Table of Contents

Candidate Choice Algorithms

The following algorithms were used on the final dataset. A total of 4 feature sets were created apart from a set containing all candidate columns.

i) Feature set obtained from the top 7 features of Random Forest Classifier.

Feature_set_1: ['transactions_amount', 'count_pay_attempt', 'nunique_beacon_type', 'count_user_stay', 'count_buy_click', 'profile_submit_count', 'sum_beacon_value']

ii) Feature set obtained from the top 7 features of Logistic Regression.

Feature_set_2: ['count_pay_attempt', 'nunique_device', 'nunique_report_type', 'count_buy_click', 'nunique_beacon_type', 'count_sessions', 'nunique_language']

iii) Feature sets obtained from Correlation Analysis.

Feature_set_3: ['sum_beacon_value', 'count_pay_attempt', 'count_buy_click', 'nunique_dob', 'nunique_language', 'nunique_report_type', 'nunique_device', 'transactions_amount']

Feature set 4: ['sum_beacon_value', 'count_pay_attempt', 'count_buy_click', 'nunique_report_type', 'nunique_device', 'transactions_amount']

iv) Feature set given by domain expert

Feature set 5: ['count_pay_attempt', 'count_buy_click', 'nunique_report_type', 'profile_submit_count']

The feature sets were scaled using **Max Absolute Scaler** which scaled each feature by its maximum absolute value and **Standard Scaler** which standardized data by removing the mean and scaling to unit variance.

Logistic Regression

Reasons: -

- The problem is a binary classification problem.
- Logistic Regression helps solve classification and **probability problems** i.e. it not only classifies the dependent but also gives us an estimated probability value of the classification belonging to the positive/negative class.
- The algorithm also yields importance scores of all features which can help us make a more efficient choice by choosing the top few features.

Random Forest

Reasons: -

- The primary reason and the one of the greatest qualities of Random Forest is that it is very easy to measure the relative **feature importance** of each feature on the prediction.
- Random Forest prevents overfitting because it creates decision tress on subset of data.
- It has several hyperparameters which one can tune to get a better predictive model.

Gradient Boosting

Reasons: -

- It is a generalised algorithm that works well for any classification task. Its predictive scores are often better than other the scores of other algorithms.
- It has several hyperparameters which can be tuned to get a better predictive model.

Stochastic Gradient Descent

Reasons: -

- Suggested by domain expert.

Conclusion based on Training various models

The following is a list of ***Top 35*** models chosen out of various experiments done with our final dataset.

model_name	feature_count	Balanced_Accuracy_test	Recall_test	Balanced_Accuracy_train	Recall_train	Fit_time	Score_time
GB feature_set_1	7	0.988	0.998	0.989	0.999	11.573	0.071
GB feature_set_1 StdScale	7	0.988	0.998	0.989	0.999	11.753	0.064
GB feature_set_1 MaxAbs	7	0.988	0.998	0.989	0.999	12.348	0.064
RF all features	13	0.986	0.994	0.996	0.999	8.9	0.364
RF feature_set_1	7	0.986	0.994	0.996	0.999	8.276	0.354
GB feature_set_4 StdScale	6	0.986	0.999	0.986	1	8.204	0.06
GB feature_set_4 MaxAbs	6	0.986	0.999	0.986	1	8.378	0.062
GB feature_set_3 StdScale	8	0.986	0.999	0.986	0.999	9.226	0.061
GB feature_set_3 MaxAbs	8	0.986	0.999	0.986	0.999	9.086	0.06
SGD feature_set_1	7	0.984	0.995	0.985	0.995	0.242	0.049
RF feature_set_3 StdScale	8	0.984	0.994	0.991	0.999	6.579	0.317
RF feature_set_3 MaxAbs	8	0.984	0.994	0.991	0.999	6.593	0.323
RF feature_set_4 MaxAbs	6	0.984	0.994	0.991	0.999	6.66	0.316
RF feature_set_4 StdScale	6	0.984	0.994	0.991	0.999	6.506	0.316
SGD all features	13	0.982	0.988	0.982	0.988	0.277	0.039
LR all features	13	0.963	0.943	0.962	0.942	24.904	0.035
LR feature_set_1	7	0.962	0.942	0.962	0.941	4.234	0.048
SGD feature_set_1 StdScale	7	0.949	0.915	0.95	0.916	0.33	0.038
SGD feature_set_3 StdScale	8	0.949	0.914	0.949	0.915	0.412	0.038
SGD feature_set_4 StdScale	6	0.948	0.914	0.948	0.914	0.329	0.037
LR feature_set_4 StdScale	6	0.941	0.898	0.941	0.898	0.448	0.035
LR feature_set_1 StdScale	7	0.941	0.897	0.941	0.897	0.553	0.036
LR feature_set_3 StdScale	8	0.941	0.898	0.941	0.898	0.51	0.035
GB feature_set_5 StdScale	4	0.93	0.888	0.932	0.89	6.352	0.063
GB feature_set_5 MaxAbs	4	0.93	0.888	0.932	0.89	6.227	0.062
RF feature_set_5 MaxAbs	4	0.93	0.891	0.936	0.896	6.107	0.336
RF feature_set_5 StdScale	4	0.93	0.89	0.936	0.896	6.002	0.333
LR feature_set_3 MaxAbs	8	0.929	0.872	0.929	0.872	1.115	0.035
LR feature_set_4 MaxAbs	6	0.929	0.872	0.929	0.872	0.907	0.035
LR feature_set_5 MaxAbs	4	0.929	0.871	0.929	0.871	0.518	0.035
LR feature_set_1 MaxAbs	7	0.929	0.871	0.929	0.871	1.046	0.036
LR feature_set_5 StdScale	4	0.928	0.871	0.929	0.871	0.27	0.035
GB feature_set_2 MaxAbs	7	0.928	0.872	0.93	0.873	6.722	0.06
GB feature_set_2 StdScale	7	0.928	0.872	0.93	0.873	6.829	0.061

To make a selection of few algorithms from the above table, we consulted a domain expert who gave us certain threshold values for the training and testing scores.

Thresholds: -

- Training Balanced accuracy = <0.96
- Training Recall = <0.97
- Testing Balanced Accuracy = >=0.92
- Testing Recall = >=0.87

Based on the above thresholds, the following sets of models were obtained: -

model_name	feature_count	bac_test	rec_test	bac_train	rec_train	time_fit	time_score
SGD feature_set_1 StdScale	7	0.949	0.915	0.95	0.916	0.33	0.038
SGD feature_set_3 StdScale	8	0.949	0.914	0.949	0.915	0.412	0.038
SGD feature_set_4 StdScale	6	0.948	0.914	0.948	0.914	0.329	0.037
LR feature_set_4 StdScale	6	0.941	0.898	0.941	0.898	0.448	0.035
LR feature_set_1 StdScale	7	0.941	0.897	0.941	0.897	0.553	0.036
LR feature_set_3 StdScale	8	0.941	0.898	0.941	0.898	0.51	0.035
GB feature_set_5 StdScale	4	0.93	0.888	0.932	0.89	6.352	0.063
GB feature_set_5 MaxAbs	4	0.93	0.888	0.932	0.89	6.227	0.062
RF feature_set_5 MaxAbs	4	0.93	0.891	0.936	0.896	6.107	0.336
RF feature_set_5 StdScale	4	0.93	0.89	0.936	0.896	6.002	0.333
LR feature_set_3 MaxAbs	8	0.929	0.872	0.929	0.872	1.115	0.035
LR feature_set_4 MaxAbs	6	0.929	0.872	0.929	0.872	0.907	0.035
LR feature_set_5 MaxAbs	4	0.929	0.871	0.929	0.871	0.518	0.035
LR feature_set_1 MaxAbs	7	0.929	0.871	0.929	0.871	1.046	0.036
LR feature_set_5 StdScale	4	0.928	0.871	0.929	0.871	0.27	0.035
GB feature_set_2 MaxAbs	7	0.928	0.872	0.93	0.873	6.722	0.06
GB feature_set_2 StdScale	7	0.928	0.872	0.93	0.873	6.829	0.061
GB feature_set_1	7	0.928	0.872	0.93	0.873	6.598	0.071
LR feature_set_1	7	0.928	0.871	0.929	0.871	1.141	0.046
LR feature_set_2 StdScale	7	0.928	0.871	0.929	0.871	0.347	0.035
LR feature_set_2 MaxAbs	7	0.928	0.871	0.928	0.871	1.226	0.037
SGD feature_set_1 MaxAbs	7	0.928	0.869	0.928	0.869	0.2	0.039
SGD feature_set_3 MaxAbs	8	0.928	0.869	0.928	0.869	0.226	0.04

Among the chosen algorithms, only SGD allows us to partial fit which will be necessary in our incremental training module. Additionally, SGD gives us a good score. Therefore, choosing models with SGD as the candidate algorithm:

model_name	feature_count	bac_test	rec_test	bac_train	rec_train	time_fit	time_score
SGD feature_set_1 StdScale	7	0.949	0.915	0.95	0.916	0.33	0.038
SGD feature_set_3 StdScale	8	0.949	0.914	0.949	0.915	0.412	0.038
SGD feature_set_4 StdScale	6	0.948	0.914	0.948	0.914	0.329	0.037
SGD feature_set_1 MaxAbs	7	0.928	0.869	0.928	0.869	0.2	0.039
SGD feature_set_3 MaxAbs	8	0.928	0.869	0.928	0.869	0.226	0.04
SGD feature_set_4 MaxAbs	6	0.928	0.869	0.928	0.869	0.201	0.037
SGD feature_set_2 StdScale	7	0.928	0.869	0.928	0.868	0.394	0.037
SGD feature_set_5 MaxAbs	4	0.928	0.868	0.928	0.868	0.173	0.038
SGD feature_set_1	7	0.928	0.868	0.928	0.868	0.502	0.048
SGD feature_set_2 MaxAbs	7	0.928	0.868	0.928	0.868	0.223	0.037
SGD feature_set_5 StdScale	4	0.928	0.868	0.928	0.868	0.291	0.037

Hyperparameter tuning of chosen Algorithms

The models chosen for hyperparameter tuning were: ***SGD feature_set_1 StdScaled, SGD feature_set_3 StdScaled, SGD feature_set_4 StdScaled, SGD feature_set_5 StdScaled***

The four models chosen after training were tested for various combinations of hyperparameters using Grid Search Cross Validation. We used the following parameter grids: -

- **param_grid_sdc** = {'penalty': ['l2', 'l1', 'elasticnet'], 'alpha': [0.00001, 0.0001, 0.01, 0.1], 'class_weight': [None, 'balanced']} for Stochastic Gradient Descent Classifier with loss 'log'.

After, running Grid Search, these are the tuned hyperparameters we obtained.

model	best hyperparameters
SGD feature_set_1 StdScaled	{'alpha': 1e-05, 'class_weight': 'balanced', 'penalty': 'l1'}
SGD feature_set_3 StdScaled	{'alpha': 1e-05, 'class_weight': 'balanced', 'penalty': 'l1'}
SGD feature_set_4 StdScaled	{'alpha': 1e-05, 'class_weight': 'balanced', 'penalty': 'l1'}
SGD feature_set_5 StdScaled	{'alpha': 0.01, 'class_weight': 'balanced', 'penalty': 'l2'}

Cross Validating the tuned models

The tuning parameters we obtained were cross validated again to obtained the following: -

model_name	feature_count	Balanced_Accuracy_test	Recall_test	Balanced_Accuracy_train	Recall_train	Fit_time	Score_time
SDC1 feature_set_1 std scaled tuned	7	0.979	0.98	0.98	0.98	5.232	1.266
SDC1 feature_set_3 std scaled tuned	8	0.976	0.977	0.976	0.977	3.122	0.109
SDC1 feature_set_4 std scaled tuned	6	0.968	0.957	0.969	0.959	2.232	0.062
SDC2 feature_set_3 std scaled tuned	8	0.956	0.93	0.955	0.929	2.796	0.094
SDC2 feature_set_4 std scaled tuned	6	0.956	0.93	0.955	0.928	2.702	0.078
SDC2 feature_set_1 std scaled tuned	7	0.955	0.927	0.956	0.927	2.749	0.109
SDC3 feature_set_4 std scaled tuned	6	0.929	0.874	0.929	0.874	0.625	0.134
SDC3 feature_set_5 std scaled tuned	4	0.929	0.874	0.929	0.874	0.578	0.094
SDC3 feature_set_3 std scaled tuned	8	0.929	0.873	0.929	0.873	0.687	0.062
SDC1 feature_set_5 std scaled tuned	4	0.929	0.872	0.928	0.871	1.734	0.078
SDC3 feature_set_1 std scaled tuned	7	0.928	0.872	0.928	0.872	0.828	0.062
SDC2 feature_set_5 std scaled tuned	4	0.928	0.871	0.928	0.872	2.499	0.109

Among the models above, we chose ***Stochastic Gradient Descent with feature set 1 Standard Scaled.*** To recall, the primary reason for not choosing the Logistic Regression model is that it doesn't allow us to perform a partial fit on the new data.

We also carry out unit tests with 2 other models to make sure that the variance in testing scores is the minimum.