Lead_Score_Model_Folder
ver. Shiyu Ma

Note: *This is a summary of the Lead Score Model's source code for personal reference. For future usage, checking the Final_Model_Logistic folder is enough.*
https://docs.google.com/presentation/d/1tFyB1KeIyU0Q52fZnW4yJhfqh6f-iLfbpiDgc1D2g1Y/edit?usp=sharing

Table of Content

1. **EDA**
    - EDA.r & EDA_Demo.rmd/html:
        - All exploratory analyses on features provided grouping's rationale
    - EDA2.r:
        - The final exploratory analysis, data cleaning and grouping being used, with the field 'industry'(df2.csv), generate df_core.csv, df_RF.csv, df_oc_category.csv, df_clean.csv
    - EDA_loc.r:
        - Conversion by State Group source code and map
2. **Logistic_all_trials**
    - Logistic.r:
        - Check logistics assumptions, apply the SMOTE oversampling method
    - Logistic_weight.r:
        - Compare the 3 sampling methods(weighted training/SMOTE/ROSE)
    - Logistic_clean.r:
        - Try 3 ways to correct Logistic assumptions and improve the weighted Logistic Model
3. **RF_all_trials**
    - RF.r:
        - All trials to solve imbalance data issue, tune parameters, generate feature importance, try Wilson Score continuity correction on avg_rating, aggregate across platforms, define customized parameter test function, and metrics plotting of RF(ROC, PR curve, TPR vs FPR curve, histogram)
    - RF_clean.r & RF_Demo.rmd:
        - final model before removing FB and aggregated features, include metrics plotting and interpretable analysis on RF(PDP plot, Most frequent deterministic feature, SHAP analysis, check Top5 FN error individual), try KNN and PCA, generate df_agg.csv
    - RF_noFB.r:
        - Try feature combinations after removing FB, try SMOTE on RF, and generate df_agg_noFB.csv
4. **Final_RF_Log_Compare**
    - RF_Log_compare.r:
        - Compare the final logistic model and random forest model with various sampling methods and features
5. **Final_Model_Logistic**
    - Final_Model:
        - The final logistic regression we applied, all the data manipulation procedures are included, and it is the main source for the presentation
    - Final_Model_impute:
        - Try to impute NA instead of replacing it as 0