Fifth Third Bank-UC Project

Customer Churn Model

April 2020

Team Members:
Aditi Singh | Anjali Gautam | Jeevisha Anandani | Pooja Sahare

Agenda

- Problem Statement & Approach
- Data Processing
- Model Comparison
- Final Recommendation
- Proposed Business Impact
- Appendix

Problem Statement & Approach

Classification models were built to predict customer churn using account balance and transaction data

Business Setting and Objective

- Checking accounts are a strong indicator of customer satisfaction
- Identify a checking account as likely to churn in the next two months

Methodology

- Built alternate linear and non-linear classifier models
- Compare models using performance metrices that are significant for business operations

Output

- 4 churn models
- Comparative analysis between the models
- Recommendation of final model based on business intuition and statistical results

Data Processing

Provided data was treated before modeling based on observations from initial data exploration

Data Filter → Additional Fields → Data Treatment

Data Filter

- Data is filtered for customer type <‘Consumer’> and Product type <‘Retail’>

Additional Fields

- Target
- Inactivity
- Recency
- Month over Month changes in account balances and transaction amounts
- % changes in the amount and quantity of transactions and balances respectively

Data Treatment

- Treated imbalanced data through under-sampling for modeling purposes

Model Comparison: Random Forest model provides better AUC and Recall

Model performance was assessed on test data (size: 750,949 accounts), and probability threshold for churn was used as 5%

Model	AUC*	Recall
Logistic Regression	0.72	0.72
Decision Tree Classifier	0.79	0.72
Random Forest	**0.78**	**0.78**
XGBoost Classifier	0.85	0.73

- **AUC** (Area Under Curve)
  - A higher AUC of a model depicts a better capability to distinguish between churners and non-churners

- **Recall**
  - It refers to the percentage of total actual churners that are correctly identified by a model
  - Recall is a significant measure of performance as it is important for the bank to be able to target as many churners as possible

* Industry benchmark for AUC is 0.7

Final Recommendation: Random Forest Model

Logistic Regression

- \+ Gives both magnitude and direction of association
- \+ Easier to implement, interpret and efficient to train
- − Lower AUC and lower recall
- − Higher number of false positives and false negatives

Decision Tree Classifier

- \+ Higher AUC
- \+ Lower number of false positives
- − Lower recall
- − Prone to model bias
- − Less stable and less accurate

Random Forest

- \+ Higher AUC and higher recall
- \+ Lower number of false positives and false negatives
- \+ Gives relative importance of input variables
- \+ Low model bias
- − Exact model structure is not known

XGBoost Classifier

- \+ Higher AUC
- \+ Lower number of false positives
- \+ Gives relative importance of input variables
- − Lower recall
- − Exact model structure is not known, prone to overfitting
- − Harder to train (more hyper-parameters to tune)

Given the comparison of models, we select Random Forest technique for predicting customer churn.

Proposed Business Impact: Minimum losses with Random Forest Model

Logistic Regression

Actual	Predicted		Total
	Not Churn	Churn	
Not Churn	71%	29%	100%
Churn	28%	72%	100%

Recall: 0.72

Decision Tree Classifier

Actual	Predicted		Total
	Not Churn	Churn	
Not Churn	78%	22%	100%
Churn	28%	72%	100%

Recall: 0.72

Random Forest

Actual	Predicted		Total
	Not Churn	Churn	
Not Churn	78%	22%	100%
Churn	22%	78%	100%

Recall: 0.78

XGBoost Classifier

Actual	Predicted		Total
	Not Churn	Churn	
Not Churn	81%	19%	100%
Churn	27%	73%	100%

Recall: 0.73

Assuming it takes $10 to solicit a customer identified as a churner while it costs ~$1500 of business value by losing a customer, we observe the below impact (dollar losses) of our models:

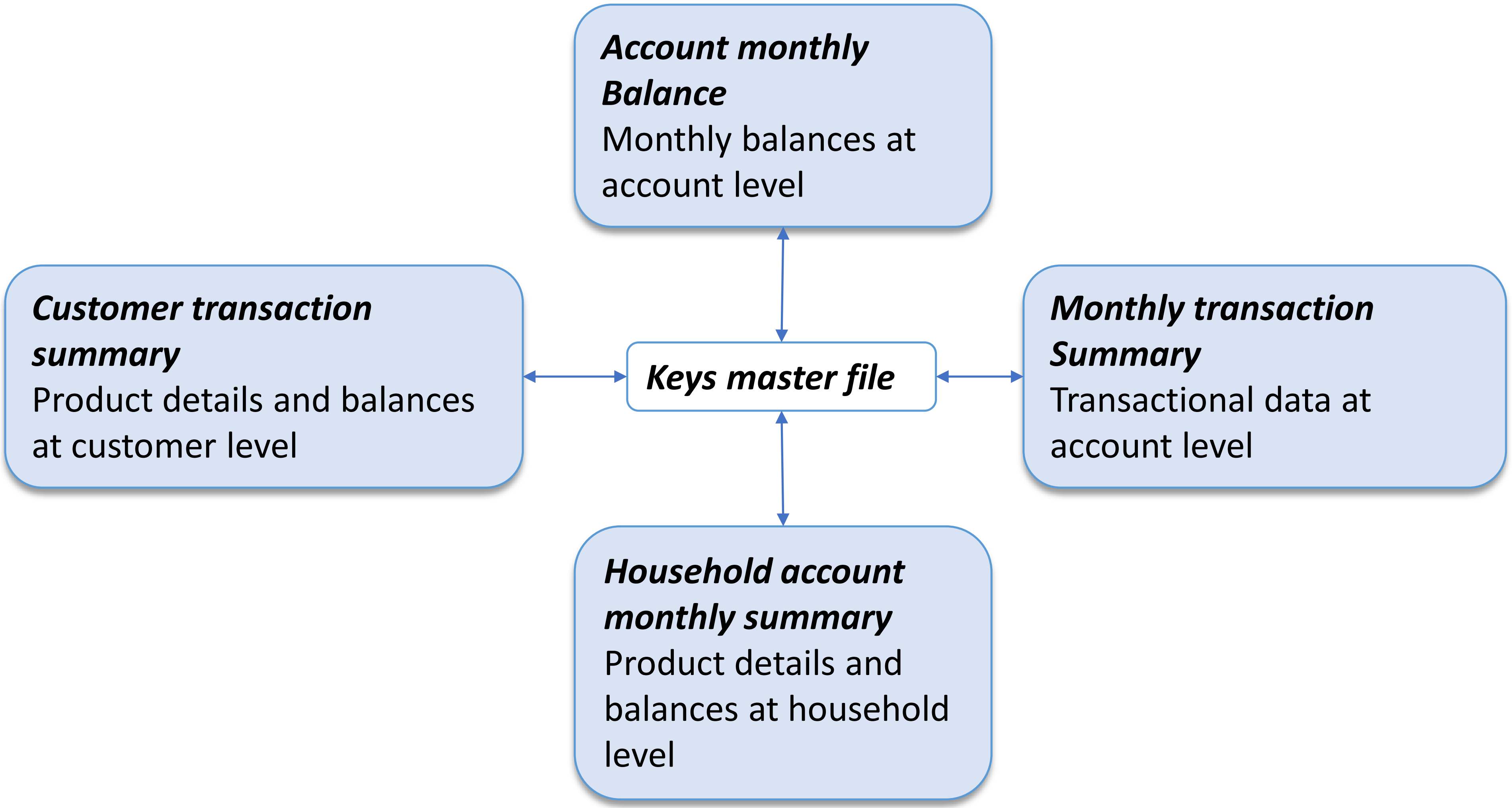
- *Loss from Random Forest Model: (22%* 7,42,015* $10) + (22%* 8,934 *$1,500) =* ***$4,580,653***
- *Loss from Logistic Regression Model: (29%* 7,42,015* $10) + (28%* 8,934 *$1,500) =* ***$5,904,124***

The recommended Random Forest technique has the least dollar losses among all models.

Appendix

Data Wrangling

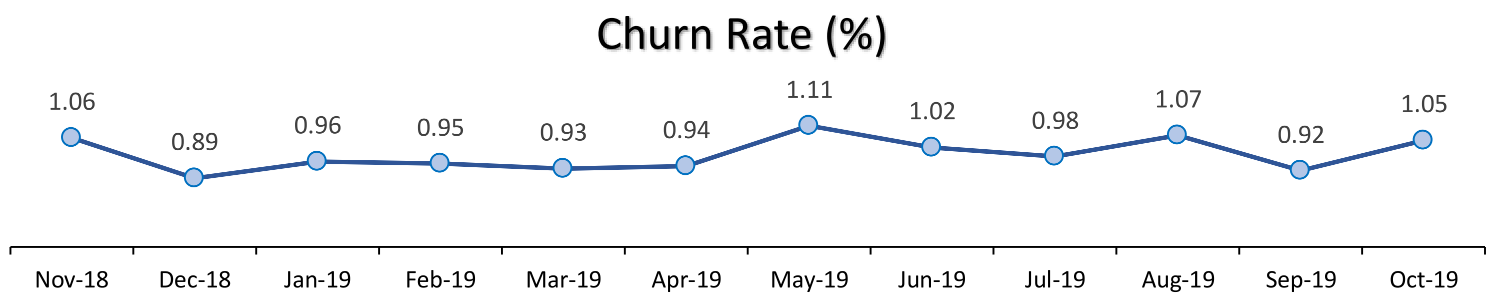
The master file is built using variables from the following files:



Data Wrangling

Target Definition

- Target population for this analysis is customers closing checking accounts with the bank
- Targets are marked using a churn window of 2 months i.e. if a customer is churned in Dec'19 it is marked as a target in Oct'19
- On an average the churn rate per month is 1% (~30,000 customers)



Under-sampling

- In order to boost target population for modeling, we have performed under-sampling of data to obtain 5% targets
- After this process, the data consists of approximately 29,767 targets against 625,107 total observations
- Further, a random sample of 80,000 records is used to train the data

Model Results: Logistic Regression

- Model Output: Churn Scores & Classification (churn/not churn)
- Hyper-parameters: Used 0.05 as Probability cut- off (same as target rate)
- Out of Sample Performance measures:
  - AUC=0.72 *
  - Recall=0.72 **
  - Confusion Matrix

Actual	Predicted		
	0	1	Total
0	**529,018**	**212,997**	742,015
1	**2,488**	**6,446**	8,934
Total	531,506	219,443	750,949

```
                              Results: Logit
=====================================================================
Model:                Logit              Pseudo R-squared:  0.053
Dependent Variable:   target             AIC:               29047.4245
Date:                 2020-04-13 19:28   BIC:               29205.3535
No. Observations:     80013              Log-Likelihood:    -14507.
Df Model:             16                 LL-Null:           -15318.
Df Residuals:         79996              LLR p-value:       0.0000
Converged:            1.0000             Scale:             1.0000
No. Iterations:       9.0000
---------------------------------------------------------------------
                       Coef.  Std.Err.    z     P>|z|   [0.025  0.975]
---------------------------------------------------------------------
ACH_IN_MTD_QTY        -0.1723   0.0121 -14.2006 0.0000 -0.1961 -0.1486
ACH_OUT_MTD_QTY       -0.0458   0.0057  -8.0598 0.0000 -0.0570 -0.0347
CHK_WRITTEN_MTD_QTY   -0.1803   0.0105 -17.1719 0.0000 -0.2009 -0.1597
DEBIT_CARD_MTD_QTY    -0.0119   0.0010 -12.4266 0.0000 -0.0138 -0.0101
MOBILE_STD_DEP_QTY    -0.4070   0.0350 -11.6409 0.0000 -0.4755 -0.3385
diff_ACH_IN_QTY       -0.1067   0.0169  -6.3035 0.0000 -0.1399 -0.0735
diff_CHECK_WRITTEN     0.0476   0.0159   2.9974 0.0027  0.0165  0.0787
diff_DEBIT_CARD_QTY   -0.0092   0.0016  -5.7638 0.0000 -0.0124 -0.0061
inactive_months       -0.2311   0.0174 -13.3142 0.0000 -0.2651 -0.1971
CONS_LOAN_WAR_PCT     -0.0146   0.0029  -5.0713 0.0000 -0.0203 -0.0090
CONS_DEPOSIT_ACCT_QTY -0.3199   0.0156 -20.5730 0.0000 -0.3504 -0.2895
%diff_AVG_MONTHLY_BAL -0.1066   0.0112  -9.5498 0.0000 -0.1285 -0.0847
%diff_LAST_STMT_BAL   -0.0082   0.0028  -2.8878 0.0039 -0.0137 -0.0026
recency               -0.0076   0.0002 -36.7549 0.0000 -0.0080 -0.0072
DIRECT_DEP_IND_Y      -0.4693   0.0425 -11.0527 0.0000 -0.5525 -0.3861
ACTIVE_CHK_IND_Y      -0.5593   0.0412 -13.5636 0.0000 -0.6401 -0.4784
HABITUAL_OD_IND_Y      1.0363   0.0403  25.7004 0.0000  0.9573  1.1153
=====================================================================
```

* Industry benchmark for AUC=0.7

** Recall=6,446/8,934

Metadata

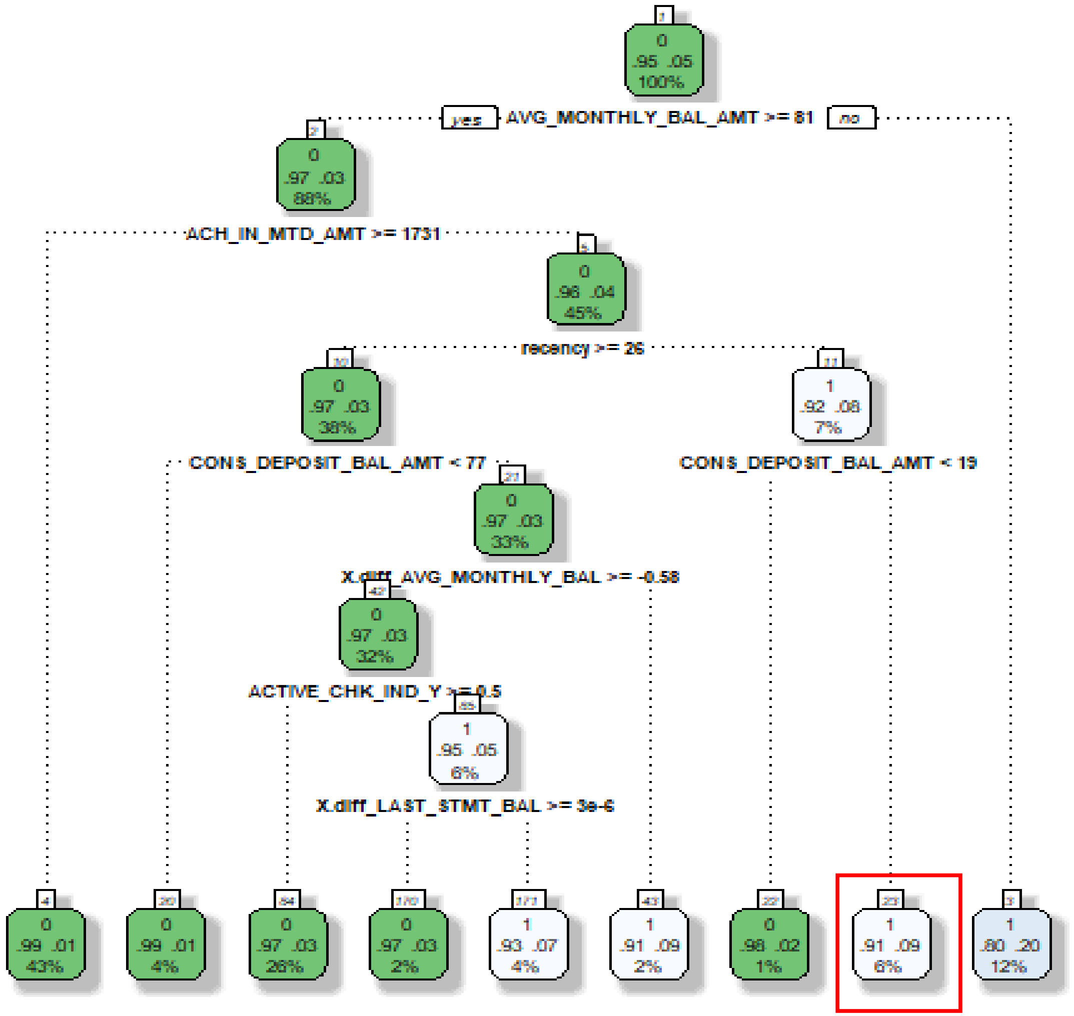
Definitions and signs of variables entering the logistic regression model

Variable Name	Definition	Sign (in final model)
ACH_IN_MTD_QTY	Number of electronic payments received till date for an account.	-
ACH_OUT_MTD_QTY	Number of electronic payments outgoing for an account till date.	-
CHK_WRITTEN_MTD_QTY	Number of cheques written till date for an account.	-
DEBIT_CARD_MTD_QTY	Number of debit card accounts till date for an account.	-
MOBILE_STD_DEP_QTY	Number of deposits through mobile till date for an account.	-
diff_ACH_IN_QTY	Difference in number of electronic payments received till date for an account for current month versus previous month.	-
diff_CHECK_WRITTEN	Difference in number of cheques written till date for an account for current month versus previous month.	+
diff_DEBIT_CARD_QTY	Difference in number of debit card accounts till date for an account for current month versus previous month.	-
inactive_months	Total number of consecutive months when the account has been inactive for the most recent period of inactivity.	-
CONS_LOAN_WAR_PCT	Consumer Loan Weighted Average Percent Rate.	-
CONS_DEPOSIT_ACCT_QTY	Total number of consumer deposit accounts for the customer.	-
%diff_AVG_MONTHLY_BAL	Percetage difference in the average monthly balance of an account for current month versus previous month.	-
%diff_LAST_STMT_BAL	Percetage difference in the last statement balance of an account for current month versus previous month.	-
recency	Difference between 'oldest open date' for a customer or month of first relationship for a customer and the observation month.	-
DIRECT_DEP_IND_Y	Indicates if a customer has a direct deposit account or not.	-
ACTIVE_CHK_IND_Y	Indicates if a customer has an active account or not (where activity is measured as per rules outlined by the bank).	-
HABITUAL_OD_IND_Y	Indicates if a customer has used overdraft facility or not (where habitual qualifies for a certain number of overdrat instances as decided by the bank).	+

Model Results: Decision Tree

- Model Output: Classification (churn/not churn)
- Hyper-parameters:
  - 0.05 as Probability cut- off (same as target rate)
  - Number of trees in forest = 100
  - Maximum depth of tree = 18
  - Minimum samples to split = 5
  - Minimum samples in leaf = 4
- Out of Sample Performance measures:
  - AUC=0.79
  - Recall=0.72

Actual	Predicted		
	0	1	Total
0	**579,879**	**162,136**	742,015
1	**2,526**	**6,408**	8,934
Total	582,405	168,544	750,949



Buckets labelled as 0 'non-churn'

Buckets labelled as 1 'churn'

Interpretation: An account having low avg. monthly balance (<$80), high automatic payment amount (> $1731), older than 26 months and low deposit balance amount (<$19) is likely to churn in next 2 months

Model Results: Advanced Models

RANDOM FOREST CLASSIFIER

- Uses average prediction of multiple decision trees built using bootstrapped sample of observations as well as features
- Model Output: Classification (churn/not churn)
- Hyper-parameters:
  - 0.05 as Probability cut- off (same as target rate)
  - Number of trees in forest = 100
  - Maximum depth of tree = 18
  - Minimum samples to split = 5
  - Minimum samples in leaf = 4
- Out of Sample Performance measures:
  - AUC=0.78
  - Recall=0.78
  - Confusion Matrix

Actual	Predicted		
	0	1	Total
0	5,76,927	1,65,088	7,42,015
1	1,977	6,957	8,934
Total	5,78,904	1,72,045	7,50,949

XGBoost CLASSIFIER

- It builds one tree at a time, where each new tree helps to correct errors made by previously trained tree.
- Model Output: Classification (churn/not churn)
- Standardized the variables and chose hyper-parameters as:
  - learning rate = 0.1
  - maximum depth of tree = 3
  - number of estimators = 100
  - 0.05 as Probability cut- off (same as target rate)
- Out of Sample Performance measures:
  - AUC=0.85
  - Recall=0.79
  - Confusion Matrix

Actual	Predicted		
	0	1	Total
0	6,01,924	1,40,091	7,42,015
1	2,400	6,534	8,934
Total	5,78,904	1,72,045	7,50,949

Relative Variable Importance

Relative Variable Importance as determined by random forest technique

Variable	Importance
CHECKING_BAL_AMT	0.119
AVG_MONTHLY_BAL_AMT	0.096
CONS_DEPOSIT_BAL_AMT	0.089
%diff_AVG_MONTHLY_BAL	0.085
LAST_STMT_BAL_AMT	0.081
recency	0.067
%diff_LAST_STMT_BAL	0.050
ACH_IN_MTD_AMT	0.035
DEBIT_CARD_MTD_AMT	0.034
%diff_DEBIT_CARD	0.032
DEBIT_CARD_MTD_QTY	0.029
LAST_DIRECT_DEPOSIT_AMT	0.028
diff_DEBIT_CARD_QTY	0.027
%diff_ACH_IN	0.024
ACH_OUT_MTD_AMT	0.024
SAVINGS_BAL_AMT	0.019
HABITUAL_OD_IND_Y	0.018
%diff_ACH_OUT	0.017
CONS_DEPOSIT_ACCT_QTY	0.015
ACH_OUT_MTD_QTY	0.014
ACH_IN_MTD_QTY	0.014
diff_ACH_IN_QTY	0.013
inactive_months	0.012
%diff_CHECK_WRITTEN	0.008
CONS_LOAN_BAL_AMT	0.008
CHK_WRITTEN_MTD_QTY	0.007
CONS_LOAN_WAR_PCT	0.007
ACTIVE_CHK_IND_Y	0.006
DIRECT_DEP_IND_Y	0.005
diff_CHECK_WRITTEN	0.005
CREDIT_CARD_BAL_AMT	0.005
CHK_WRITTEN_per_trans	0.004
MOBILE_STD_DEP_QTY	0.002
MORTGAGE_BAL_AMT	0.001

Model Assumptions & Codes

Additional Modeling Assumptions

- Target: The accounts which are not present in the monthly account summary file after two months from the observation month
- Weights associated with false positive and false negative used in the classification model is taken to be 19:1 (FN:FP)

Statistical Assumptions

- 250,000 accounts per month including targets and non targets sufficiently represent actual population
- Oversampling of targets will be required for model accuracy

Codes for Churn Models

Logistic Regression.ipynb

Random Forest.ipynb

XGBoost.ipynb

Decision Tree Classifier.R