AFCU 2019
Churn Analytics Competition

Team Number: 6

Yuan Du, Jianbin Zhu

Department of Statistics & Data Science

OUTLINE

- **Problem Statements**
- **Data Analysis**
- **Model and Approach**
- **Results**
- **Recommendations**
- **Conclusions**

Problem Statements

- **The goal of this competition is to develop a model that can correctly predict if a member is about to churn by using AFCU customer historical data.**
- **We designed two objectives:**

  1) To predict customer churn as a given date (Model-1)

  7/31/2019 in this problem

  2) To predict customer churn in the early date (Model-2)

  At month 3, 6 and 9 after the open date

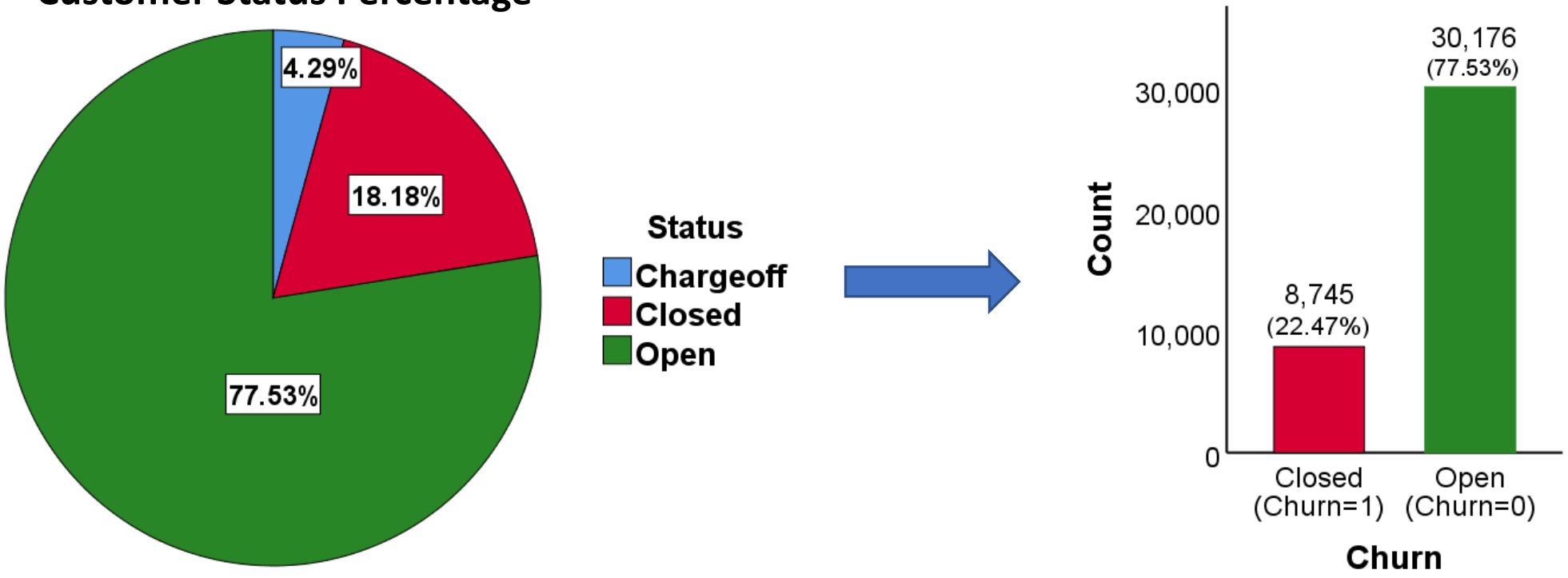
Data Analysis - Datasets

	Name	Dimensions (Rows x Cols)	Variables
Training datasets (4)	Summary	**38,921** x 12	Memberid, **Status**, OpenDate, Age, Tenure,…
	Online Banking	220,511 x 3	Memberid, MonthEndDate, + 1 acctivity
	Loan Transaction	228,104 x 5	Memberid, MonthEndDate, + 2 acctivities
	Transaction	482,826 x 10	Memberid, MonthEndDate, + 8 acctivities
Test datasets (4)	Summary	**9730** x 9	The same as training dataset except Churn Status (Target variable)
	Online Banking	54,388 x 3	
	Loan Transaction	56,522 x 5	
	Transaction	121,264 10	

Data Analysis – Define Churn

- **Target variable:** { Closed & Charge Off: Churn =1 ; Open: Churn=0 }

Customer Status Percentage

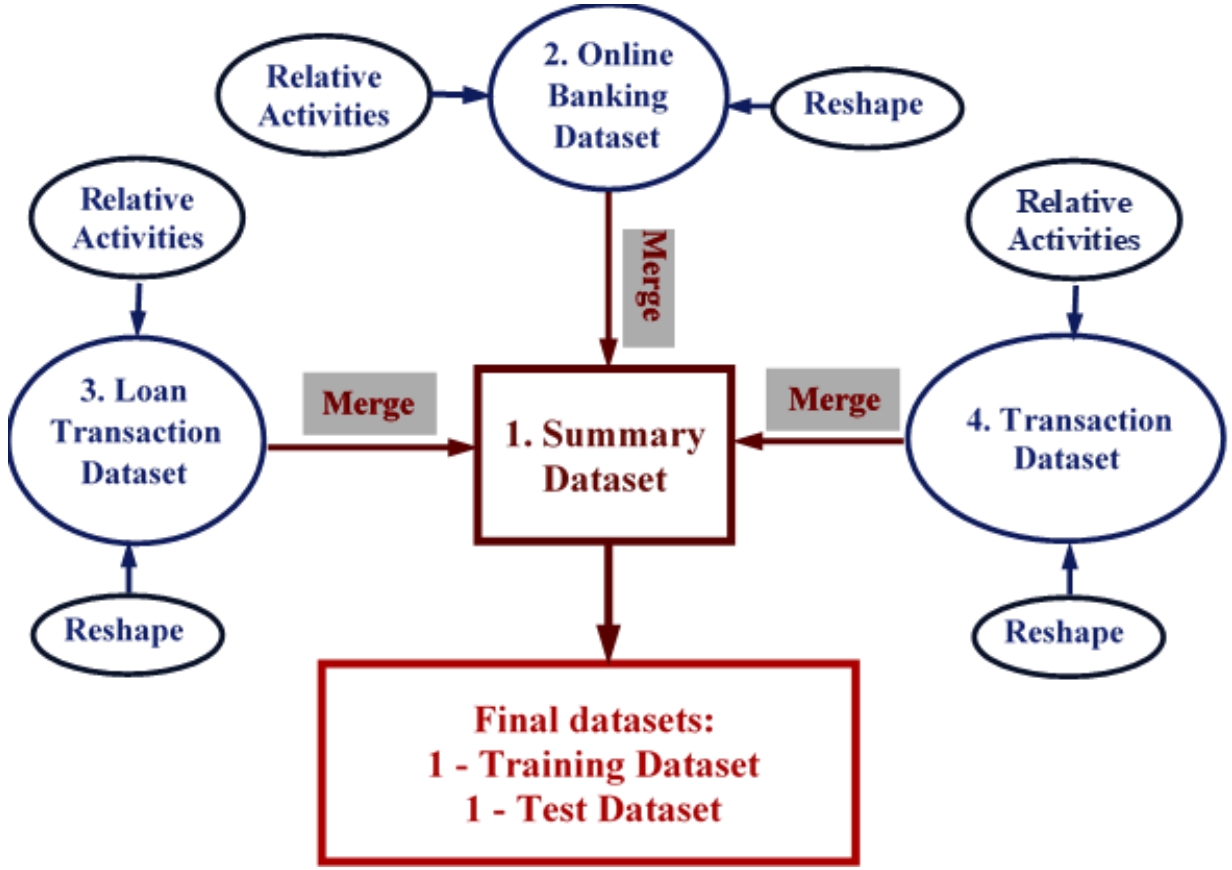


Data Analysis – Data Merging(1/2)

- **Data merging plan**

- Develop an R function to fulfill this plan
- Primary Key: Memberid

- Final dataset:
  - Includes all variables from Summary Dataset.
  - New variables from three Transaction Datasets.

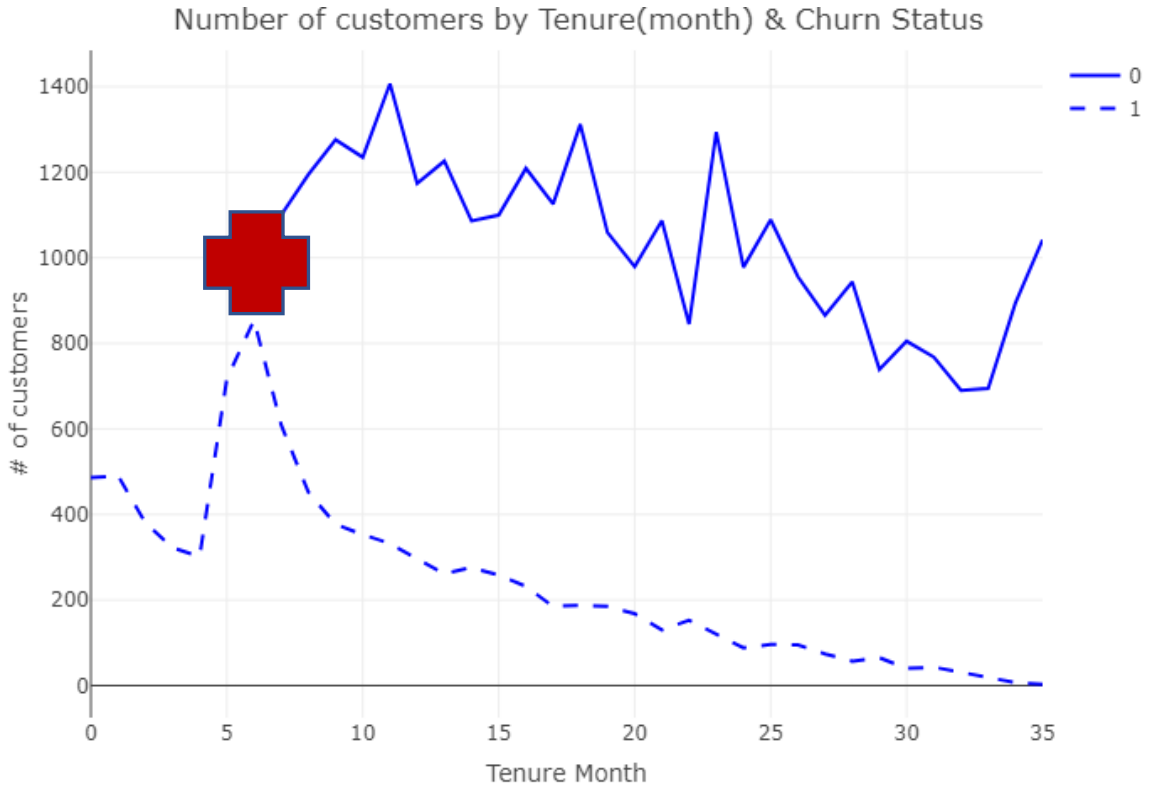


Data Analysis – Data Merging(2/2)

- **Data merging training cut-off**

- Churn happens on the first month and pick at month 6

- Customer whose account is open ranges from 7 to 35 months.



Data Analysis – New Variables

- **Model 1 :** To create three sets of segments variables at 1, 4,7 (11 new variables in each set)

Given date

Month 1 Month 4 Month 7

Memberid	Online0	Online1	Online2	Online3	Online4	Online5	Online6	Online7
4444R4MGG	0	0	0	0	1	0	0	0
444F4RR4K	0	0	0	0	0	0	2	1
4458RZK85	4	6	13	0	8	9	7	9

Last open date: 12/31/2018

- **Model 2:** To create three sets of segments variables at Month 3, 6 and 9 (11 new variables in each Month) and predict using these three segments variables separately.

Open date

Month 3 Month 6 Month 9

Memberid	Online0	Online1	Online2	Online3	Online4	Online5	Online6	Online7	Online8	Online9
4444R4MGG	0	0	3	0	1	0	0	0	0	1
444F4RR4K	0	3	0	1	0	0	3	1	0	0

Model and Approach – Model Plan

- **Model-1:**

  - Include all variables from Summary Dataset and three sets of new variables.
  - Predict customer churn by using these three segments variables separately and compare to determine which set is the best predictors.
  - Obtain one result of churn probability as the given date for the test dataset.

- **Model-2:**

  - Include all variables from Summary Dataset and all new variables of Month 3, 6 and 9.
  - Predict customer churn by using these three segments variables separately.
  - Obtain three results of churn probabilities for month 3, 6, and 9 for the test dataset.

Model and Approach – Variable Selection

- **Model 1:**

  Target variable: Churn
  Predictors: 16
  "Age","Tenure","DBINDICATOR","CCINDICATOR", "NumberofSavingsProducts", "INDIRECT", "Online_sum4","LoanNum_sum4", "LoanAmount_sum4", "NDirDep_sum4", "SumDirDep_sum4", "NBillPay_sum4", "NBranch_sum4", "FeeCharge_sum4", "SumFee_sum4", "NTotTrans_sum4"

  Drop variables: "CountofLoans" , high correlation with LoanNum_sum4 (0.828)
  " NDebit_sum4 “, high correlation with NTotTrans_sum4 (0.912)

- **Model 2:**

  Include the same types of variables.

Model and Approach – Data Partition (70/30)

- **Logistic Regression**
- **Decision Tree**
- **Random Forest**
- **XGBoost**

Results – Model-1 (1/3)

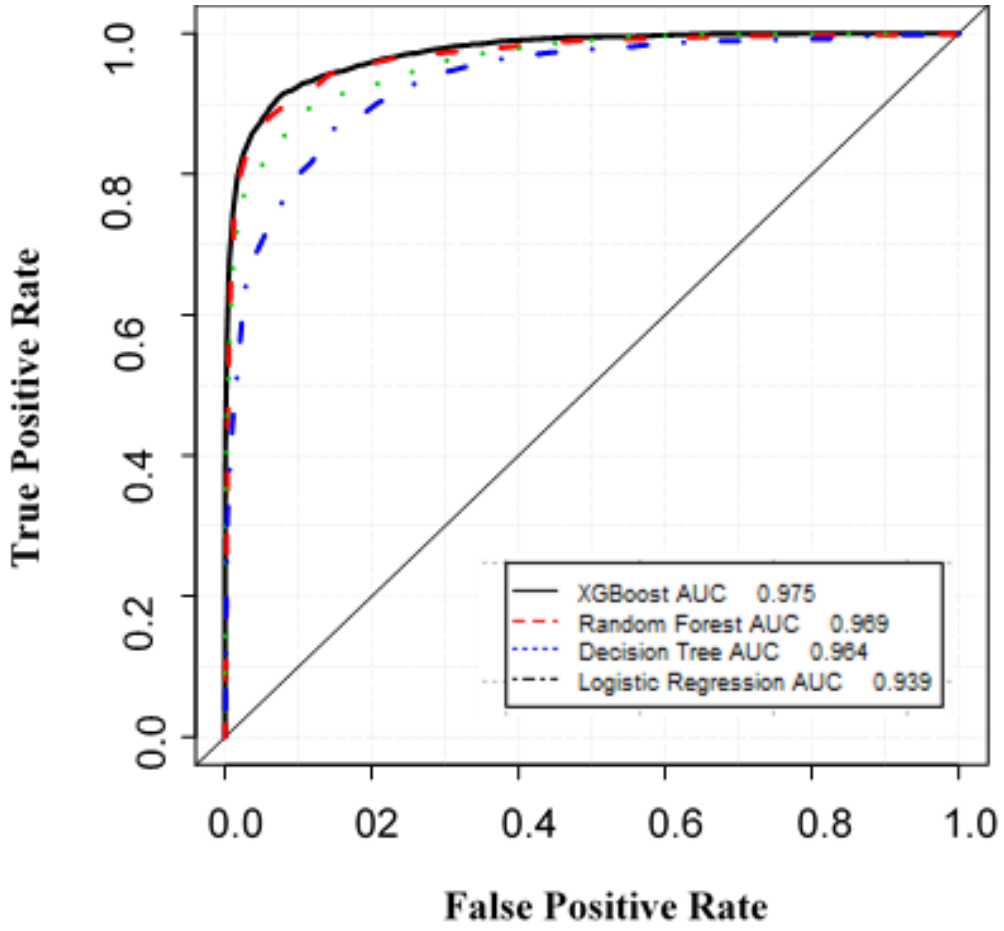
- **Predictive Model Performances**

	Logistic Regression		Decision Tree		Random Forest		XGBoost	
	Train	Validation	Train	Validation	Train	Validation	Train	Validation
Accuracy	88.5%	88.8%	93.7%	93.6%	96.2%	94.4%	96.8%	94.6%
Misclassification	11.5%	11.2%	6.3%	6.4%	3.8%	5.6%	3.2%	5.4%
True Positive Rate	73.3%	75.3%	83.0%	83.1%	91.7%	84.2%	93.0%	85.1%
False Positive Rate	7.1%	7.3%	3.3%	3.3%	2.4%	2.7%	2.1%	2.7%
Specificity	92.9%	92.7%	96.7%	96.7%	97.6%	97.3%	97.9%	97.3%
Precision:	74.9%	75.0%	88.0%	87.7%	91.4%	90.0%	92.9%	90.1%
Prevalence	22.5%	22.5%	22.5%	22.5%	22.5%	22.5%	22.5%	22.5%

Results – Model-1(2/3)

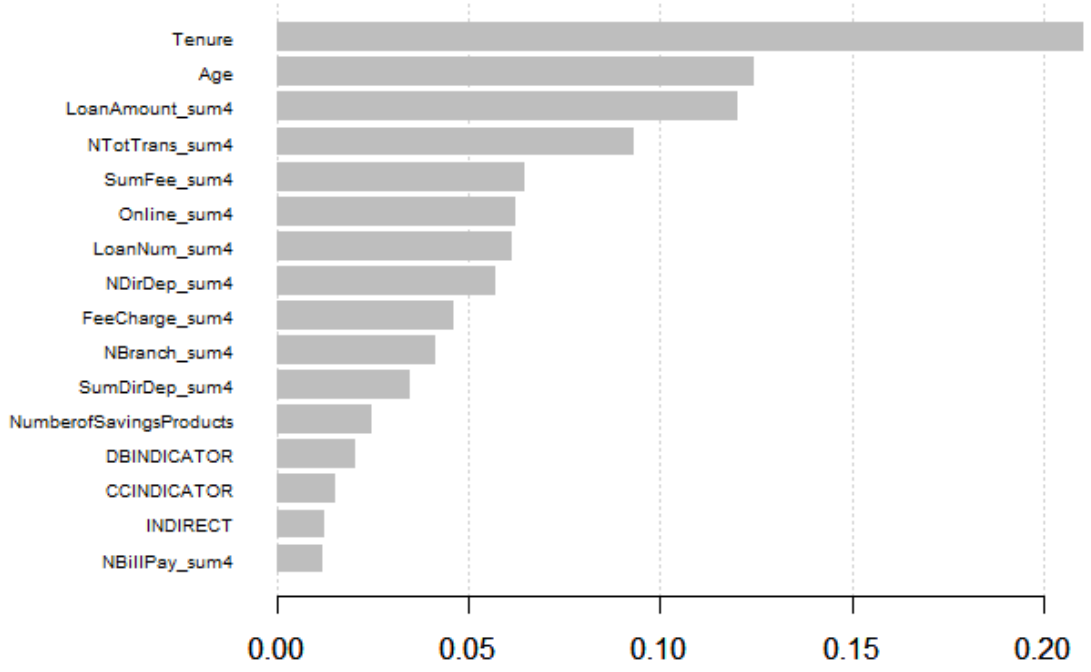
- **AUC of ROC Curves**

Approach	AUC	
	Train	Validation
XGBoost	0.994	0.975
Random Forest	0.986	0.969
Decision Tree	0.967	0.964
Logistic Regression	0.937	0.939



Results – Model-1(3/3)

- **Variable Importance**



Results – Model-2

- **Predictive model performances and AUC**

Model Approach	Months	Accuracy		AUC	
		Train	Validation	Train	Validation
XGBoost	Three	95.0%	87.0%	0.937	0.932
	Six	96.4%	88.0%	0.967	0.964
	Nine	97.3%	88.5%	0.986	0.969
Logistic Regression	Three	84.9%	85.1%	0.812	0.816
	Six	85.2%	85.6%	0.820	0.822
	Nine	85.5%	85.8%	0.828	0.832

Results – Prediction on test dataset

- **Confusion Matrix and Accuracy (Model-1 & Model-2)**

	Model -2						
Model-1	**Three Months**		**Six Months**		**Nine months**		
	Closed	**Open**	**Closed**	**Open**	**Closed**	**Open**	Total
Closed	**1,181**	**877**	**1,268**	**790**	**1,310**	**748**	2,058
Open	**273**	**7,399**	**246**	**7,426**	**232**	**7,440**	7,672
Total	**1,454**	**8,276**	**1,514**	**8,216**	**1,542**	**8,188**	**9,730**
Accuracy	88.18%		89.35%		89.92%		

Suggestions

- **Month 6 is the most critical point for customer churn.**
- **Predict individual customer churn probability at month 3, 6, 9 is recommended**
- **Rank customer based on probability and develop retention strategies such as easy hanging fruits on reducing number/amount of fee charges, promotions, etc.**
- **Collect additional social economics data to better develop retention strategies based on member's features.**

Conclusions

- We explored the data sets with data visualizations, designed the data merging plan and identified the members' activities.
- We developed two models, which can predict not only the probabilities of members' churn status as a given date, but also the probabilities of members' status of churn in three months, six month and nine months after the open date of members' accounts.
- Both models have higher accuracies and AUCs of ROC curve.
- The comparison of model approaches shows that XGBoost method is the best one.
- Our models can help AFCU to detect the early signs of members' churn. Thus, AFCU can take specific actions to prevent churn and dramatically improve the success of the retention offers to the potential churners.

Thank You!

By Team Number: 6