

Appendix

Performance Comparison of Models Trained with Various Churn Definitions

The area under ROC curve scores of the churn prediction models generated with respect to churn definitions are listed in Table 1.

Table 1. Predictive model scores for various churn definitions.

Label	Label Set	Data Set A1			Data Set A2			Data Set B1			Data Set B2		
		DMG	STC	ALL	DMG	STC	ALL	DMG	STC	ALL	DMG	STC	ALL
<i>inac-full</i>	SB	0.51	0.78	0.79	0.52	0.79	0.81	0.54	0.80	0.81	0.53	0.80	0.82
<i>inac-3m</i>	SB	0.53	0.76	0.78				0.54	0.77	0.78			
<i>inac-11m</i>	SB	0.54	0.74	0.76	0.53	0.75	0.77	0.54	0.74	0.75	0.51	0.77	0.79
<i>inac-12m</i>	SB	0.54	0.75	0.77	0.53	0.77	0.78	0.54	0.76	0.77	0.51	0.77	0.79
<i>inac-13m</i>	SB	0.53	0.76	0.77				0.54	0.76	0.79			
<i>inac-1m</i>	SB				0.53	0.74	0.77				0.51	0.76	0.77
<i>inac-2m</i>	SB				0.54	0.77	0.78				0.50	0.78	0.79
<i>che-low13m</i>	CA				0.59	0.74	0.75				0.57	0.73	0.74
<i>cc-zerostmtl3m</i>	CC				0.53	0.67	0.69				0.54	0.67	0.69
<i>cc-ltal3m-ccowned</i>	CC				0.54	0.71	0.72				0.53	0.71	0.72
<i>cc-lowstmtl3m</i>	CC				0.55	0.67	0.69				0.55	0.69	0.71
<i>cc-inacl3m</i>	CC				0.54	0.76	0.76				0.53	0.77	0.78
<i>cc-inacl3m-ccowned</i>	CC				0.53	0.73	0.73				0.51	0.74	0.75

Predictive model scores (area under ROC curve) for various churn definitions and data sets are presented. DMG, STC, and ALL stand for demographic feature set, spatio-temporal and choice behavior feature set, and combination of all the sets, respectively. Labels are grouped by the data source based on which they are generated. SB, CA, and CC stand for segmentation-based, checking account usage-based, and credit card usage-based label sets. Listed under the *Label* column are the abbreviations for the churn definitions. Not all the labels could be generated for each of the data sets due to different labeling window sizes of the data sets. For such label-data set combinations, the score fields are left blank.

Pair-wise Comparisons of Models

5

Table 2. Pair-wise comparison results for various churn definitions for data set A1.

Label	Label Set	DMG and STC			DMG and ALL			STC and ALL		
		t-score	df	p value	t-score	df	p value	t-score	df	p value
<i>inac-full</i>	SB	33.15	7	$p^{***} < 0.001$	36.20	7	$p^{***} < 0.001$	1.38	7	$p > 0.1$
<i>inac-3m</i>	SB	28.82	7	$p^{***} < 0.001$	29.13	7	$p^{***} < 0.001$	1.77	7	$p < 0.1$
<i>inac-11m</i>	SB	32.64	7	$p^{***} < 0.001$	36.61	7	$p^{***} < 0.001$	2.83	7	$p^* < 0.05$
<i>inac-l2m</i>	SB	28.24	7	$p^{***} < 0.001$	30.44	7	$p^{***} < 0.001$	2.15	7	$p^* < 0.05$
<i>inac-l3m</i>	SB	24.52	7	$p^{***} < 0.001$	25.14	7	$p^{***} < 0.001$	1.68	7	$p < 0.1$

Pair-wise prediction score comparisons of feature sets generated based on data set A1 are presented. DMG, STC, and ALL stand for demographic feature set, spatio-temporal and choice behavior feature set, and combination of all the sets, respectively. For each comparison pair, t-score, degree of freedom (df), and significance levels (p value) are listed. p^* , p^{**} , and p^{***} represent values smaller than 0.05, 0.01, and 0.001, respectively.

Table 3. Pair-wise comparison results for various churn definitions for data set A2.

Label	Label Set	DMG and STC			DMG and ALL			STC and ALL		
		t-score	df	p value	t-score	df	p value	t-score	df	p value
<i>inac-full</i>	SB	15.38	7	$p^{***} < 0.001$	18.36	7	$p^{***} < 0.001$	1.05	7	$p > 0.1$
<i>inac-11m</i>	SB	16.23	7	$p^{***} < 0.001$	18.47	7	$p^{***} < 0.001$	2.27	7	$p^* < 0.05$
<i>inac-l2m</i>	SB	20.90	7	$p^{***} < 0.001$	22.23	7	$p^{***} < 0.001$	1.04	7	$p > 0.1$
<i>inac-1m</i>	SB	25.38	7	$p^{***} < 0.001$	31.49	7	$p^{***} < 0.001$	4.88	7	$p^{***} < 0.001$
<i>inac-2m</i>	SB	17.84	7	$p^{***} < 0.001$	17.70	7	$p^{***} < 0.001$	0.92	7	$p > 0.1$
<i>cc-inacl3m</i>	CC	49.92	7	$p^{***} < 0.001$	50.71	7	$p^{***} < 0.001$	1.46	7	$p < 0.1$
<i>cc-inacl3m-ccowned</i>	CC	31.16	7	$p^{***} < 0.001$	35.33	7	$p^{***} < 0.001$	0.79	7	$p > 0.1$
<i>cc-ltal3m-ccowned</i>	CC	30.46	7	$p^{***} < 0.001$	34.59	7	$p^{***} < 0.001$	1.95	7	$p^* < 0.05$
<i>cc-zerostmt-l3m</i>	CC	16.25	7	$p^{***} < 0.001$	18.84	7	$p^{***} < 0.001$	2.54	7	$p^* < 0.05$
<i>cc-lowstmt-l3m</i>	CC	16.75	7	$p^{***} < 0.001$	21.72	7	$p^{***} < 0.001$	2.99	7	$p^* < 0.05$
<i>che-lowl3m</i>	CA	26.89	7	$p^{***} < 0.001$	27.78	7	$p^{***} < 0.001$	1.88	7	$p < 0.1$

Pair-wise prediction score comparisons of feature sets generated based on data set A2 are presented. DMG, STC, and ALL stand for demographic feature set, spatio-temporal and choice behavior feature set, and combination of all the sets, respectively. For each comparison pair, t-score, degree of freedom (df), and significance levels (p value) are listed. p^* , p^{**} , and p^{***} represent values smaller than 0.05, 0.01, and 0.001, respectively.

Table 4. Pair-wise comparison results for various churn definitions for data set B1.

Label	Label Set	DMG and STC			DMG and ALL			STC and ALL		
		t-score	df	p value	t-score	df	p value	t-score	df	p value
<i>inac-full</i>	SB	21.03	7	$p^{***} < 0.001$	23.82	7	$p^{***} < 0.001$	1.24	7	$p > 0.1$
<i>inac-3m</i>	SB	18.79	7	$p^{***} < 0.001$	19.84	7	$p^{***} < 0.001$	1.26	7	$p > 0.1$
<i>inac-11m</i>	SB	18.14	7	$p^{***} < 0.001$	21.14	7	$p^{***} < 0.001$	1.05	7	$p > 0.1$
<i>inac-12m</i>	SB	20.40	7	$p^{***} < 0.001$	25.04	7	$p^{***} < 0.001$	1.01	7	$p > 0.1$
<i>inac-13m</i>	SB	17.84	7	$p^{***} < 0.001$	21.13	7	$p^{***} < 0.001$	1.38	7	$p > 0.1$

Pair-wise prediction score comparisons of feature sets generated based on data set B1 are presented. DMG, STC, and ALL stand for demographic feature set, spatio-temporal and choice behavior feature set, and combination of all the sets, respectively. For each comparison pair, t-score, degree of freedom (df), and significance levels (p value) are listed. p^* , p^{**} , and p^{***} represent values smaller than 0.05, 0.01, and 0.001, respectively.

Table 5. Pair-wise comparison results for various churn definitions for data set B2.

Label	Label Set	DMG and STC			DMG and ALL			STC and ALL		
		t-score	df	p value	t-score	df	p value	t-score	df	p value
<i>inac-full</i>	SB	18.66	7	$p^{***} < 0.001$	22.91	7	$p^{***} < 0.001$	1.59	7	$p < 0.1$
<i>inac-11m</i>	SB	19.01	7	$p^{***} < 0.001$	22.30	7	$p^{***} < 0.001$	1.21	7	$p > 0.1$
<i>inac-12m</i>	SB	18.43	7	$p^{***} < 0.001$	28.23	7	$p^{***} < 0.001$	1.16	7	$p > 0.1$
<i>inac-1m</i>	SB	21.17	7	$p^{***} < 0.001$	23.14	7	$p^{***} < 0.001$	0.85	7	$p > 0.1$
<i>inac-2m</i>	SB	19.26	7	$p^{***} < 0.001$	24.00	7	$p^{***} < 0.001$	0.56	7	$p > 0.1$
<i>cc-inacl3m</i>	CC	26.13	7	$p^{***} < 0.001$	26.57	7	$p^{***} < 0.001$	0.43	7	$p > 0.1$
<i>cc-inacl3m-ccowned</i>	CC	23.56	7	$p^{***} < 0.001$	23.06	7	$p^{***} < 0.001$	0.12	7	$p > 0.1$
<i>cc-ltal3m-ccowned</i>	CC	36.58	7	$p^{***} < 0.001$	35.34	7	$p^{***} < 0.001$	1.52	7	$p < 0.1$
<i>cc-zerostmt-13m</i>	CC	14.27	7	$p^{***} < 0.001$	17.35	7	$p^{***} < 0.001$	1.84	7	$p < 0.1$
<i>cc-lowstmt-13m</i>	CC	17.68	7	$p^{***} < 0.001$	18.14	7	$p^{***} < 0.001$	3.50	7	$p^{**} < 0.01$
<i>che-lowl3m</i>	CA	45.40	7	$p^{***} < 0.001$	42.25	7	$p^{***} < 0.001$	4.12	7	$p^{**} < 0.01$

Pair-wise prediction score comparisons of feature sets generated based on data set B2 are presented. DMG, STC, and ALL stand for demographic feature set, spatio-temporal and choice behavior feature set, and combination of all the sets, respectively. For each comparison pair, t-score, degree of freedom (df), and significance levels (p value) are listed. p^* , p^{**} , and p^{***} represent values smaller than 0.05, 0.01, and 0.001, respectively.

Churn Definitions

inac-full The customer is labeled as churner if he/she was evaluated to be in segment *inactive* by the bank for all of the months during the labeling window of the data collection period.

inac-3m The customer is labeled as churner if he/she was evaluated to be in segment *inactive* by the bank for any three of the months of the labeling window of the data collection period.

inac-11m The customer is labeled as churner if he/she was evaluated to be in segment *inactive* by the bank for the last month of the labeling window of the data collection period.

inac-12m The customer is labeled as churner if he/she was evaluated to be in segment *inactive* by the bank for the last two months of the labeling window of the data collection period.

inac-13m The customer is labeled as churner if he/she was evaluated to be in segment

inactive by the bank for the last three months of the labeling window of the data collection period. 20

inac-1m The customer is labeled as churner if he/she was evaluated to be in segment *inactive* by the bank for any of the months of the labeling window of the data collection period. 21

inac-2m The customer is labeled as churner if he/she was evaluated to be in segment *inactive* by the bank for any two of the months of the labeling window of the data collection period. 22

che-lowl3m The customer is labeled as churner if monthly checking account statements with low balance have been issued for him or her during the labeling window of the data collection period. 23

cc-zerostmtl3m The customer is labeled as churner if monthly credit card statements with zero balance have been issued for him or her during the labeling window of the data collection period. 24

cc-ltal3m-ccowned The customer is labeled as churner if he/she owned a credit card and made purchases with very low amounts during the labeling window of the data collection period. 25

cc-lowstmtl3m The customer is labeled as churner if monthly credit card statements with low balance have been issued for him or her during the labeling window of the data collection period. 26

cc-inacl3m The customer is labeled as churner if he/she never used credit card, or his/her credit card was inactive during the labeling window of the data collection period. 27

cc-inacl3m-ccowned The customer is labeled as churner if he/she never used credit card, and his/her credit card was active during the labeling window of the data collection period. 28

Parameter Values Applied to Prediction Model 29

Random Forest Parameters 30

The Random Forest classification model has been implemented by employing the SciKit-learn library. The parameters are listed (and abbreviated) as in the library's documentation in order to support straightforward reproduction of our results. 31

Number of Estimators (`n_estimators`) The number of trees in the forest. Applied Value: 500 32

The Quality of Split Criterion (`criterion`) The function to measure the quality of a split. Applied Criterion: Gini Impurity 33

Maximum Number of Features (`max_features`) The number of features to consider when looking for the best split. Applied Value: 2 34

Maximum Depth of the Tree (`max_depth`) The maximum depth of the tree. Applied Value: No specific limit has been put on the depth of the tree leading to the expansion of the nodes until either all leaves are pure or all leaves contain less than *min_samples_split* samples. 35

Minimum Number of Samples for Split in Internal Node (`min_samples_split`) The minimum number of samples required to split an internal node. Applied Value: 2 36

Minimum Number of Samples at Leaf Node (`min_samples_leaf`) The minimum number of samples required to be at a leaf node. Applied Value: 1

Limit on the Leaf Node Number (`max_leaf_nodes`) Grow trees with *max_leaf_nodes* in the best-first fashion. Best nodes are defined as relative reduction in impurity. Applied Value: No limit has been applied.

Minimum Impurity for Split (`min_impurity_split`) Threshold for early stopping in tree growth. A node will split if its impurity is above the threshold, otherwise it is a leaf. Applied Value: 10^{-7}

SVM-SMOTE Parameters

Synthetic minority over-sampling has been implemented with Imbalanced-learn library by Lemaître et al. We considered it useful to report over-sampling parameters with their abbreviations as listed in the library’s documentation to support reproducibility of our study.

Minority and Majority Class Ratio (`ratio`) The number of samples in minority class over the number of samples in majority class. Applied Value: 0.25

Number of Nearest Neighbors (`k_neighbors`) Number of nearest neighbors that are used to construct synthetic samples. Applied Value: 5

Number of Nearest Neighbors for Danger Detection (`m_neighbors`) Number of nearest neighbors that are used to determine if a minority sample is in danger. Applied Value: 10

Extrapolation Step Size (`out_step`) The step size used in the extrapolation step. Applied Value: 0.5