Logo

Description automatically generated

Assignment

**Advanced Business Analytics and Visualization**

Student Name : Yap Jia Xin

Student ID : TP055804

Module Code & Title : CT045-3-M-ABAV

Intake Code : APUMP2310DSBA (DE) - ODL

Name of Lecturer : Mr. Raheem Mafas

Table of Contents

[Introduction 3](#_Toc167656229)

[Business Case / Domain Knowledge 5](#_Toc167656230)

[Problem Statement / Business Goals 5](#_Toc167656231)

[Aim & Objectives 6](#_Toc167656232)

[Scope 7](#_Toc167656233)

[Methodology Selection 8](#_Toc167656234)

[Data Modelling Techniques 9](#_Toc167656235)

[Dataset Preparation & EDA 9](#_Toc167656236)

[Model Construction, Optimization and Validation 21](#_Toc167656237)

[Critical Interpretation of Outcomes (model output) in achieving the defined Business Goals 32](#_Toc167656238)

[References 35](#_Toc167656239)

List of Figures

[Figure 1: File Import – Edit Variables 9](#_Toc167656311)

[Figure 2: Output of File Import 10](#_Toc167656312)

[Figure 3: Output of StatExplore 1 11](#_Toc167656313)

[Figure 4: Output of StatExplore 2 12](#_Toc167656314)

[Figure 5: Output of StatExplore 3 12](#_Toc167656315)

[Figure 6: Output of StatExplore 4 12](#_Toc167656316)

[Figure 7: Output of StatExplore – Variable Worth 13](#_Toc167656317)

[Figure 8: Output of StatExplore - Chi-Square 15](#_Toc167656318)

[Figure 9: Drop - Edit Variables 16](#_Toc167656319)

[Figure 10: Sample Node - Edit Properties 17](#_Toc167656320)

[Figure 11: Output of Sample Node 18](#_Toc167656321)

[Figure 12: Data Partition Node – Edit Properties 19](#_Toc167656322)

[Figure 13: Output of Data Partition Node 20](#_Toc167656323)

[Figure 14: Output of Model Comparison (Decision Tree) 1 21](#_Toc167656324)

[Figure 15: Output of Model Comparison (Decision Tree) 2 22](#_Toc167656325)

[Figure 16: Output of Decision Tree Comparison - Misclassification Rate 22](#_Toc167656326)

[Figure 17: Output of Decision Tree Comparison - ROC Chart 23](#_Toc167656327)

[Figure 18: Output of Decision Tree Comparison - Cumulative Lift 23](#_Toc167656328)

[Figure 19: Output of Decision Tree Comparison - Gain 24](#_Toc167656329)

[Figure 20: Output of Model Comparison (Logistic Regression) 1 25](#_Toc167656330)

[Figure 21: Output of Model Comparison (Logistic Regression) 2 25](#_Toc167656331)

[Figure 22: Output of Model Comparison (Gradient Boosting) 1 27](#_Toc167656332)

[Figure 23: Output of Model Comparison (Gradient Boosting) 2 27](#_Toc167656333)

[SAS Enterprise Miner Node 1: File Import 9](#_Toc167656344)

[SAS Enterprise Miner Node 2: StatExplore 11](#_Toc167656345)

[SAS Enterprise Miner Node 3: Drop 16](#_Toc167656346)

[SAS Enterprise Miner Node 4: Sample 17](#_Toc167656347)

[SAS Enterprise Miner Node 5: Data Partition Node 19](#_Toc167656348)

[SAS Enterprise Miner Node 6: Decision Tree 21](#_Toc167656349)

[SAS Enterprise Miner Node 7: Logistic Regression 25](#_Toc167656350)

[SAS Enterprise Miner Node 8: Gradient Boosting 27](#_Toc167656351)

# Introduction

Customer churn is a term for the phenomenon where customers of any business decide to discontinue their relationship with the business or organization. In the banking industry specifically, the phenomenon of customer churn occurs when individuals or customers close their bank accounts, switch to other financial institution, or when they stop using or purchasing certain bank services. Analyzing the customer churn is crucial for the decision makers in financial institutions because it is effective for understanding the reason of why customers leaving or discontinue the relationship with them. Other than that, it allows them to take proactive measures and make informed decisions to retain valuable customers based on the results of the analysis. Also known as customer attrition, customer churn poses critical challenges for banks, it does not only impact the profitability of the business, it also lead to drop of market share (Saxena et al., 2023).

# Business Case / Domain Knowledge

## Problem Statement / Business Goals

The future bank will be evolving from the current banking model that we are exposed to nowadays under the influence of competition, complementarity, and co-evolution. The implementation of financial technology is reshaping banking model, it enables third-party data processing and online financial services such as robo-advirosy, peer-to-peer lending, and crowdfunding. Commercial pressures have led banks to cut cost by closing their branches and increasing automation in the business processes. This can be seen by looking at how online transactions becoming a new norm for the public. The nature of money is changing with the rise of digital wallets, some countries such as Sweden are replacing the very nature concept of money with digital money. The banks must adapt these changes to maintain their positions among other competitors in the same industry(Broby, 2021).

The organization in financial industry can utilize the power of data analytics to solve business problems and further achieve business goals. With customer retention and churn reduction as the major challenges in business, implementation of predictive analytics is capable of assisting the decision makers to identify at-risk customers for targeted customer retention strategies. Leveraging analytics helps organizations to improve decision-making process, operational efficiency, customer satisfaction, and ultimately driving success to the business (Delgosha, 2020).

## Aim & Objectives

The aim of the assignment is to build a data analytical solution in order to help decision makers within banking industry to identify the factors contributing to customer churn.

The objectives are:

1. To build an interactive dashboard using tableau which can reveal the action insights.
2. To identify and analyse key demographic and financial factors influencing customer churn in a bank.
3. To create a comprehensive report summarizing the findings and recommendations for reducing customer churn in bank industry.

## Scope

The dataset of this assignment was sourced from Kaggle, accessible [here](https://www.kaggle.com/datasets/gauravtopre/bank-customer-churn-dataset/data). The dataset contains 12 attributes across 10,000 instances, capturing detailed customer profile of ABC Multistate bank. The notable features in the dataset include demographic data such as gender, age, and geography; account details of customers such as credit score, balance, and number of products. The target variable, churn indicates whether the customer has churned (value is 1), or stays (value is 0). Additional relevant attributes such as tenure, credit card status, and estimated salary are involved as well. The metadata summary is provided in below table:

Table 1: Variables Summary

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **Description** | **Data Type** | **Length** | **Sample Data** |
| **customer\_id** | Account number | numeric | 8 | 15634602, 15647311 |
| **credit\_score** | Credit score | numeric | 8 | 300, 500, … |
| **country** | Country of residence | Character | 7 | France, Spain, Germany |
| Gender | sex | character | 6 | Male, female |
| Age | age | numeric | 8 | 32, 33, 36, … |
| Tenure | Tenure of customers in the bank | numeric | 8 | 1, 2, 3, … |
| Balance | Account balance | numeric | 12 | 5600, 9983, 10444, … |
| Product\_number | Number of product from bank | numeric | 8 | 1, 2, 3, … |
| Credit\_card | Credit card flag | numeric | 8 | 1; 0 |
| Active\_member | Active member flag | numeric | 8 | 1; 0 |
| Estimated\_salary | Estimated salary | numeric | 8 | 2000, 3000, 5000, … |
| churn | Churn flag | numeric | 8 | 1; 0 |

## Methodology Selection

In this project, the methodology is a combination of exploratory analysis (EDA) and machine learning techniques. This methodology is chosen to be applied to understand the factor influencing customer churn in the banking sector and provide intelligence to reduce the churn rate. The dataset was analyzed using a visualization tool known as Tableau to identify patterns, trends, and relationships among the variables. This step is crucial in banking, as the understanding of customer behavior and demographic are keys to address churn customers. Subsequently, machine learning algorithms such as decision trees, logistic regression, and gradient boosting were applied using SAS Enterprise Miner. This approach leverages a robust predictive modeling capabilities to proactively identify potential churn customers, this enabling the construction of customer retention strategies. The incorporation of Tableau for data visualization and SAS Enterprise Miner for data modeling promotes comprehensiveness and effectiveness in analyzing data in the banking domain exclusively.

# Data Modelling Techniques

## Dataset Preparation & EDA

**File Import Node**

A close-up of a sign

Description automatically generated

SAS Enterprise Miner Node 1: File Import

[File Import] node was selected to import the dataset into SAS Enterprise Miner. After dropping the node, click on the “Import File” option from the properties panel to insert file path.

A screenshot of a report

Description automatically generated

Figure 1: File Import – Edit Variables

The role of the independent variable, churn was updated to "Target," and its level was adjusted to "Binary."

A screenshot of a document

Description automatically generated

Figure 2: Output of File Import

Figure above is the output of File Import Node. Dataset containing 12 variables were being imported to the SAS Enterprise Miner, there is 1 ID variable, 8 interval input variables, 2 nominal input variables, and 1 binary target variable indicating churn flag. The dataset comprises 10 thousand observations. The key attributes such as type, length, format, and informat of each variable were identified by running this node. This summary provides a clear understanding of the data structure, which makes it essential for the following analysis steps.

**StatExplore Node**

**A close-up of a sign

Description automatically generated**

SAS Enterprise Miner Node 2: StatExplore

A screenshot of a computer

Description automatically generated

Figure 3: Output of StatExplore 1

A screenshot of a computer

Description automatically generated

Figure 4: Output of StatExplore 2

A screenshot of a computer screen

Description automatically generated

Figure 5: Output of StatExplore 3

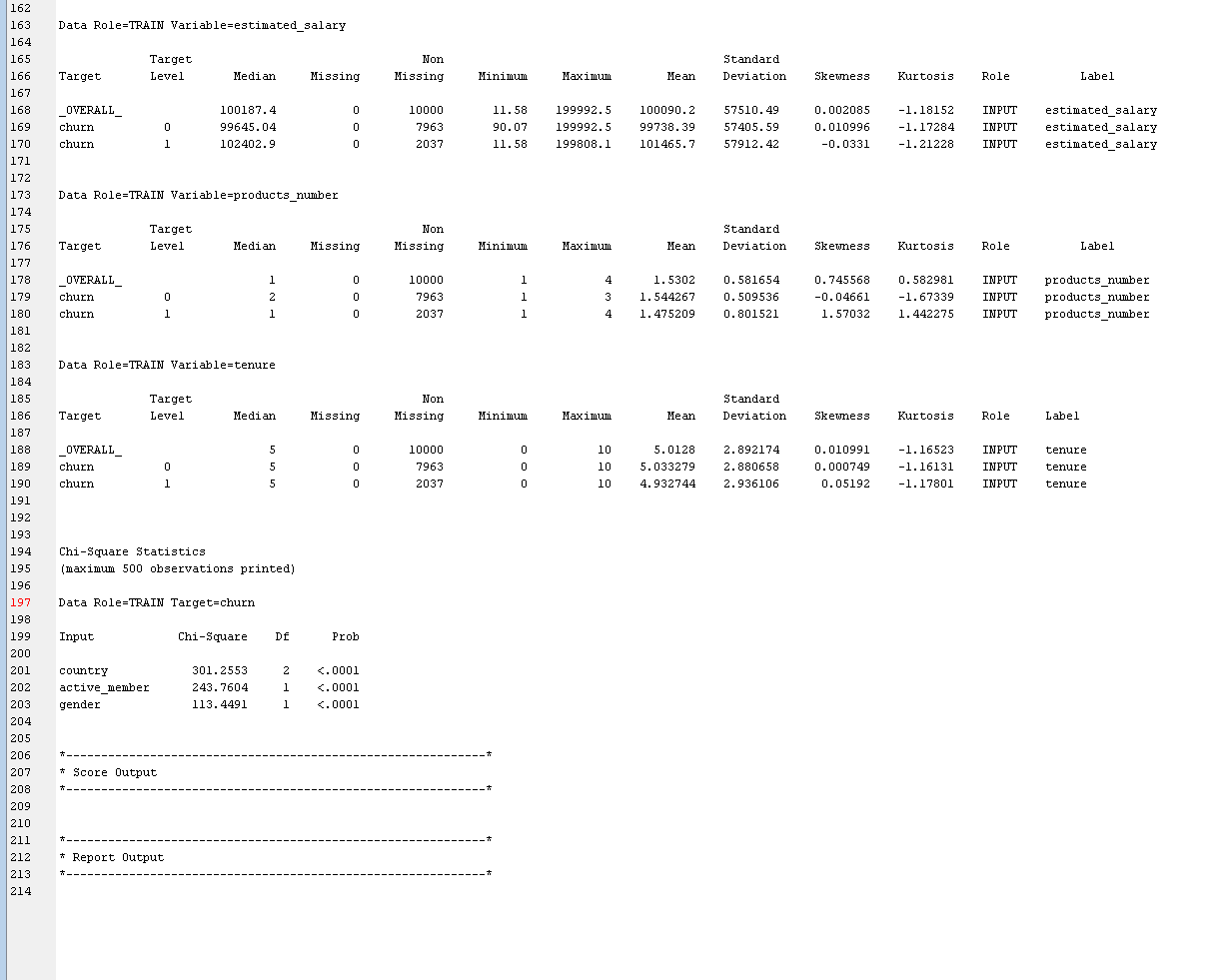


Figure 6: Output of StatExplore 4

A screenshot of a computer

Description automatically generated

Figure 7: Output of StatExplore – Variable Worth

The results from running the StatExplore node in SAS Enterprise Miner provide detailed insights into the dataset and its variables:

Variable Summary

The variables are categorized by their respective roles (such as ID, input, target) and measurement levels (interval, binary, nominal). There is one ID variable, **customer\_id**, it is categorized as an interval variable. Other than that, the input variables consist of 1 binary variable, which is **active\_member**; 7 interval variables, which are **age**, **balance**, **credit\_score**, **estimated\_salary**, **products\_number**, **tenure;** 2 nominal variables, which are **country** and **gender**. There is one binary variable, which is **churn**, serving as the target variable.

Summary Statistics

According to the Summary Statistics, there is 51.51% of active customers and 48.49% of inactive customers with the mode set to 1. In terms of country, majority of customers are from France, which contributes to 50.14% of the customer base, following by Germany (25.09%). The number of male customers (54.57%) is slightly more than female customers (45.43%). 20.37% of the customers are churn while 79.63% of them remain.

Interval Variable Summary Statistics

Based on the statistics, the mean **age** of the customers is 38.92. The mean **balance** is around 76k while the mean **estimated\_salary** is nearly 100k. Mean balance of 0.7055 for **credit\_card** indicating that majority of the customers have credit card. The mean value for **product\_number** is 1.53, which tells that majority of the customer has around 1 to 2 products purchased with the bank. The mean **tenure** is 5.01 years.

Class Variable Summary Statistics by Class Target

Majority of the loyal customers are being active, and majority of churn customers are inactive. The churned customers are equally distributed between Germany (39.96%) and France (39.76%). There are more female customers (55.92%) who churned as compared to male customers (44.08%).

Interval Variable Summary Statistics by Class Target

Churned customers are older by looking at the mean value, with the mean age of 44.84, as compared to non-churned customers, with the mean age 37.41. The churned customers have higher balance comparing to the loyal customers. Mean value of churn customers are lower when it comes to credit card and credit score. The estimated salary of churn and loyal customers is showing similar mean value. Churned customers have lesser number of products with (mean value = 1.48) compared to loyal customers (mean value = 1.54). For tenure, there is slightly lower mean value (4.94) among churn customers when comparing with loyal customers (mean value = 5.03).

A screenshot of a graph

Description automatically generated

Figure 8: Output of StatExplore - Chi-Square

Chi-Square Statistics

Based on the Chi-Square chart, the most significant variables associated with the target variable are country, active\_member, and gender. These 3 variables are noteworthy for the customer churn prediction.

**Active\_member** is showing a significant difference in distribution between churn and loyal customers. Other than that, it can be seen that **country** and **gender** show significant chi-square statistics, which makes them valuable for the analysis. Older customers are more likely to churn, making **age** an important variable as well. The account **balance** also suggests importance in the model as higher balance is associated with higher churn rates. The **credit\_card** only showing moderate difference, at the same time, **estimated\_salary** is similar between two groups of customers, both of these variables could be useful, but they are less critical than other variables. **Credit\_score** has some predictive worth as it is shown that the churn customers have slightly lower credit score than loyal customers. **Products\_number** is significant for the analysis as fewer products correlate with higher churn rate. Among churn customers, the tenure is slightly lower, which makes **tenure** contribute modestly to the prediction.

**Drop Node**

***A close-up of a button

Description automatically generated***

SAS Enterprise Miner Node 3: Drop

A screenshot of a computer

Description automatically generated

Figure 9: Drop - Edit Variables

During the phase of data exploration, the analysis of variables importance has revealed that 3 variables (credit\_card, estimated\_salary, and tenure) exhibited lower variable significance comparatively. Therefore, these 3 variables were bring eliminated using the [Drop] node in SAS Enterprise Miner. By eliminating unwanted or less significant variables, the runtime of the model can be reduced, thus enhancing the overall performance. With the model’s efficiency being optimized, the accuracy of the prediction will be potentially improved.

Class imbalance

**Sample Node**

A close-up of a blue label

Description automatically generated

SAS Enterprise Miner Node 4: Sample

A screenshot of a computer

Description automatically generated

Figure 10: Sample Node - Edit Properties

A screenshot of a computer

Description automatically generated

Figure 11: Output of Sample Node

There was 79.63% of churn customers and 20.37% of loyal customers in the original dataset. After sampling, the resulting dataset contains total of 4074 observations, the churn, and loyal customers each constituting 50% (2037 observations). This balanced sample is aimed to mitigate the class imbalance to enhance the performance of the model by ensuring that the predictive model able to accurately identify and distinguish between churn and loyal customers in the dataset.

Data Splitting

**Data Partition Node**

A close-up of a sign

Description automatically generated

SAS Enterprise Miner Node 5: Data Partition Node

A screenshot of a computer

Description automatically generated

Figure 12: Data Partition Node – Edit Properties

A screenshot of a document

Description automatically generated

Figure 13: Output of Data Partition Node

The [Data Partition] node was applied to split the dataset into training set (70%) and testing set. After splitting the data with ratio of 70:30, the churn distribution remains balanced across the dataset, with approximately 50% of customer for each churn and loyal customers group.

## Model Construction, Optimization and Validation

Decision Tree

Selection of decision tree

A diagram of a tree model

Description automatically generated

SAS Enterprise Miner Node 6: Decision Tree

Tree comparison and selection summary:

Based on the results provided, among the [Decision Tree] node, (Optimal Tree Ass - Entropy) performed the best due to following reasons:

A close-up of a number

Description automatically generated

Figure 14: Output of Model Comparison (Decision Tree) 1

A screenshot of a computer

Description automatically generated

Figure 15: Output of Model Comparison (Decision Tree) 2

A screenshot of a computer

Description automatically generated

Figure 16: Output of Decision Tree Comparison - Misclassification Rate

It has the lowest misclassification rate on the validation set, which falls on 0.2396 as shown as figure above. It indicated that this tree is better at the target variable prediction as compared to other decision trees.

A screenshot of a computer

Description automatically generated

Figure 17: Output of Decision Tree Comparison - ROC Chart

As shown as figure above, this Tree has a high ROC index of 0.83 on the validation set. The ROC index measures the model's ability to distinguish between classes, and a higher value indicates better performance of the model.

A screenshot of a computer

Description automatically generated

Figure 18: Output of Decision Tree Comparison - Cumulative Lift

A screenshot of a computer

Description automatically generated

Figure 19: Output of Decision Tree Comparison - Gain

This tree has a cumulative lift of 1.87 and a gain of 86.76 on the validation set. These metrics indicate that it is effective at identifying the positive class and provides a good lift over random guessing.

Below figures is the property setting of the selected Decision Tree model, the Nominal Target Criterion is set to “Entropy”, with “Assessment” as method, and “Misclassification” as Assessment Measure.

A close-up of a sign

Description automatically generated

A screenshot of a computer

Description automatically generated

Logistic Regression

Selection of logistic regression

A diagram of a comparison

Description automatically generated

SAS Enterprise Miner Node 7: Logistic Regression

The best logistic regression model can be determined by examining key metrics such as the misclassification rate, average squared error, ROC index, and the Kolmogorov-Smirnov statistic for both the training and validation datasets. The summary of performance metrics for each model is as follow:

A screenshot of a computer program

Description automatically generated

Figure 20: Output of Model Comparison (Logistic Regression) 1

A white paper with black text

Description automatically generated

Figure 21: Output of Model Comparison (Logistic Regression) 2

The misclassification rate of both Logistic Regression (None) and (Backward) is 0.30581, slightly better than Logistic Regression with (Forward) and (Stepwise) as model selection. The average squared error of Logistic Regression (None) and (Backward) is alai lower than the other two models.

Considering the slightly better performance in terms of misclassification rate and average squared error, the Logistic Regression model chosen is LR (None). Even though LR (Backward) has identical performance metrics, LR (None) is being selected as it is much more simpler and the variable selection steps can be avoided.

Below figures is the property setting of the selected Logistic Regression model, the selection model is set to “None”

A close-up of a blue rectangle

Description automatically generated

A screenshot of a computer

Description automatically generated

Gradient Boosting model selection

A diagram of a model comparison

Description automatically generated

SAS Enterprise Miner Node 8: Gradient Boosting

A white text with black text

Description automatically generated with medium confidence

Figure 22: Output of Model Comparison (Gradient Boosting) 1

A screenshot of a computer

Description automatically generated

Figure 23: Output of Model Comparison (Gradient Boosting) 2

Gradient Boosting model with leaf fraction of 0.001is selected after comparing with same models but with different leaf fraction, such as 0.005, 0.010, and 0.015. Gradient Boosting model with the least leaf fraction has the lowest misclassification rate in both training and validation sets, which represent its high performance. Similarly, it maintains the lowest average squared error in training and validation sets.

Thus, the preferred gradient boosting model is the one with leaf fraction of 0.001 as it outperforms others consistently in terms of its low misclassification rate and average squared error.

Figure below is the property setting of Gradient Boosting model, the Leaf Fraction is set to 0,001:

A close-up of a blue rectangle with black text

Description automatically generated

A screenshot of a computer

Description automatically generated

Fit-Statistics Comparison Table

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Model Description** | **Train Misclassification Rate** | **Train Average Squared Error** | **Valid Misclassification Rate** | **Valid Average Squared Error** |
| Boost2 | Gradient Boosting 0.001 | 0.22649 | 0.14778 | 0.21326 | 0.15864 |
| Tree5 | Optimal Tree Ass (Entropy) | 0.23957 | 0.15328 | 0.21045 | 0.17051 |
| Reg3 | LR None | 0.30581 | 0.19129 | 0.28411 | 0.20025 |

Event Classification Table for Validation Data

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Model Description** | **Data** | **Target** | **False Negative** | **True Negative** | **False Positive** | **True Positive** |
| Tree5 | Optimal Tree Ass (Entropy) | Validate | Churn | 141 | 470 | 152 | 460 |
| Reg3 | LR None | Validate | Churn | 180 | 431 | 194 | 418 |
| Boost2 | Gradient Boosting 0.001 | Validate | churn | 117 | 494 | 160 | 452 |

 **True Negative (TN)**: (Correctly predicted as Loyal)

 **False Negative (FN)**: (Incorrectly predicted as Loyal, actually Churn)

 **False Positive (FP)**: (Incorrectly predicted as Churn, actually Loyal)

 **True Positive (TP)**: (Correctly predicted as Churn)

|  |  |  |
| --- | --- | --- |
| Optimal Tree Ass (Entropy) | Loyal (Predicted) | Churn (Predicted) |
| Loyal (Actual) | TN (470) | FP (152) |
| Churn (Actual) | FN (141) | TP (460) |

|  |  |  |
| --- | --- | --- |
| Logistic Regression (None) | Loyal (Predicted) | Churn (Predicted) |
| Loyal (Actual) | TN (431) | FP (194) |
| Churn (Actual) | FN (180) | TP (418) |

|  |  |  |
| --- | --- | --- |
| Gradient Boosting (0.001) | Loyal (Predicted) | Churn (Predicted) |
| Loyal (Actual) | TN (494) | FP (160) |
| Churn (Actual) | FN (117) | TP (452) |

A black text on a white background

Description automatically generated

Above is the formula to calculate precision. Precision is a parameter to inform “how many selected data items are relevant”, it is used to represent the actual positive observations which the algorithm predicted to be positive(Ahmed et al., 2023).

A math equations with black text

Description automatically generated with medium confidence

Above is the formula to calculate recall. Recall is a representation of “amount of relevant data being selected”. In the other words, it represents the amount of actual positive observations which have been predicted by the selected algorithm(Ahmed et al., 2023).

A math equation with black text

Description automatically generated

Above is the formula to calculate F1-score, it takes both precision and recall into consideration while evaluating the algorithm’s performance(Ahmed et al., 2023).

A black lines with black text

Description automatically generated

Above is the formula to calculate accuracy, it is the most known parameter and is often the first choice for algorithm performance evaluation in most of the classification problems. It can be defined as ratio of accurately or correctly classified observations to the total amount of observations. This parameter is suggested for the situations where target variable classes are balanced in the dataset (Ahmed et al., 2023).

Summary of model evaluation

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Precision** | **Recall** | **F1-score** | **Accuracy** |
| Optimal Tree Ass (Entropy) | 0.7516 | 0.7654 | 0.7584 | 0.7609 |
| Logistic Regression (None) | 0.6830 | 0.6989 | 0.6908 | 0.6941 |
| Gradient Boosting (0.001) | 0.7386 | 0.7944 | 0.7656 | 0.7737 |

Three algorithms, decision tree, logistic regression, and gradient boosting are being compared in this bank customer churn analysis. The table above is summary of the model evaluation based on their performance metrics. Gradient boosting outperforms other models across several“key metrics as below:

**Recall** - With a recall of 0.7944, gradient boosting identifies 79.44% of actual churn cases effectively, it is crucial for the problem of customer churn prediction where missing any potential churn case can cause significant business implications to the bank.

**F1-Score** - Gradient boosting has the highest F1-Score (0.7656), it is reveals a sound balance between precision (0.7386) and the recall (0.7944). This signifies the capabilities of this model to identify churn occurrences while maintaining a high level of precision.

**Accuracy** - Gradient boosting has the highest accuracy score (0.7737) among other models, this accuracy value means that 77.37% of overall instances were correctly predicted, indicating the reliability of this model.

Though the precision of gradient boosting model is slightly lower than the decision tree’s, the significant gain in recall and F1-score still makes it a more robust and effective model for customer churn prediction. The higher recall ensures fewer churn cases to be missed, and the higher F1-score symbolized a well-balanced accomplishment between identifying true positives and avoiding the false positive cases.

In a nutshell, gradient boosting model offers the best combination of precision, recall, F1-score, and accuracy, which promote it to be the most suitable model for this customer churn analysis task. “

## Critical Interpretation of Outcomes (model output) in achieving the defined Business Goals

Variable importance of selected model

A screenshot of a computer

Description automatically generated

Evaluation of factors affecting churn based on model output based on variable importance.

A screenshot of a computer

Description automatically generated

**Age** - According to the variable importance table of model (Gradient boosting 0.001), age has the highest importance score of (1.000). This indicates that age is the most significant predictor of customer churn analysis in this model.

**Products Number** - The financial products that a customer purchases from a bank is highly predictive, with importance value of 0.784. According to the analysis, customers with more products indicate a lower likelihood of churn, as higher product amount leads to higher dependency or engagement to services of the bank. Therefore, the decision makers in the bank should develop marketing strategies such as bundles offers or cross-selling to encourage customers to use more various products and services. This suggestion is aimed to increase customers’ commitment to the organization and thus reducing risk of churning.

**Active Member** - Importance value of active member is 0.486, active members are predicted to be less likely to churn when comparing to customer with inactive status. The bank should come out with strategies to boost customer’s activity while using their products or services of the bank. These strategies could be loyalty program, regular check-ins, or offering usage incentives.

**Country** - Country contributes 0.382 of importance value, which might reflect the difference among cultural or regional perspectives in service satisfaction. Customer retention strategies or marketing strategies should be customized according to the regional preferences and needs. By localizing the service and support style for customers from each country, the bank will be able to improve customer satisfaction.

**Balance** - The account balance has an importance value of 0.358, it indicates financial engagement or capability of the customers. Hence, the bank should constantly monitor the account balance of their customers and offer financial advising services or flexible payment options like installment plan to customers with low balance in order to prevent churn occurrences.

**Gender** - Gender has an importance value of 0.227 to the bank customer churn analysis, it indicates potential differences in how different genders interact with the products or service of the bank. Therefore, the marketing team should consider gender-specific marketing campaign and support strategies to cater needs or requirements of both groups.

**Credit Score** - Credit score has the lowest importance value (0.122) among other variables, yet still impact the churn pattern of a customer. This is because credit score potentially reflects financial stability of a customer. Financial planning resources and support should be provided to the customers with lower credit score in order to avoid the churn occurrences.

Discussion and Conclusion

The analysis reveals that age is the most significant predictor of customer churn (importance value = 1.000), followed by number of products (0.784), customer activity status (0.486), country (0.382), account balance (0.358), gender (0.227), and lastly credit score (0.122). The variable importance value score provides a quantitative basis for comprehending the influential factors in predicting customer churn.

The most critical factor, age, suggests that there is variety in service preferences among respective group. Younger customers probably are more inclined towards digital engagements while customers from older age group may prefer personalized services from the bank. Customers with more product numbers are less likely to churn due to the engagement and dependency with the organizations. Customers who are actively using the products or services of the bank are more likely to stay with the bank. Country plays a significant role in this analysis because it reflects regional difference which assist to suggest service localization to improve customer retention. Account balance, gender, and credit score are less significant variables, however, they still provide insights into behavioral differences of customers, and financial stability, as well as engagement to respective products and services.

“With the objective to reduce occurrence of customer churn the decision makers in the bank should implement strategies tailored to the insights provided by the selected Gradient Boosting model. Age-specific engagement approaches is beneficial for addressing distinct needs of respective age groups to enhance the customer satisfaction level. Encouragement of using multiple products from the bank through bundles offers and cross-selling can boost customer commitment to the business and thus reduce the risk of churn. Maintaining high level of customer activity by launching loyalty programs is also an advantageous suggestion. Additionally, the organization should initiate support based on regional preferences to increase customer’ satisfaction. Financial health monitoring program for customers with low balance and credit scores can help to mitigate churn by addressing their financial stability. Last of all, gender-specific strategies can cater to individual preferences of male and female customers to further enhancing the effort of customer retention programs. The aforementioned targeted strategies are aligned with the business goal of reducing churn, and lead to larger customer base with high engagement and loyalty.”

# References

Vakili, M., Ghamsari, M., & Rezaei, M. (2020). Performance Analysis and Comparison of Machine and Deep Learning Algorithms for IoT Data Classification. *ArXiv:2001.09636 [Cs, Stat]*. <https://arxiv.org/abs/2001.09636>

Soltani Delgosha, M., Hajiheydari, N., & Fahimi, S. M. (2020). Elucidation of big data analytics in banking: a four-stage Delphi study. *Journal of Enterprise Information Management*, *ahead-of-print*(ahead-of-print). <https://doi.org/10.1108/jeim-03-2019-0097>

Broby, D. (2021). Financial Technology and the Future of Banking. *Financial Innovation*, *7*(1). <https://doi.org/10.1186/s40854-021-00264-y>

Saxena, A., Singh, A., & Govindaraj M. (2023). Analyzing customer churn in banking: A data mining framework. *Multidisciplinary Science Journal*, *5*, 2023ss0310–2023ss0310. <https://doi.org/10.31893/multiscience.2023ss0310>