TO: Banking Managers
FROM: Charlie Evert
DATE: July 19, 2021
RE: Predicting & Preventing Churn

Introduction

The banking industry is a global leader in using business analytics, and [preventing along with predicting customer churn] is a great usage of the large volumes of data that banks collect. "Churn" is a term that describes when customers stop using a banks' services; this is particularly damaging to banks' bottom lines considering banks' entire business model involves investing customers' money. Without customers, there would be no money to invest, and therefore no profits. Thus, predicting and preventing churn is very integral to the success of banks, and the following analysis will help clarify ways to predict and reduce churn given accessible data.

The following write-up will summarize data cleaning prior to analysis. After this, I will describe how to predict and then reduce churn. These insights will be arrived upon via the CC Churn Final document, as analyzed in JMP.

Data Cleaning

Given raw and unfiltered data, some transformation was necessary to prepare the data for analysis. The steps taken to clean the data are as follows:

1. Marital Status has a single typographical error, accounting for 0.002% of all data. This error was that one respondent was misclassified as "N" instead of married or unmarried. Given that it is a fair assumption that "N" stood for "no" if asked if the respondent was married, I fixed this issue by reclassifying the "N" as a "U," indicating that the respondent is unmarried.

2. Age of Accounts has a large amount of missing data, amounting to 17,853 missing cells and approximately 38.93% of the data. To fix this issue, I imputed all 17,853 missing values with the mean of Age of Accounts; this value amounts to 28.652834.

3. Age Group has 27,640 missing values, amounting to approximately 60.28% of the data. Due to the facts that imputation has the potential to skew insights and that the majority of the data was missing, it seems illogical to include Age Group in further analysis. Thus, I discarded this variable and did not use it to find how to predict and reduce churn.

4. Gender has a large amount of missing values, amounting to 22,500 missing cells and approximately 49.07% of the data. Unlike the second step that I took in imputing missing values for the Age of Accounts, imputation will not work for this qualitative binary measure. Thus, I decided that it would be best to omit Gender from any analysis aimed at predicting and reducing churn.

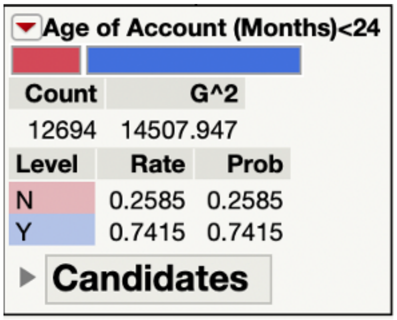
5. Occupation Group has 36,534 errors, accounting for approximately 79.67% of the data. OTHER and OTHERS were listed as separate Occupation Groups; to fix this issue, I recoded all cells listed as OTHERS to OTHER.

Predicting Churn

I omitted the Customer ID (a nominal variable that only serves as a primary key and not an attribute) column as well as Gender and Age Group from the decision tree since these values would only serve to skew the data.

The following leaves are divided into those that are likely to churn, and those that are unlikely to churn.

Likely to Churn



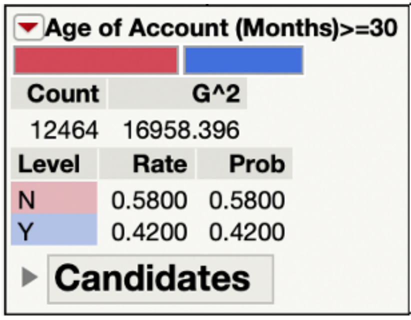
Age of Account (Months)<24

Count	G^2
12694	14507.947

Level	Rate	Prob
N	0.2585	0.2585
Y	0.7415	0.7415

1. If customers have accounts less than 24 months old, they are 48.3% more likely to churn (74.15%) than not churn (25.85%).

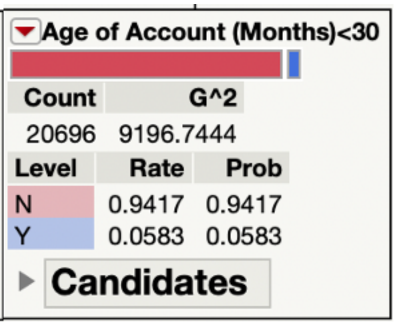
Unlikely to Churn



Age of Account (Months)>=30

Count	G^2
12464	16958.396

Level	Rate	Prob
N	0.5800	0.5800
Y	0.4200	0.4200



Age of Account (Months)<30

Count	G^2
20696	9196.7444

Level	Rate	Prob
N	0.9417	0.9417
Y	0.0583	0.0583

1. If customers have accounts greater than or equal to 30 months old, they are 16% more likely to not churn (58%) than churn (42%).

2. If customers have accounts less than 30 months old, they are 88.34% more likely to not churn (94.17%) than churn (5.83%).

Recommendations to Reduce Churn

The following two recommendations have the potential to reduce churn based upon insights derived from the decision tree:

1. Incentivize customers with accounts under 24 or over 30 months old to stay with the bank. Specifically, better promotions should be available for new accounts under 24 months old since this demographic is vastly more likely to churn. For some unknown reason, these account age groups are more likely than account age groups between 24-30 months to leave for another bank; a solution to this would be to offer these groups something that other banks do not provide such as bonuses for consistent monthly direct deposits, special discounts or better rates of return for accounts. Perhaps incentives aligned specifically with new customers (up to 24 months) and older customers (after 30 months) should be considered.

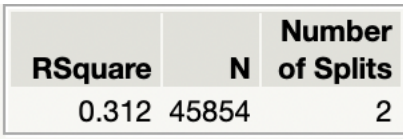
2. Focus targeted advertising on people with accounts under 24 or over 30 months old. Since these demographics are most likely to leave, and their accounts can only make banks money if they stay open, customers that fit these demographics should be bombarded with targeted ads. The promotions mentioned in the first recommendation should be directly advertised to current customers (in those demographics) as often and as effectively as possible to preserve them as customers.

Conclusion

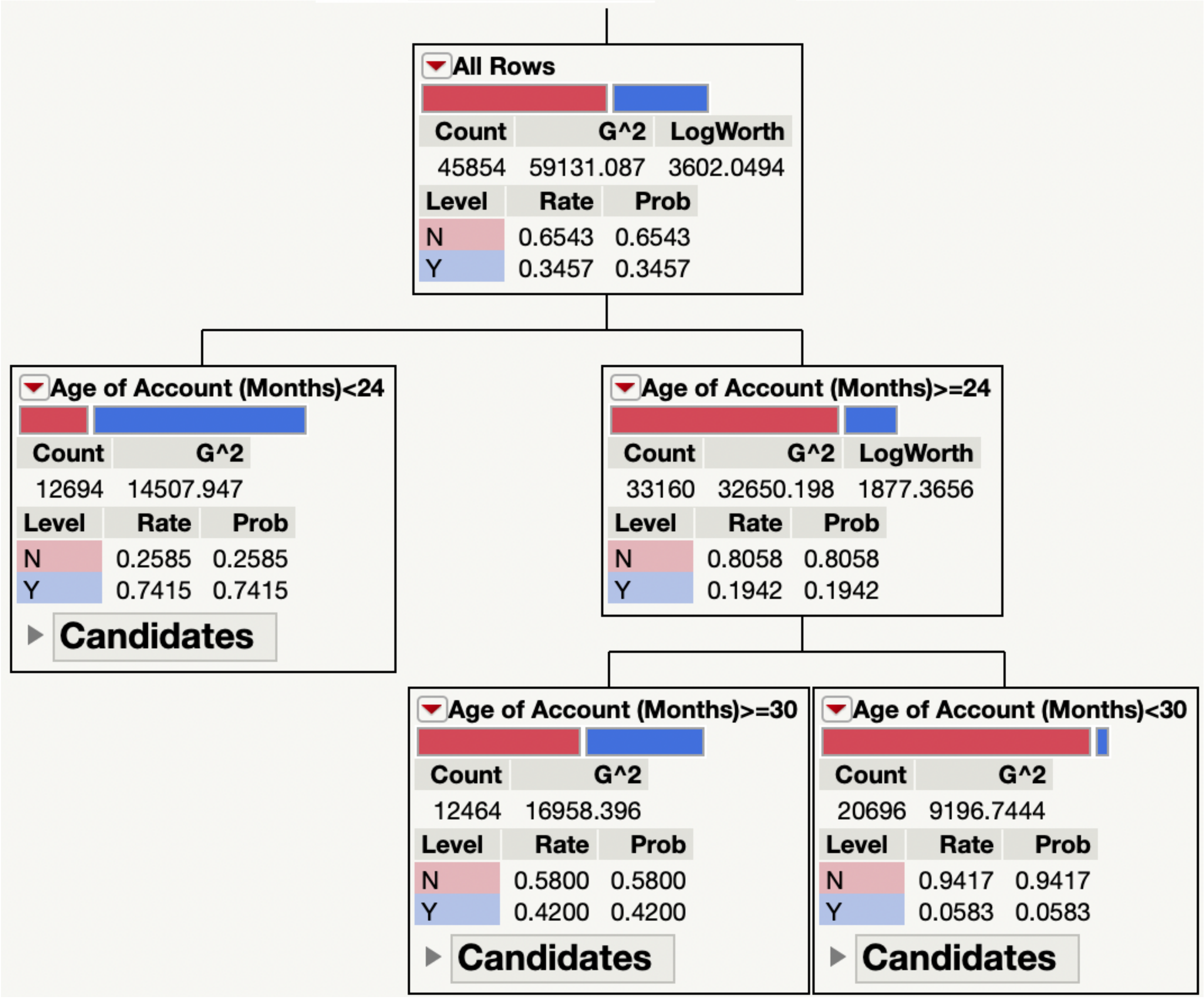
More data must be collected to determine alternate paths towards prosperity in retaining account age groups likely to churn. Similar analysis to the following appendices must be conducted with more data for there to be more actionable results. There were limitations of actionable insights that could be takes based exclusively on age groups; more reliable data that reflects different attributes would help to gather more complex, more actionable information.

The following appendices contain each aspect of the decision tree model used for analysis:

Appendix A: Decision Tree & R^2 Table



RSquare	N	Number of Splits
0.312	45854	2



All Rows

Count	G^2	LogWorth
45854	59131.087	3602.0494

Level	Rate	Prob
N	0.6543	0.6543
Y	0.3457	0.3457

Age of Account (Months)<24

Count	G^2
12694	14507.947

Level	Rate	Prob
N	0.2585	0.2585
Y	0.7415	0.7415

▸ Candidates

Age of Account (Months)>=24

Count	G^2	LogWorth
33160	32650.198	1877.3656

Level	Rate	Prob
N	0.8058	0.8058
Y	0.1942	0.1942

Age of Account (Months)>=30

Count	G^2
12464	16958.396

Level	Rate	Prob
N	0.5800	0.5800
Y	0.4200	0.4200

▸ Candidates

Age of Account (Months)<30

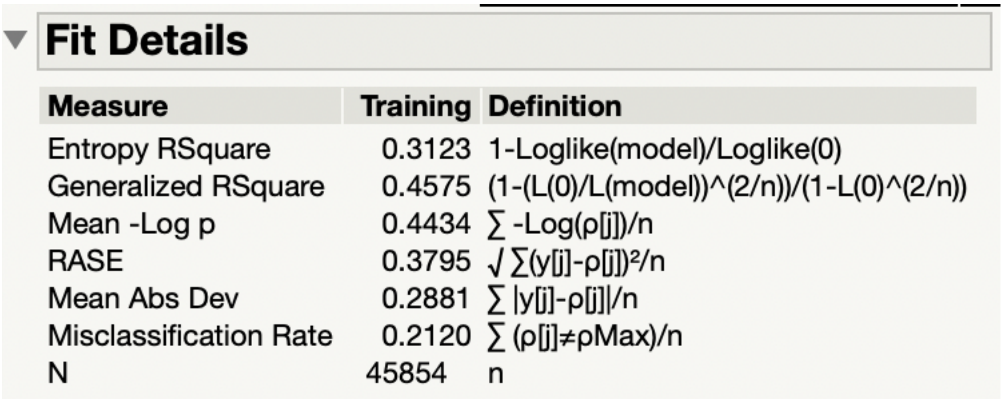
Count	G^2
20696	9196.7444

Level	Rate	Prob
N	0.9417	0.9417
Y	0.0583	0.0583

▸ Candidates

R^2 for this decision tree is moderately low. This suggests that each independent variable (in this case different account ages) weakly correlates with the dependent variable (churn). This decision tree was split twice, and has 3 alternative scenarios in total (as covered in the "Predicting Churn" section). There are 44,854 observations covered by this model.

Appendix B: Fit Details

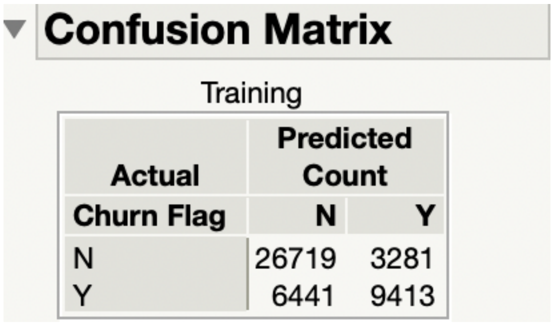


Fit Details

Measure	Training	Definition
Entropy RSquare	0.3123	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.4575	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.4434	∑ -Log(ρ[j])/n
RASE	0.3795	√ ∑(y[j]-ρ[j])^2/n
Mean Abs Dev	0.2881	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.2120	∑ (ρ[j]≠ρMax)/n
N	45854	n

The misclassification rate is very low at approximately 21%, and the accuracy rate is very high at approximately 79%. This suggests that the model is very reliable in predicting churn.

Appendix C: Confusion Matrix



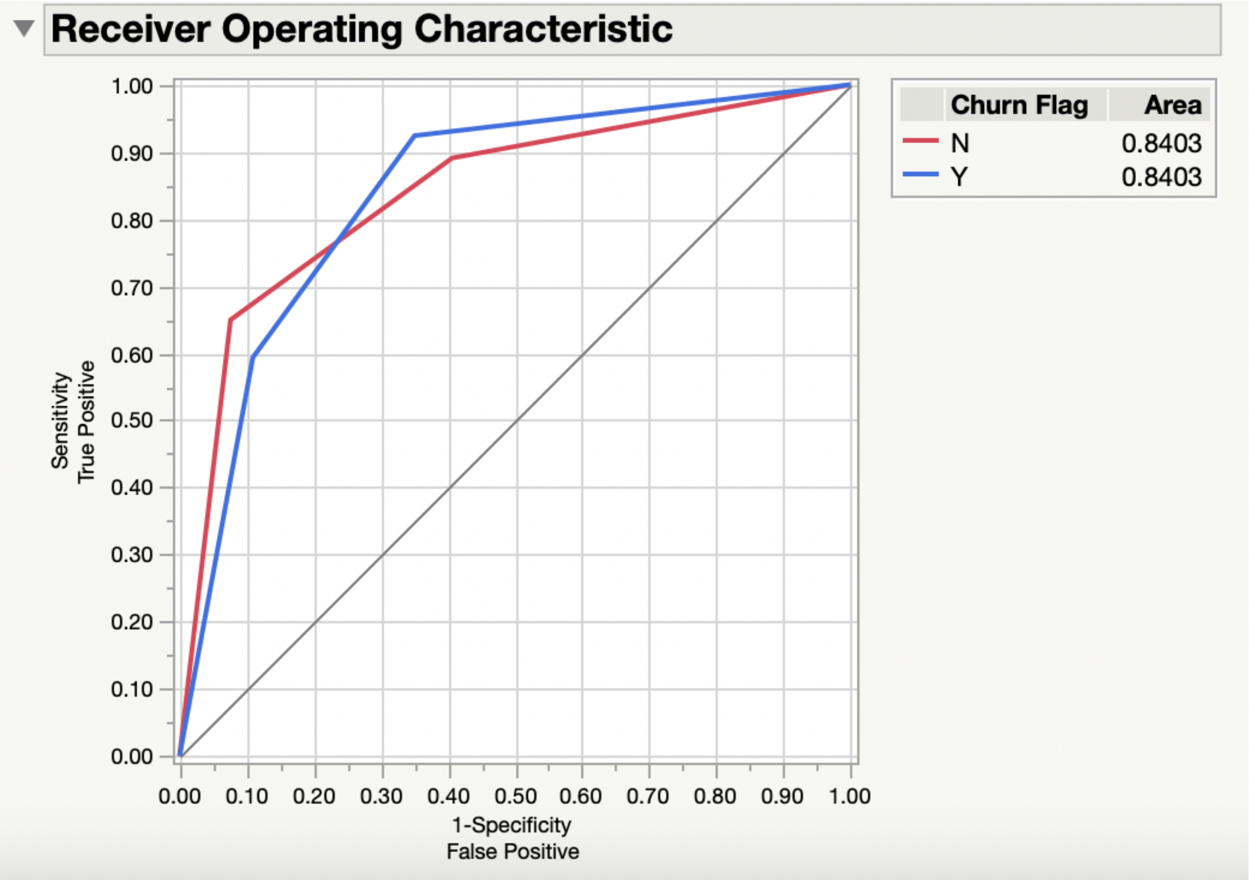
Confusion Matrix

Training

Actual	Predicted Count	
Churn Flag	N	Y
N	26719	3281
Y	6441	9413

This confusion matrix demonstrates that the model has 36,132 accurate predictions and 9,722 inaccurate predictions. Since N is 45,854, this reflects the accuracy rate of approximately 79% that was previously covered.

Appendix D: ROC Curve



	Churn Flag	Area
🟥	N	0.8403
🟦	Y	0.8403

The ROC Curve is leftward of the random guess line by a significant margin, indicating that the model is much more accurate of predicting whether customers will churn than a random guess. There are significantly more accurate predictions based on this model than randomly guessing.