AFCU 2019 Customer Churn Analytics Competition

Project Report

Team ID_6

SECTION 1: EXECUTIVE SUMMARY

The goal of this project in AFCU 2019 Customer Churn Analytics Competition is to develop a model that can correctly predict if a customer is about to churn. Four training datasets, including one summary dataset of customers' information with their churn status and three transaction datasets of customers' activities in online banking, loans and accounts, were provided for training. And another four test datasets including the same information except the churn status were provided for prediction.

In our project, we developed not only a model to predict the probabilities of customers' status of churn as a given date in the problem(07/31/2019), but also built another model to predict the probabilities of customers' status of churn in three months, six month and nine months after the open date of customers' accounts. We summarize two models in the followings.

In the first model, the purpose is to predict if a customer is about to churn as a given date. We focused on customers' recent activities and used these activities as predictors. To join the four datasets, firstly, we used the given date as a beginning date and identified customers' recent activities backward until 7 months in each transaction dataset. Secondly, we aggregated these activities in different months and reshaped the dataset to the transposed structure that each member has a row with multiple variables with the activities of months. After that we joined the three transposed transaction datasets with the summary dataset and then grouped monthly activities variables to different segments. To build the model, we chose three cases of customers segments with 1 month, 4 months, and 7 months as our predictors to compare the modeling performances with different approaches like Logistic Regression, Decision trees, Random Forest and XGBoost. We found the case of the predictors of 4 months with XGBoost method was the best one. The model performance was excellent and has a higher accuracy of 94.6% and AUC of ROC curve with 97.4%. The most important ten factors are Tenure, Age, Sum of Payment in loan, Number of Total Transactions, Sum of Fees Charged, Number of Days accessing account in online banking,

Number of Payments in loan, Sum of Direct Deposits, Num of Fee Charged and Number of Transactions Conducted in Branch. In the last, we predicted the test data with this case and obtained the probabilities of customers' churn and churn status.

In the second model, the purpose is to predict the probability if a customer is about to churn at 3rd , 6th and 9th month after the open date, respectively. We focused on customers' open activities and used these activities as predictors. The data processing is the same as that of the first model, but we used the open date as a beginning date and identified customers' open activities forward until 9 months in each transaction dataset. In the modeling, we build three models with customers segments at 3rd , 6th and 9th month as our predictors with different approaches like Logistic Regression, and XGBoost, respectively. We found XGBoost method had high performance. The three models have accuracies with 87.0%, 88.0%, 88.5% and AUC of ROC curve 0.932, 0.964 and 0.969 respectively. We found the important factors are the same as our first model. Three predictions of the probabilities of customers' churn in 3, 6 and 9 months after open date of their accounts for the test data were reported.

Our models can help AFCU to detect the early signs of customers' churn. Thus, AFCU can take specific actions to prevent churn and dramatically improve the success of the retention offers to the potential churners. The following report summarizes data analysis and predictive modeling methods conducted to fulfill the objective of the project.

SECTION 2: DATA ANALYSIS

2.1 Problem Statements and Datasets

The goal of this project in AFCU 2019 Customer Churn Analytics Competition is to develop a model that can correctly predict if a customer is about to churn. In this study, we set two analysis questions in our modeling prediction: (1) what the predicted churn status of a customer member at a given date, 07/31/2019; (2) Can we early predict the probability of a customer's churn in three months, six months and nine months after the open date of the account. The two questions are the directions of our model building.

We used the datasets that AFCU provided. Four training datasets, including one summary dataset of customers' information with their churn status and three transaction datasets of customers' activities in online banking, loans and accounts, were provided for training. And another four test datasets including the same information except the churn status were provided for prediction.

2.2 Data Processing

2.2.1 Data exploration

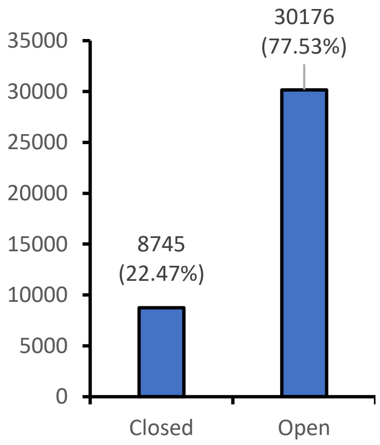
The first step we analyzed the data was to explore the four datasets. In summary dataset, each member has a row information of ID, Age, Tenure, OpenDate, Indicators of Debit card and credit card, Numbers of saving and Loan products, and indictor of indirect membership. Target variable - Status is in training Summary data. There are 1, 2 and 8 variables of transaction activities in the Online Banking, Loan Transaction and Transaction datasets, respectively. Each member has multiple rows of these activities at the date of MonthEndDate. The structures of four datasets of training and test are listed in Table 1.

Table 1: Data structures

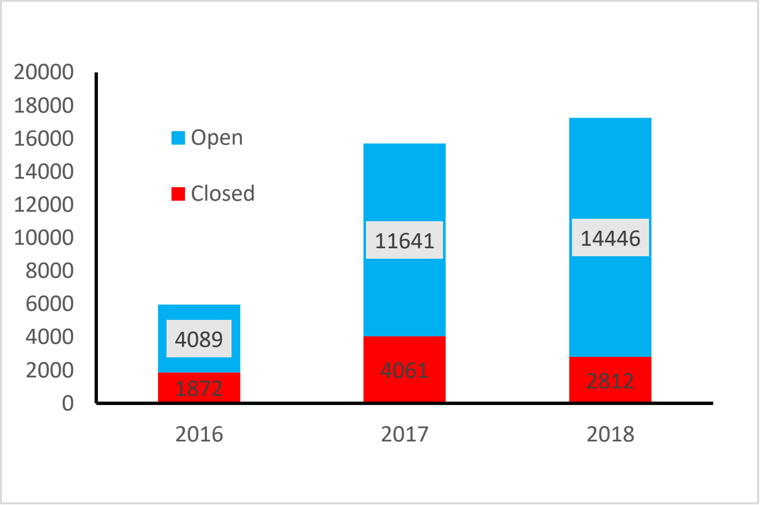
	Dataset Name	Obs	Variables number	Variables name
Training	Summary	38,921	12	Memberid, Status, OpenDate, CloseDate, ChargedOffDate, Age, Tenure, DBINDICATOR, CCINDICATOR, NumberofSavingsProducts, CountofLoans, INDIRECT
	Online Banking	220,511	3	Memberid, Count, MonthEndDate
	Loan Transaction	228,104	5	Memberid, NumberofPayments, SumofPaymentAmount, MonthEndDate, LoanCollateralCodeDesc
	Transaction	482,826	10	Memberid, NumberofDirectDeposits, SumofDirectDeposits, NumberofBillPayTransactions, numberofDebitCardTransactions, NumberofTransactionsConductedinBranch, FEESCHARGED, SUMofFEESCHARGED, NumberofTotalTransactions, MonthEndDate
Test	Summary	9730	9	Memberid, OpenDate, Age, Tenure, DBINDICATOR, CCINDICATOR, NumberofSavingsProducts, CountofLoans, INDIRECT
	Online Banking	54,388	3	The same as Traning
	Loan Transaction	56,522	5	The same as Traning
	Transaction	121264	10	The same as Traning

Here are some examples of figures of the target, customer summary and transaction activities in Figure 1 – Figure 5. Additional figures are in appendix.

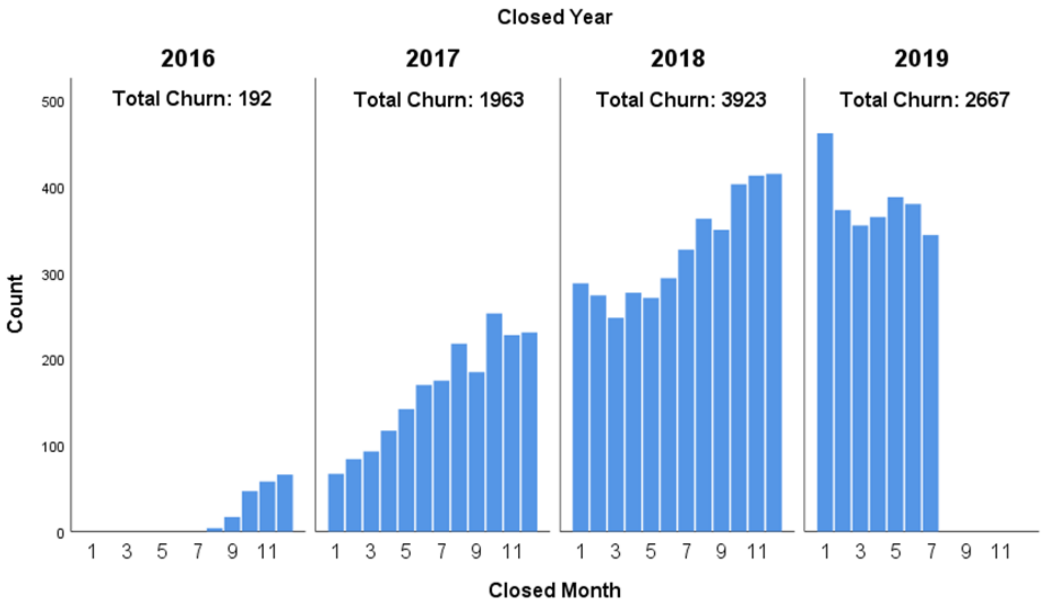
Target Variable: We created a new variable of Churn with the values of 0 and 1, which 0 is not Churn and 1 is Churn by grouping two levels of "charge-off" and "closed" from the Status variable Figure 1 (a) shows the frequency and percentage of the target variable - churn status. Figure 1 (b) and Figure 1(c) show the distribution of members' churn state with open year and close year respectively. As we can see, the total churn percentage of members is 22.47%. The churn number has an upward trend with the years.



(a)



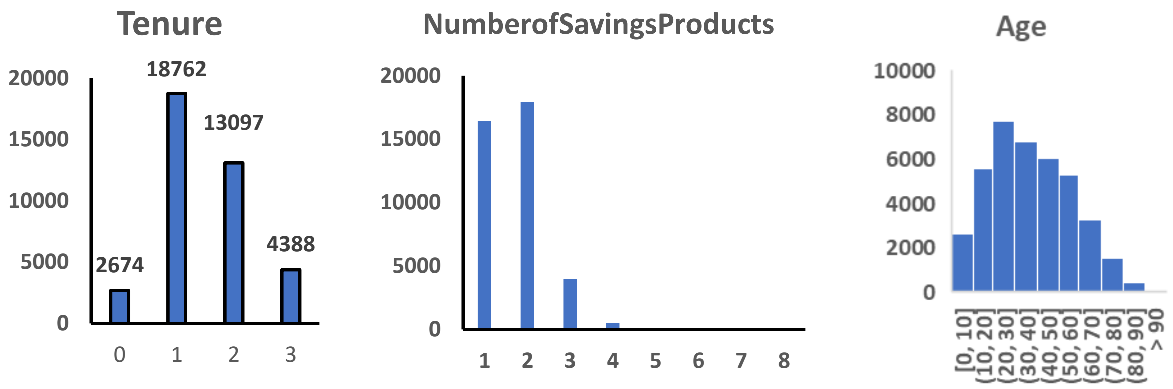
(b)



(c)

Figure 1: Target variable: Status (a) Frequency(%); Distribution with (b) open year; (c) close year.

Figure 2 shows two examples of bar charts of variables: Tenure and Number of Saving Products and an example of histogram of variable Age. Age has missing values which are replaced by median value of age.



(a) Bar charts (b) Histogram

Figure 2: Some examples of bar chats and histogram plot

Figure 3 shows summaries of customers' activities in online banking, Figure 4 shows summaries of customers' activities in Loan account and Figure 5 shows summaries of customers' activities in Total Transaction, respectively.

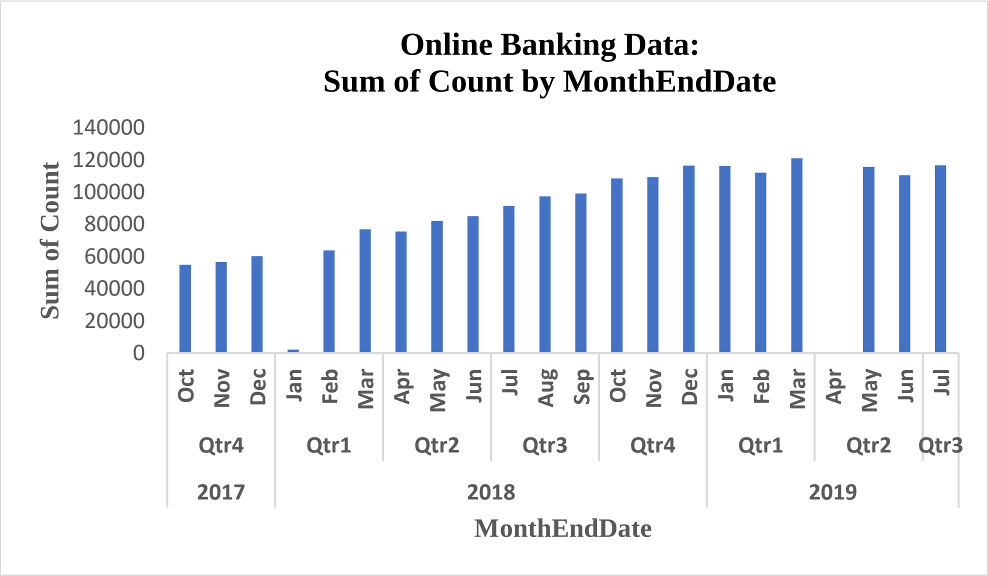


Figure 3: The summary of customers' activities of online banking

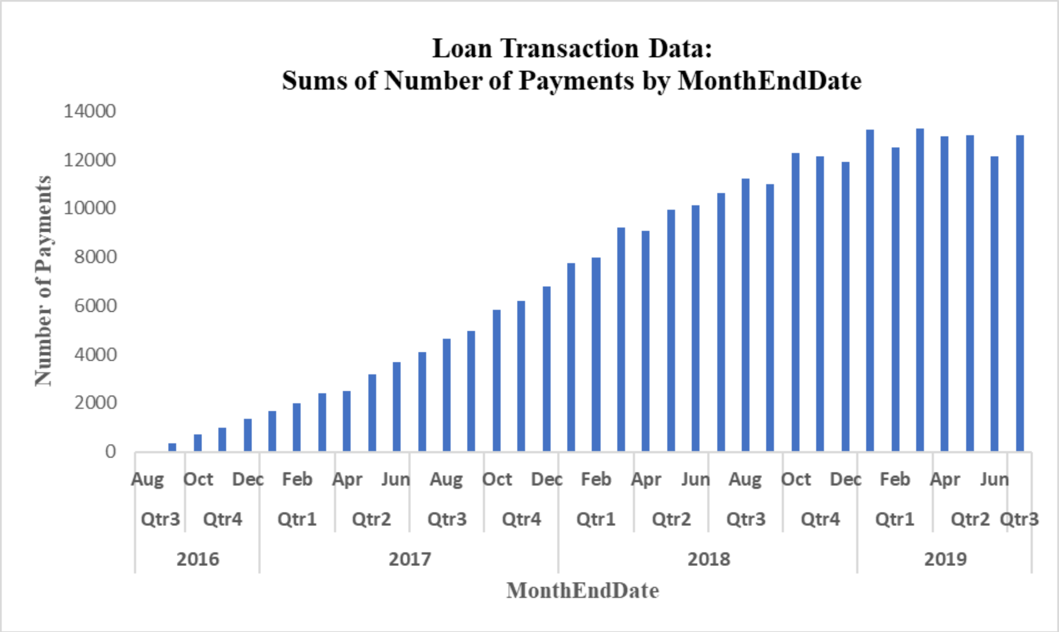


Figure 4: The summary of customers' activities of Loan Transaction data

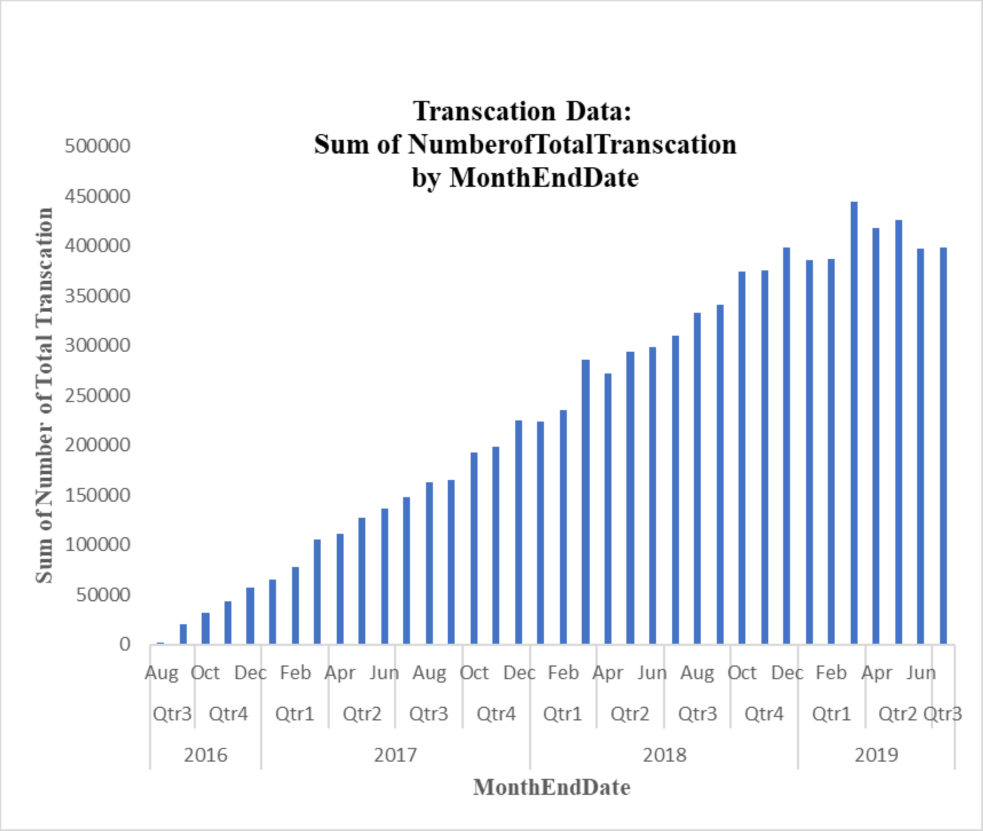


Figure 5: The summary of customers' activities of Transaction data

2.2.2 Data merging

After data exploration, the next step is to merge the four datasets. This is very challenging because what segments of the activities we need to consider is important to answer our analysis questions. For this step, we designed our data join plan as shown in Figure 6. It includes these steps: (1) Summary data is the primary dataset, three transaction datasets join the summary data by using the primary key of Memberid; (2) In each transaction data, first we identified customers' activities we considered, which will be shown details in next section, and then aggregated these activities in month. Then we reshaped the dataset to the wider transposed structure. Thus we obtained the transposed dataset in the structure that each member has one row with multiple variables of the monthly activities. (3) After that we joined the three transposed transaction datasets with the summary dataset and then grouped monthly activities variables to different segments for modeling building. (4) The final data is one train dataset and one test dataset. A R function was developed to carry out these steps.

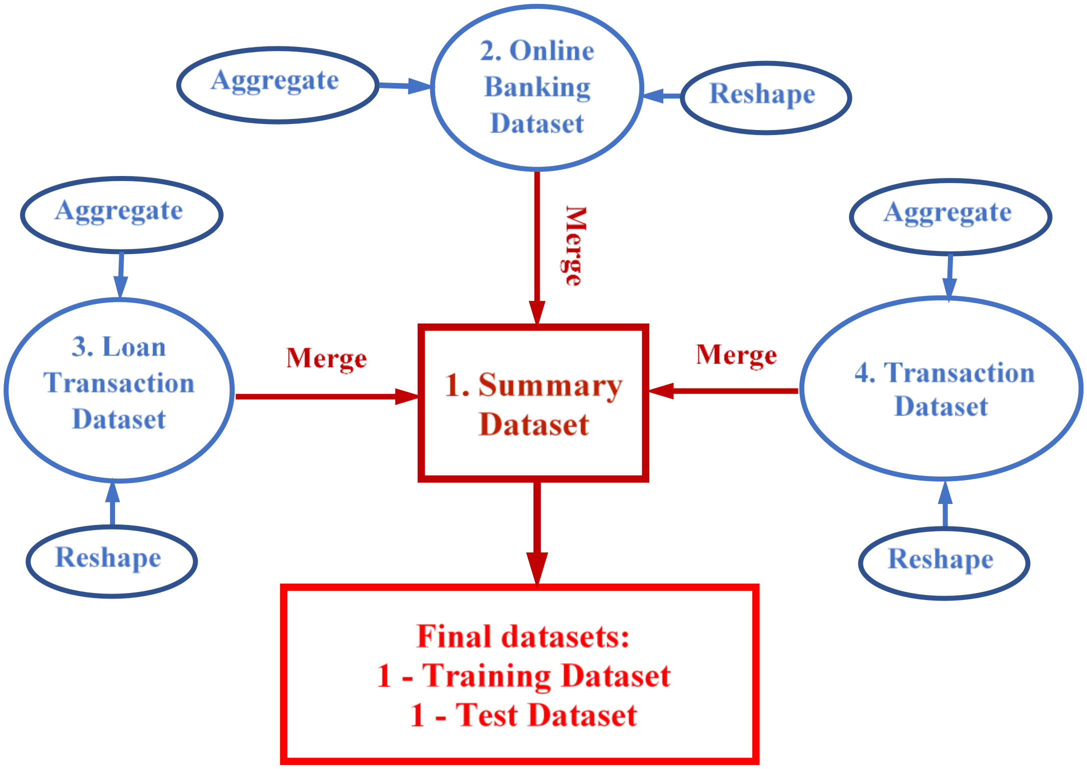


Figure 6: The data merging plan

2.2.3 New Variables

As described in Section 2.2.2, we have new segments variable for customers' activities after data merge. For our two analysis questions, we designed two sets of variables for customer segments. To answer the first question to predict the churn status as a given date, we focused on customers' recent activities and used these activities as predictors. We used the given date (7/31/2019) as a follow-up/churn date and identified customers' recent activities backward until 7 months in each transaction dataset. After finished the data merge, we chose three cases of customers segments with 1 month, 4 months, and 7 months as our predictors for our model comparison. The purpose is to identify which segments is the best one. For each case, the new variables for transaction activities are 11. The total variables is 17 (including the target variable) by adding other variables from summary data.

To answer the second question, to predict the churn status at 3rd , 6th and 9th month after the open date, we focused on customers' open activities and used these activities as predictors. The data processing is the same, but we used the open date as a beginning date and identified customers' open activities forward until 9 months in each transaction dataset. After finished the data merge, we build three sets of news variables with customers segments of at 3rd, 6th and 9th month. For each set, the new variables for transaction activities are 11. The total variable is 17 (including the target variable) by adding other variables from summary data.

Section 3: Model and approach

We developed two sets of predictive models. The first model is to predict the probabilities of customers' churn status, given date (07/31/2019) in the problem. The second model is to predict the probabilities of customers' churn status at 3rd, 6th and 9th month after the open date of customers' accounts. The relative predictors we used were described in last section. We also excluded some variables that had high correlations each other. The data partition is 70% train and 30% validation.

We used four modeling approaches. We started with two tradition methods - Logistic Regression and Decision Tree. We found out that decision tree was the better one. And then we focused on hypermeter tree methods of Radon Forest and XGBoost to see if we can improve the modeling performance. We tuned the parameters for both methods and carried out them for our modeling. The results show both hypermeter tree method can improve modeling prediction and

XGBoost is the best one. The modeling accuracy and AUC of ROC curve were used for modeling performance evaluation. We obtained the four predictions of test, the probabilities of churn status at 3rd, 6th and 9th month after open date and as date of 07/31/2019.

Section 4: Results

First Model: The model is to predict the probabilities of customers' churn by given date (7/31/2019). We found the predictors in 5 months before the given date were the best case. Table 2 shows predictive model performances. All the models perform excellent. The logistic Regression has 88.8% accuracy in validation. Decision tree can perform better than Logistic Regression with the accuracy of 93.6%. Random Forest and XGBoost can further improve the model performance with accuracy of 94.4% and 94.6%, respectively. XGboost is the best one in the accuracy.

Table 2: Predictive model performance

	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%

Table 3 shows the statistics of AUC of ROC curve for each model. Logistic Regression model has AUC 0.932, Decision three has AUC 0.964, Random Forest has AUC 0.969 and XGboost has AUC 0.975. XGBoost is also the best one in AUC. It can be sure that XGBoost model is suitable for churn prediction. The ROC curve for both validation data are shown in Figure 7.

Table 3: Statistical result summary – AUC of ROC Curves

Predictive model methods	AUC	
	Train	Validation
Logistic Regression	0.937	0.932
Decision Tree	0.967	0.964
Random Forest	0.986	0.969
XGBoost	0.994	0.975

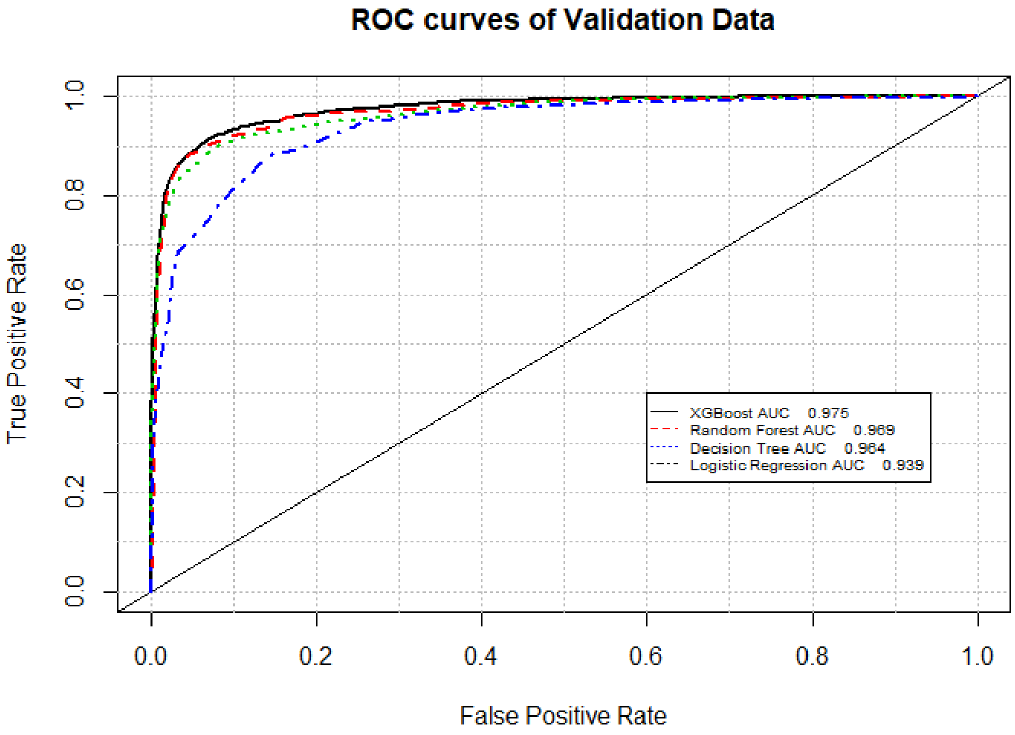


Figure 7: ROC curves

Second model: The model is to predict the probabilities of customers' churn at 3rd , 6th and 9th month after the open date. Table 4 shows predictive model performance and Statistics of AUC of ROC curves in three sets of models with XGBoost method. The accuracy at 3rd , 6th and 9th are 87.%, 88.0% and 88.5% in validation, respectively and AUC at 3rd , 6th and 9th are 0.932, 0.964 and 0.969 in validation, respectively. It shows the model have the excellent performance in early predictions.

Table 4: Predictive model performances and AUC for early prediction

XGBoost model methods	Accuracy		AUC	
	Train	Validation	Train	Validation
Three months	95.0%	87.0%	0.937	0.932
Six months	96.4%	88.0%	0.967	0.964
Nine months	97.3%	88.5%	0.986	0.969

An example of variables importance is shown in Figure 8. As we can see, the most important ten factors are Tenure, Age, Sum of Payment in loan, Number of Total Transactions, Sum of Fees Charged, Number of Days accessing account in online banking, Number of Payments in loan, Sum of Direct Deposits, Num of Fee Charged and Number of Transactions Conducted in Branch.

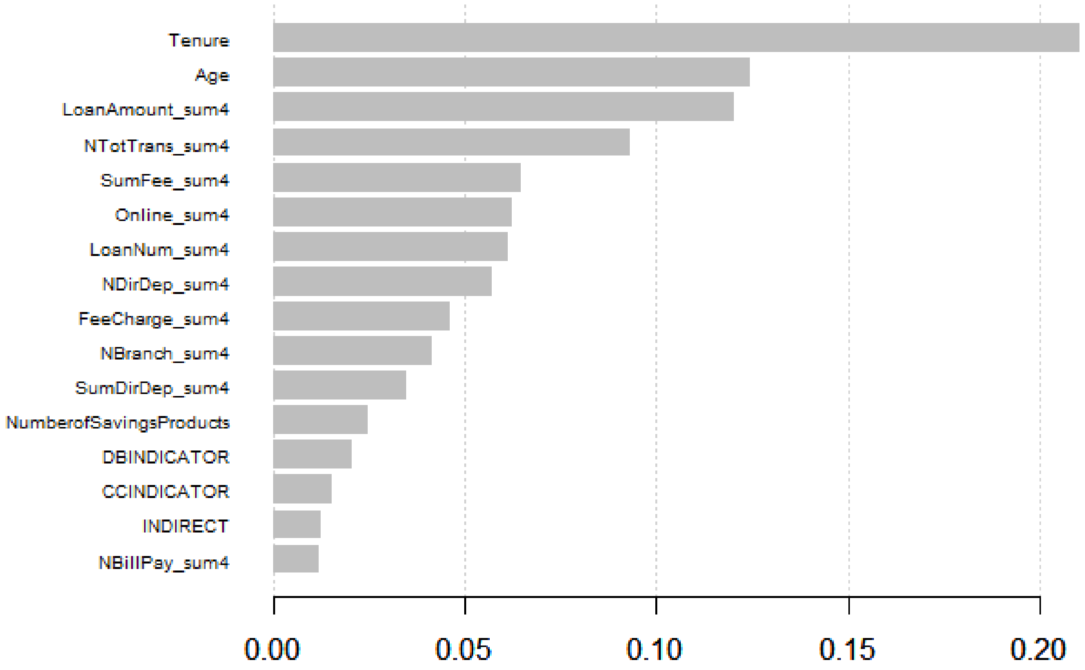


Figure 8: Variable importance

Prediction: For the test data, we have four predictions: One is the prediction of churn status as the given date (7/31/2019) and other three predictions are the churn status at 3rd, 6th, and 9th month after the open date. If we suppose the prediction as of the given date (7/31/2019) is the "True" churn status, we can get the confusion matrixes for the three different month cases. Table 5 shows the three confusion matrixes and accuracies. As we can see, the three models predict excellent by using early transaction activities with accuracies of 88.18%, 89.35% and 89.92% if we use the prediction model as the “True” churn status. It shows again that the model has the excellent performance in early prediction.

Table 5: Confusion matrixes among three early predictions and prediction in 07/31/2019 (Total 973 in test data)

	Model2						
Model 1	3rd month		6th month		9th month		
	Closed	Open	Closed	Open	Closed	Open	Total
Closed	1181	877	1268	790	1310	748	2058
Open	273	7399	246	7426	232	7440	7672
Total	1454	8276	1514	8216	1542	8188	9730
Accuracy	88.18%		89.35%		89.92%		

Section 5: Conclusion

The conclusions are summarized as below:

(1) We explored the data sets with data visualizations, identified the customers’ activities, and designed the data merging plan.

(2) We developed two models, which can predict not only the probabilities of customers’ churn status as of a given date, but also the probabilities of customers’ churn status at 3rd, 6th, and 9th month after the open date of customers’ accounts.

(3) Both sets of models have high accuracies and AUCs of ROC curves. XGBoost method is the best one.

(4) Our models can help AFCU to detect the early signs of customers’ churn. Thus, AFCU can take specific actions to prevent churn and dramatically improve the success of the retention offers to the potential churners.

Section 6: Appendix

Additional figures (9-13) are in this section：

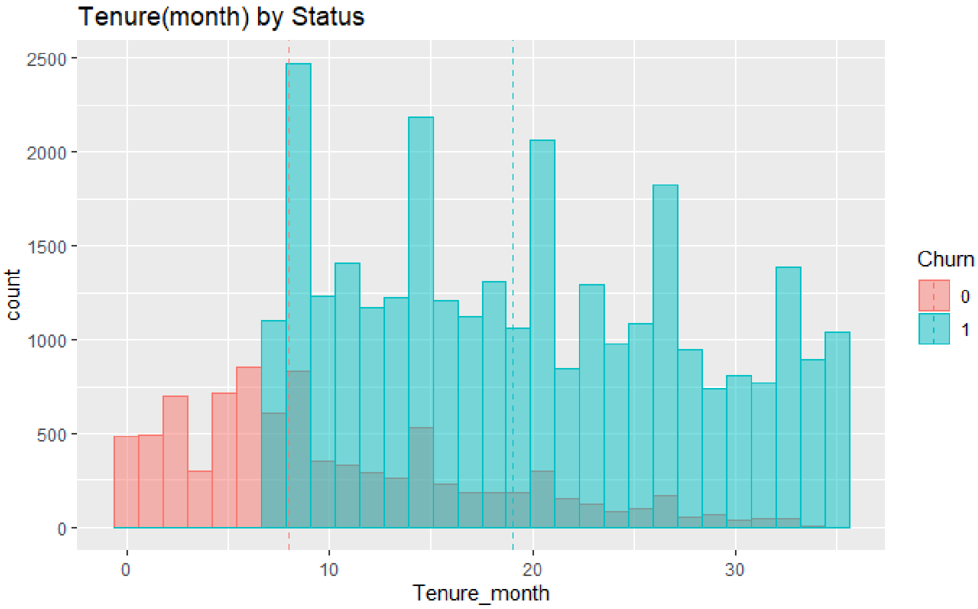


Figure 9: Tenure(month) by Status histogram chart

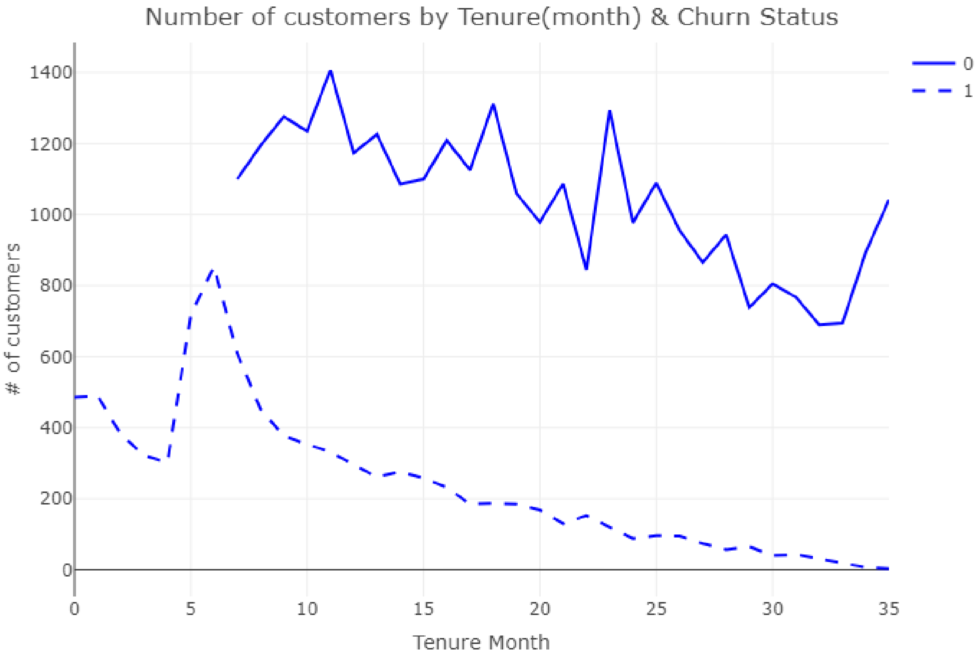


Figure 10: Tenure(month) by Status line chart

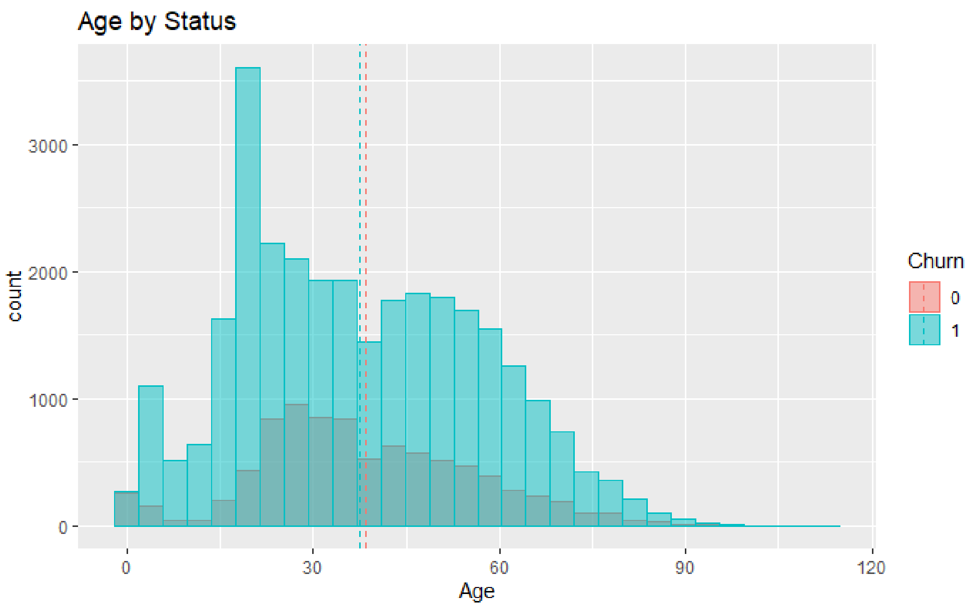


Figure 11: Age by Status

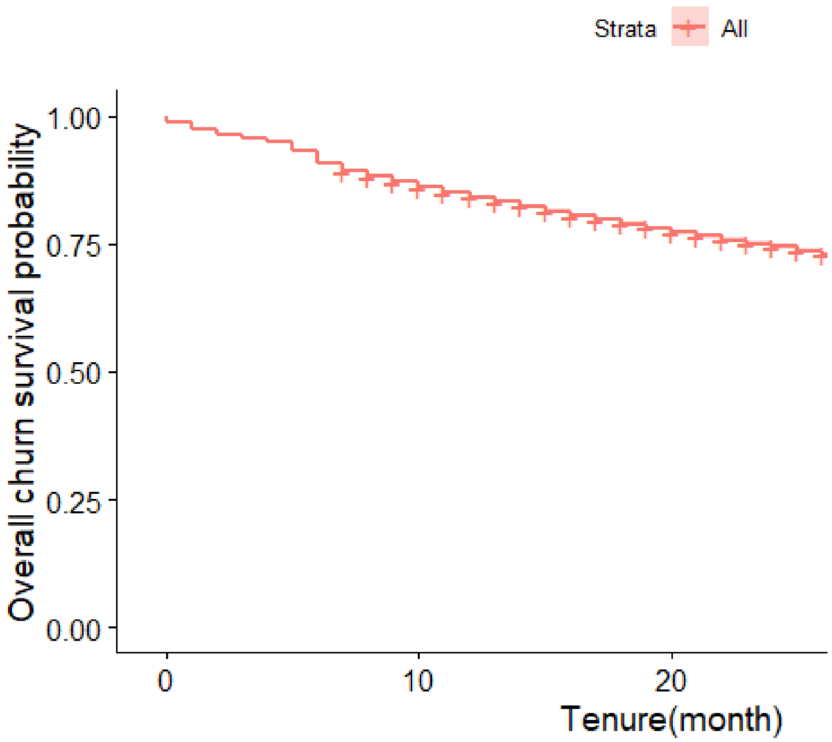


Figure 12: Overall churn survival curve

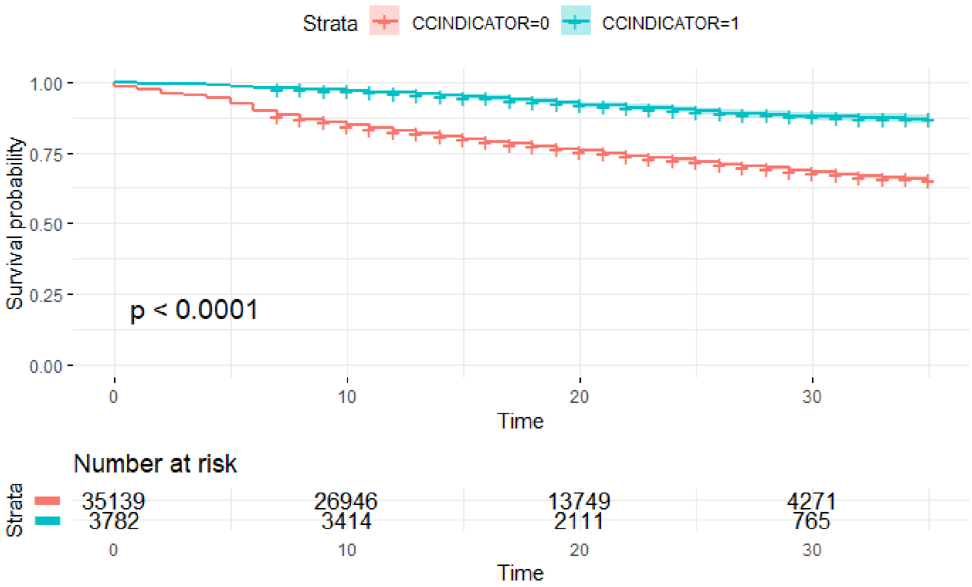


Figure 13: Churn survival curve by CCINDICATOR