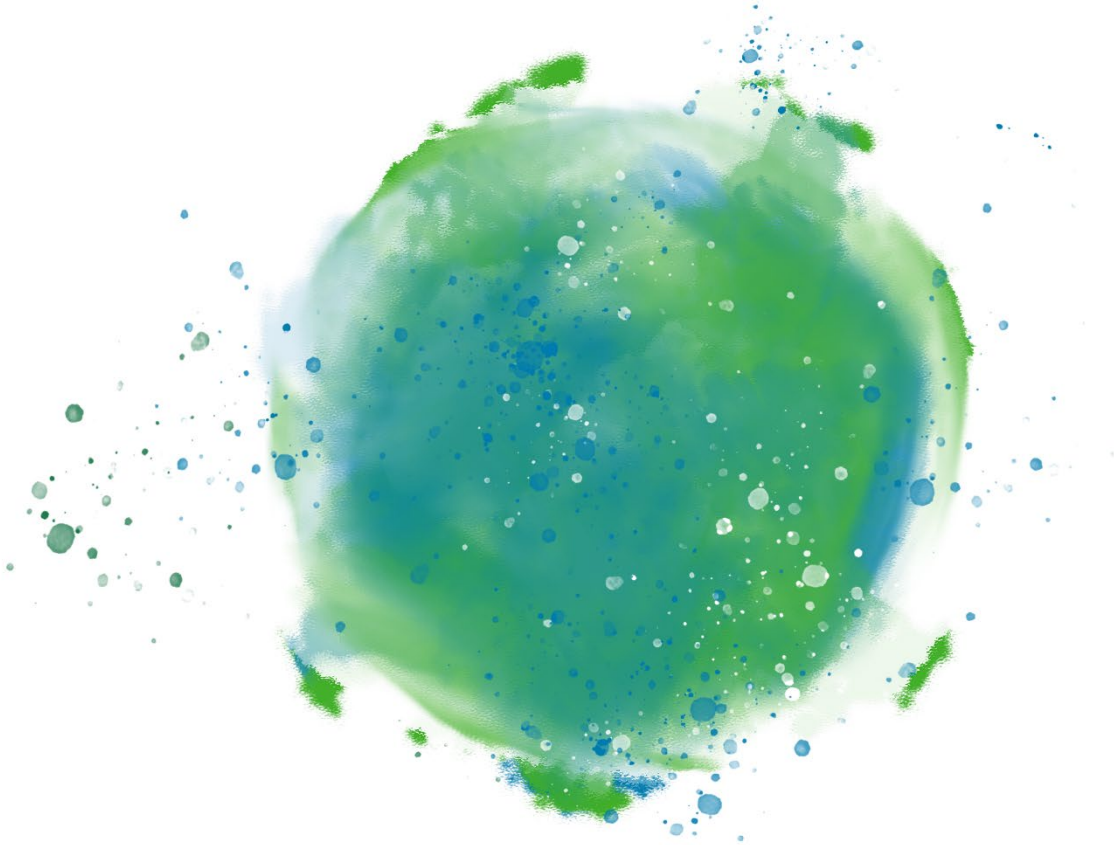


Capstone Group 2
Credit Card Churn Analysis

Capstone Group-Two Team Members

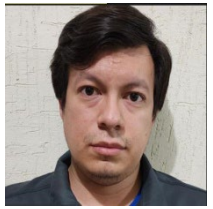
Quinn Sencenbaugh
Advisory Analyst

Aidan Surowiec
Consulting Analyst

Paniz Herrera
Advisory Analyst

Paola Malagon
Consulting Analyst

Max Schliesman
Consulting Analyst



Juan José González
Advisory Analyst

Contents

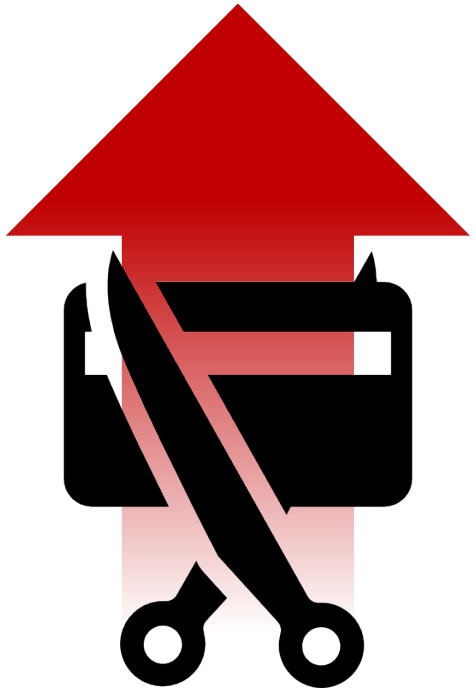
Introduction

Project Overview

A Business manager at Credit Card Incorporated (CCI) bank hired our team to help formulate a strategy to mitigate their pressing business problem.

High Customer Attrition

The bank has an issue with retaining customers, high numbers of customers are cancelling their credit cards.



Why is this a Problem?

Reduced Customer Lifetime Value

Customers who churn will reduce the overall potential revenue generated.

Increased Customer Acquisition Cost

High churn rates can increase the cost of acquiring new customers – resulting in higher marketing and sales costs.

Churn Hurts Company Valuation

High churn rates indicate that business has difficulty retaining customers, resulting in a less attractive business opportunity.

Churned Clients can be Vocal

Negative reviews for why customers stopped services could impact ability to land new clients.

Business Understanding

Layout of the business question, what this solution means for CCI, and an overview of our solution approach.

Key Business Question:

Credit Card Churn Prediction aims to predict if a customers will cancel their card.

Additionally, what factors are the strongest predictors for a customer to cancel.

Stakeholder Understanding:

A business manager from Credit Card Incorporated (CCI) bank is trying to reduce his attrition rates. Lower attrition means more people are using their credit card for longer, increasing bank profits.

Our Solution

Our team will utilize AI and ML models to analyze the credit card data. We will determine which factors are the strongest predictors of attrition and which model will provide the most accurate results. With an understanding of what leads to attrition and which model to use, we can guide our recommendations to CCI.

Data Understanding

Data Understanding

The data source we used was from Kaggle. The dataset is from a bank manager trying to reduce attrition, the same problem we are trying to solve.

Credit Dataset

Rows: 10,127

Columns: 19*

Null Values: 0

Variable Type

Demographic
6 Variables

- *4 variables are demographic and categorical*
- *Example columns:*
  - ***Gender***
  - ***Education Level***
  - ***Marital Status***

Financial
13 Variables

- *1 variable is categorical and a financial / account variable:*
  - ***Card Category:*** *Credit card tier of customer*
  - *Blue, Silver, Gold, Platinum*

Categorical Variables *5 Variables*

- *2 variables are demographic and numeric:*
  - ***Customer Age***
  - ***Dependent Count***

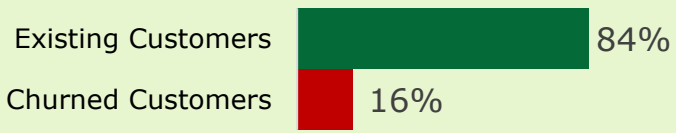
Numeric Variables *14 Variables*

- *12 variables are financial and numeric*
- *Example columns:*
  - ***Credit Limit***
  - ***Total Transaction Amount***
  - ***Average Utilization Ratio***

Data Exploration

Class Imbalance

Percentage of Observations by **Attrition Flag**



High Correlation

Some variables are highly correlated, such as:

- **Average Utilization Ratio** vs. **Total Revolving Balance**: +0.62
- **Card Category** vs. **Credit Limit**: +0.49

Target Relation

Our team also examined which variables are closely related with attrition, some examples:

- **Number of Transactions**: -0.37
- **Total Revolving Balance**: -0.26

**The original dataset has 23 columns; the analysis is moving forward with 19 of these*

Data Preparation

To properly analyze the data and make recommendation to reduce attrition, our team took several steps to clean the data.

Unclean Data

Dropped Irrelevant Columns

Three columns dropped:
Client number (ID) was dropped, ID numbers should have no influence on our predictions.

Two additional columns were dropped due to the author note from the dataset, these were calculated values.

Evaluated "Unknown" Values

No Null Values, "Unknown" Values Present:
After evaluating value counts, it was clear the three categorical columns had observations with "Unknown" string values.

Income Category, **Education Level**, and **Marital Status** had unknowns – only seven observations had "Unknown" as the value for all three, equal to 0.07% of the data.

The small number of instances with multiple unknowns was unlikely to skew results.

Encoded Categorical Variables

Dummy Variables:
Categorical variables that had no natural order or rank were encoded using dummy variables.

Gender and **Marital Status** were the two categorical variables that fit this description. The encoding of these variables resulted in six new columns.

Label Encoder

Ranked Encoding:
The remaining three categorical columns did have a natural order, these were manually encoded for each value to sequentially represent a rank.

Card Category was ranked as Blue being the lowest and Platinum being the highest.

Education Level and **Income Category** both had "Unknown" values present, those were encoded as 0.

Ready for Modeling

Solving the Problem

Based on the data, how can we approach the customer attrition issue?

The goal is to predict the **attrition flag** of any given customer based on the input data.

The **attrition flag** is an identifier for whether a customer cancelled their credit card or not, since the outcome is either yes (1) or no (0), this is a ***binary classification*** problem.

To solve the ***binary classification*** task, **four models** were tested and developed.

A **prediction model** enables CCI to identify at-risk groups, understand key factors for customers who churn, and make necessary strategic recommendation.

Modeling

Modeling Approach

The binary classification task can be approached in many ways, multiple models were tested in efforts to offer the optimal solution for this case.

Which models were tested and why?

Logistic Regression: Simple but effective approach.

Decision Tree: Alternative, highly interpretable classification model.

Random Forest: Complex model that typically give more accurate results.

XGBoost: Highly complex model that works well with unbalanced data.

The final model will ideally have high ***recall*** – we would like to detect as many exiting customers as possible.

It will be better for the model to predict incorrectly that a customer will leave (*False Positive*), rather than predicting incorrectly they will stay (*False Negative*).

Logistic Regression

Logistic regression is a statistical model that is often used for classification and predictive analytics.

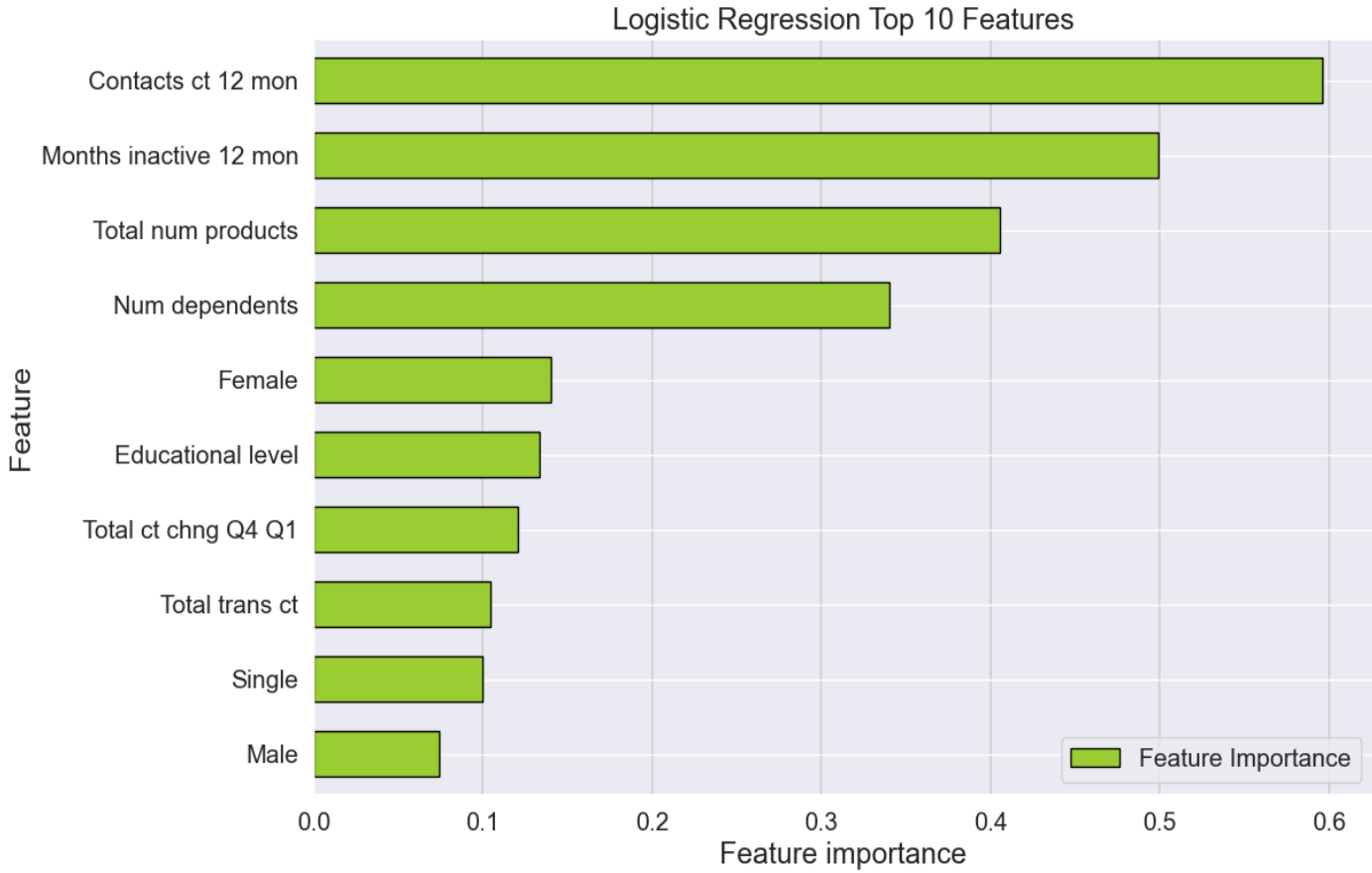
Modeling Logistic Regression

Feature Importance

- Feature Importance is an indicator of what variables are the *strongest predictors for customer attrition*.
- The Feature with the highest correlation is the *number of contacts over the past year.*
- The *number of months inactive over the past year*, and the *total relationship count* are also Features that are highly correlated.

Performance Metrics

- The accuracy of the logistic regression was able to reach 89%.
- Recall and F-1 score are going to give a better sense of the performance for our situation. Recall was 74% and the f-1 had a .77. This is significantly lower performance than the accuracy would indicate.



The logistic model performed well with the best model reaching 74% recall and .77 f-1 score.
Because this is such a simple model, we are confident that increasing model complexity will allow us to develop a model better suited for production.

Decision Tree

Decision Tree is a supervised machine learning model, are used to classify by partitioning recursively the sample space until we get a discrete class.

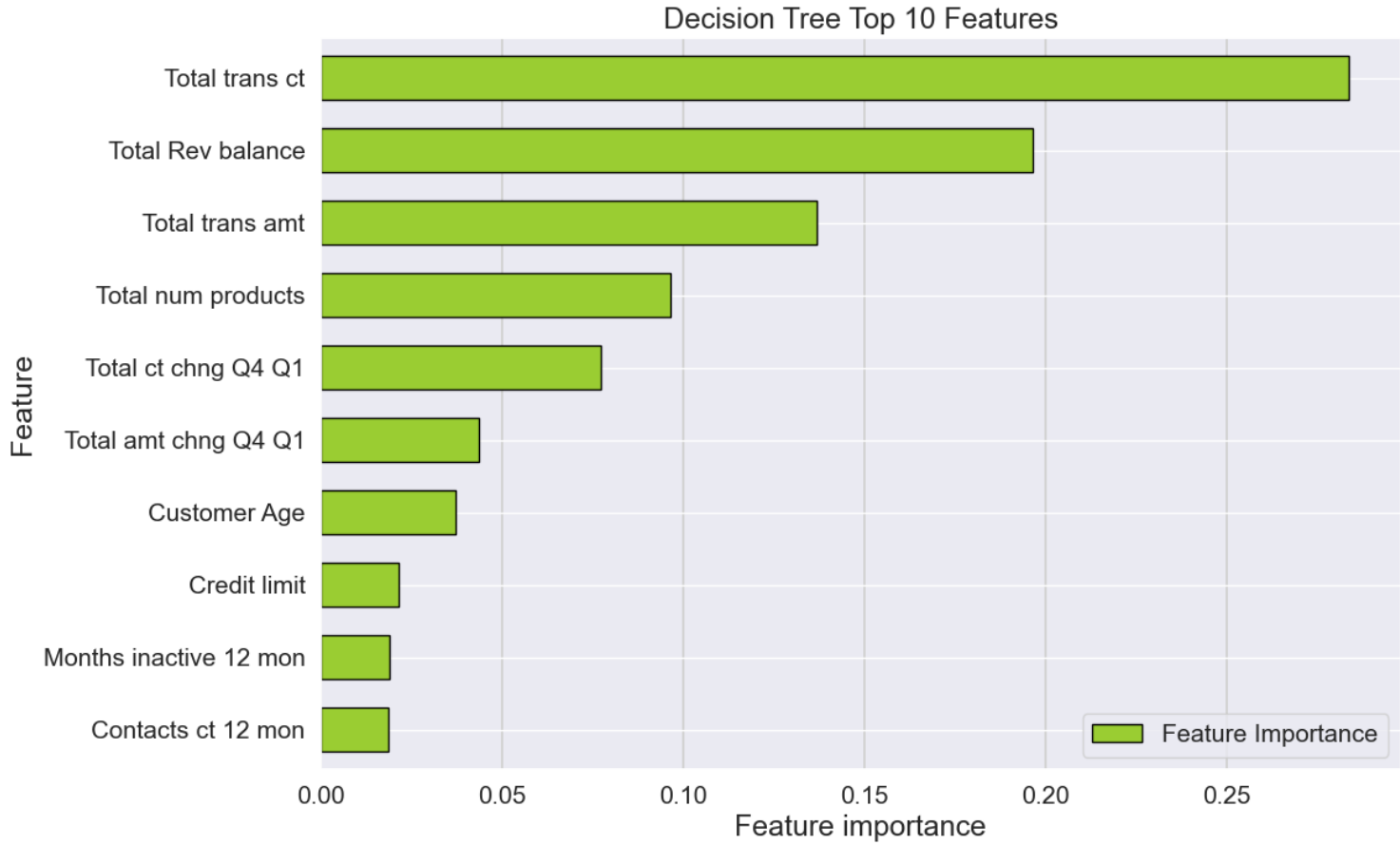
Modeling Decision Tree

Feature Importance

- The Feature with the highest performance is the amount of times that *a customer uses their card*.
- The *amount of transactions* and the *revolving balance a customer has on their card* are also very strong predictors.

Performance Metrics

- Decision tree reached an accuracy of 94%.
- The recall score was an 89% and the f-1 score was .90. This is a significant improvement over the logistic regression model.



The decision tree achieved a better recall than the logistic regression with a 15% increase in score. The problem with this model is a relatively high number of false negatives. Our team wants to reduce the false negative even further.

Random Forest

Random Forest is a supervised learning algorithm that makes use of decisions trees and can be used for classification and regression tasks.

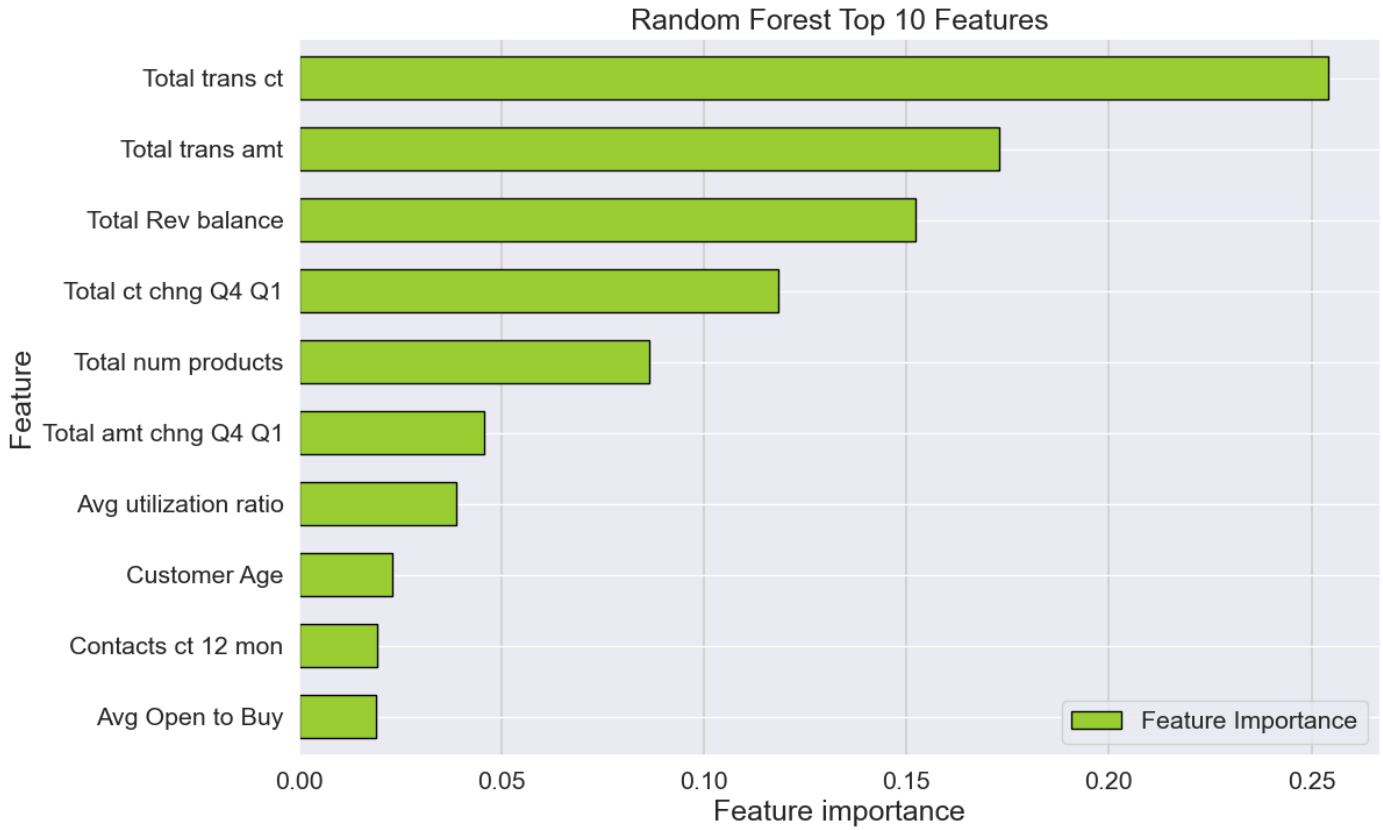
Modeling Random Forest

Feature Importance

- Feature Importance is an indicator of what variables are the *strongest predictors for customer attrition*.
- The Feature with the highest performance is the *amount of times that a customer uses their card*.
- The *Total Revolving Balance* and the *Total Amount of Transactions* are also important predictors of attrition.

Performance Metrics

- The default setting for a random forest reached an accuracy of 90.6%, after tuning the hyperparameter an accuracy of 96% was achieved.
- Accuracy is a crude metric, so for a better idea of performance we can use an f-1 score and recall. The random forest had a macro f-1 score of .93 and a recall of 91%.



The random forest performed the best so far, however, the recall rate on the test data still has room for improvement. By using a complex model that better handles unbalanced data, we may be able to predict less false negatives.

XGBoost

XGBoost is a machine learning algorithm which combines the predictions of multiple other models to create strong predictive models.

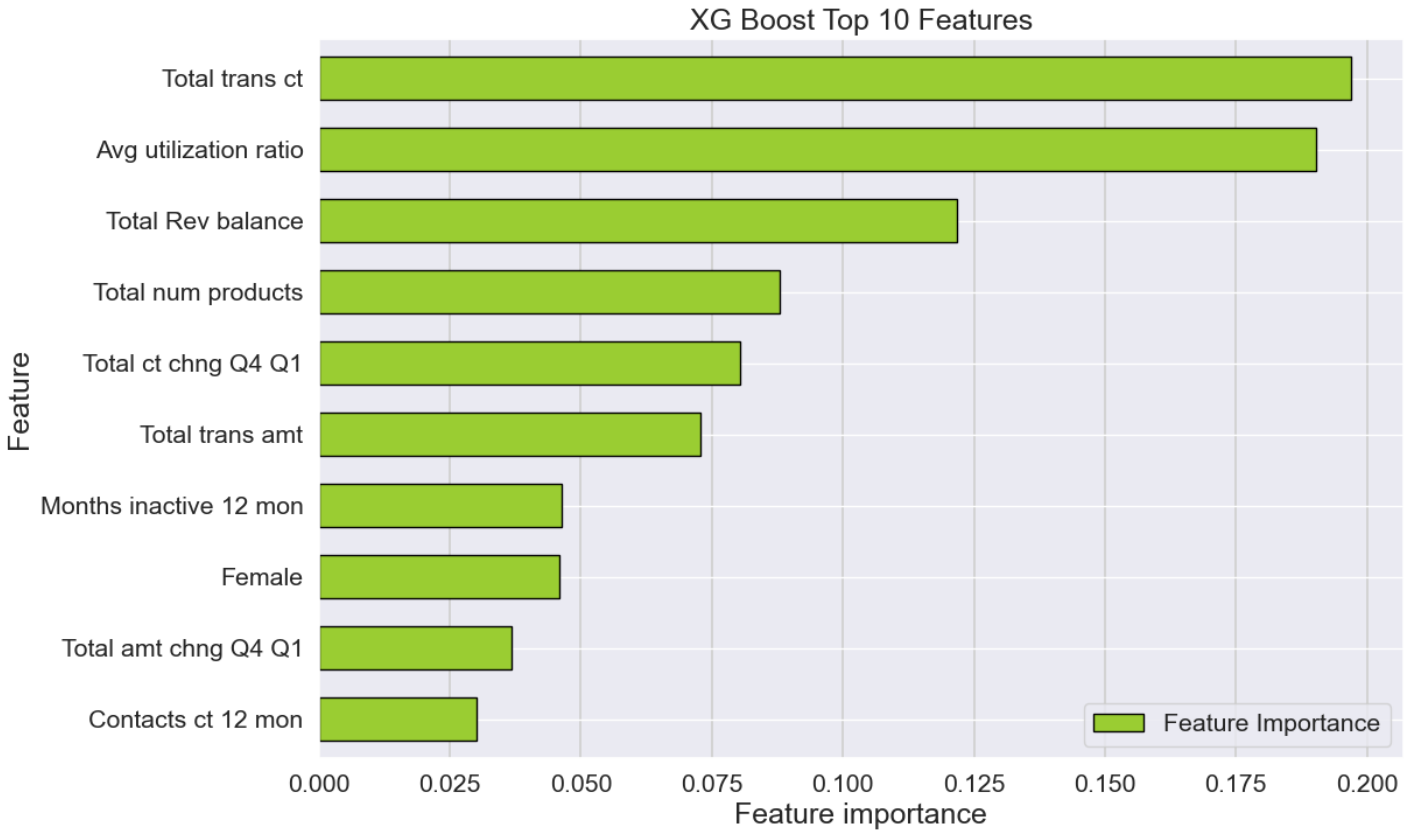
Modeling XGBoost

Feature Importance

- Feature Importance is an indicator of what predictors are the *strongest for customer attrition*.
- The variable with the highest Feature Importance is *Total Transaction Count*.
- 9 out of 10 of the most important Features are *financial variables*, demographic variables do not appear to impact customer attrition as much.

Performance Metrics

- After hyperparameter tuning, the model scored a high accuracy of 94.9%.
- Class imbalance was addressed in hyperparameters.
- The recall and F1 scores on the test data were 0.96 and 0.86 respectively.
- The optimal model predicted 17 false negatives.



The XGBoost model resulted in the highest recall score and lowest false negatives out of all the models. The precision and F1 scores suffered but this model was great at achieving optimization in the metrics we deemed most necessary.

Model Evaluation

Comparison of each model's performance across various metrics on the testing data.

Finding the Optimal Model

High Priority Metrics:

Recall: Best metric for evaluation when there is a high cost of False Negatives.

False Negatives: Predicting customers will not cancel and they do - low score is ideal.

F1: Metric that combines Precision and Recall.

Low Priority Metrics:

Accuracy: The number of correct predictions by the model.

Precision: Best metric for evaluating when there is a high cost of False Positives.

Model Scores

Model	Accuracy	Precision	Recall	F1	False Positives	False Negatives	True Positives	True Negatives
Logistic Regression	89%	81%	74%	.77	81	186	198	2067
Decision Tree	94%	90%	89%	.90	81	67	351	2033
Random Forest	**96%**	**95%**	**91%**	**.93**	**84**	**24**	**348**	**2076**
XGBoost	95%	79%	96%	.87	106	17	402	2007

The Solution Model

Deep dive for choosing the ideal model for providing a solution to the customer attrition issue.

Chosen Model:

Random Forest Classification Model

- Can be tuned to handle unbalanced data, should be able to adapt to a production environment well.
- Highly interpretable model relative to XGBoost.
- F1: 0.93 (*1st*),
- Accuracy: 96% (*1st*),
- Recall: 91% (*2nd*),
- False Negatives: 24 (*2nd*).

Using the Model:

Applications

- Random Forest will be able to compute churn predictions in real time and can flag new customers as churn risks.
- Give insights into which factors influence churn the most and relay how this changes over time.
- Able to introduce parameter values that deal with natural data imbalance.

Our team recommends for CCI to implement a **Random Forest** model to reduce their attrition rates and increase business performance.

Conclusion

Solving the Problem and Recommendations

The identified ideal model for this classification problem is Random Forest – how will this be used to solve the problem?

Strategic Impact

Exiting Factors

By understanding elements associated with exiting customers, CCI could begin testing targeted incentivized solutions to mitigate typical churn risk areas.

Shift Focus

Adjust the marketing strategy to focus primarily on the groups that are not predicted to be at-risk. CCI could confidently change the targeting tactics for new customers.

Support High Risk

Reach out to the high-risk customers and provide support or gauge where there is a gap in needs being met.

Re-evaluate Larger Initiatives

The modeling results can invite conversation for re-evaluating larger retention initiatives such as credit card interest rates, benefits, etc.

Anticipated Implementation Outcomes

Increase Customer Lifetime Value

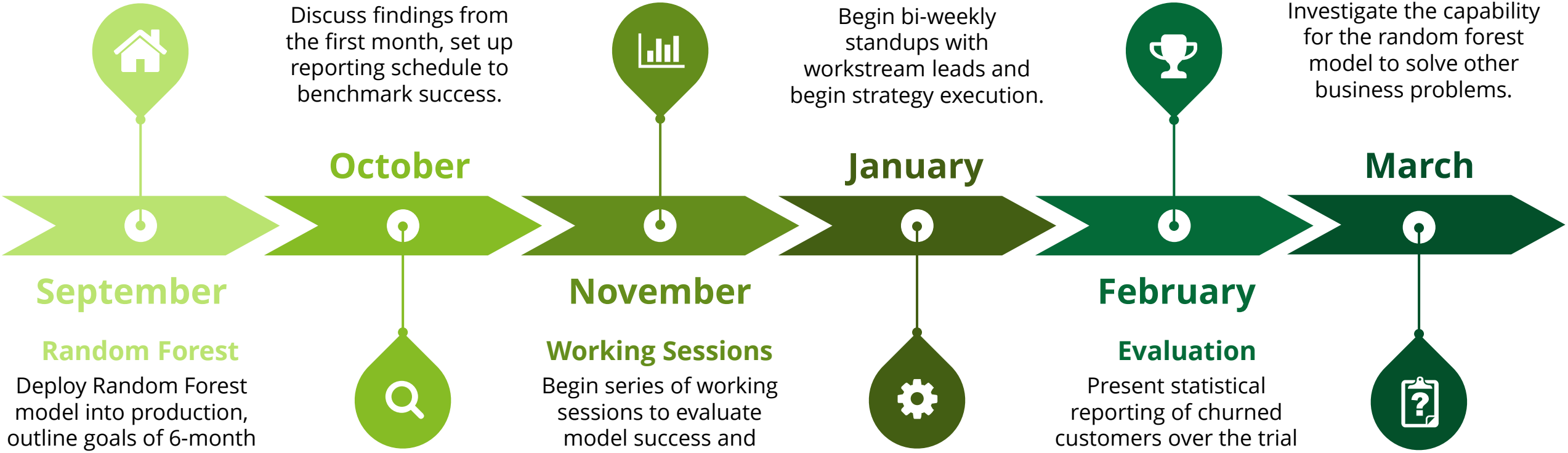
Decrease Customer Acquisition Costs

Improve Company Valuation

Satisfied Customers

The Road Ahead

The Random Forest model can provide more than our time-of-analysis suggestions, here is how we think the next six months should look for CCI – if this approach is working, CCI should consider expanding their use of AI solutions.



Thank You!

Questions?

Appendix