Analytics Practicum II: Propensity to Lapse Model Building Exercise

Lefteris Souflas (Student ID: f2822217)

Monday, 31 July 2023

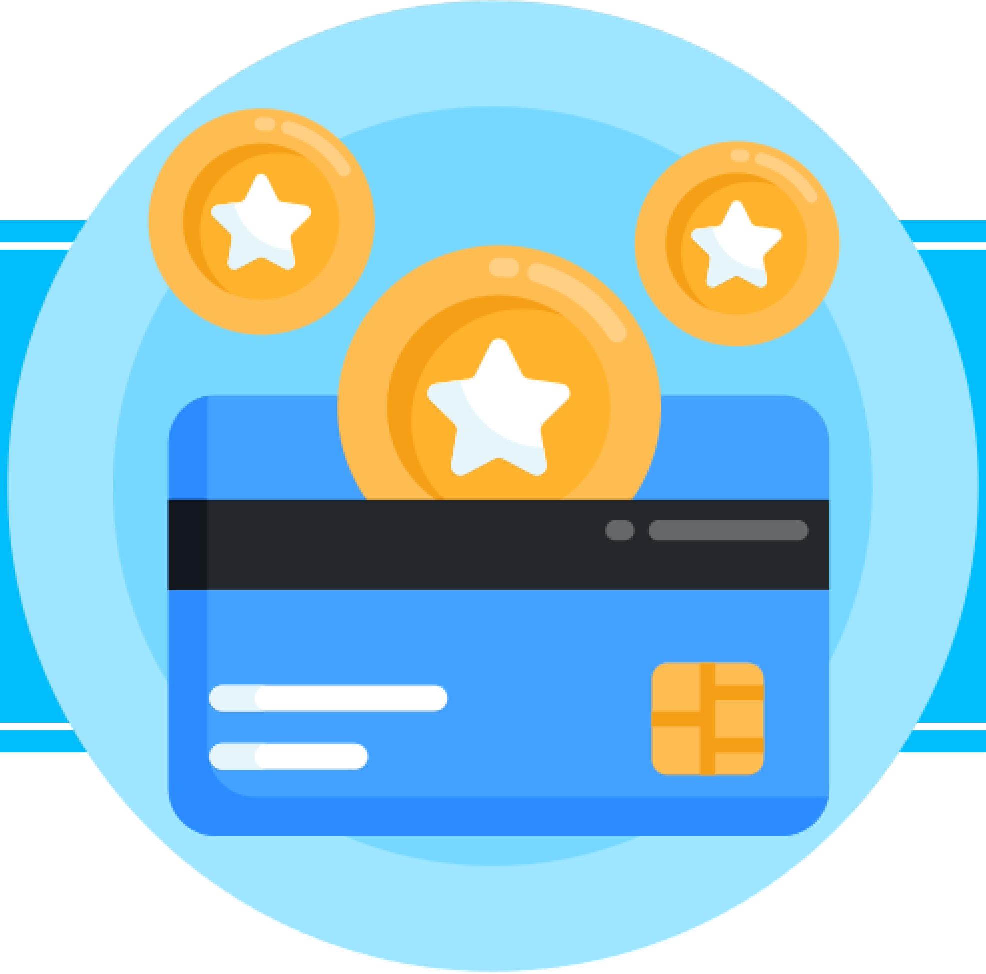
Contents

Introduction

Background

❖ Loyalty business partner with a reward system for customers to collect and redeem points.

❖ Customer lapsing status determined by 12 consecutive months without collections or redemptions.



Problem

Need to predict which customers are likely to become inactive (lapse) in the future.

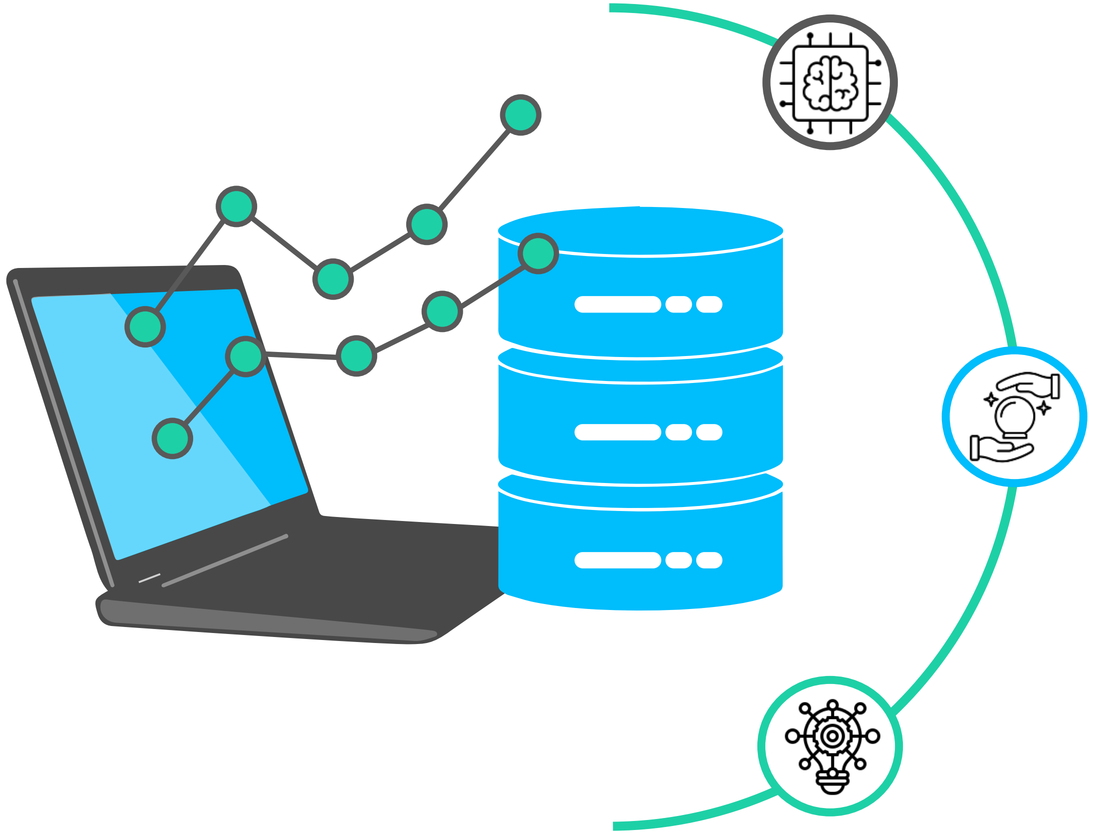
The business partner faces challenges with customer churn (lapse).

Importance

Customer retention is crucial for the success of the reward system.

Identifying potential lapsers allows for targeted retention strategies.

Goal



Machine Learning

Develop a Propensity to Lapse Model using Machine Learning.

Predictions

Predict customers likely to lapse in the near future based on their transaction history.

Insights

Provide insights to the business partner for effective churn prevention.

Data

Dataset Overview

<u>Dataset Size</u>: 5000 Observations,
each consisting the <u>following features</u>:

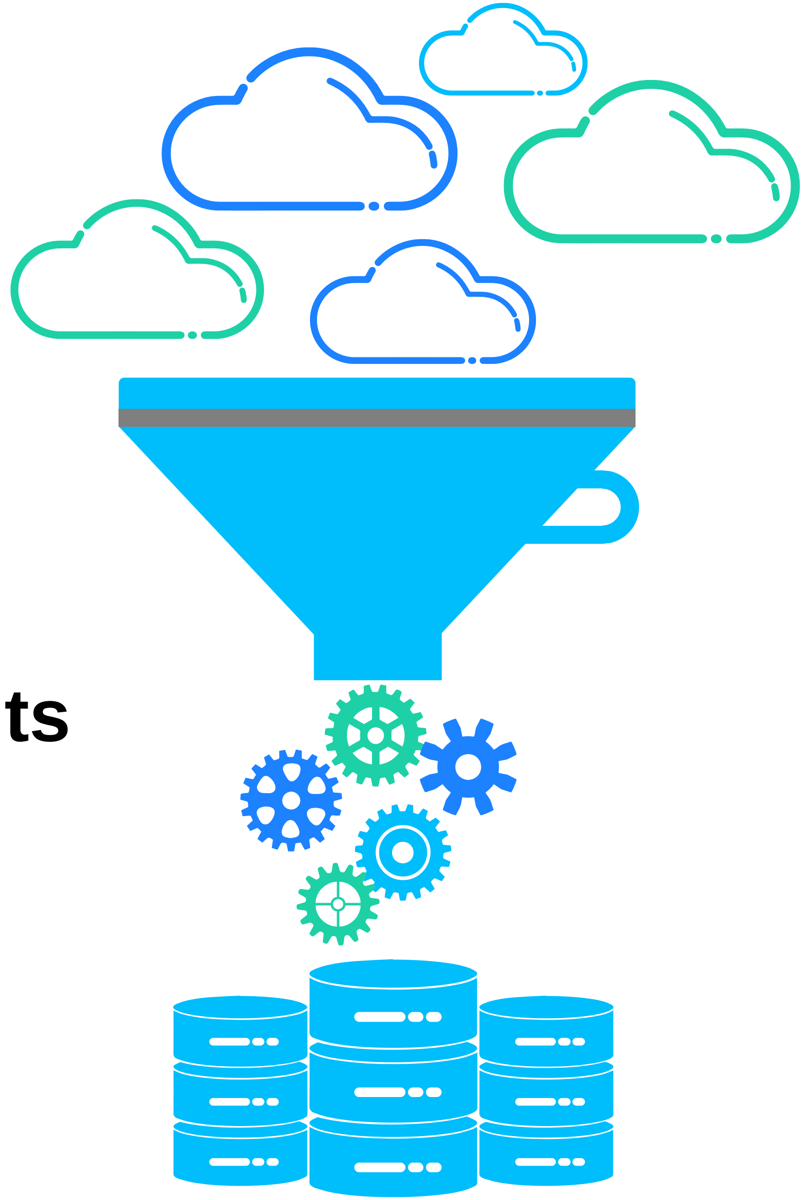
State (trying to predict given the rest)
Lapsed status: Active=0, Lapsed=1.

Sum collect, Sum redeem
How many times a customer collected or redeemed.

Sum collect points, Sum redeem points
Total points collected or redeemed.

Years in the program
Years since customer's registration to the program.

Months since last transaction
Months passed since the customer's last action.



Data Preparation

Feature Engineering

Creation of relevant features, such as the frequency of collections and redemptions, time since last collection/redemption, total points collected, total points redeemed, etc. (already implemented in the given dataset).

Data Preprocessing

Raw transaction data cleaning, handling and imputing missing values, data transformations, etc. (already implemented in the given dataset).

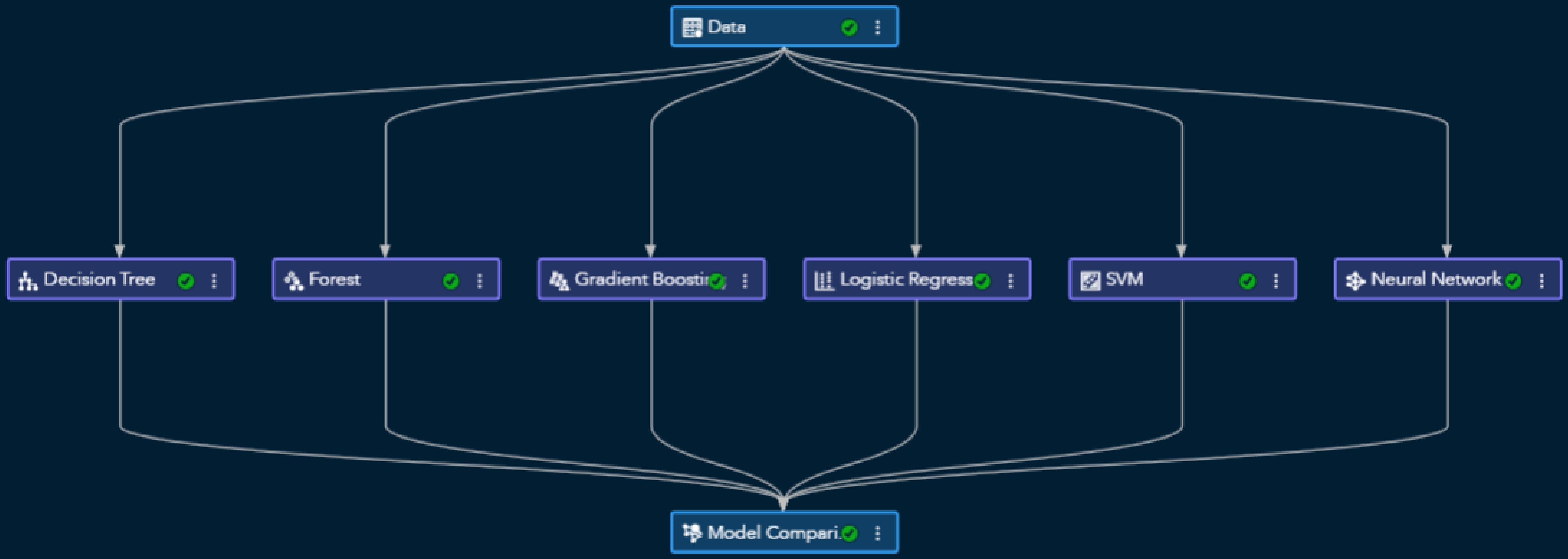
Data Split

Division of the dataset into stratified training (70%) and testing (30%) sets:

- ➢ The training set will be used for training the machine learning models, while the testing (unseen) set will be used for evaluating their performance.
- ➢ The sets were stratified to preserve the percentage of the two sub-groups included in the original dataset (Active: 50.1%, Lapsed: 40.9%) to the resulting datasets.

Model

Model Selection



SAS VDMML

Model Studio in SAS Visual Data Mining and Machine Learning (VDMML) was implemented for selecting appropriate Machine Learning algorithms for binary classification (Active vs. Lapsed) and performing Model Building and Evaluation.

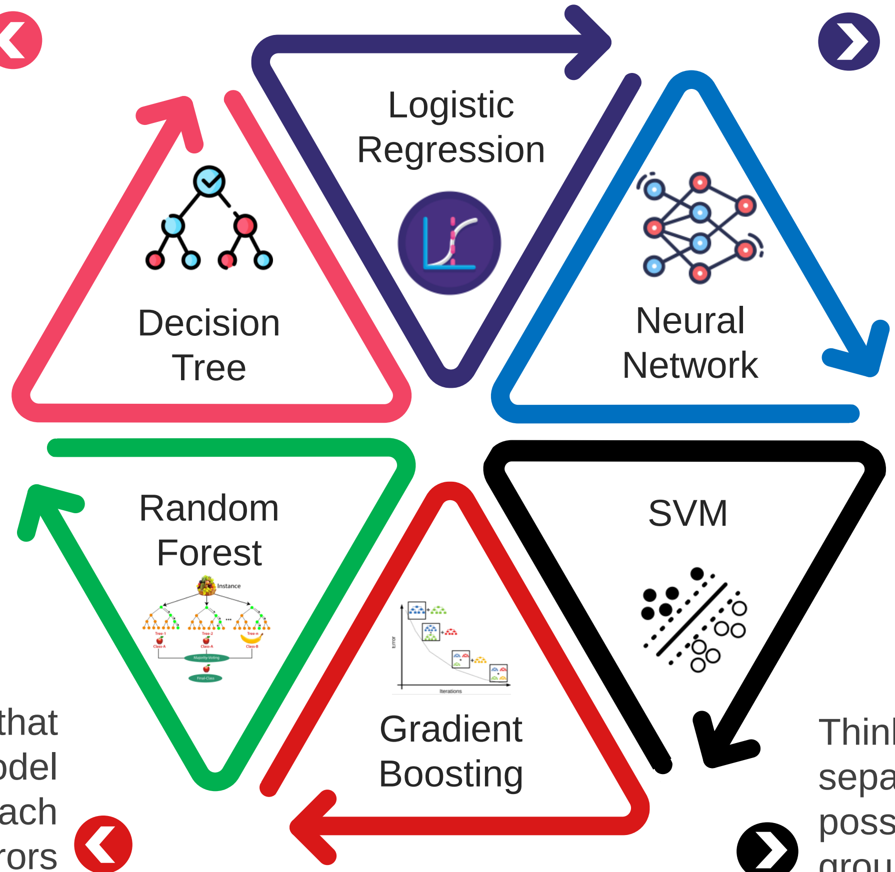
Model Selection

(Explanation in Non-Technical Terms)

Imagine Decision Tree as a flowchart that helps make decisions. Each branch represents a choice based on specific criteria, and each leaf (end of a branch) is a final decision or outcome.

Think of Random Forest as a team of Decision Trees working together. Instead of relying on a single Decision Tree, multiple trees are built, each with a slightly different perspective. The final decision is a combination of the predictions from all the trees, resulting in a more robust and accurate model.

Gradient boosting is like a learning process that improves over time. It starts with a simple model and focuses on correcting its mistakes. In each step, it builds a new model to capture the errors of the previous one. This process continues until the model becomes highly accurate.



Logistic Regression is like a detective trying to solve a case by analyzing clues. It helps predict a binary outcome, like whether a customer will lapse or not. The algorithm examines various factors (clues) and calculates the likelihood of an event occurring per case.

Neural Network is inspired by the human brain's structure and function. It consists of interconnected nodes that process information and learn from data. Just like we learn from experiences, neural networks learn patterns in data to make predictions.

Think of Support Vector Machine (SVM) as a line that separates two groups of customers. It finds the best possible line to maximize the distance between these groups. This line helps classify customers into different categories, such as active vs. lapsed customers.

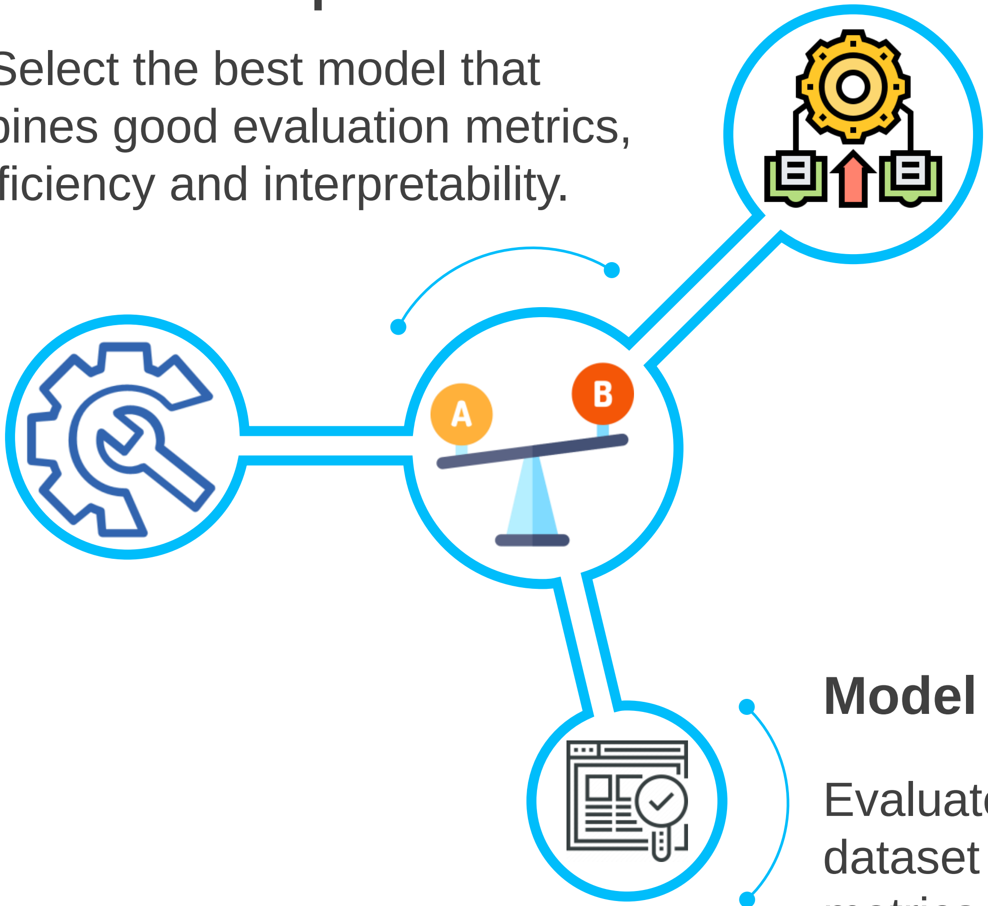
Model Building

Model Comparison

Select the best model that combines good evaluation metrics, efficiency and interpretability.

Model Training

Fit the selected models on the training dataset.



Model Optimization

Tune the models' hyperparameters for optimal performance.

Model Evaluation

Evaluate the fitted models on the testing dataset using relevant evaluation metrics.

Model Evaluation

Models / Metrics	Logistic Regression	Decision Tree	SVM	Neural Network	Random Forest	Gradient Boosting
Accuracy	0.8233	0.8580	0.8193	0.8233	0.8567	0.8600
Precision	0.8793	0.8284	0.8782	0.8678	0.8608	0.8598
Recall	0.7490	0.9025	0.7410	0.7623	0.8505	0.8598
F1-Score	0.8089	0.8639	0.8038	0.8117	0.8556	0.8598
AUC-ROC	0.9010	0.8885	0.8369	0.8719	0.9259	0.9259
Average	83.23%	86.83%	81.58%	82.74%	86.99%	87.31%

Accuracy

Accuracy measures how well the model correctly predicts both active and lapsed customers.

Precision

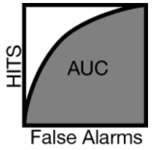
Precision evaluates how many of the customers predicted to be lapsed by the model actually lapsed.

Recall

Recall evaluates how many of the actually lapsed customers were predicted to lapse by the model.

F1-Score

F1-Score is the harmonic mean of precision and recall, providing a balanced evaluation of the model's performance.



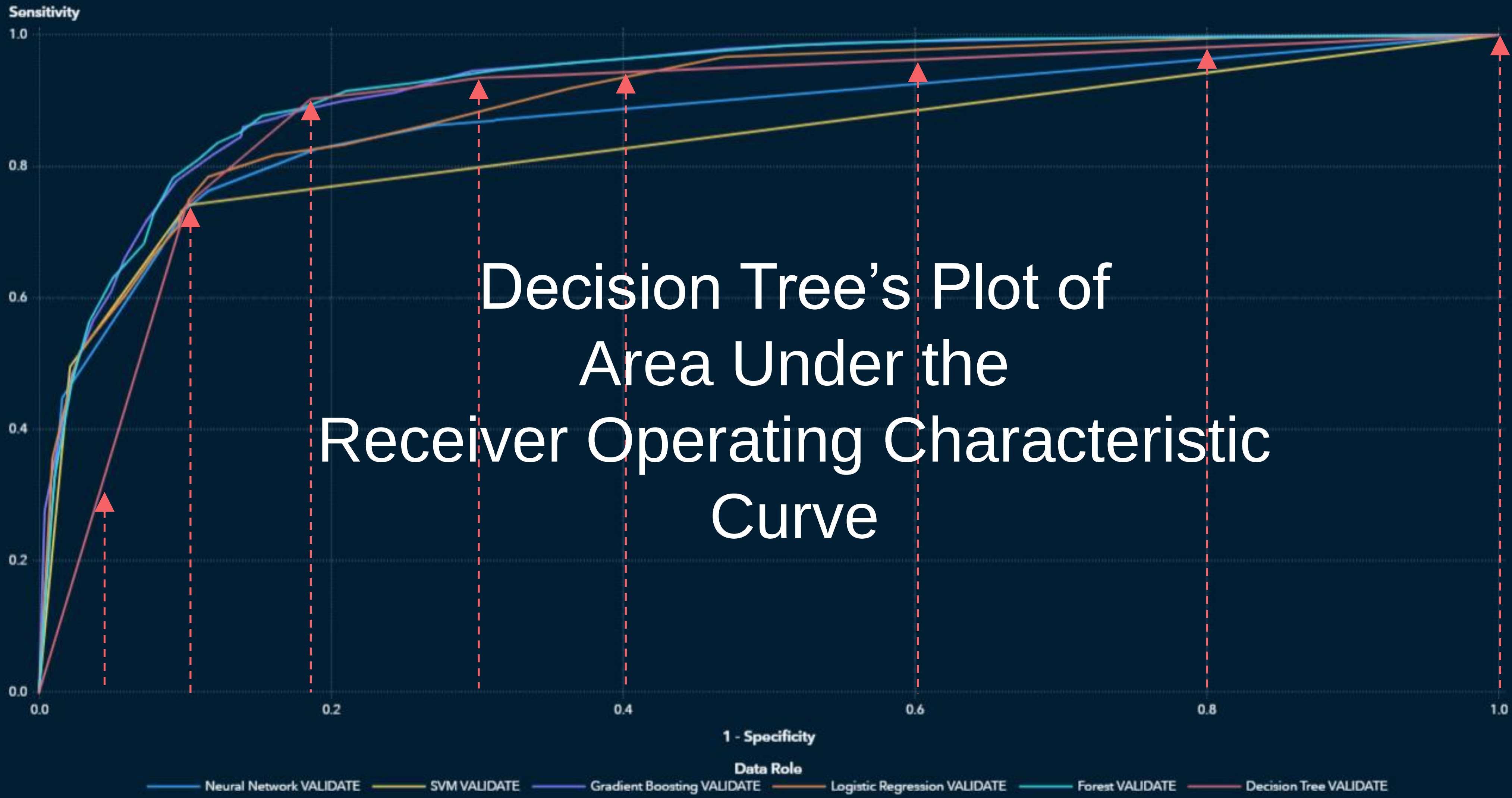
Area Under the ROC Curve

A higher AUC-ROC indicates a better-performing model with improved discriminatory power between active and lapsed customers.

Model Evaluation (cont.)

Area Under the ROC Curve

The ROC curve plots the true positive rate (recall or sensitivity) against the false positive rate (1-specificity) at various classification thresholds. The AUC-ROC score represents the area under this curve, ranging from 0 to 1. The greater the area, the better. A higher AUC-ROC score signifies that the model can differentiate between lapsed and active instances effectively, leading to fewer misclassifications and better overall performance.



Decision Tree’s Plot of Area Under the Receiver Operating Characteristic Curve

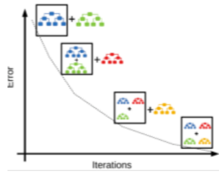
Model Evaluation (cont.)

Random Forest

#2

Random Forest is the second-best model of all the models implemented in terms of performance with less than 0.4% on average difference from the best.

Good Performance

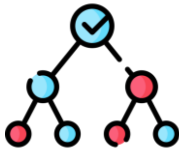


Gradient Boosting

#1

Gradient Boosting is the winner of all the models implemented in terms of performance, as it has on average the best evaluation metrics.

Best Performance

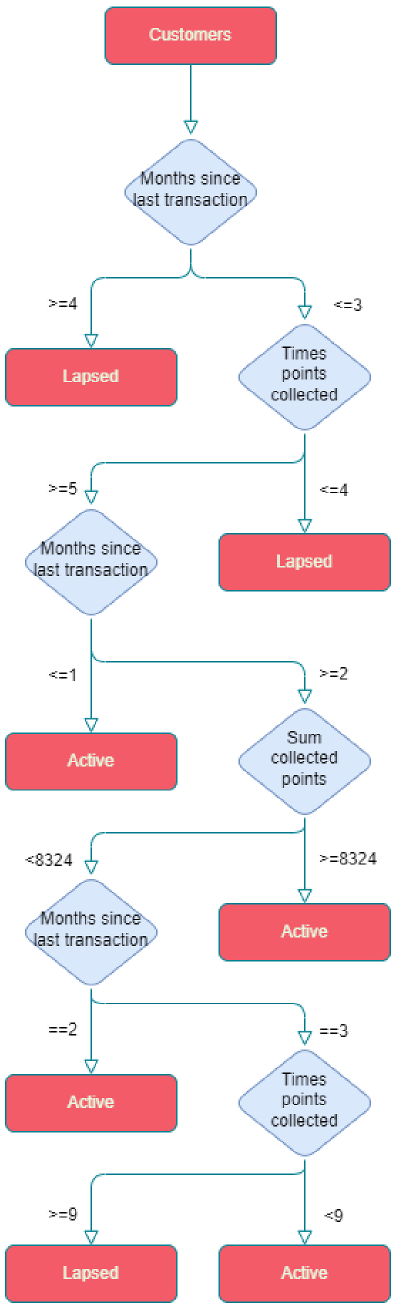


Decision Tree

#3

Decision Tree is the third best model in terms of performance with less than 0.5% on average difference from the best. We selected this model, as it has the best interpretability.

Best Interpretability



Model Interpretation

- Customers whose last action (points collection or redemption) was 4 or more months ago must be considered as potential lapsed customers.
- Customers that 3 or less months passed since their last action but have collected points only 4 or less times overall, must also be considered as potential lapsed customers.
- Customers that have collected points 5 or more times and their last action was 1 month ago or less must be considered as active customers.
- Customers that their last action was between 2 and 3 months ago and have collected points 5 or more times whose total collected points is 8324 or more must be considered as active customers.
- Customers that their last action was 2 months ago and have collected points 5 or more times whose total collected points is less than 8324 must be considered as active customers.
- Customers that their last action was 3 months ago whose total collected points is less than 8324 must be considered as:
  - Active, if they have collected points between 5 and 8 times.
  - Lapsed, if they have collected points 9 or more times.

Findings

Feature Importance

Months since last transaction

Months passed since customer's last action (collection or redemption) is, by far from the second, the most important aspect in classifying a customer as probable to lapse or not, with the threshold of 4 months since last transaction playing a vital role for a customer to churn.

Times points collected

How many times a customer collected points is the second most important factor in categorizing a customer as probable to lapse or not, with the threshold of 5 times points collected being a key turnover for a customer not to churn.

Sum collected points

How many points a customer has collected in total is the third most important factor in classifying a customer as lapsing or active, with the value of 8324 points being a cutoff point for a customer to decide not to churn.

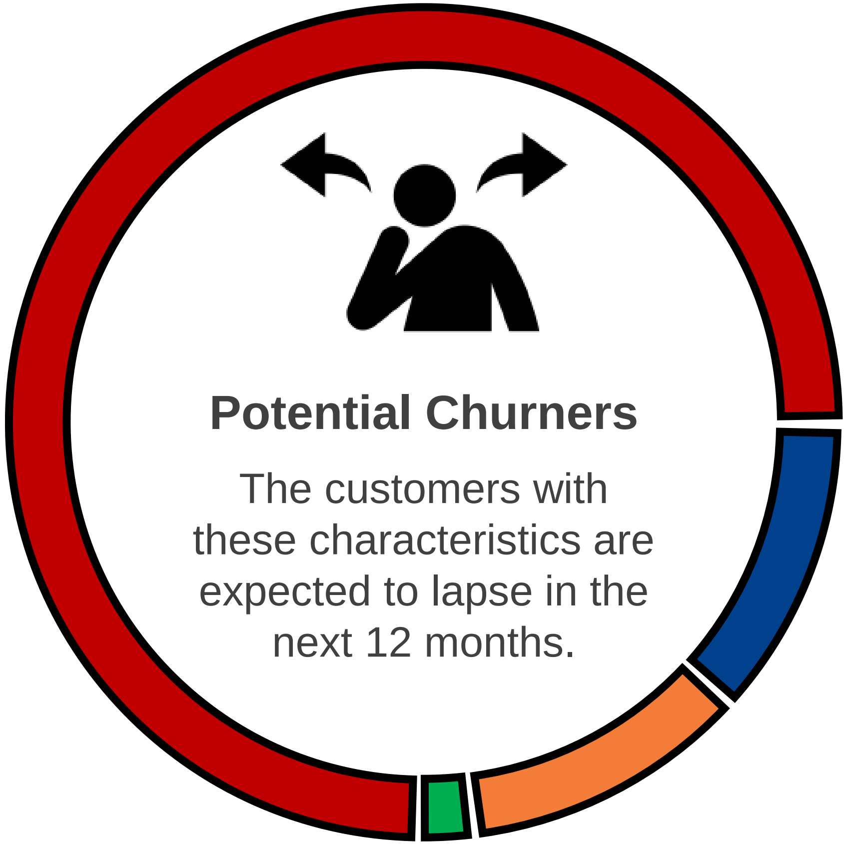
Churners Distribution

Four-Months Inactive

Customers whose last action (collection or redemption of points) was 4 or more months ago consist of approximately **73%** of the whole population of lapsed customers. This characteristic seems to be inherent of lapsed customers, as approximately only 10% of active customers seem to be dormant for more than 3 months.

Nine-Times Points Collectors

Customers that their last action was 3 months ago, their total collected points is less than 8324 and have collected points 9 or more times must be considered as lapsed customers. However, this portion of customers consist of only **2%** of the total lapsed customers.



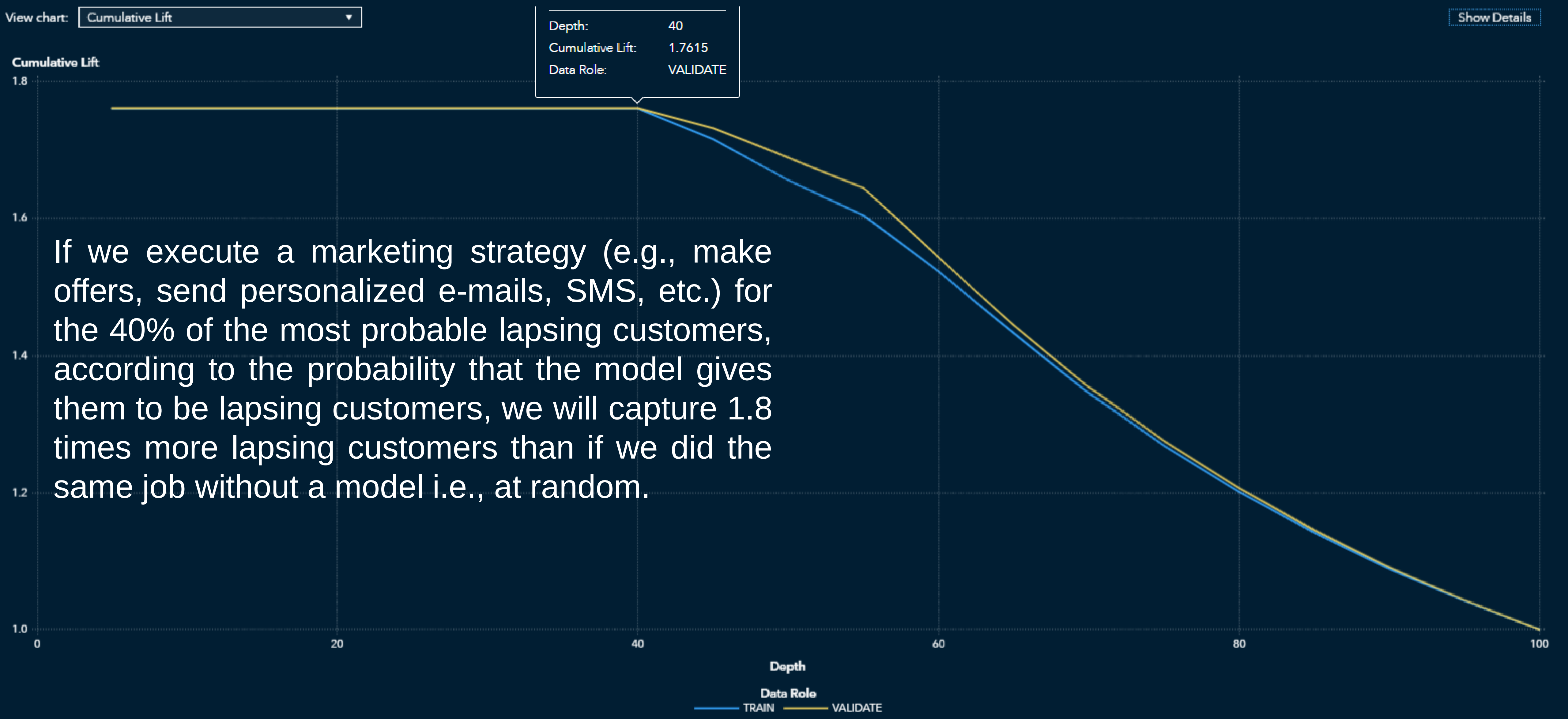
Four-Times Points Collectors

Customers that do not fall into the majority group of churners but have collected points only for up to 4 times, consist of approximately **13%** of the whole population of lapsed customers. However, this characteristic seems to be inherent not only of lapsed customers, as approximately more than 10% of active customers also present this feature, and therefore it should not be seriously taken into consideration.

Rest

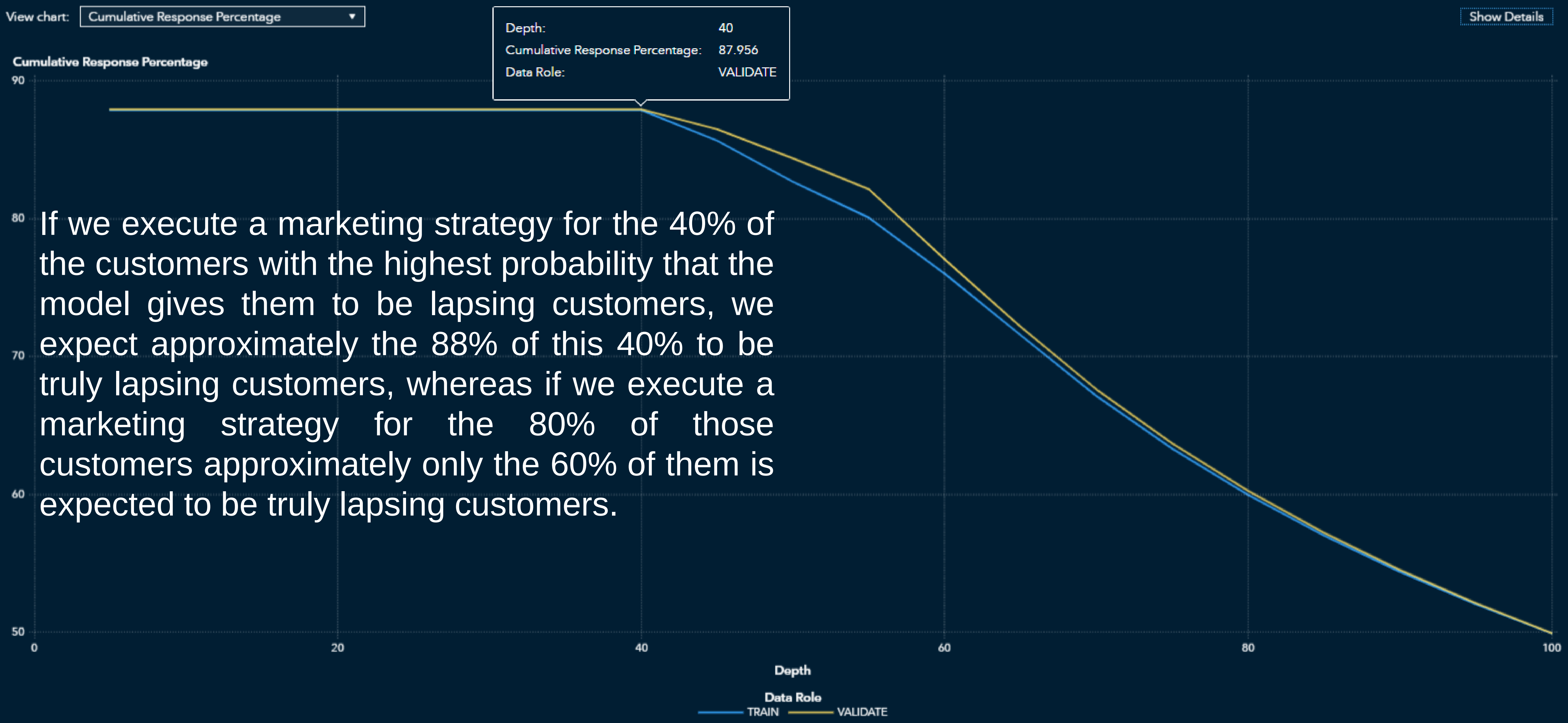
There exist approximately **12%** of the total Lapsed Customers that do not fall into any of the forementioned categories but share common features with the features of the groups of the Active Customers.

Model-based Marketing



If we execute a marketing strategy (e.g., make offers, send personalized e-mails, SMS, etc.) for the 40% of the most probable lapsing customers, according to the probability that the model gives them to be lapsing customers, we will capture 1.8 times more lapsing customers than if we did the same job without a model i.e., at random.

Marketing Efficacy



If we execute a marketing strategy for the 40% of the customers with the highest probability that the model gives them to be lapsing customers, we expect approximately the 88% of this 40% to be truly lapsing customers, whereas if we execute a marketing strategy for the 80% of those customers approximately only the 60% of them is expected to be truly lapsing customers.

Targeted Marketing

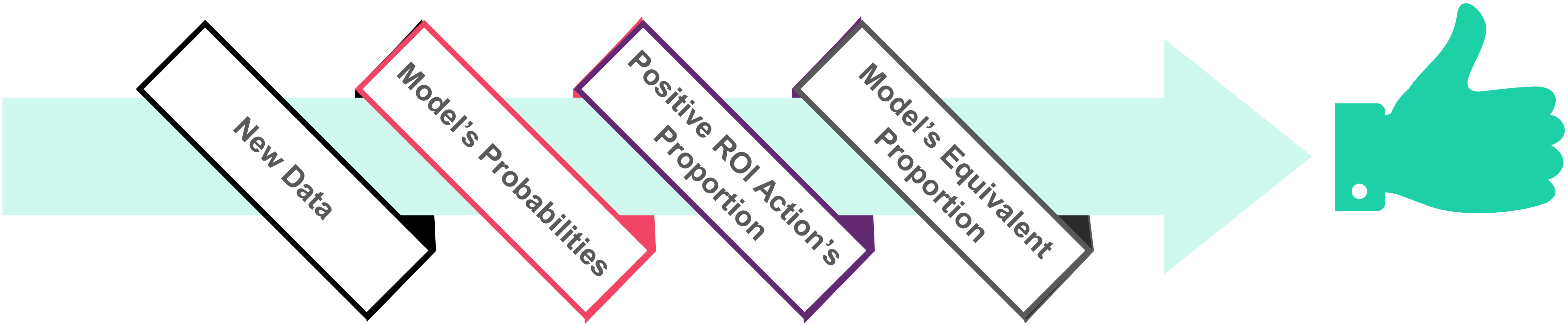
If we want to capture approximately the 90% of all future lapsing customers, we must execute a marketing strategy for the 55% of the most probable lapsing customers, according to the probability that the model gives them to be lapsing customers.

Suggestions

Data-Driven Actions

New Data
Pass through the model the current customer data with the features it was trained on.

Positive ROI Action’s Proportion
Calculate the proportion of potential lapsing customers that, if they respond positively to a specific offer/action, it will worth it (Return On Investment).



Model’s Probabilities
Sort customers in descending order according to the probabilities that the model gives them to be potential lapsing customers.

Model’s Equivalent Proportion
Find the model’s required proportion of most probable lapsing customers that is needed for the positive ROI action’s proportion to be captured.

Reduce Churn Rates

1	Identify customers whose last action (collection or redemption) was 4 or more months ago and proactively reach out to them with personalized offers and incentives to encourage them to re-engage with the loyalty program. These customers represent a significant portion of the lapsed customer population and can be effectively targeted for retention efforts.	**Proactive Outreach for Dormant Customers**
2	Implement targeted engagement strategies for customers who have collected points only for up to 4 times and are inactive for up to 3 months. While this group includes both lapsed and active customers, a focus on engaging them with relevant rewards and promotions may help prevent potential churn.	**Engagement Strategies for Low-Activity Customers**
3	Offer tier-based rewards and recognition for customers who have collected points 5 or more times and have been active recently (1 month ago or less). Enhance their experience by providing exclusive benefits, early access to rewards, or priority customer support to foster loyalty and retention.	**Tier-Based Rewards and Recognition**
4	For customers, whose last action was between 2 and 3 months ago and have collected points 5 or more times, consider offering additional benefits, discounts, or bonus points if their total collected points are at least 8324. This strategy can incentivize them to continue engaging with the loyalty program.	**Enhanced Loyalty Program Benefits**
5	Design targeted reactivation campaigns for customers whose last action was at least 12 months ago and thus are categorized as lapsers. Offer attractive rewards or promotions to entice them back into the loyalty program.	**Reactivation Campaigns**

Reduce Churn Rates (cont.)

6	Utilize customer segmentation techniques, such as Recency-Frequency-Monetary (RFM) Segmentation, to tailor communication campaigns. Send relevant and personalized messages to VIP customers based on their predicted likelihood of lapsing, ensuring that the right offers reach the right customers.	**Targeted Communication for Specific Segments**
7	Introduce gamification elements to make the loyalty program more interactive and engaging. Offer exclusive events, experiences, or early access to new rewards for loyal customers to create a sense of exclusivity and encourage ongoing participation.	**Gamification and Exclusivity**
8	Use real-time data to engage customers with timely offers and promotions based on their recent behavior and preferences. Promptly address any issues or concerns raised by customers to enhance satisfaction and loyalty.	**Real-Time Engagement and Feedback**
9	Regularly monitor the effectiveness of the implemented strategies and analyze customer feedback. Use insights from ongoing data analysis to fine-tune and optimize the loyalty program for better customer retention and engagement.	**Continuous Monitoring & Optimization**

QUESTIONS???

THANK YOU