



DECODING CUSTOMER CHURN

A Data-Driven Approach in the
Banking Industry

Presented By: Team 18

Cuthbert Liaw, Yali Chen, Nayeema Nonta



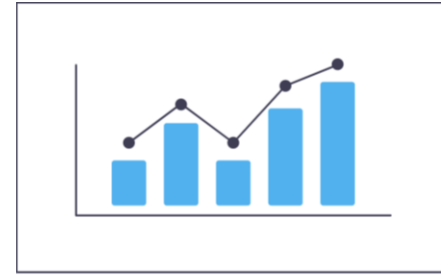
Why is Bank Churn Important?

High customer churn rates impacts regulatory compliance and profit maximization for banks. With the rise of digital banks offering competitive rates, bank churn has become a pressing issue (Marous, 2023).

How can we proactively **anticipate customer churn risks**, ultimately allowing banks to **deploy countermeasures** in ensuring deposit retention in the evolving banking industry?

Descriptive Problem



What can we learn about the dataset? Where is the dataset typically distributed?



Age, Gender, and Balance of Payments are Major Attributes of the Dataset

Age	Gender	Dependents	Occupation	City	Branch Code	Current Balance	Average Monthly Balance (prev. Q)	Current Month Credit	Previous Month Credit	Current Month Debit	Date of Last Transaction
66	Male	0	self-employed	187	755	1458.71	1458.71	0.2	0.2	0.2	2019-05-21
35	Male	0	self-employed		3214	5390.37	7799.26	0.56	0.56	5486.27	2019-11-01
31	Male	0	salaried	146	41	3913.16	4910.17	0.61	0.61	6046.73	NaT
42	Male	2	self-employed	1494	388	927.72	1643.31	0.33	714.61	588.62	2019-11-03
.....	and more										

Table 1.1., Selected Attributes Shown from Kaggle Data



Statistical insights show data centers around 48 y.o., with 60.9% & 60.8% of self-employed and men, respectively.

Data Cleaning



We cleaned the given data, removing any columns with blank values.



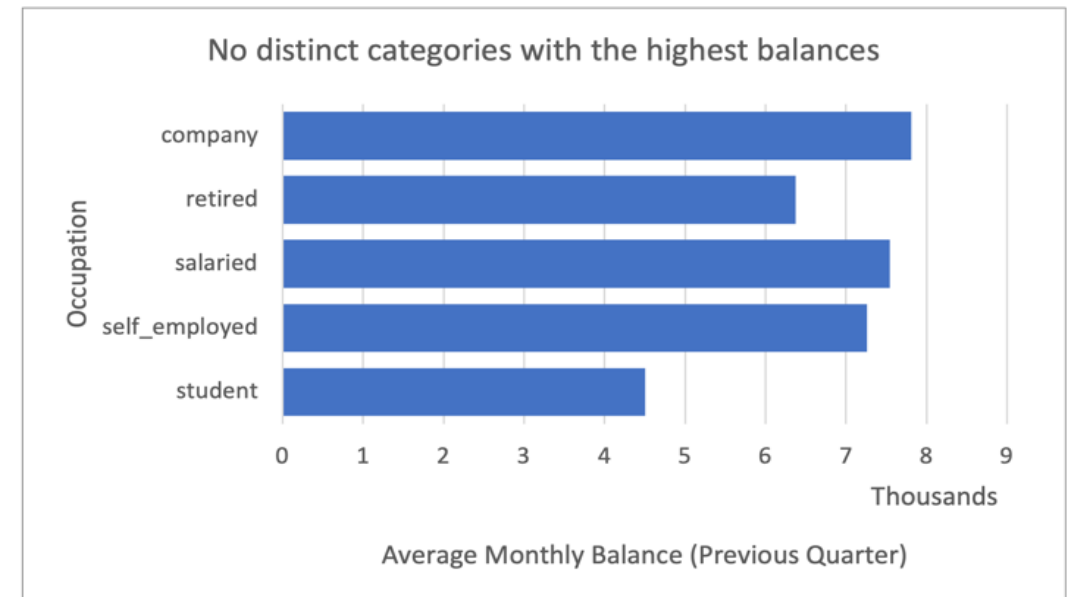
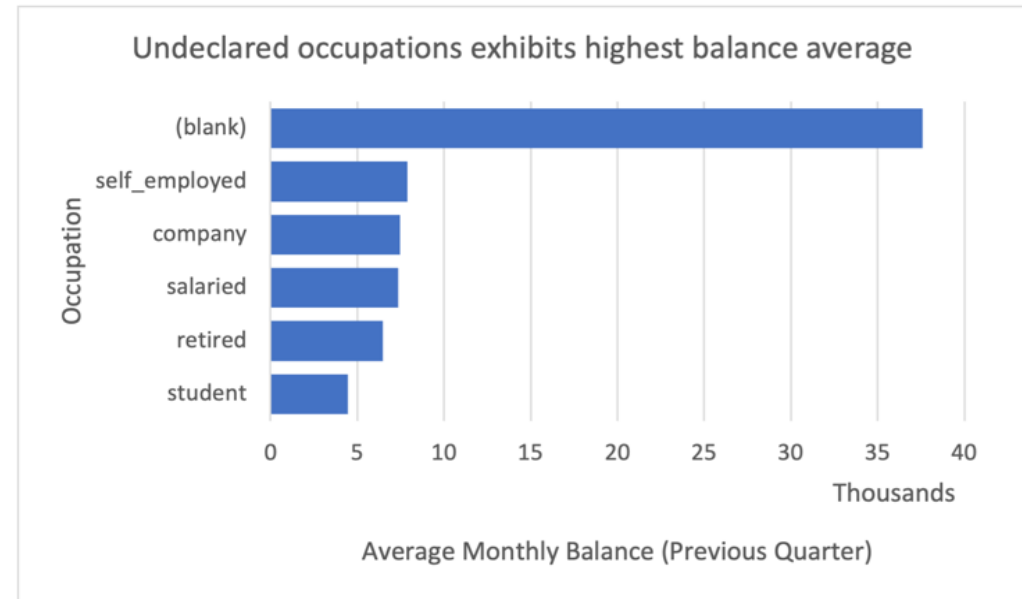
Exemplified on the chart, **undeclared occupation** exhibits highest balance average.



We also removed any **outliers with credit percentage above 700%**.



Removing outliers did not impact feature distribution*

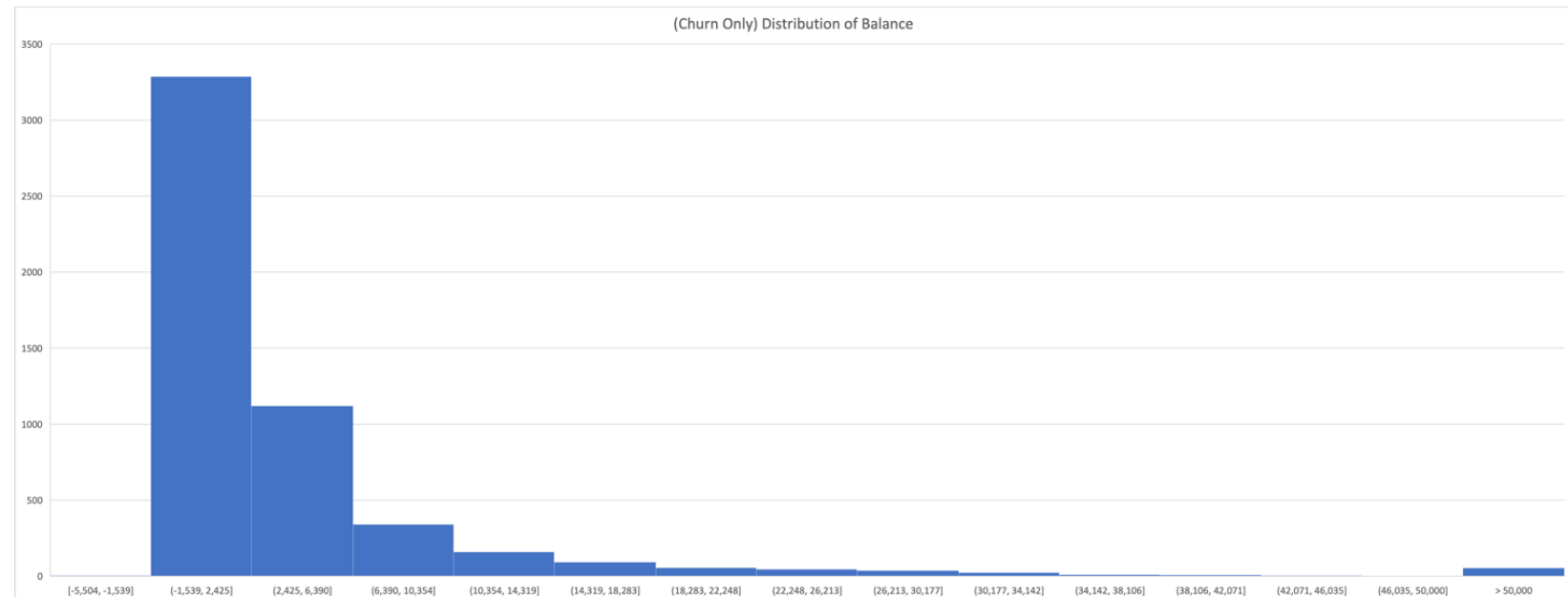
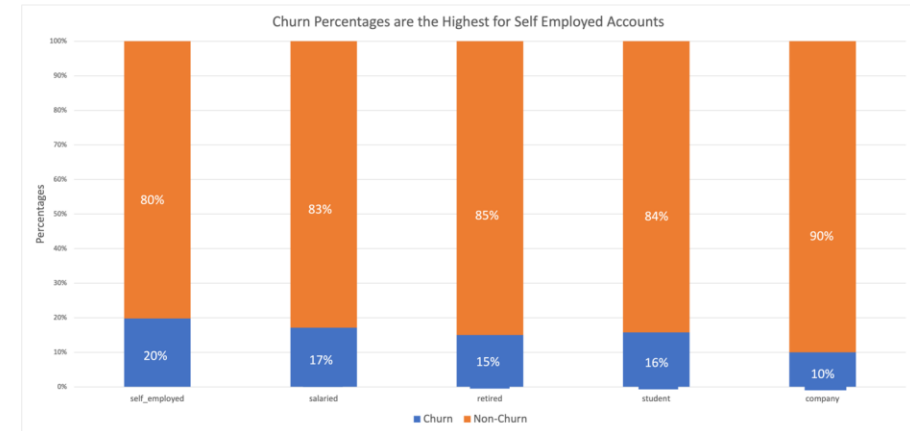
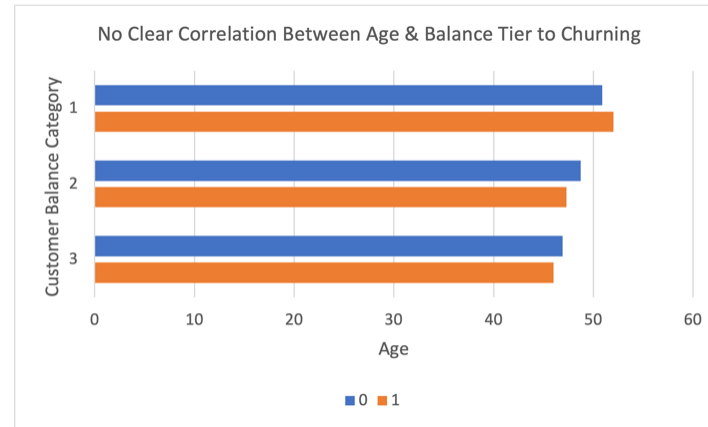


*Based on analysis done on % of balance churned, and balance tier to churning statistics

Data Analytics

Churn percentages are highest for self-employed, with **23% of all self-employed accounts churning**.

While most churning happens on accounts with balances < \$2500, there are **no clear correlation between age & balance tier to churning**, and instead are only proportional to the total balance.



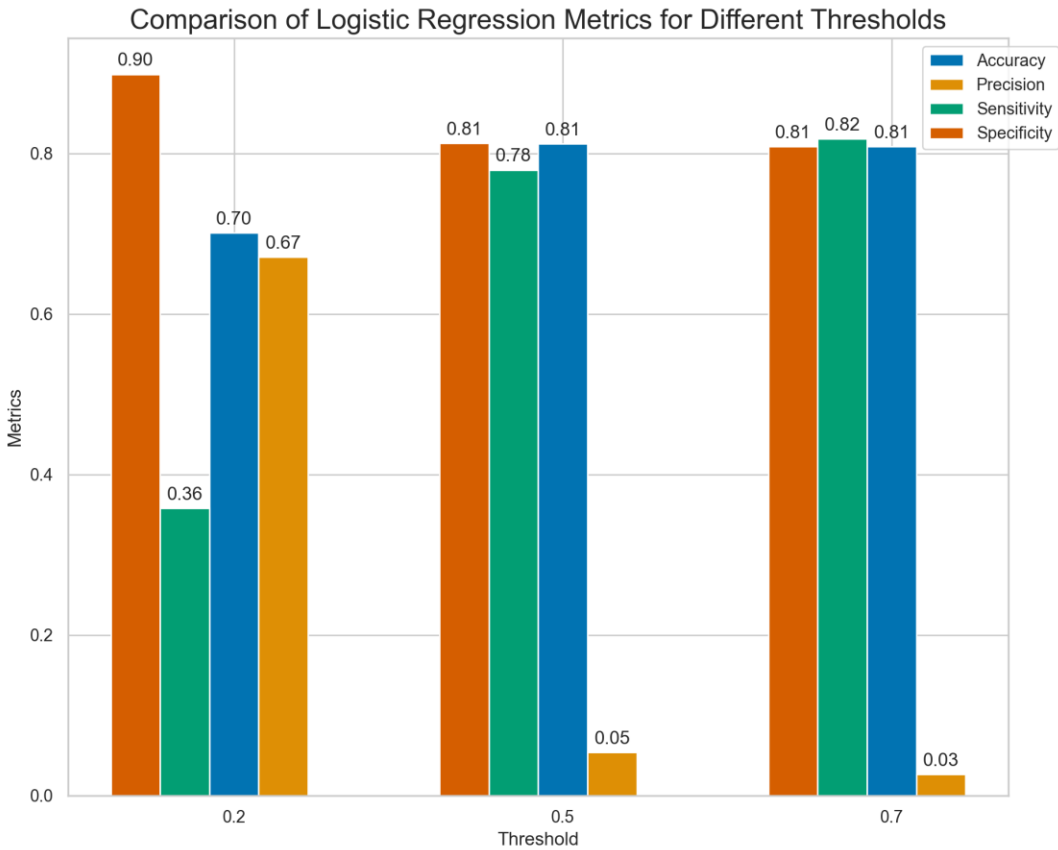
*All visualizations are performed on cleaned data set

Predictive Problem

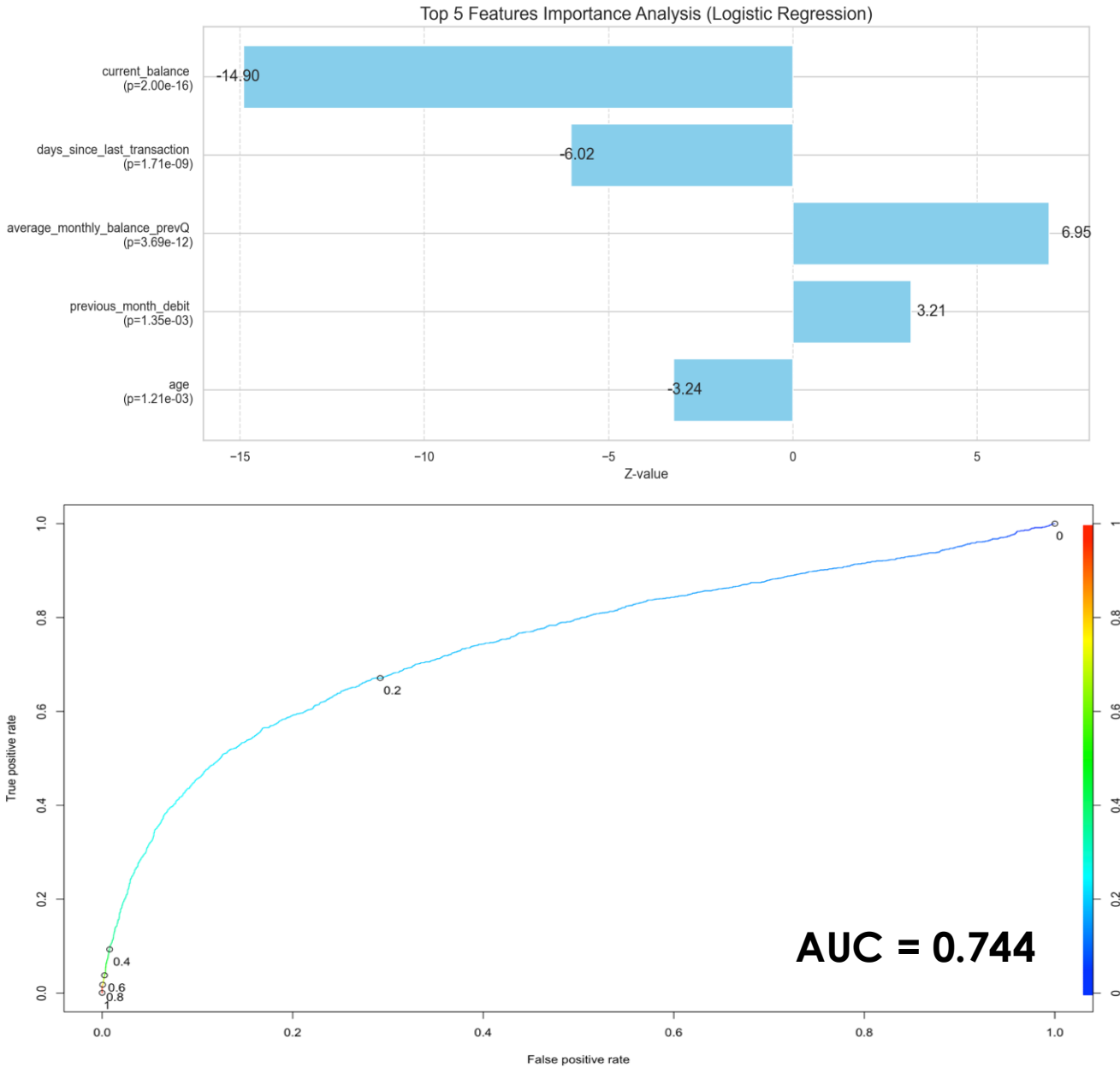
How can we use features to predict churning potential?



Logistic Regression



Accuracy	Sensitivity (TPR)	Specificity (TNR)
0.701	0.358	0.899



Decision Tree



Decision tree first splits on balances to achieve the most information gain, to then use credit balance.



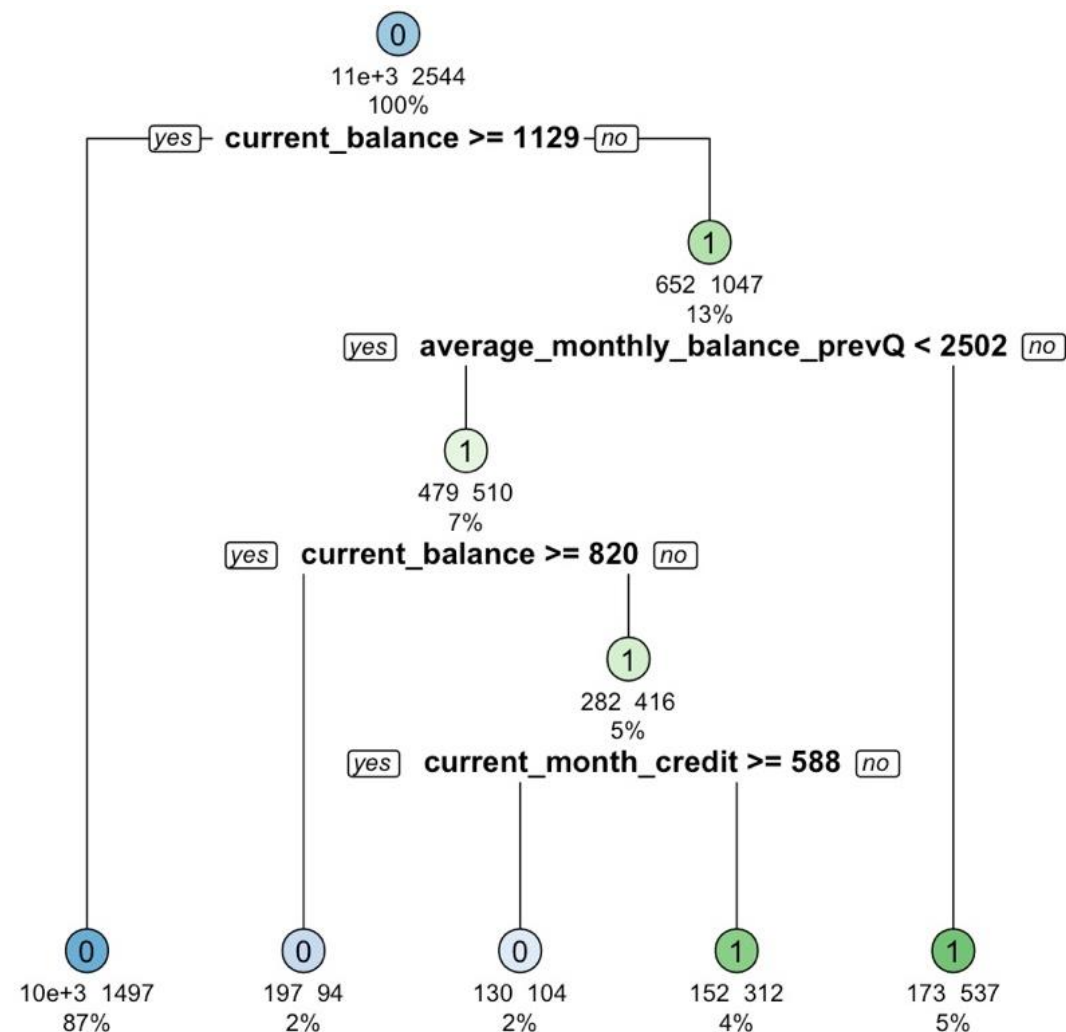
There is **87% accuracy** simply by having **balance below 1129**. This signifies current balance as the most significant predictor for decision tree.



Decision Tree has an overall accuracy of 0.837. It has a higher accuracy, but **lower AUC compared to logistic regression**.



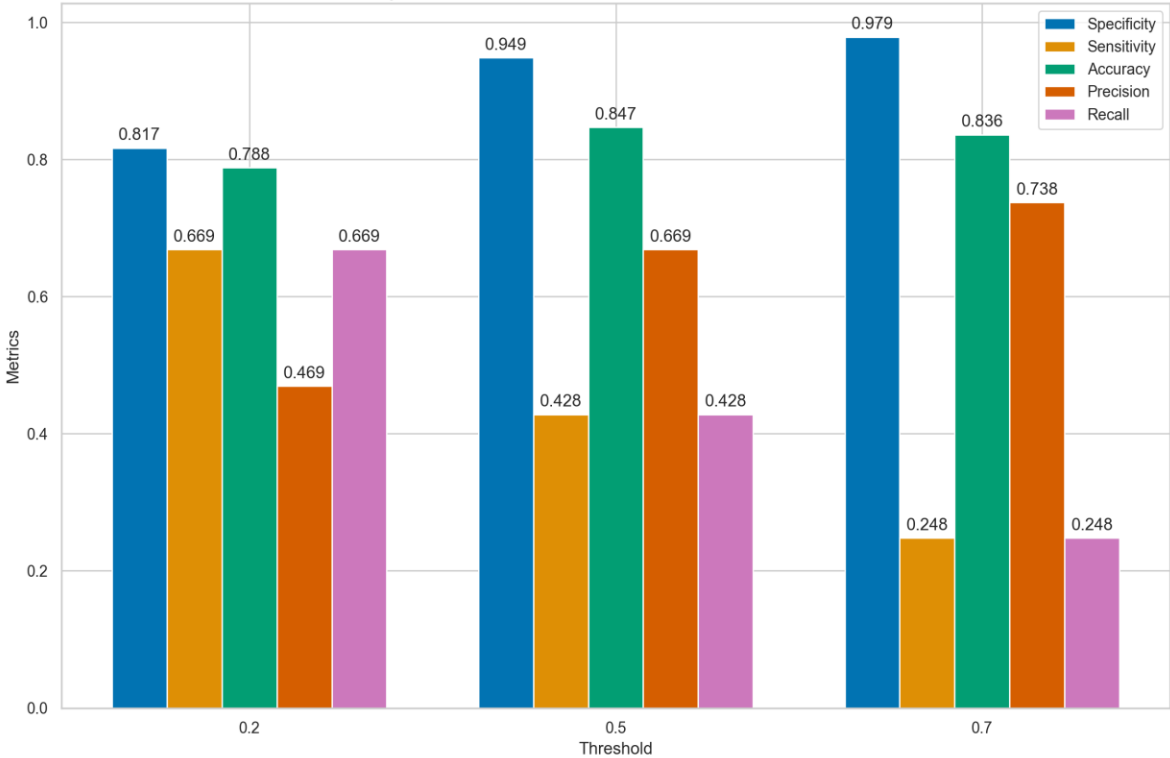
This might imply overfitting, which is prone in decision tree models.



AUC = 0.671

Neural Network & SVM (Support Vector Machine)

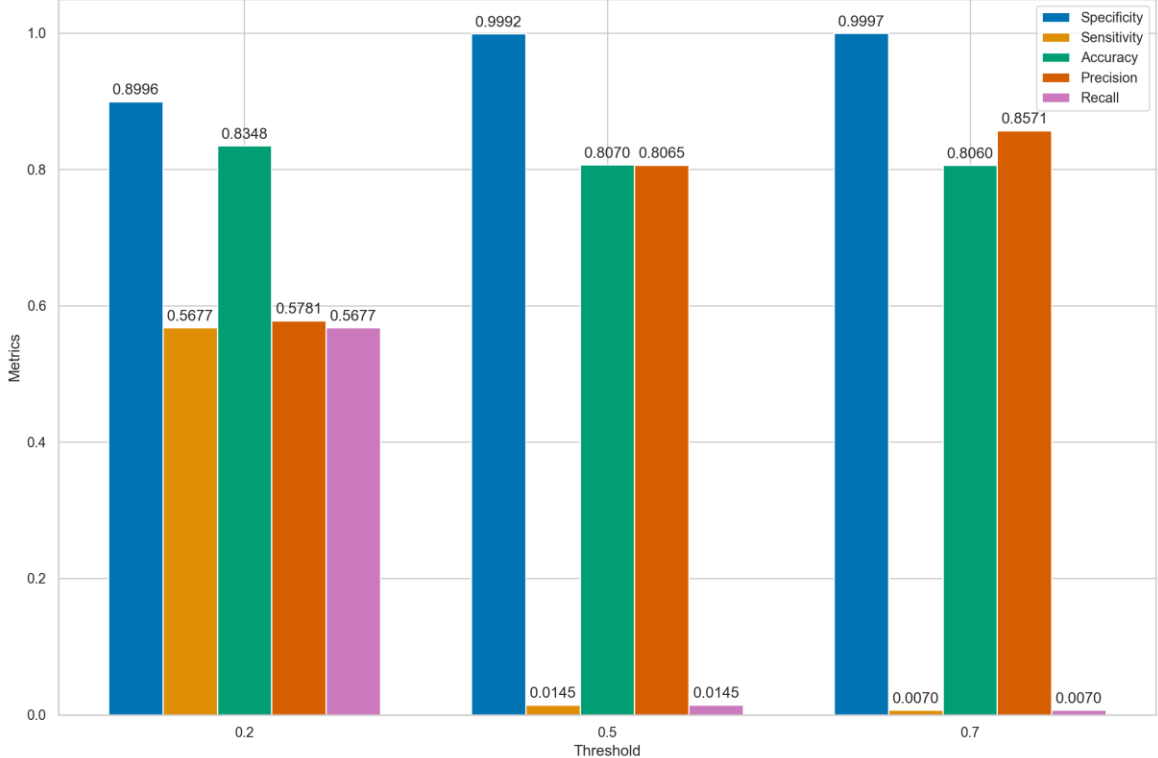
Comparison of NN Metrics for Different Thresholds



AUC = 0.80

Accuracy	Sensitivity (TPR)	Specificity (TNR)
0.788	0.668	0.8167

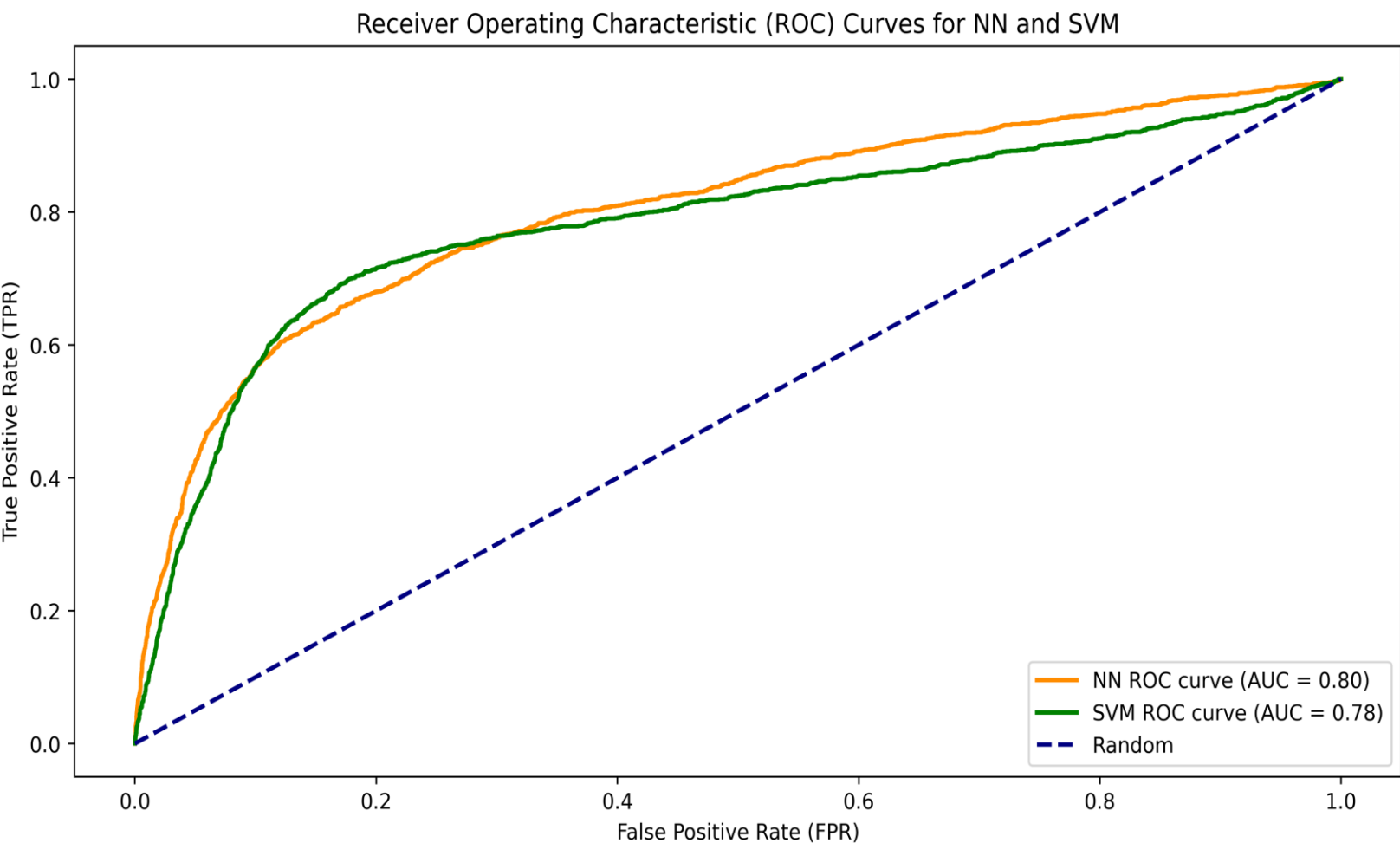
Comparison of SVM Metrics for Different Thresholds



AUC = 0.78

Accuracy	Sensitivity (TPR)	Specificity (TNR)
0.834	0.5677	0.8996

Neural Networks & SVM produced the best AUC at 0.80 and 0.78, indicating a powerful distinguishing capability.



Balance & its derivations acts as strongest predictors, with credit balance acts as a strong second predictor.

SVM Top 5 Most Important Features	Feature Coefficient
current_balance	-0.26136
average_monthly_balance_prevQ	0.199397
current_month_balance	0.029027
current_month_credit	-0.017453
current_month_debit	0.013953

Comparison Between Models

Logistic Regression tends to favor sensitivity, (TPR) while SVM excels in specificity (TNR), especially at higher thresholds. Neural Networks provide a balanced approach with relatively stable accuracy across thresholds.

The choice between both comes to **client’s choice in having a high FP or FN rate**. SVM with higher specificity will desirable if the cost of FN is extremely high (for example, in asset management firms)

Note: Higher thresholds result in models predicting churn more conservatively, leading to higher specificity but lower sensitivity.

Model (Threshold 0.2)	Accuracy	Sensitivity (TPR)	Specificity (TNR)
Logistic Regression	0.701	0.358	0.899
Neural Network	0.788	0.668	0.8167
SVM	0.834	0.5677	0.8996

Model (Threshold 0.5)	Accuracy	Sensitivity (TPR)	Specificity (TNR)
Logistic Regression	0.812	0.7797	0.8182
Neural Network	0.8471	0.4282	0.9486
SVM	0.807	0.0145	0.9992

Model (Threshold 0.7)	Accuracy	Sensitivity (TPR)	Specificity (TNR)
Logistic Regression	0.809	0.8182	0.8088
Neural Network	0.8361	0.4282	0.9786
SVM	0.806	0.007	0.9997

Prescriptive Problem

How to use churning potential to deploy appropriate countermeasures?



Basic Optimization Model

Single Tier Promotion Model



The objective is to minimize the overall churn rate, given a finite budget.



Get churning probability per customer from the logistic regression model, then multiply that based on promotion classification.



In the basic model, **each customer** irrespective of churning probability/balance **gets the same promotion cost**.



The constraint ensures that total cost of promotions does not exceed the maximum budget.

Parameters

- p_i : Probability of customer i churning, as predicted by the logistic regression model.
- B : Promotion budget per customer.
- B_{\max} : Maximum budget for promotions.

Decision Variables

- x_i : Binary decision variable indicating if a customer receives a promotion.
 - $x_i = 1$ if customer i receives a promotion.
 - $x_i = 0$ otherwise.

Objective Function

$$\text{Maximize } \sum (p_i \cdot x_i)$$

This objective seeks to prioritize promotions for customers with a higher probability of churning.

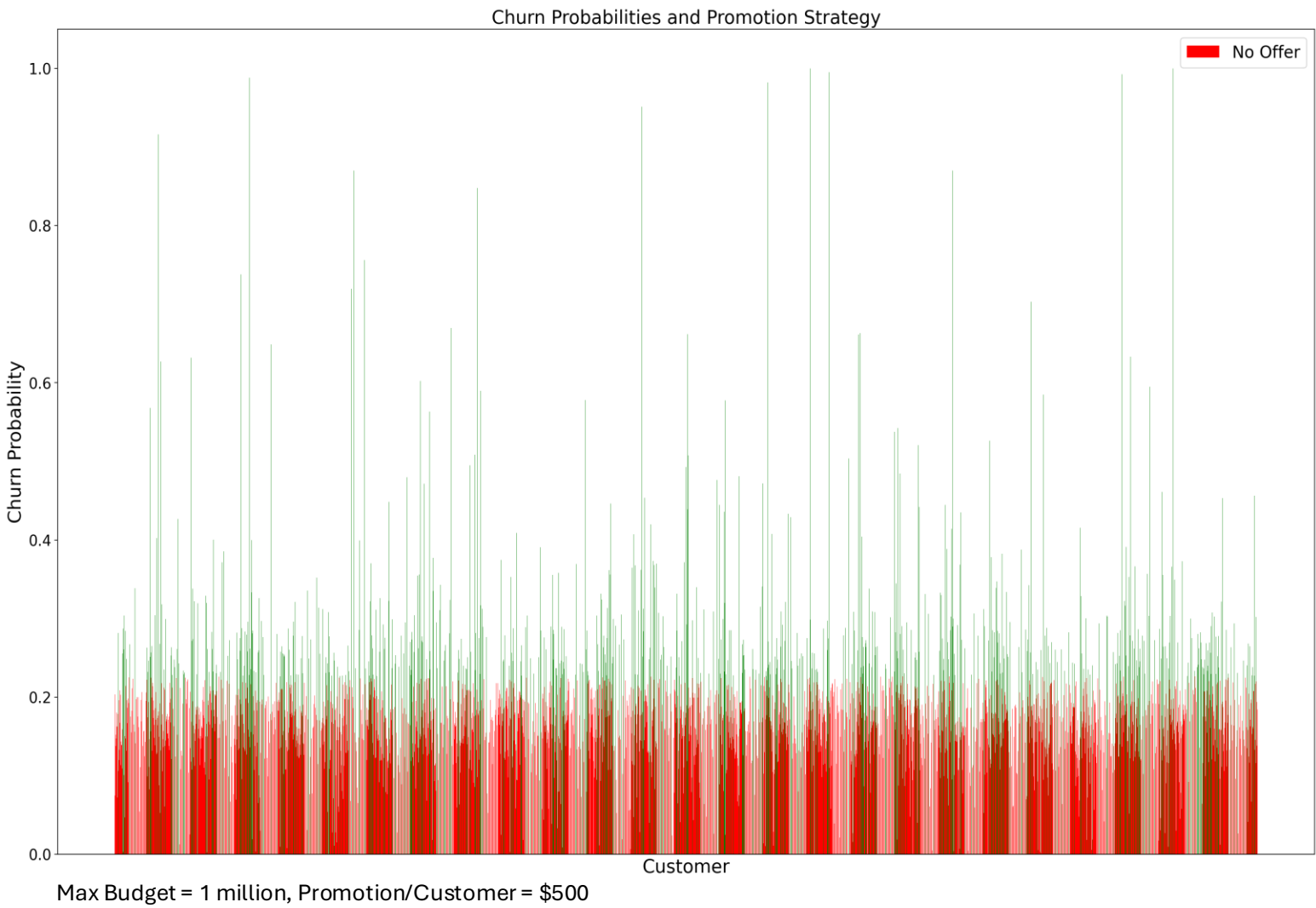
Constraints

Budget constraint

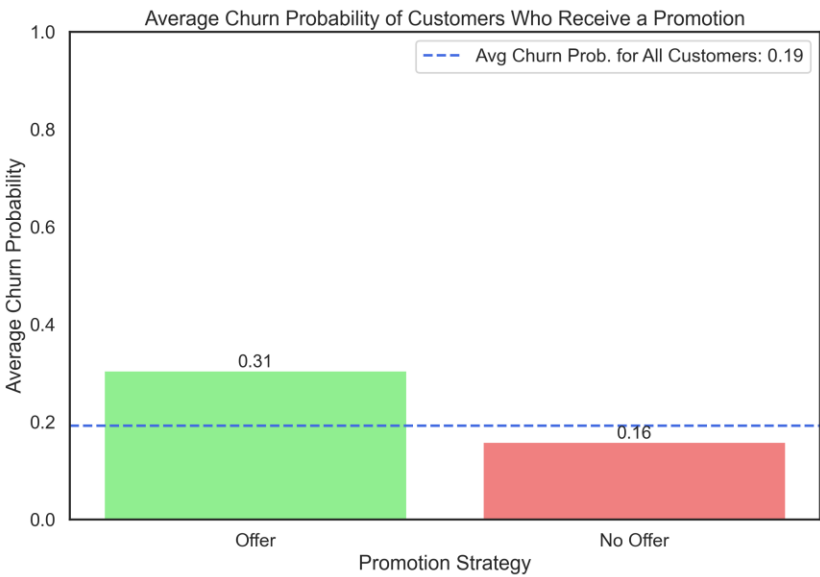
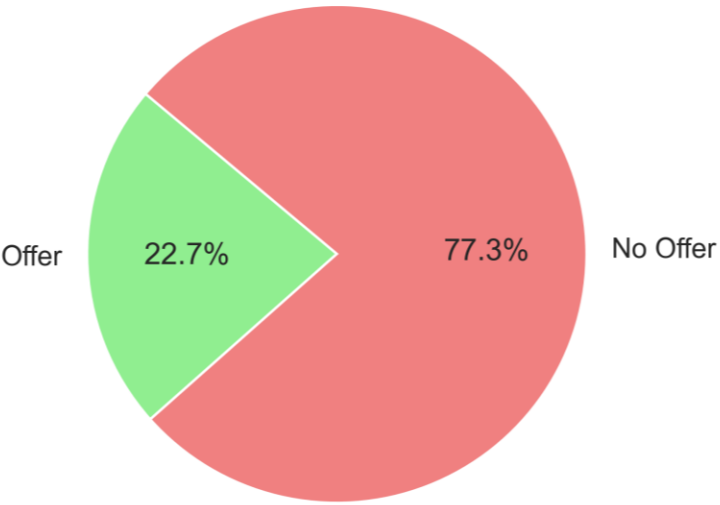
Ensure that the total cost of promotions does not exceed the maximum budget.

$$\sum (B \cdot x_i) \leq B_{\max}$$

Basic model typically ‘cuts-off’ promotion if churning probability is below 0.2



Proportion of Customers Offered a Promotion



Implemented Optimization Model

Multiple Tiers Promotion Model

Parameters & Variables

With added **trade-off parameters** w and **promotion tiers** $\{\$100, \$200\}$, this model aims to minimize the churn rate along with promotion budget spent.

Model Objective

The objective function has two parts:

1. the first part penalizes the churn probability of customers not receiving promotions, weighted by w
2. while the second part represents the cost of promotions relative to the budget, weighted by $1 - w$

Constraints

Constraints ensure that the total promotion cost does not exceed the budget and that each customer receives **at most one promotion**.

Model implemented in Google Colab with Gurobi Academic License

Parameters

- B : Total promotion budget.
- p_i : Probability of customer i churning (from predictive analysis).
- c : Cost of promotion $\in \{100, 200\}$.
- w : Trade-off parameter between reducing churn and optimizing the budget.

Decision Variables

- $x_{i,100} = 1$: customer i receives a promotion 100.
- $x_{i,200} = 1$: customer i receives a promotion 200.

Objective Function

Minimizing Churn Probability for Unpromoted Customers & Minimize Total Budget Spent for Customers Receiving Promotions.

$$\min w \sum_i p_i (1 - x_{i,100} - x_{i,200}) + (1 - w) \frac{\sum_i (x_{i,100} \cdot c_{100} + x_{i,200} \cdot c_{200})}{B} \quad (1)$$

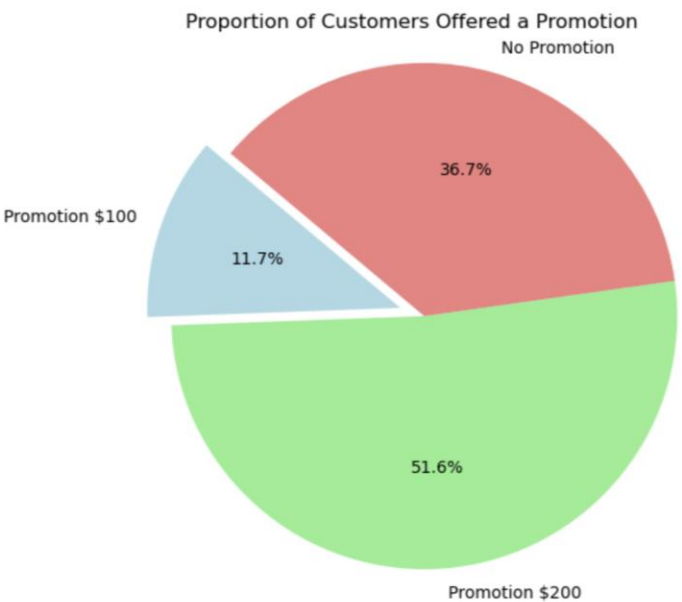
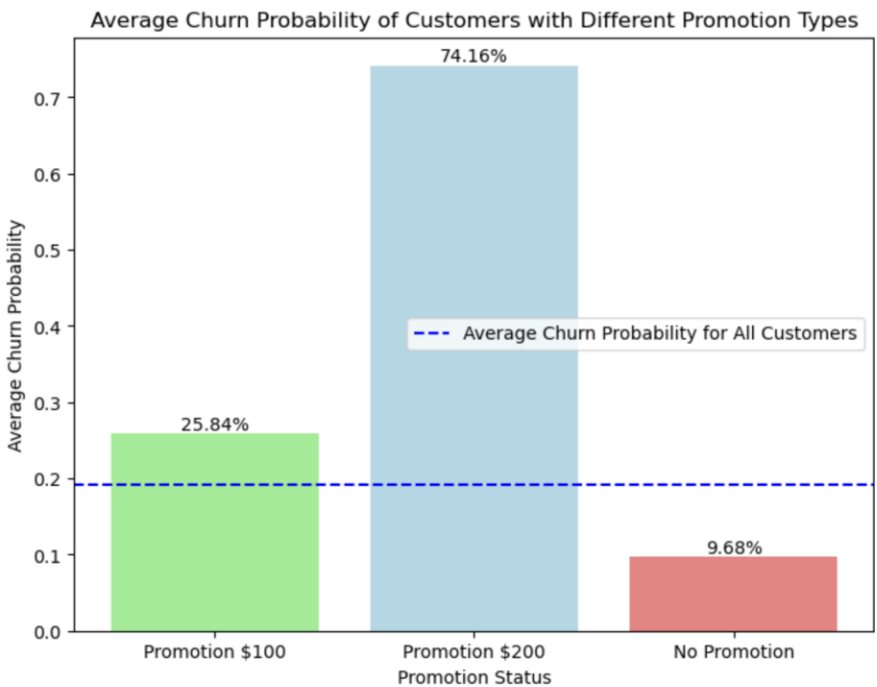
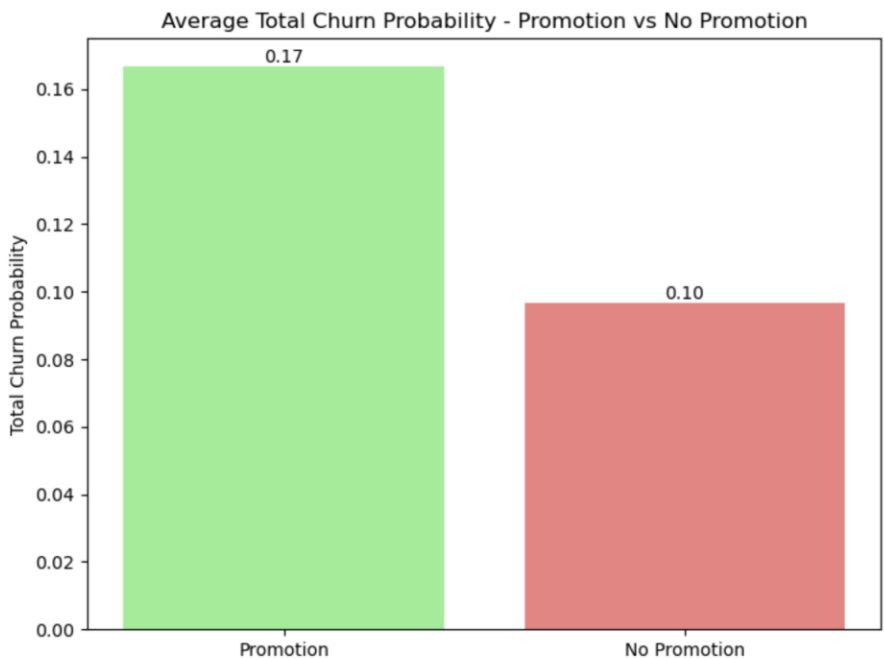
Constraints

$$\sum_i (x_{i,100} \cdot c_{100} + x_{i,200} \cdot c_{200}) \leq B \quad (\text{Total Budget}) \quad (2)$$

$$x_{i,100} + x_{i,200} \leq 1 \quad (\text{No more than 1 promotion per customer}) \quad (3)$$

$$0 \leq w \leq 1, \quad x_{i,100} \in \{0, 1\}, \quad x_{i,200} \in \{0, 1\} \quad (4)$$

Implemented model is set with a lower threshold at 0.2 for promotion \$100 and upper threshold at 0.4 for promotion \$200

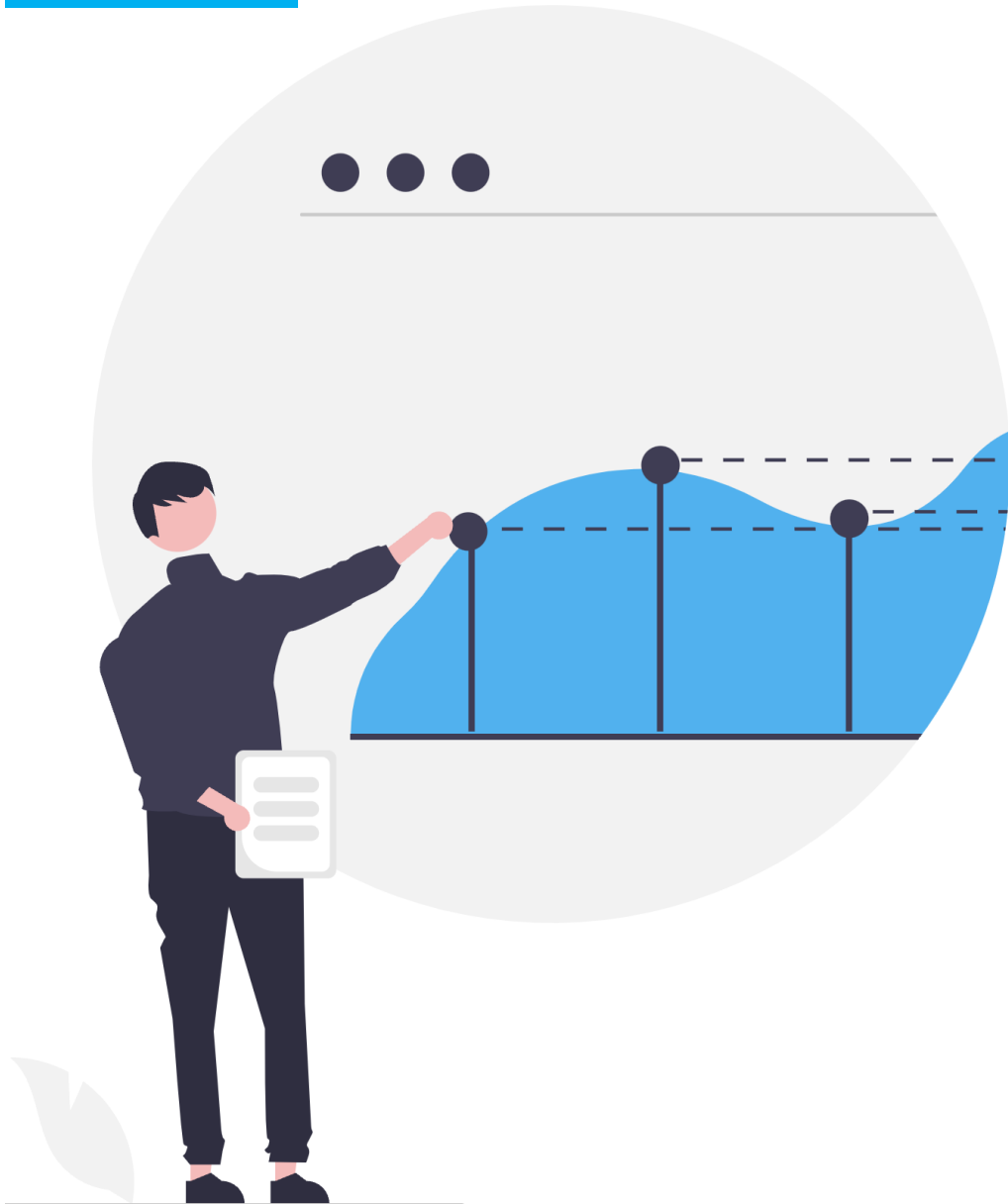


Number of customers with churn rate Less than 0.2: 5594
Number of customers with churn rate 0.2 – 0.4: 3010
Number of customers with churn rate 0.4 – 0.6: 133
Number of customers with churn rate Greater than 0.6: 83

Max Budget = 1 million, Promotion/Customer = {\$100, \$200}

Next Steps

- 1) Conduct sensitivity analysis on our models
- 2) Iterate on our prescriptive analysis models





References

Kishore, P. K. (n.d.). *Bank Customer Churn Data*. Wwww.kaggle.com. Retrieved February 29, 2024, from https://www.kaggle.com/datasets/pentakrishnakishore/bank-customer-churn-data/data?select=churn_prediction.csv

Marous J. *Recognizing “Silent Attrition” Is Key to Maintaining Loyalty in Banking*. (2023, April 23). Recognizing “Silent Attrition” Is Key to Maintaining Loyalty in Banking; The Financial Brand. <https://thefinancialbrand.com/news/customer-experience-banking/silent-attrition-key-to-customer-loyalty-in-banking-161500/>
