# **Executive Summary**

This analysis study is regarding bank customer churn, utilizing advanced technologies such as Apache Spark, Python, Hive, and Hadoop, combined with integrated learning methods, to identify key factors leading to customer churn. the study further emphasizes the significance of early intervention to retain at-risk customers. Banks able to recognize patterns, such as reduced activity and demographic factors, enabling them to design targeted retention strategies by analysing historical data. Moreover, the study will focus on optimizing predictive models like, Logistic Regression, Naive Bayers, Decision Tree, Linear SVC, FM Classifier, Random Classifier, and GBT Classifier while finding the most efficient model to improve efficiency and accuracy, ensuring competitive stability, minimizing revenue loss, and enhancing customer lifetime value. The goal is to perform successful big data analytical study that improves satisfaction and loyalty while addressing emerging challenges in a highly competitive market.

# **Introduction**

This section introduces the background, problem definition, purpose, aim, data source (including dataset details), and research questions, which presents an in-depth foundation for the Bank Customer Churn Analysis study. This report has been produced as a part of a study to underscore the factors influencing customer churn in banking while facing limitations, such as the restricted scope of the dataset that impact the results. Moreover, the report scope is outlined by summarizing the content covered, the structure of information presentation, and the methodology for data collection, generally using Kaggle as a secondary data source.

# **Background**

Customer churn or attrition take place when a bank and its client cancel their relationship, with closing client account and terminating all business interactions. This often considered a significant loss for the banking sector since customer retaining is generally more cost-effective than obtaining new ones. Customer acquisition can cost 5 to 25 times more than retention, according to the studies (Eva Ascarza, 2017).

Banks can implement a set of tailored strategies to reduce churn and improve customer satisfaction by identifying behaviours such as decreased engagement, demographic factors, and reduced account activity. Effective retention strategies are efficient for banks in saturated markets to maintain competitive stability, minimize revenue loss, and improve customer lifetime value.

Singapore is a global financial hub well known for its advanced financial ecosystem and competitiveness although it faces persistent challenges with customer churn. Customer attrition remains a significant issue despite their strong market position, as retaining existing customers proves more cost-effective and impactful for profitability than acquiring new ones (Asian Insights Office, 2019).

# **Problem**

Banks that fail to reach customer demands risk revenue losses and reduced lifetime value. Nearly 50% of Gen-Z fond of online banking, highlighting the importance of digital services in Singapore underlining the importance of digital services (Fintech News Singapore, 2024). In addition, 49% of Singaporean adults have closed old accounts when opening new ones, indicating significant customer churn (Forrester, 2024). According, to historical data analyzation, it illustrates factors affecting satisfaction and loyalty is crucial for effective churn prediction and mitigation (Wiley Online Library, 2024).

# **Purpose**

With a model capable of identifying key factors determining customer attrition and predicting potential churn movement, Singapore banks can improve foster loyalty, customer satisfaction, and mitigate churn effectively as it leverages predictive analytics and customer-centric retention strategies. Therefore, there's a need of a study focused to develop a potential big data analysis model specialized for the Singapore banking sector to manage the constant customer churn challenge.

# **Aim**

The goal of this research experiment is to analyse datasets to identify patterns in customer behaviour, engagement, and preference. Additionally, the model enables banks to stay competitive in a saturated market by bridging gaps like, competitive interest rates, unmet demand for seamless digital experiences, and personalized services.

# **Data Source**

Figure -Data Source Information

# **Research Questions**

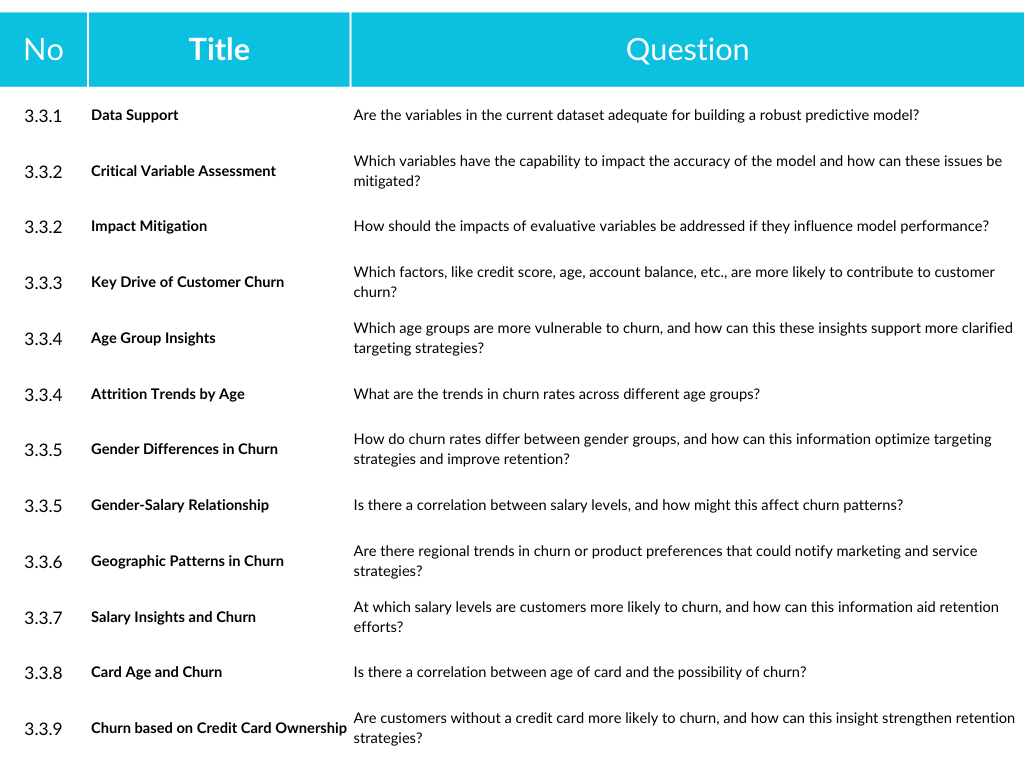


Figure -12 Research Questions

# **Methodology**

This section acknowledges Data Information, Research Design, Exploratory Data Analysis, Data Pre-processing, Model Training, Model Evaluation, Results, and Model Optimization. To ensure clarity and adaptability throughout the process, CRISP-DM methodology was chosen for its structured approach to machine learning projects (Suhendar, M.T.H. and Widyani, Y., 2023).

# **Data Information**

The data was queried via Spark SQL (Beyer, T., 2016) and the results were converted to local data frames using Pandas (McKinney, W., 2021) as shown below (Figure 3):

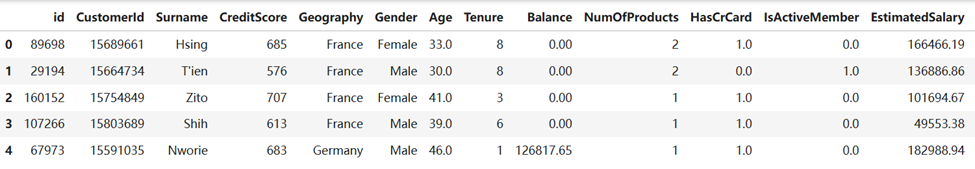


Figure -Data Frame

# **Research Design**

Including CRISP-DM theory (Martínez-Plumed et al., 2021), we integrated theory into the research, effectively bridging principles with practical implementation. The following figure illustrates our process of implementation and the corresponding steps in a structured manner (Figure 4):

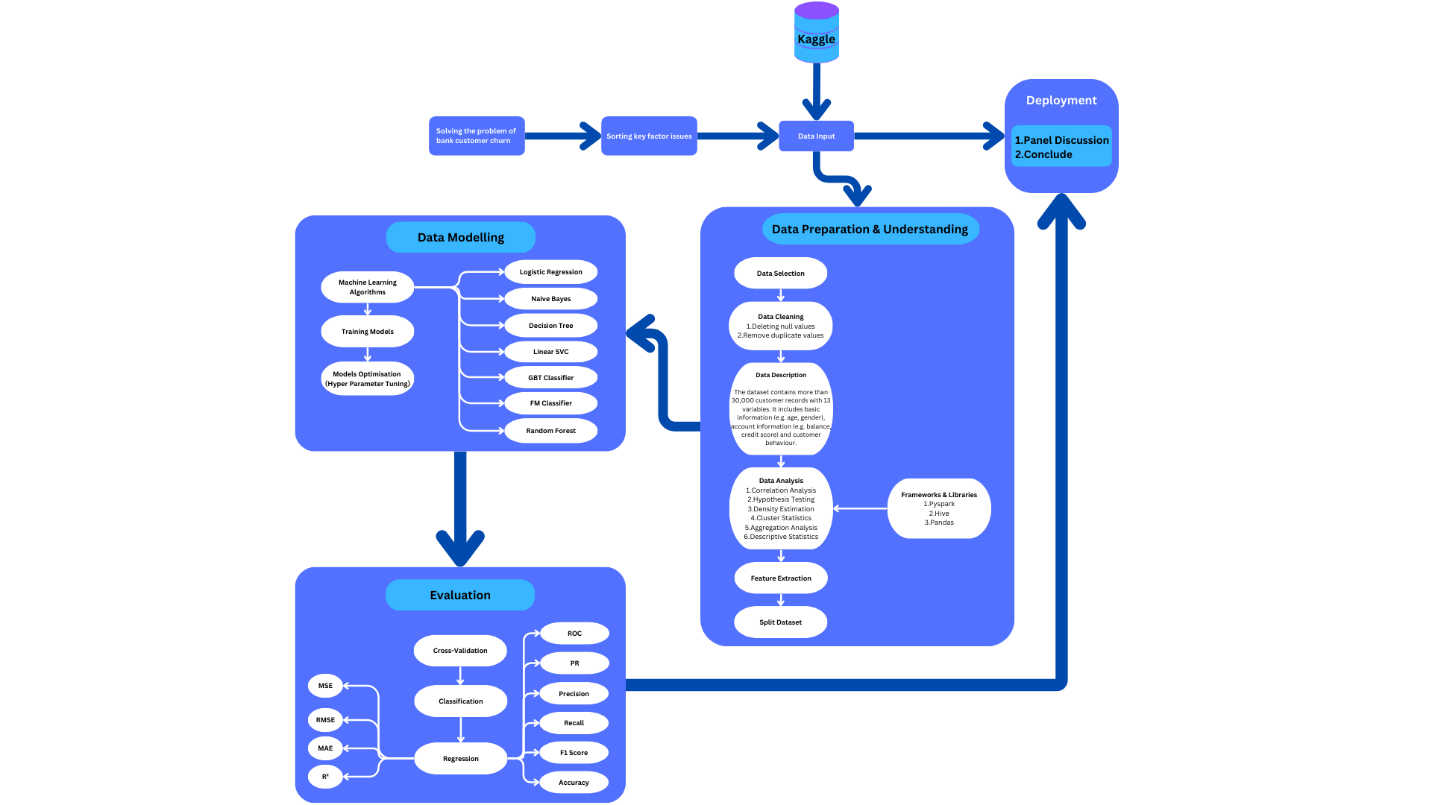


Figure -CRISP DM Methodology

# **Data Cleansing**

The data cleansing process involved removing null values to mitigate biases and eliminating duplicates to ensure dataset uniqueness, leading to a cleaner and more accurate foundation for analysis (Figure 7).

|  |  |
| --- | --- |
| Explanation | Code and Results |
| Remove Duplicate values: As shown (Figure 5) from the analysis results, there are no duplicate values in the data. | Figure -Remove duplicate values |
| Remove Duplicate Values: As shown (Figure 6) from the analysis results, the data does not have duplicate values. | Figure -Remove duplicate values: Grouped |

Figure -Data Cleansing

# **Exploratory Data Analysis**

The end goal of this study is to identify factors influencing bank customer churn through multidimensional data analysis (Myles & Gallo, 2021). Hive was used for large-scale data processing (Jain, 2018), while visualization tools like Seaborn (Proquest, 2020) and Matplotlib (Wicks, 2020) were used to create plots, charts, tables, and pie charts for effective result presentation.

|  |  |  |
| --- | --- | --- |
| Name | Question | Code and Illustration |
| **2.3.1. Correlation Matrix** | **Data Support**: Are the variables in the current dataset adequate for building a robust predictive model? | Figure -Correlation Matrix  The correlation matrix was used to analyse the correlation between different characteristics and customer churn. The figure (Figure 8) shows that age has a medium strength positive correlation (0.38) with customer churn, indicating that age may influence churn. Account balance is weakly correlated with churn (0.13), while estimated salary and active membership have weak correlations of -0.20 and -0.23, respectively, and the remaining variables are not significantly worthy of attention. However, the overall variables still provide a basis for constructing a predictive model for customer churn. |
| **2.3.2. Descriptive Statistics** | **Critical Variable Assessment**: Which variables have the capability to impact the accuracy of the model and how can these issues be mitigated? | Figure -Credit Score  According to the findings of correlation matrix results, 4 variables were selected.  **Credit Score (Figure 9)**: Ranges from 350 to 850 with a mean of 655.14 and a standard deviation of 80.33. While comparatively centralized, lower scores may introduce risks that affect predictions (Downey, 2014).    Figure -Age  **Age (Figure 10)**: Ranges from 18 to 92 with a mean of 38 and a standard deviation of 8.86. The main customer group is middle-aged, but the wide age span increases complexity with a certain distribution range (Downey, 2014).    Figure -Account Balance  **Account Balance (Figure 11):** Ranges from 0.0 to 250,898.09 with a mean of 54,953.60 and a standard deviation of 62,644.87, reflecting high variability. Outliers may skew performance (Downey, 2014).    Figure -Estimated Income  **Estimated Salary (Figure 12):** Ranges from 11.8 to 199,992.48 with a mean of 112,140.42 and a standard deviation of 50,360.84. The significant dispersion (income) indicates considerable economic differences among customers (Downey, 2014).  High variability of these variables can cause uneven gradient descent updates in algorithms sensitive to feature scales, leading to slower convergence and reduced accuracy.  **Mitigation (Figure 13):**    Figure -Mitigation |
| **2.3.2. Descriptive Statistics** | **Impact Mitigation:** How should the impacts of evaluative variables be addressed if they influence model performance? | Figure -Statistical Analysis    Figure -Statistical Analysis Code  The variability (Figure 14) in key evaluative variables (Credit score, Age, Balance, income), can disrupt model performance. The dispersion must be mitigated to maintain accuracy.  **Solution (Figure 16)**:  Figure -Solution |
| 2.3.3. Correlation Analysis | **Key Drive of Customer Churn**: Which factors, like credit score, age, account balance, etc., are more likely to contribute to customer churn? | Figure -Correlation Analysis Code  Focus of this correlation analysis is to identify key variables influencing customer churn (Tabachnick & Fidell, 2013). According to the illustration (Figure 18),  **Credit Score**: Correlates with churn at -0.031, indicating little to no effect on churn likelihood.  **Age**: Shows a moderate positive correlation of 0.340 with churn, suggesting older customers are more likely to churn.  **Account Balance**: Correlates with churn at 0.132, indicating a weak but noticeable connection.  **Estimated Income**: Correlation is 0.023, signifying minimal relevance to churn.  **Gender**: Correlates with churn at 0.16, reflecting a slight positive influence with limited impact.  Age appears as a key determiner for customer churn based on the findings.    Figure -Correlation Analysis |
| 2.3.4. Hypothesis Testing | **Age Group Insights**: Which age groups are more vulnerable to churn, and how can this these insights support more clarified targeting strategies? | Based on hypothesis testing results (Deng, 2013):    **Under 30 Years (Figure 19)**:  Total customers: 4,673; churned customers: 379; churn rate: 5.5%.    Figure -Under 30  **30–50 Years (Figure 20)**:  Total customers: 25,295; churned customers: 4,983; churn rate: 17.0%.    Figure -Age 30~50  **50+ Years (Figure 21)**:  Total customers: 3,038; churned customers: 1,593; churn rate: 53.4%.    Figure -Age above 50  The age group above 50 years represents the highest churn rate, significantly surpassing other age groups, suggesting that older customers are more vulnerable to churn, possibly due to difficulties adapting to services, evolving preferences, or reduced engagement.  Targeting Strategies:  **50+ age group**: Suggest developing tailored strategies such as personalized support, simplified interfaces, or loyalty programs to address specific challenges and retain customers.  **30–50 age group:** Suggestfocusing on enhancing service value through promotions and engagement campaigns, as their churn rate is moderate but influential due to the large customer base. |
| 2.3.4. Hypothesis Testing | **Attrition Trends by Age:** What are the trends in churn rates across different age groups? | Hypothesis testing results (Deng, 2013) reveal an intelligible trend with age increasing Churn rates (Figure 22):    Figure -Churn Rate    Figure -Histogran of analysed results for different age groups |
| 2.3.5. Density Estimation | **Gender Differences in Churn**: How do churn rates differ between gender groups, and how can this information optimize targeting strategies and improve retention. | Figure -Density Estimation Code  Density estimation analysis (Downey, 2014) reveals the following churn rates by gender (Figure 25):  **Male Clients**: Total: 17,886; churned: 2,354; churn rate: 13.16%.  **Female Clients**: Total: 13,522; churned: 3,473; churn rate: 25.68%.  Male clients exhibit a lower churn rate while, female clients contribute more to revenue with a higher average income of $112,724.58, compared to $111,768.02 for males. This highlights the importance of focusing on retention strategies for female customers, particularly high-income earners, who are more likely to churn.  Targeting Strategies:   * Develop gender-specific retention plans, recognizing the behavioural differences in churn tendencies between male and female customers. * Implement personalized engagement for high-income female clients, such as premium customer care and exclusive offers. |
| 2.3.5. Density Estimation | **Gender-Salary Relationship:** Is there a correlation between salary levels, and how might this affect churn patterns? | This analysis shows (Figure 26) a relationship between salary levels and churn pattern, with female clients having slightly higher average incomes yet higher churn rates. This propose that salary level alone may not be a cause to mitigate churn rather, could correlate with customer expectations.  Suggestions (Figure 25):    Figure -Suggestions    Figure -Histogram of gender analysis results |
| 2.3.6. Cluster Statistics | **Geographic Patterns in Churn**: Are there regional trends in churn or product preferences that could notify marketing and service strategies? | Figure -Subgroup Statistical Codes  This analysis conducted using subgroup statistics (Müller, 2021) to find geographic trend (Figure 28), highlights notable variations in customer distribution and churn rates across different countries (Figure 29). For instance:   * Countries with higher churn rates require targeted attention to address customer dissatisfaction and reduce attrition effectively. * Countries with the largest customer bases do not necessarily exhibit the highest churn rates, indicating diverse regional dynamics ().     Figure -Distribution rates by country    Figure -Attrition rates by country  Insights for Service and Marketing Strategies (Figure 30):    Figure -Strategies |
| 2.3.7. Aggregation analysis | **Salary Insights and Churn**: At which salary levels are customers more likely to churn, and how can this information aid retention efforts? | Figure -Aggregation Analysis Code  An aggregation analysis of customer salary levels (Müller, 2021) reveals (Figure 33) significant differences in churn rates across various salary brackets. Notably:     * Higher salary levels correspond to higher churn rates, suggesting that high-income customers exhibit greater sensitivity to service quality. * These customers often have elevated expectations of banking services and are more likely to switch to competitors following negative experiences or unmet expectations.   Retention Strategy Recommendations (Figure 32):      Figure -Recommendations    Figure -Histogram of salary analysis results |
| 2.3.8. Descriptive Statistics | **Card Age and Churn**: Is there a correlation between age of card and the possibility of churn? | Figure -Descriptive statistics code  The descriptive statistics (Downey, A. B., 2014) examined the relationship between the length of time customers have used their bank cards and their possibility of churning. The results reveal (Figure 35):  Higher churn rates are observed among users with less than one year of card usage.  Trend suggests a correlation between shorter card tenure and increased likelihood of churn, displaying the importance of retaining new customers during the staring period of engagement.    Figure -Histogram of the results of bank cards vs. life span  Recommendations for Retention Strategy (Figure 36):    Figure -Strategy |
| 2.3.9 Subgroup Statistics | **Churn based on Credit Card Ownership:** Are customers without a credit card more likely to churn, and how can this insight strengthen retention strategies? | Figure -Subgroup Statistical Codes  This subgroup statistics (Müller, M., 2021) revealed a significant variation in churn rates between customers with and without credit cards (Figure 39):   * Customers who hold credit cards exhibit lower churn rates, indicating a stronger sense of engagement or dependency on the bank's services. * On the other hand, customers without credit cards show higher churn rates, suggesting a lack of deeper integration into the bank's ecosystem.   Recommendations for Retention Strategy (Figure 38):    Figure -Strategy    Figure -Credit Card Holding Analysis Bar Chart |

Figure -Exploratory Data Analysation

## Data Pre-processing

The Spark MLlib Pipeline framework (Géron, A., 2022; Vohra, R., 2020) is used for data preprocessing, integrating missing value handling, feature transformation, and normalization into a unified workflow (Figure 41) during the study, for consistent and efficient dataset processing.

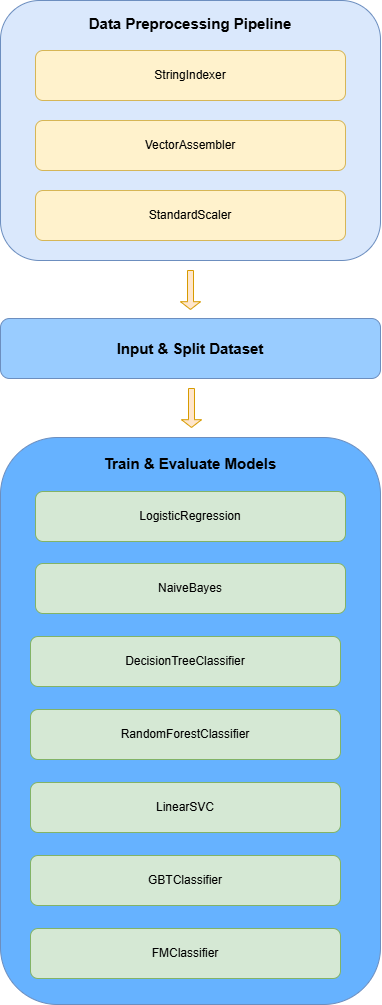
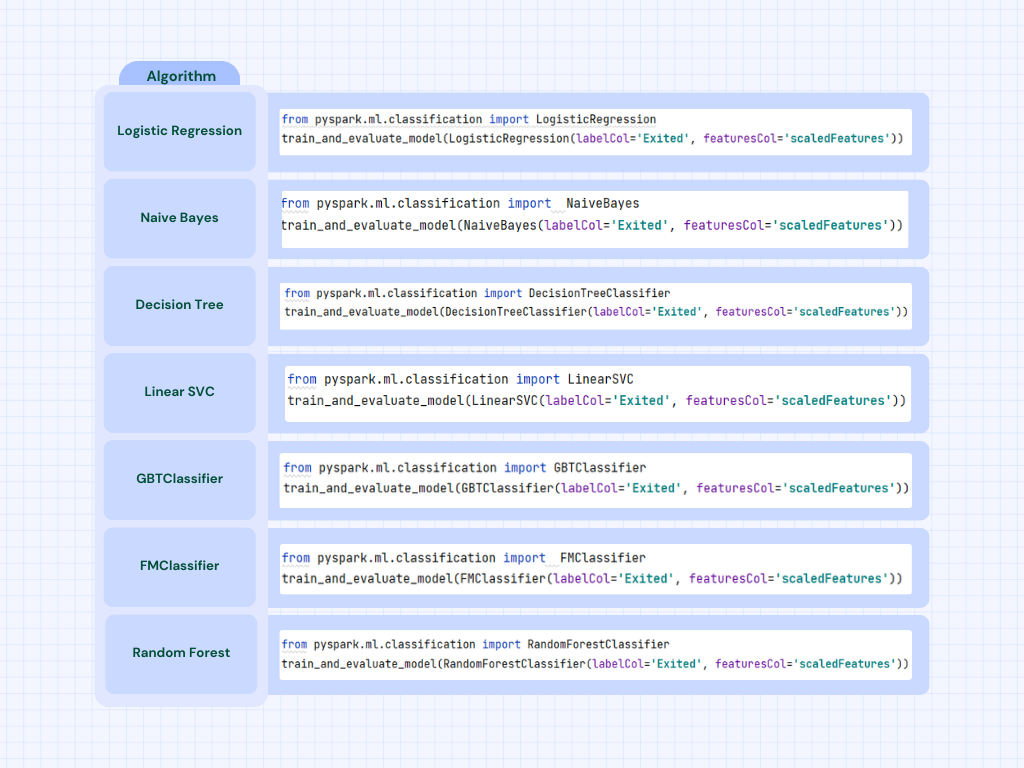


Figure -Preprocessing Pipeline Diagram

## Model Training

To ensure a fast modelling and data processing, this study uses Python and the Spark framework (Vohra, R., 2020) with Hadoop (Kumar, S., 2017) which makes efficient customer churn prediction and leveraging distributed computation. Various classification algorithms are engaged, including ML models like, logistic regression (Cox, D. R., 1958), Naive Bayes (Lewis, D. D., 1998), decision trees (Quinlan, J. R., 1986), linear SVMs (Cortes & Vapnik, 1995), factorization machines (Rendle, S., 2010), ensemble models such as random forests (Breiman, L., 2001) and gradient boosting trees (Friedman, J. H., 2001). Furthermore, while normalized scaledFeatures serve as input, Exited is the target label and automated model training is performed using the unified train\_and\_evaluate\_model function.

Figure -Model List

As the final model the gradient boosting tree (Friedman, J. H., 2001) was selected for its superior performance in capturing nonlinear relationships, feature interactions, and robustness to noise and feature scales. The gradient boosting tree capability to hold complex feature relationships while decreasing redundant information allowed optimal prediction results for this dataset.

## Model Evaluation

This is the section that focus on key components such as Indicators for Classification Tasks, Visualization of Accuracy and Recall, Critical Outcome Metrics, and the Confusion Matrix to comprehensively assess model performance.

## 2.6.1. Indicators for Classification Tasks

The Gradient Boosted Trees (GBT) Classifier showcases (Figure 43) a strong performance out of all the models across all evaluation metrics. It achieves an ROC score of 92.5%, along with a 91.0% score for accuracy and recall. Other algorithms, like Naive Bayes, record a significantly lower ROC of 44.8% and Precision-Recall (PR) of 17.1%, while GBT Classifier defines a clear lead, particularly in prediction accuracy and reliability. Contrarily, the FM Classifier presents an impoverished performance across all metrics, with an accuracy level of 74.6%, which shows its inadequacy for this classification task.

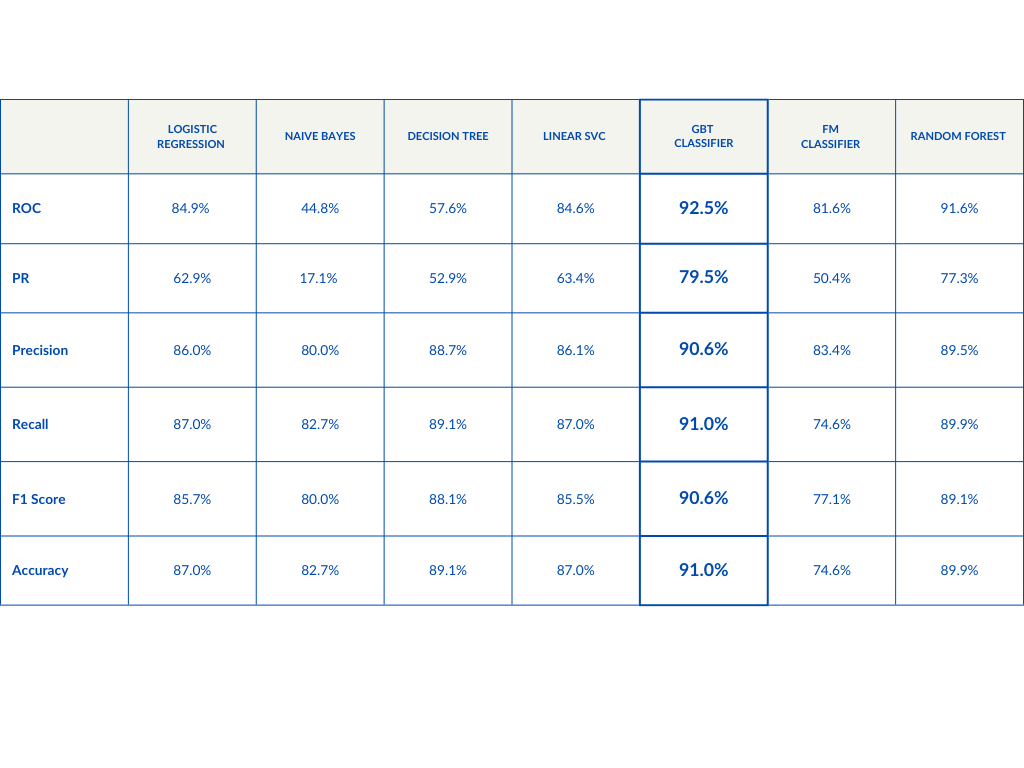


Figure -Indicators & Classifications

## 2.6.2. Accuracy and Recall Visualization

The visualisation results present the recall (Figure 45) and accuracy (Figure 44) comparison of the algorithms we analysed more intuitively. Among all the algorithms, the GBT Classifier and Random Forest Classifier have the highest accuracy and recall (0.91 and 0.90 for both), while the FM Classifier has significantly lower recall and accuracy than the other models with only 0.75. This helped us to finalise that GBT Classifier is the best choice for this task.

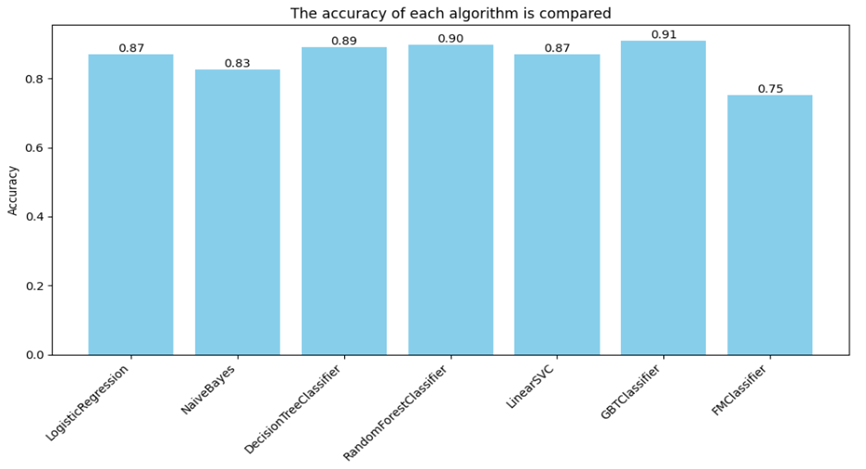


Figure -Accuracy Comparison

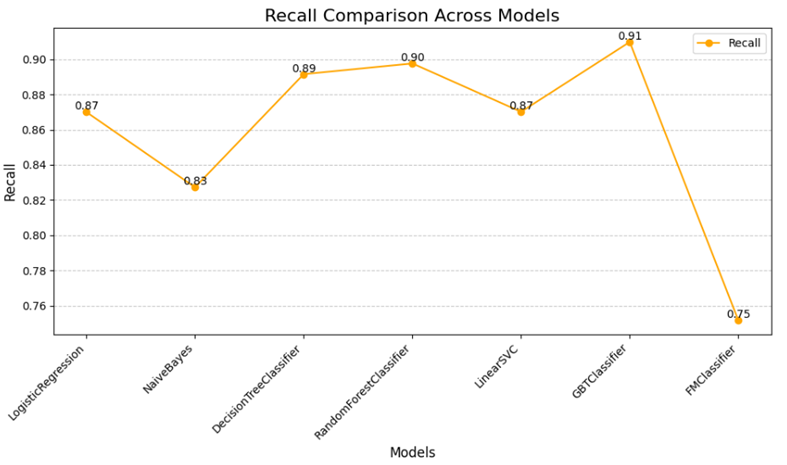


Figure -Recall rate Comparison

## 2.6.3. Key Indicators for Outcomes

As the final model selection, gradient boosting tree achieved the following results (Figure 46): MSE (Géron, A., 2022) of 0.0912, specifying minimal squared error; RMSE (Myles & Gallo, 2021) of 0.3020, indicating an average deviation of 0.3; MAE (Géron, A., 2022) of 0.0912, returning stable and low error; and R² (Chatterjee & Hadi, 2012) of 0.3977, describing 39.8% of data variation. These metrics recommending reasonable predictive ability with potential for optimization.

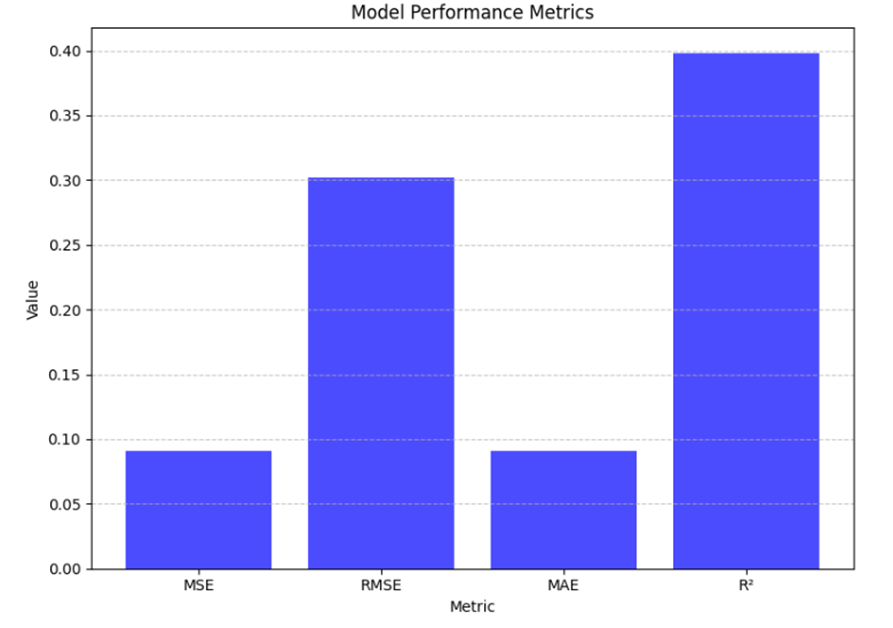


Figure -Performance Matrices

## 2.6.4. Confusion Matrix

According to the illustration below (Figure 47),

* Upper left corner: A high value (with 7345) here indicates that the model we finalised, is good at correctly identifying customers who stay.
* Upper right corner: The number of customers who did not churn, but the model incorrectly predicted them to churn (with 275).
* Bottom left corner: A value of 579 indicates that the model is missing many customers who will leave, which is an issue because the model is failing to identify customers at risk of churn.
* Bottom right corner: A higher value (with 1163) here indicates that the model is effective in identifying customers who will leave.

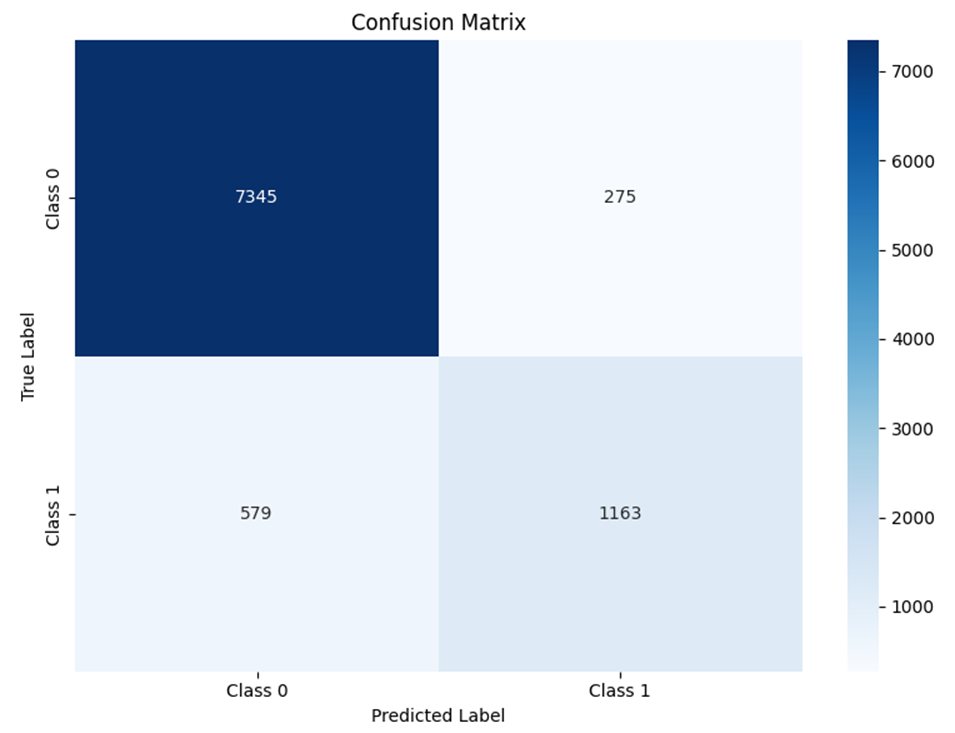


Figure -Confusion Matrix

## 2.7. Results and Model Optimization

Current confusion matrix reveals areas for improvement, especially where customers at risk of churning are not being identified (false negatives: 579). By addressing this issue, we can enhance the model's sensitivity and ensure more effective retention strategies, reducing customer loss.

## 2.7.1. Cross-Validation

The model parameters were optimized using grid search with maxDepth, maxIter, and stepSize, merged with five-fold cross-validation. As for optimal settings, maxDepth = 5, maxIter = 50, and stepSize = 0.1, that achieved a 92.5% AUC on the test set's ROC curve.

## 2.7.2. Hyperparameter Tuning

The gradient boosting model achieved maximum performance with differing train-test split ratios: 90.4% accuracy at 6:4, 90.7% at 8:2, and 91% at 7:3. The 7:3 ratio was selected as the final division (Figure 48).

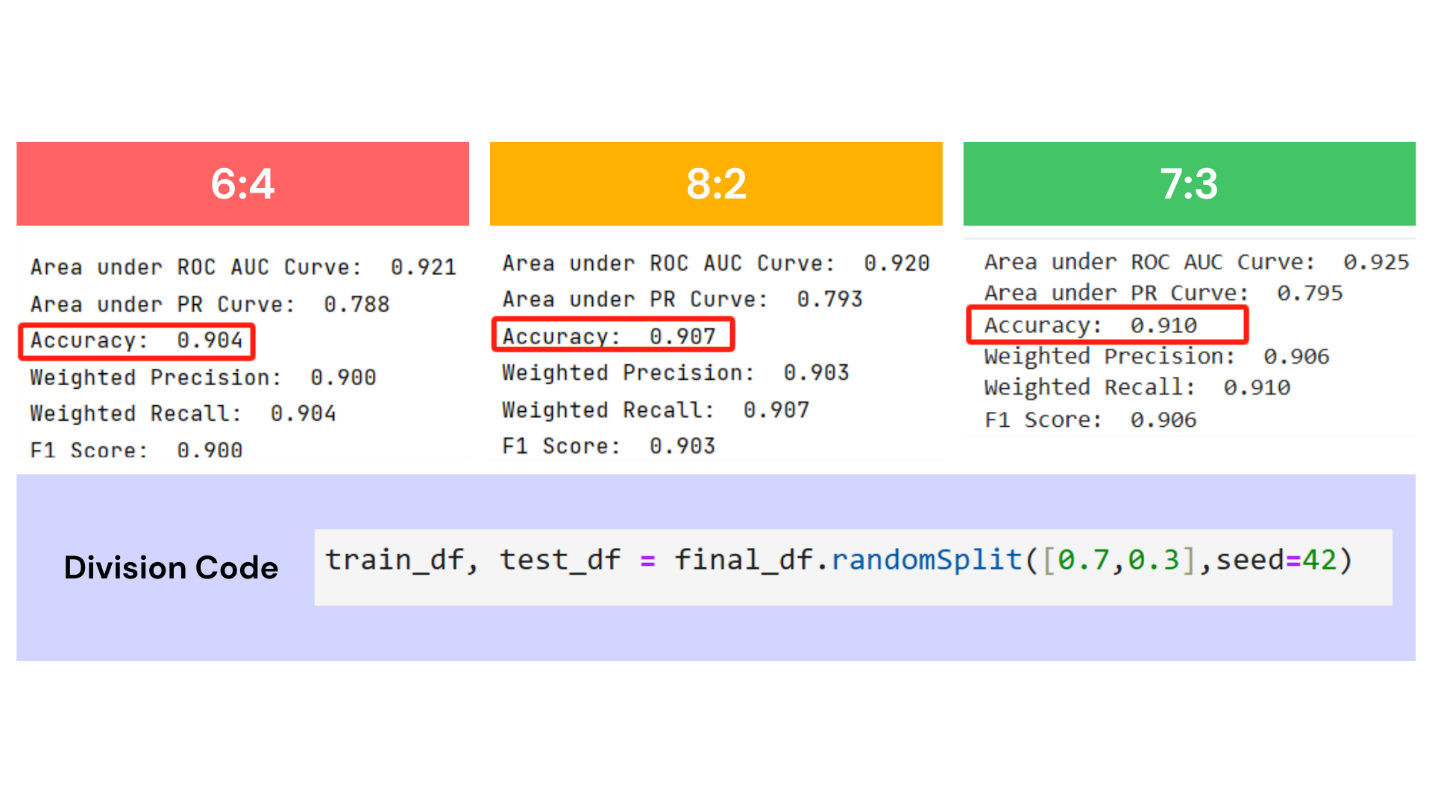


Figure -Hyperparameter Ratio

## 2.7.3. Limitations

Few limitations were identified further attempts to optimisation (Figure 49),

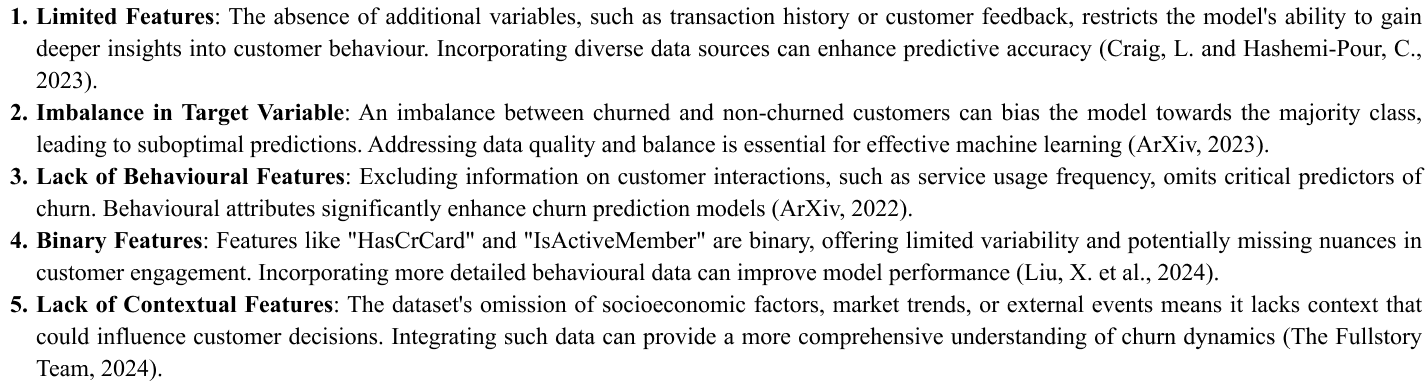


Figure -Limitations

## 2.7.4. Future Enhancements and Research Questions

Future enhancement plan (Figure 50),

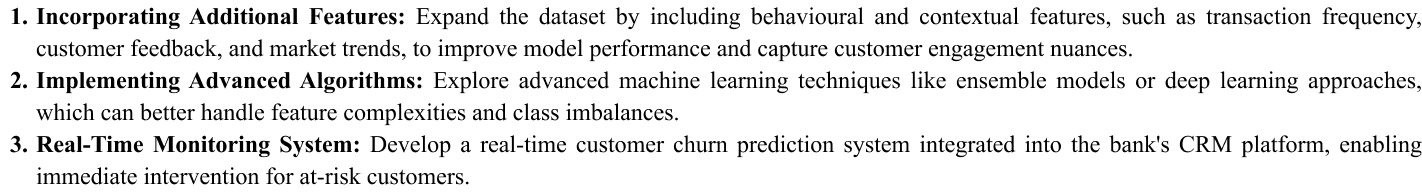


Figure -Enhancements

Expecting future research questions (Figure 51),

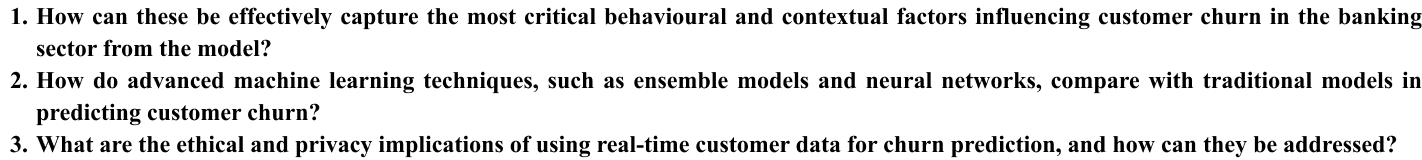


Figure -Future Research Questions

## 3. Ethical, Legal, and Professional Considerations (Figure 52)

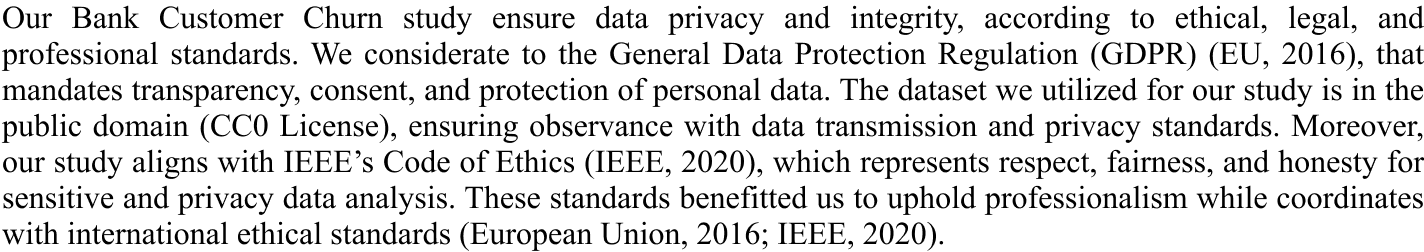


Figure -Considerations

## 4. Conclusion

In conclusion, implementing early-warning systems is vital for a bank to identify at-risk customers. For a bank, success relies on satisfaction, engagement, and service quality which gained by retaining existing customers. Banks can design targeted retention strategies based on factors like reduced activity and demographics, analysing historical data. According to our study, we have identified that merging advanced technologies and predictive analytics with a powerful algorithm like, Gradient Boosting Tree will improve accuracy, reduce churn, and enhance customer lifetime value, ensuring competitive stability and long-term success. Moreover, as a group we learnt that the quality and features of the dataset is parallelly important as the model we’re using to predict the outcome.