Predicting Customer Churn

Provident Credit Union

Hasib Azami - Duc Turney - Johan Sjoden - Elizabeth Romero - Kevin Eddy

Introduction

Provident Credit Union Churn Prediction

OUR TEAM

Hasib Azami

Duc Greenwell

Johan Sjoden

Elizabeth Romero

Kevin Eddy

Agenda

- Background
- Data Processing/ Exploration
- Models
- Application & Discussion
- UI & Reports

Background

Business Problem

Acquiring New Customers VS. Retaining Existing Customers

Predicting Churn Within The Banking Industry

Many contributing factors to cause a customer to churn

Objectives

- Identify main causes of why member churn
- Predict who will churn
- Predict when they will churn

Provident Credit Union

Brief History:

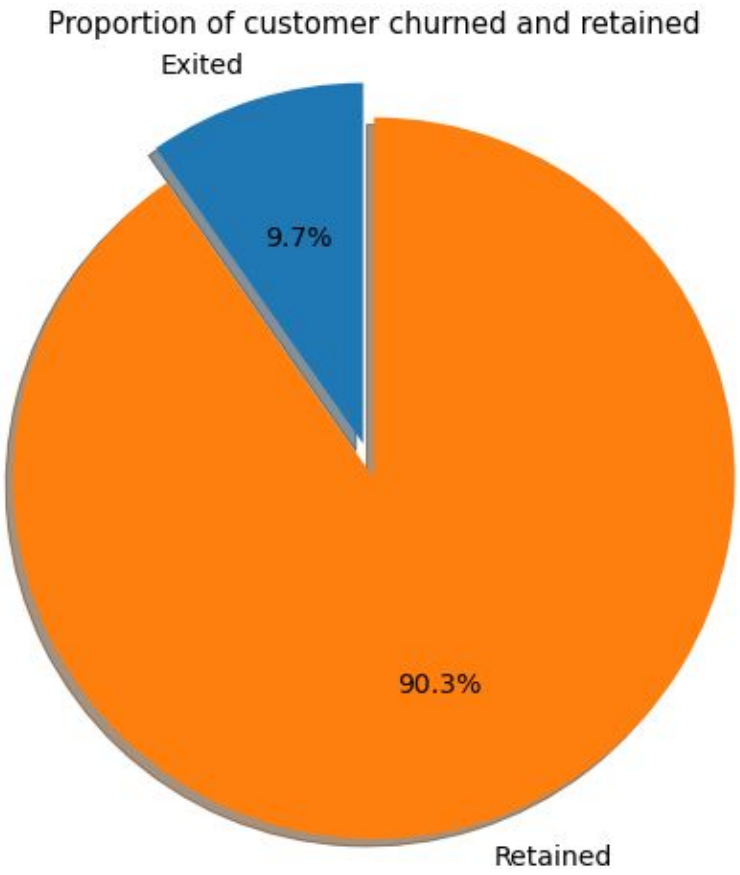
- Established in 1950 to serve the California's Teacher Association
- Headquartered in San Francisco
- 20th largest credit Union in California and 108th in the U.S.

Fun Fact:

- PCU President, Jim Ernest, is a Saint Mary's alumni

Understanding the Data

Data Exploration



- **9.7% Customer Churn**
- **90.3% Customers Retained**

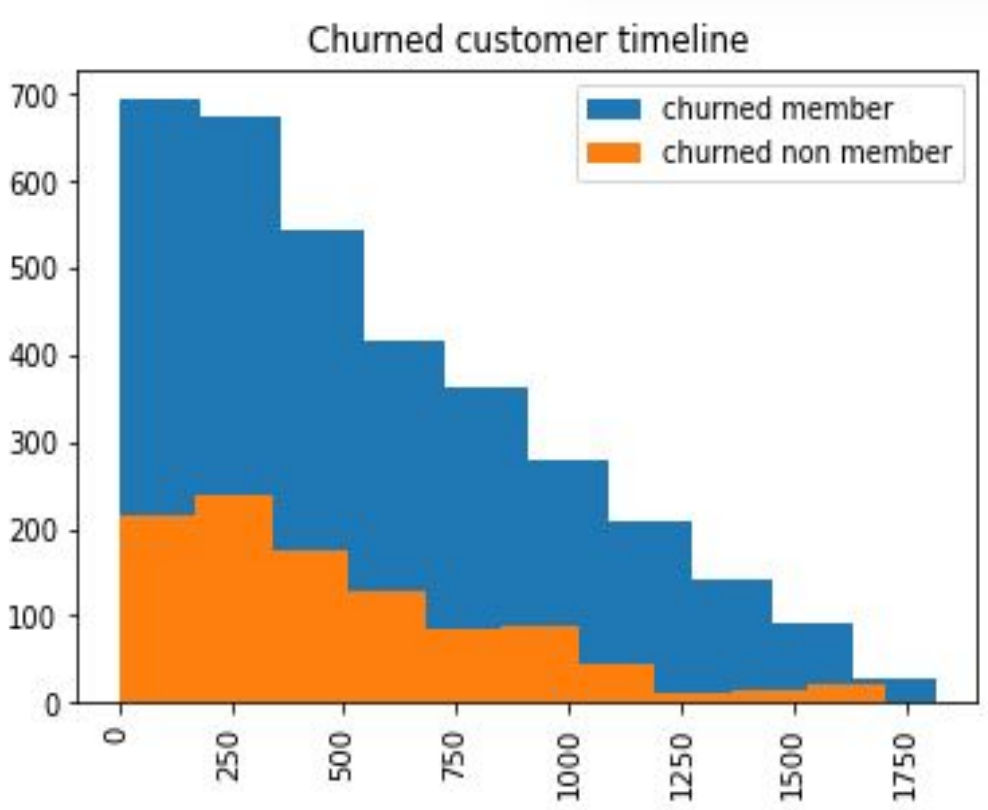
- Churned Non-members, overwhelmingly hold either both a **savings and checking account** or **1 savings account**.
- Churned Members hold similar accounts- however, are more likely to **have a loan as well**.

Data Exploration

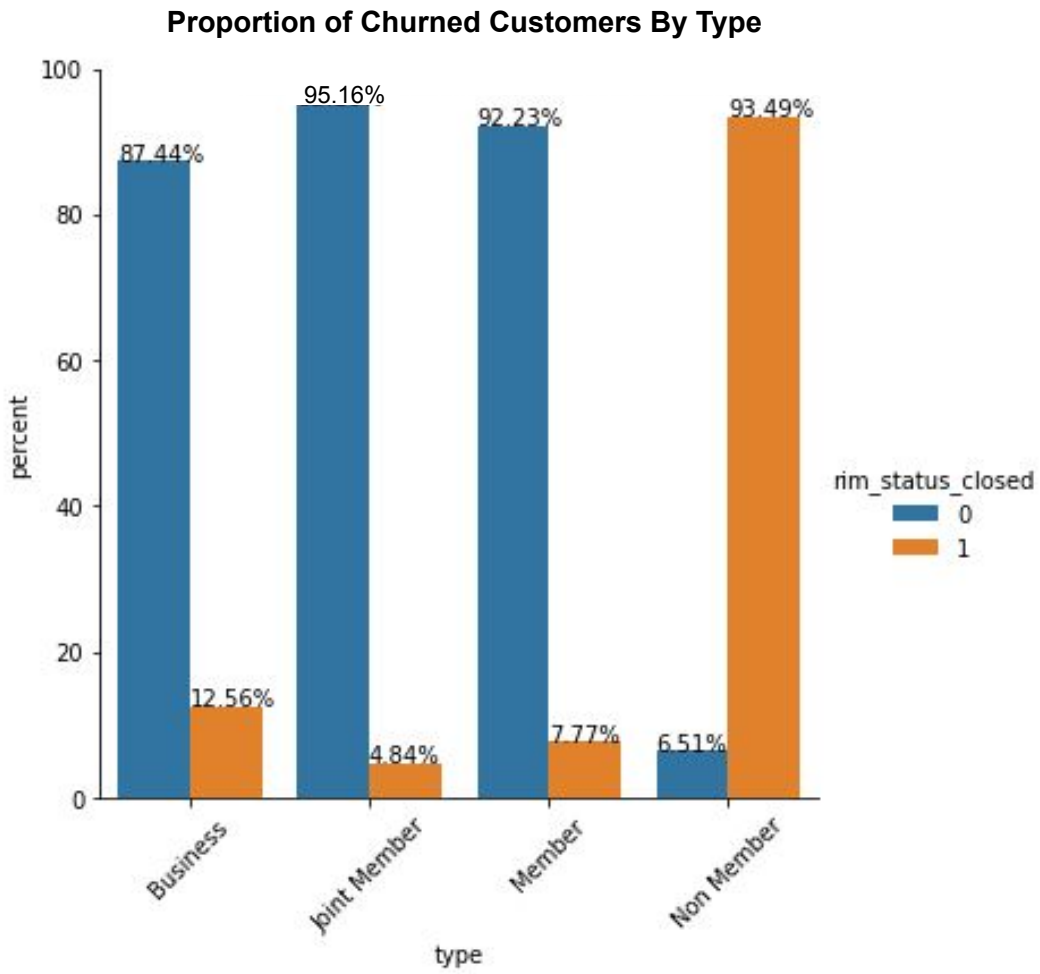
- **68% of Members churn within 2yrs**

- **77% of Non-Members churn within 2yrs**

Members and non-members typically churn quicker - closing their account within the first 300 days of opening.



Data Exploration

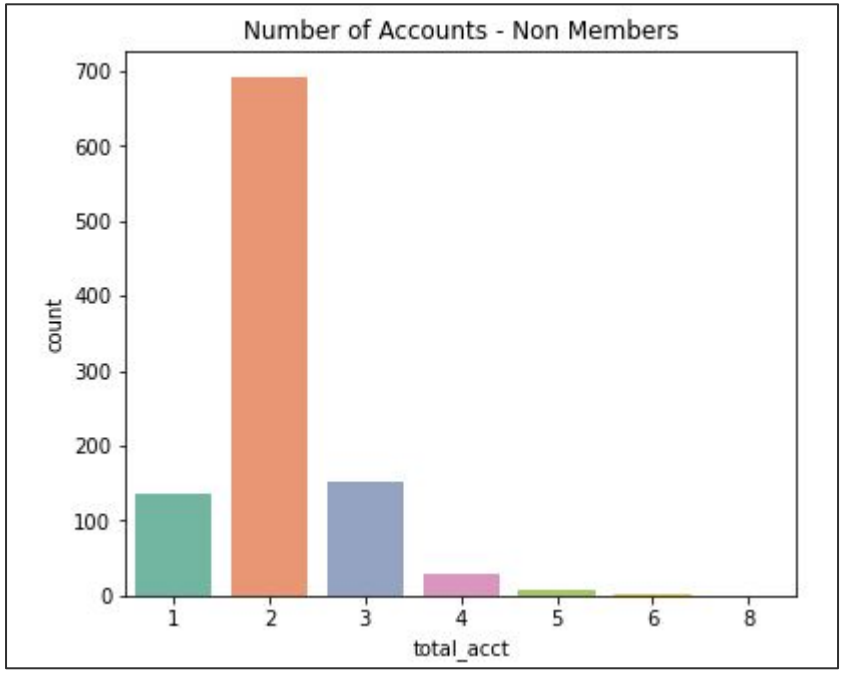


Account Types Closed

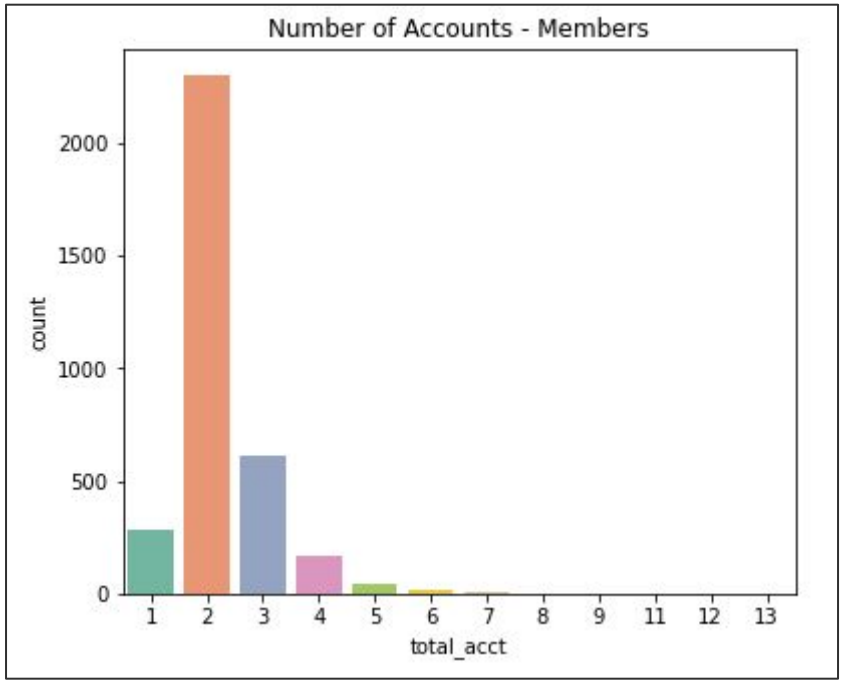
- **Business: 12.56%**
- **Joint-Member: 4.84%**
- **Member: 7.77%**
- **Non-Member: 93.49%**

Non-Member accounts are more likely to close their accounts than other account types.

Non - Member



Member



Number of Accounts - Member / Non-Member

The total number of accounts between members and non-members vary, where **non-members** typically hold **between 1 and 6 accounts with 96% having 3 or less**. **Members** tend to hold more accounts, typically **between 1 and 7, with 93% having 3 or less accounts**.

What This Could Mean

Non-Members:

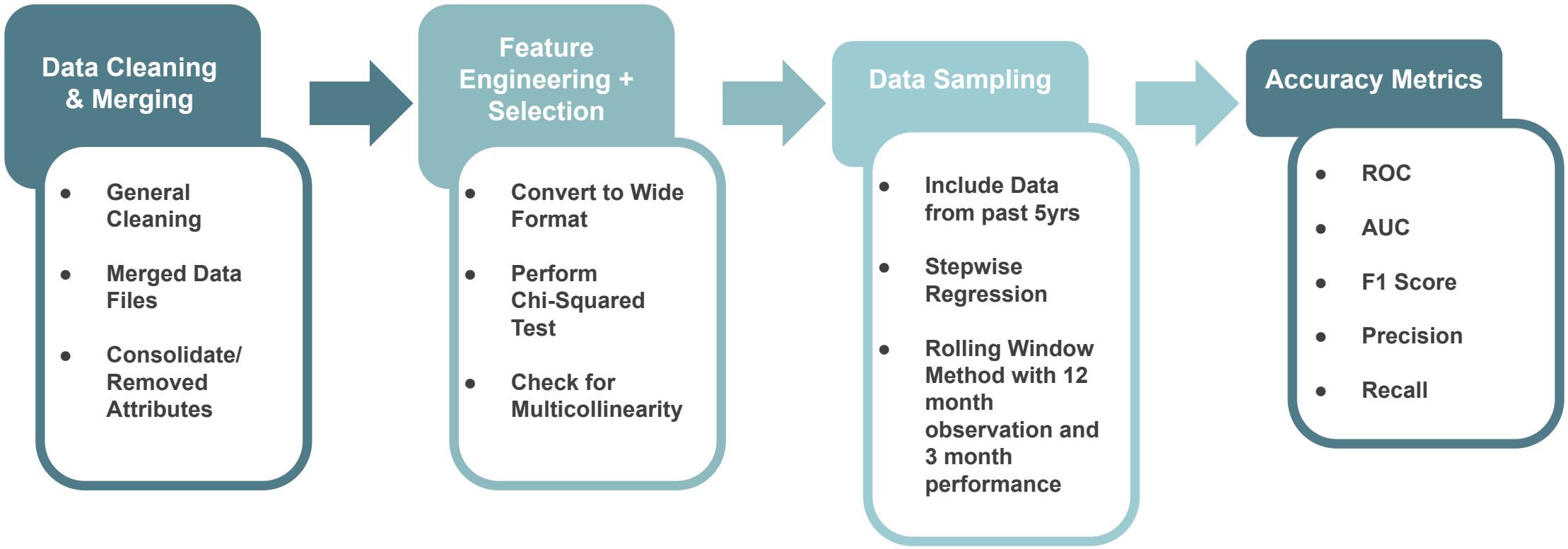
May indicate that PCU is not their main bank

Members

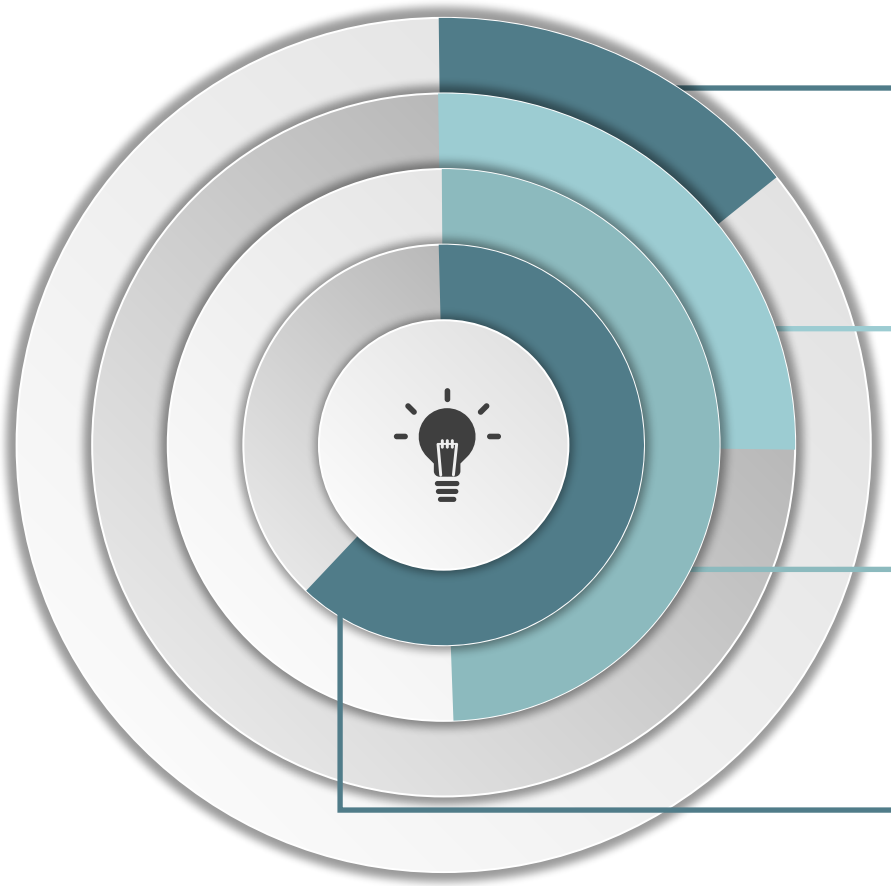
May indicate PCU plays a larger part in their overall finances

Methodology

Data Processing



Models



Logistic Regression

Used to evaluate each dataset from under,over, both, and ROSE sampling methods.

Survival Models

Used the Kaplein Meyer Model and Cox Proportional Model to evaluate customer's time with Provident Credit Union.

Decision Trees

Used Decision Trees Model in order to predict the value of a target variable by learning simple decision rules inferred from data features

Random Forest

Created Random Forest Classifier in order to improve predictive accuracy and control overfitting

Logistic Regression Model

Advantages

- **Easier to:**
  - Implement
  - Interpret
  - Efficient To Train
- **Makes No Assumptions**
- **Performs Well**

Disadvantages

- **May Lead To Overfitting**
- **Constructs Linear Boundaries**
- **Requires Average to No Multicollinearity Between Independent Variables**

Random Forest & Decision Tree

Advantages Random Forest

- Less Interpretation on the model but want better accuracy
- Random Forest will reduce variance part of error rather than bias part

Advantages Decision Tree

- **Non Parametric Model**
- **When you want your model to be simple and explainable**

Logistic Regression

After balancing the data using under, over, both, and ROSE sampling techniques we implemented Logistic Regression to evaluate each dataset.

From here we identified that under-sampling was the better approach based on the better performing scores that were returned.

Logistic Regression	Recall	F1 score	AUC
Unbalanced original dataset	0.067	0.058	0.943
Over-sampling train dataset	0.958	0.028	0.894
Under-sampling train dataset	0.955	0.028	0.894
Both-sampling train dataset	0.959	0.027	0.896
ROSE-sampling train dataset	0.964	0.027	0.893

Random Forest & Decision Tree

The most optimal model Random Forest compared to Decision Tree because model is overfitting due to high recall

Models	Recall	F1 score	AUC
Random Forest	0.926	0.032	0.905
Decision Tree	1.00	0.025	0.811

Survival Modal

```
Concordance= 0.947  (se = 0.001 )
Likelihood ratio test= 13580  on 19 df,   p=<2e-16
Wald test            = 42923  on 19 df,   p=<2e-16
Score (logrank) test = 26780  on 19 df,   p=<2e-16
```

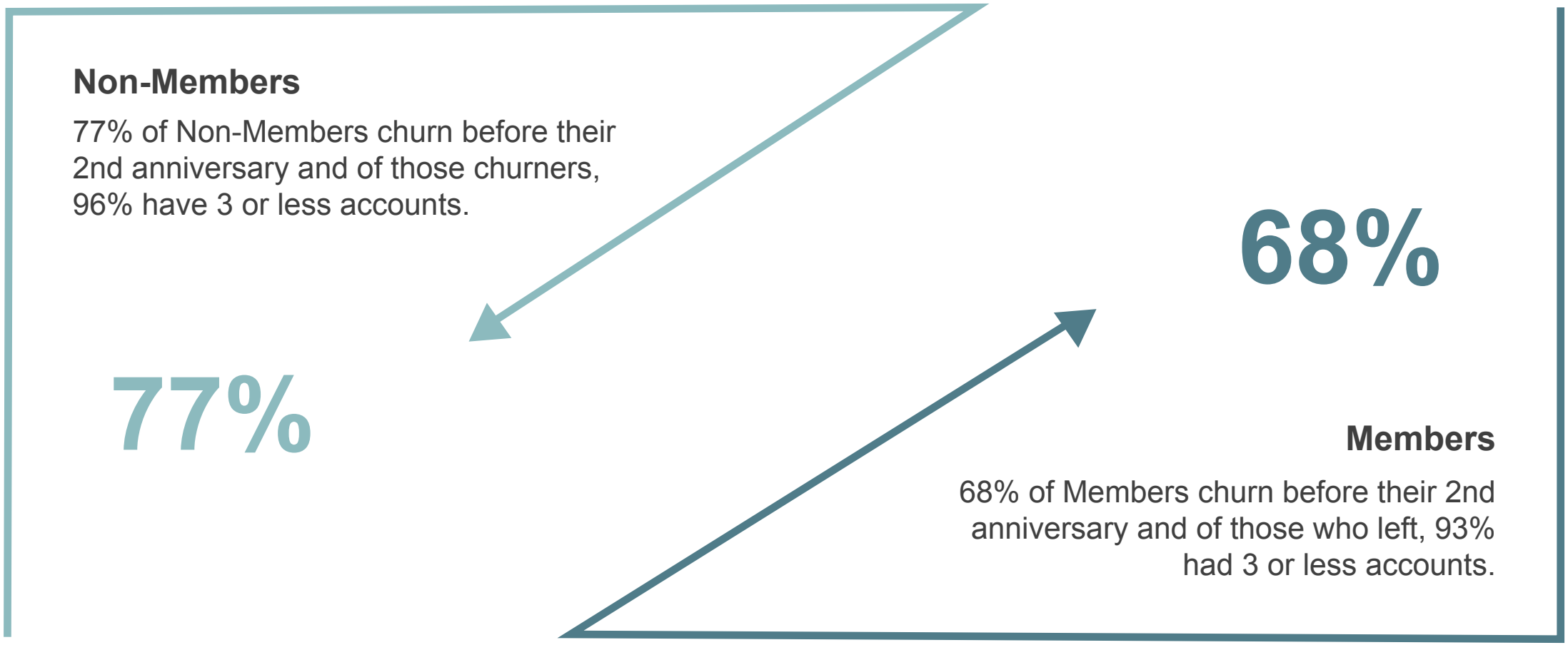
```
coxph(formula = Surv(rim_duration, rim_status_closed) ~ age +
    type + moving_time + total_acct + CK + SV + CD + IL + ML +
    VISA + mrm + trm + orig_rate + acc_duration + branch_rating +
    member_rating + contract_rating, data = dat_survival)

  n= 425297, number of events= 4212

                      coef  exp(coef)   se(coef)       z Pr(>|z|)
age               1.967e-02  1.020e+00  7.862e-04  25.022  < 2e-16 ***
typeJoint Member -2.242e+00  1.063e-01  1.106e-01 -20.278  < 2e-16 ***
typeMember       -1.128e+00  3.238e-01  3.475e-02 -32.455  < 2e-16 ***
typeNon Member    1.069e+00  2.913e+00  3.803e-02  28.111  < 2e-16 ***
moving_time      -2.211e-01  8.017e-01  2.024e-02 -10.922  < 2e-16 ***
total_acct       -1.513e+00  2.203e-01  1.067e-02 -141.755  < 2e-16 ***
CK                1.675e+00  5.341e+00  1.860e-02  90.075  < 2e-16 ***
SV                1.709e+00  5.524e+00  2.098e-02  81.460  < 2e-16 ***
CD                1.708e+00  5.516e+00  3.287e-02  51.949  < 2e-16 ***
IL                8.022e-01  2.230e+00  4.396e-02  18.246  < 2e-16 ***
ML                2.783e-01  1.321e+00  1.074e-01   2.592  0.00953 **
VISA              1.528e+00  4.609e+00  4.181e-02  36.546  < 2e-16 ***
mrm               1.609e+01  9.745e+06  7.641e+01   0.211  0.83320
trm               4.204e-03  1.004e+00  2.071e-04  20.301  < 2e-16 ***
orig_rate        -8.103e-01  4.447e-01  6.494e-02 -12.478  < 2e-16 ***
acc_duration     -1.630e-03  9.984e-01  5.233e-05 -31.146  < 2e-16 ***
branch_rating    -1.547e-01  8.566e-01  2.128e-02  -7.272 3.53e-13 ***
member_rating     2.070e-01  1.230e+00  2.690e-02   7.695 1.41e-14 ***
contract_rating   2.612e-01  1.299e+00  5.185e-02   5.039 4.69e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Bringing It Together

Customers Who Churn



Customers Who Stay

Members and non-members who stay longer than 2 years were found to have similar characteristics.

01

Members and non-members who had a loan in addition to a savings and checking account were more likely to stay long-term.

02

Members and non-members who indicated that PCU was their main bank were also more likely to stay longer.

Applications

Churn prediction has become very important for businesses across industries. And even more so for professional service based industries where acquiring new customers is both timely and costly.

How Churn Prediction Is Implemented

Reduce Costs + Increase Profits

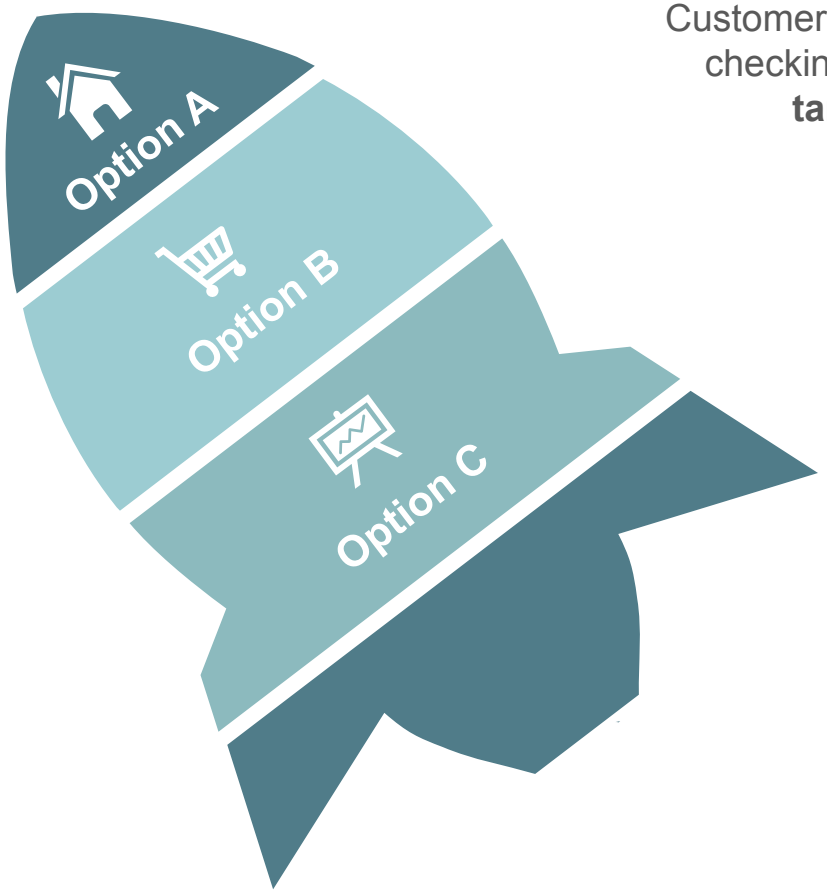
Group Customers Based on Characteristics

The core objective to churn analysis is to identify unique characteristics that may suggest future churn. Businesses can create customer segments and engage with them in a personalized way that may reduce churn and improve their lifetime value.

Define A Roadmap For New Customers

Churn analysis allows businesses to be proactive in retaining new and future customers. By knowing the characteristics of past customers, businesses can match similar behaviors and interject by offering special promotions, services, or incentives.

Next Steps | Recommendation



Loan Conversion

Customers are significantly less likely to churn if they have a loan in addition to a checking and savings account. **Our primary recommendation is for PCU to target customers without loans but do have a checking and savings account with either auto, home, business, ect loans.**

Customer Mapping

Knowing the potential actions of new customers is powerful. **We recommend that PCU uses past customer data to place new customers in cohorts based on similar characteristics**. Doing so will allow them to engage with customers and make more personalized and relevant product recommendations when it matters most to each group.

Incentives

In a highly competitive market, incentives are a must. **We recommend that PCU partners incentives with the previous two recommendations to keep current customers happy and engaged** at important periods throughout their customer lifecycle.

UI & Reports

- **Components: UI & Server**
- **Reports**

Thank You

Insert the Sub Title of Your Presentation

Preferences:

	Total Membership	No Churn	Churn	% Churn	Note
Original 5-year Dataset	119,994	109,002	10,992	9.16%	Churn anytime
Rolling -window data	425,297	421,085	4,212	0.99%	Churn within 3 months
After removing outliers	425,297	421,085	4,212	0.99%	Churn within 3 months
Train dataset	297709	294,760	2,949	0.99%	Churn within 3 months
Test dataset	127588	126,325	1,263	0.99%	Churn within 3 months
Undersampling train	5898	2,949	2,949	50.00%	Churn within 3 months
Total	538773		3099	0.005751958617	

	0	1	
Predict 0	3	6	
1	5	4	

	Precision	Recall	Sensitivity	Specificity	Accuracy	F1	AUC
Original	0.333	0.001	0.0008	1	0.9901	0.001	0.893
Over Sampling	0.028	0.958	0.958036	0.672036	0.6749	0.028	0.894
Under Sampling	**0.029**	**0.955**	**0.954869**	**0.676224**	**0.679**	**0.028**	**0.894**
Both	0.028	0.959	0.958828	0.67073	0.6736	0.027	0.893
ROSE	0.028	0.964	0.963579	0.663732	0.6667	0.027	0.893

Models (with under-sampling train dataset)	Accuracy	Sensitivity (True Positive Rate)	Specificity (True Negative Rate)	precision (how precise the model is)	recall (how robust the model is)	F1 score	AUC
Logistic Regression	0.679	0.955	0.676	0.029	0.955	0.028	0.894
Decision Tree	0.625	1.000	0.621	0.026	1.000	0.025	0.811
Random Forest	0.737	0.926	0.735	0.034	0.926	0.032	0.904

Logistic Regression	Accuracy	Sensitivity (True Positive Rate)	Specificity (True Negative Rate)	precision (how precise the model is)	recall (how robust the model is)	F1 score	AUC
Unbalanced original dataset	0.990	0.067	0.999	0.393	0.067	0.058	0.943
Over-sampling train dataset	0.675	0.958	0.672	0.028	0.958	0.028	0.894
Under-sampling train dataset	0.679	0.955	0.676	0.029	0.955	0.028	0.894
Both-sampling train dataset	0.674	0.958	0.670	0.028	0.959	0.027	0.896
ROSE-sampling train dataset	0.667	0.964	0.664	0.028	0.964	0.027	0.893