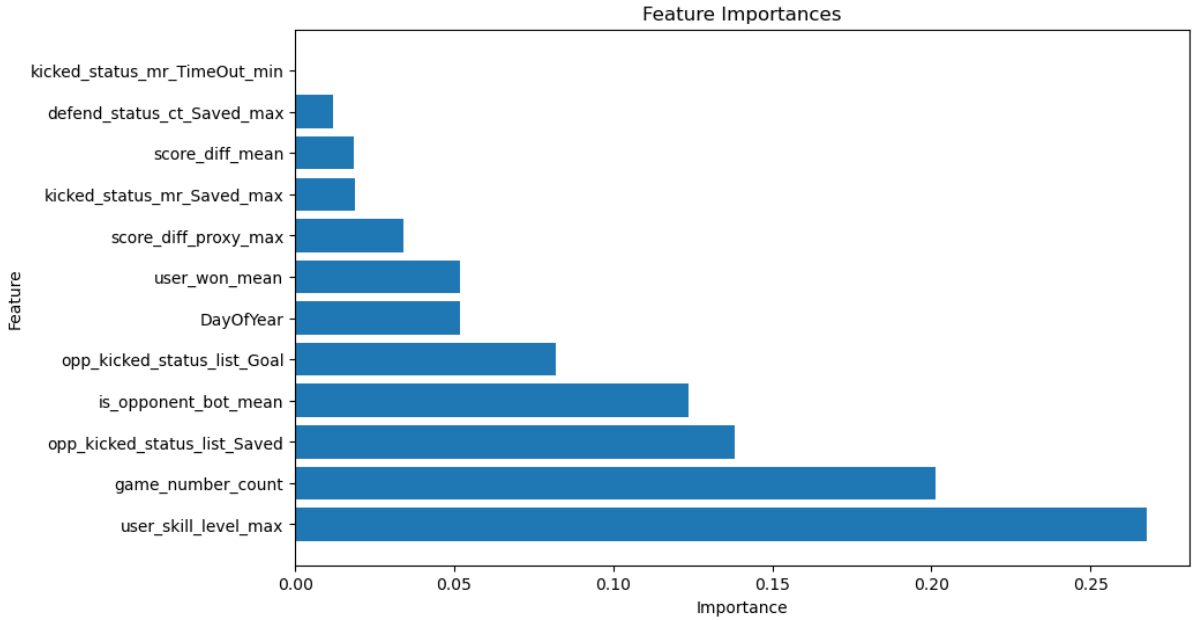
EDA:



Using this pairplot we can clearly see that there is no linear relationship between the target variable and independent variable.Hence, we should build non- linear model for classification task.However we can observe various relationships between variables which are as follows.

As the game_number_count increase there is a high chance that the user is going to play the game next day.
High game number count , High opp_kicked_status_list_Goal leads next_day_play.
Opp_kicked_status_list_saved also contributes to 1(next_day_label)

Feature contribution to the classification task:



Feature: user_skill_level_max, Importance: 0.2678077466107468
Feature: game_number_count, Importance: 0.20131234506816104
Feature: opp_kicked_status_list_Saved, Importance: 0.1383294958522232
Feature: is_opponent_bot_mean, Importance: 0.12358069159843554
Feature: opp_kicked_status_list_Goal, Importance: 0.08181902972508104
Feature: DayOfYear, Importance: 0.05196476957829885
Feature: user_won_mean, Importance: 0.05176570200819753
Feature: score_diff_proxy_max, Importance: 0.03391955706464053
Feature: kicked_status_mr_Saved_max, Importance: 0.018937127539994037
Feature: score_diff_mean, Importance: 0.018613797908437785
Feature: defend_status_ct_Saved_max, Importance: 0.01187175825252516
Feature: kicked_status_mr_TimeOut_min, Importance: 7.797879325870532e-05