Group 1 Presents:

Predicting the Attrition of Credit Card Customers

Alex Delacruz, Bennett Northcutt, Emily Neaville, Hays Kronke, and Stephen Mims

Project 4 Proposal

Data Source: https://zenodo.org/record/4322342#.Y8OsBdJBwUE

Prediction of Churning Credit Card Customers

Objective: This dataset will be used to build a machine learning model that can accurately predict customer attrition. This dataset contains customer information ranging from demographic (age, gender, education) to financial data (income bracket, card history, credit limit). Using these predictions, we should be able to better predict customers who are at high risk of churning.

Bank Churn Dataset

Our dataset features 13 columns and thousands of rows of data.

Our first step was cleaning our data and have it in a more usable form for machine learning and visualizations.

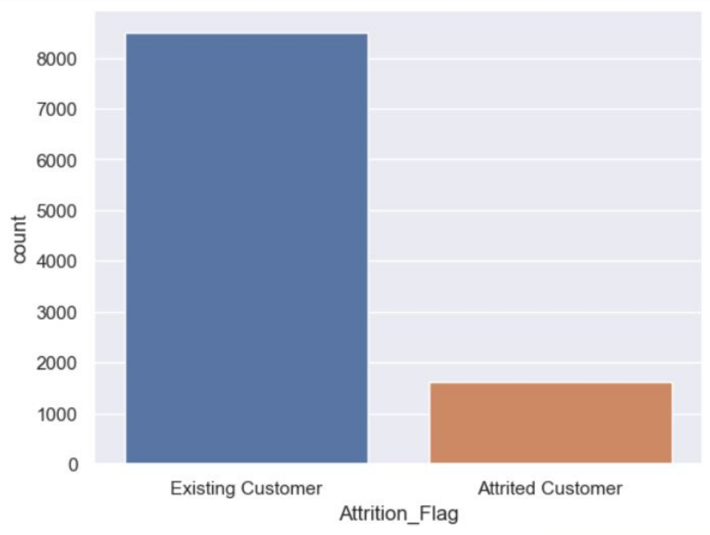
Our target for our models is the attrition flag.

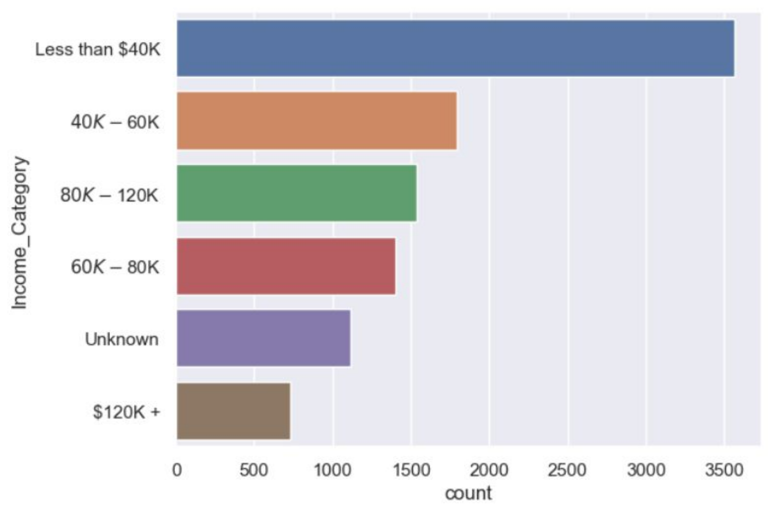
CustomerId	Customer ID: A unique identifier for each customer
Surname	Surname: The customer's surname or last name
# CreditScore	Credit Score: A numerical value representing the customer's credit score
Geography	Geography: The country where the customer resides (France, Spain or Germany)
Gender	Gender: The customer's gender (Male or Female)
# Age	Age: The customer's age.
# Tenure	Tenure: The number of years the customer has been with the bank
# Balance	Balance: The customer's account balance
# NumOfProducts	NumOfProducts: The number of bank products the customer uses (e.g., savings account, credit card)
# HasCrCard	
# IsActiveMember	
# EstimatedSalary	
# Exited	

Bank Churn Dataset

In our exploratory analysis we discovered that there was a large class imbalance between attritted and existing customers.

This imbalance was something we had to keep in mind when developing our models.

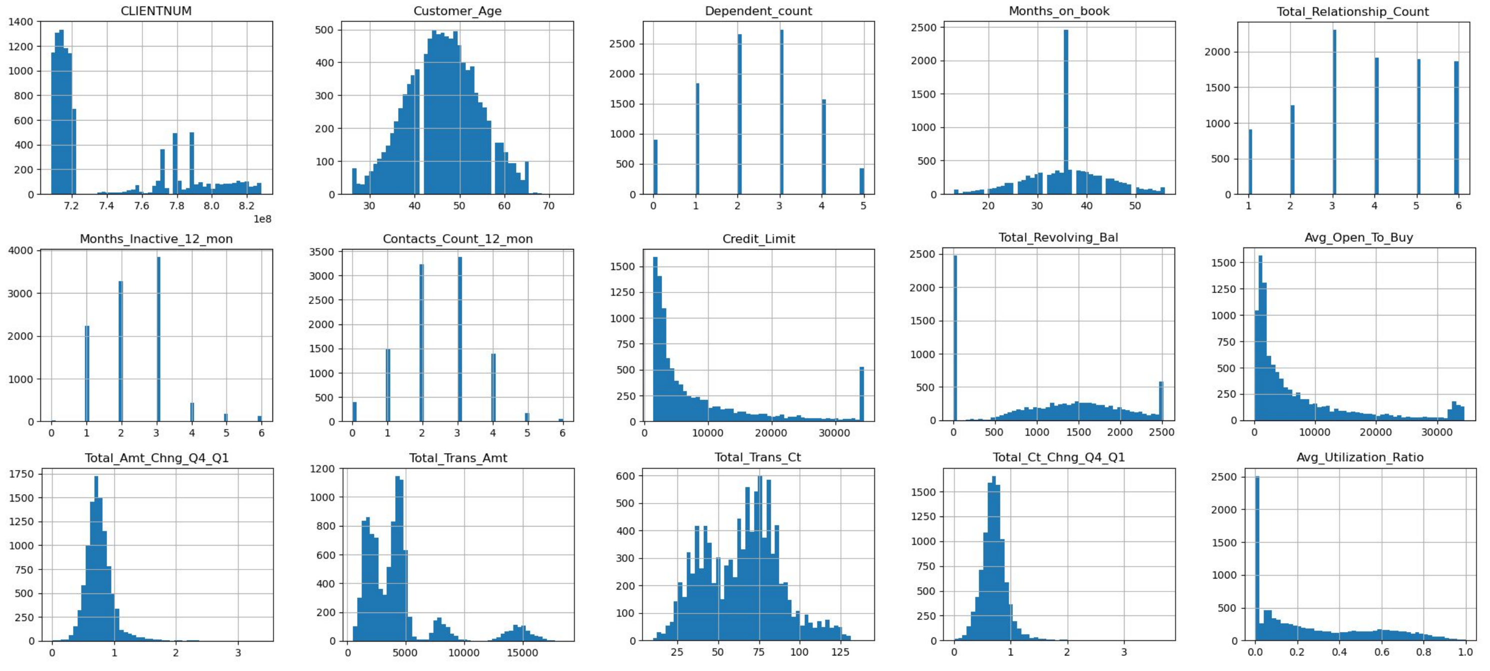




Histograms

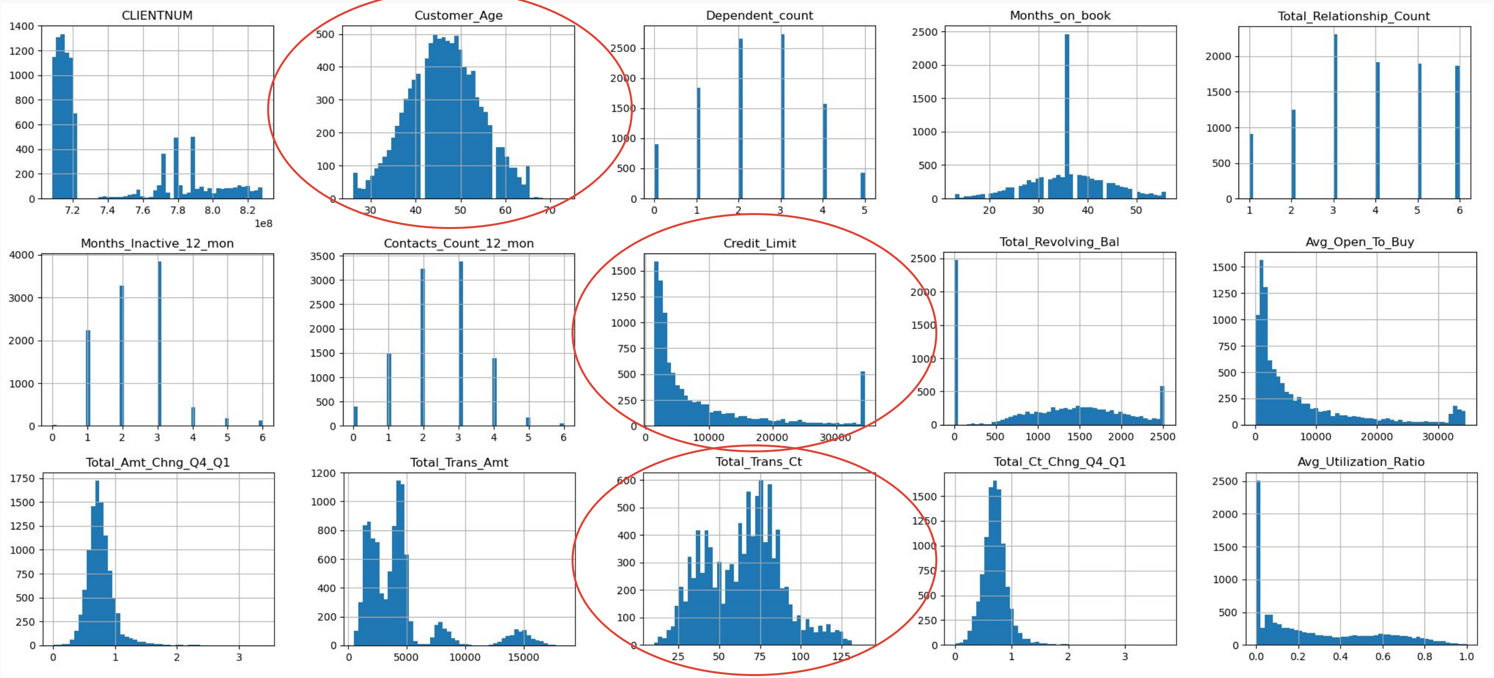
After cleaning up the data, we can create preliminary visualizations of the different metrics.

Seeing the distribution of these stats, we can begin to estimate our customer base!



Histograms

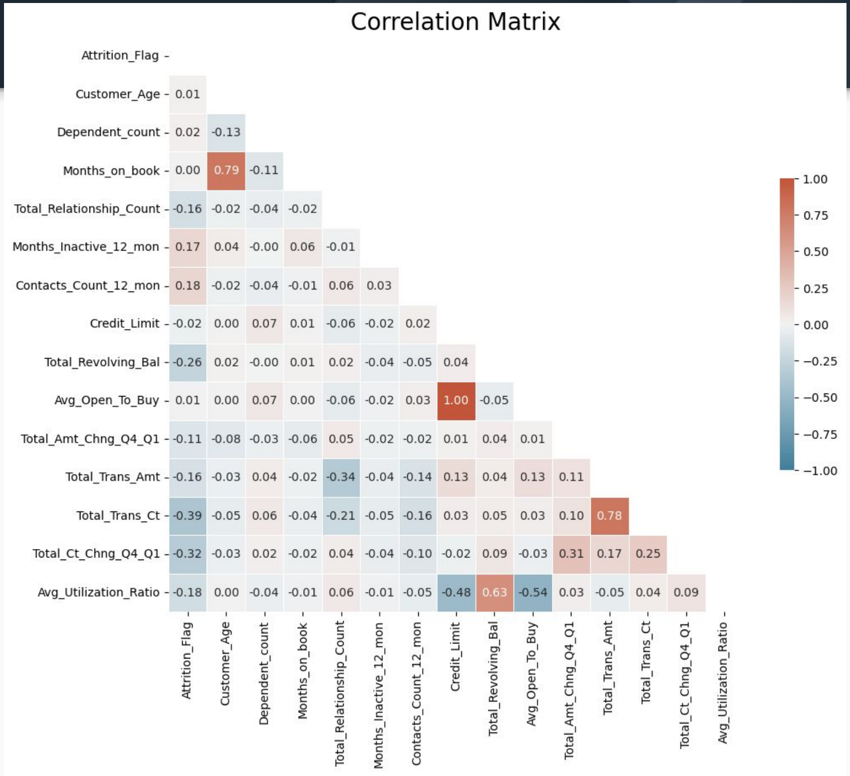
On just a cursory glance, we can already get an idea of distribution of the customer age, credit limit, and transaction counts!



Correlation Matrix

Our dataset has an “Attrition Flag” that marks each client as an “Existing Customer” or an “Attrited Customer”.

We created a correlation matrix to visualize the relationships between the flag and **the other demographics provided**.



Logistic Regression

Using logistic regression, we can begin to correlate the attrition flag with the other demographics.

We can see our accuracy score is looking pretty good for our first model!

```
Logistic Regression Classification Report
                   precision    recall  f1-score   support

Attrited Customer       0.75      0.55      0.63       309
Existing Customer       0.91      0.96      0.94      1554

         accuracy                           0.89      1863
        macro avg       0.83      0.76      0.79      1863
     weighted avg       0.89      0.89      0.89      1863

Accuracy Score: 0.8942565754159957
```

Logistic Regression

To better visualize the logistic regression, we create a confusion matrix.

While a decent score, we want to try and improve the numbers, so next we will use adjusted weights to rerun the model.

Logistic Regression

Logistic Regression Classification Report

	precision	recall	f1-score	support
0	0.97	0.84	0.90	1554
1	0.52	0.87	0.65	309
accuracy			0.84	1863
macro avg	0.74	0.85	0.77	1863
weighted avg	0.89	0.84	0.86	1863

Accuracy Score: 0.8427267847557702

Using adjusted weights for the dataset, we have improved the accuracy score! But precision is not good

Perhaps using other methods, we can raise the number even more? Next we will try k-nearest neighbors (KNN).

KNN Analysis

Accuracy dipped a little!

We are still in a better place than our first Logistic Regression, but there's still room for improvement!

If nothing else, it reinforces our results since we got to similar places by different routes.

KNN Classification Report

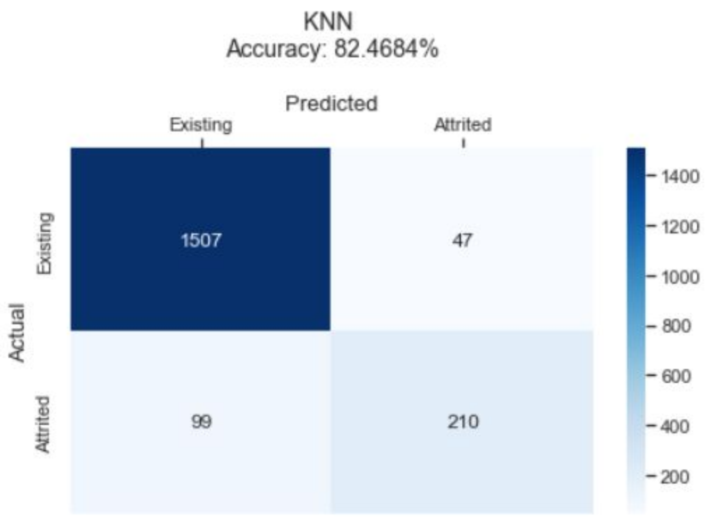
	precision	recall	f1-score	support
0	0.94	0.97	0.95	1554
1	0.82	0.68	0.74	309
accuracy			0.92	1863
macro avg	0.88	0.82	0.85	1863
weighted avg	0.92	0.92	0.92	1863

Accuracy Score: 0.8246835601204534

KNN Analysis

While looking very similar to our original logistic regression, we continue to see a how weighted our dataset is towards existing customers.

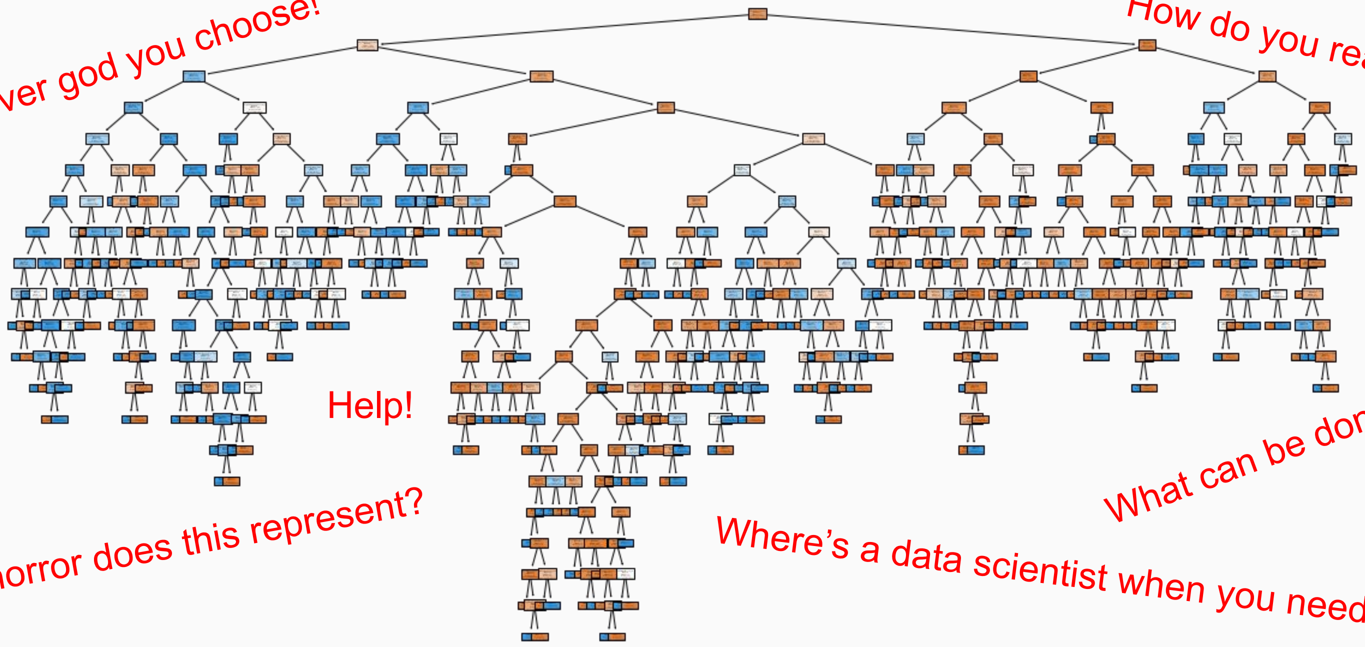
Let’s try another method to better understand our data, how about we view a decision tree!



The Decision Tree from Hell

Pray to whatever god you choose!

How do you read this thing?



Help!

What can be done to contain it?

What horror does this represent?

Where's a data scientist when you need one?

The Decision Tree from Hell

Okay, no. This doesn't really tell us anything. Let's move on to another analysis method: Random Forest!

Random Forest Analysis

- Balanced accuracy score: 91%
- Precision: 95%
  - When our model predicts a customer as churned, it is right 95% of the time
- Recall: 84%
  - Our model correctly identifies 84% of all churned customers

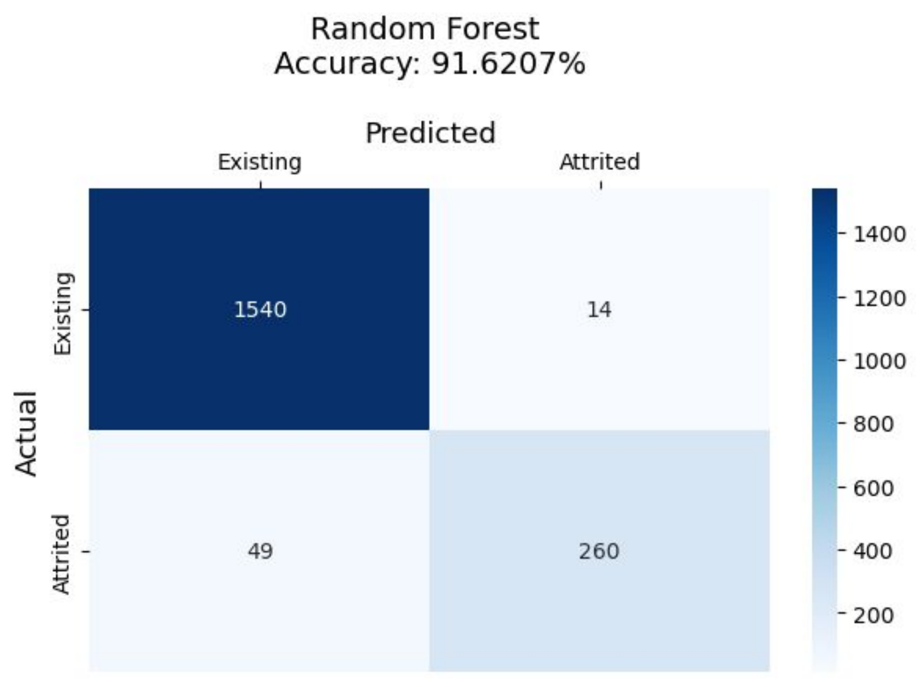
Random Classification Report

	precision	recall	f1-score	support
0	0.97	0.99	0.98	1554
1	0.95	0.84	0.89	309
accuracy			0.97	1863
macro avg	0.96	0.92	0.94	1863
weighted avg	0.97	0.97	0.97	1863

Accuracy Score: 0.9162074696055278

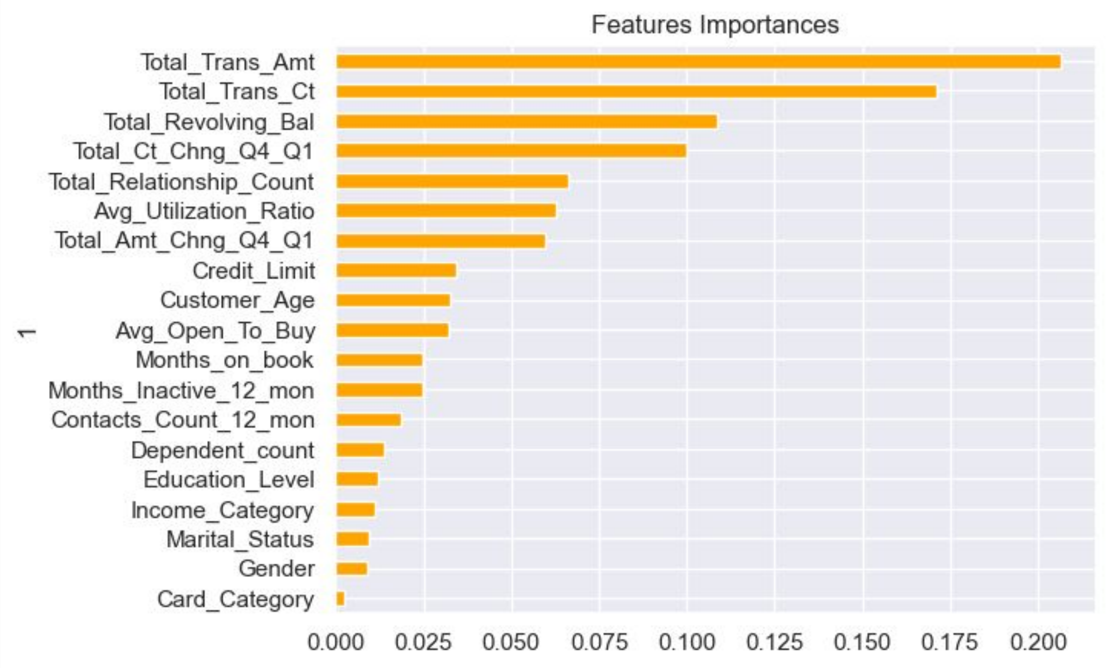
Random Forest Analysis

- Visualize model performance using confusion matrix
- Better understand the recall and precision of the model
- Better visualize the imbalanced class distribution



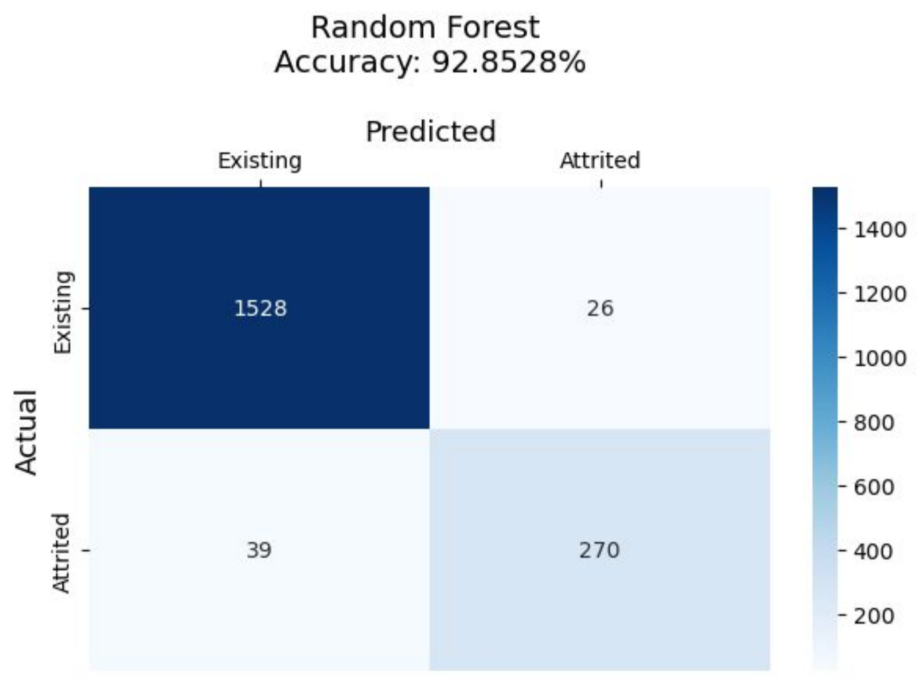
Optimizing with Feature Importance

- Utilized random forest model to identify the most important features
- Rebuilt models and trained only using selected features
- Trained using features from Total Transaction Amount to Average Open to Buy



Optimized Models

- Retrained all models using only selected features
- Random Forest model still our most accurate
- Slight improvements in accuracy and recall, slight decrease in precision



Thank you!

Questions?