**ANALYSIS OF BANK CHURNERS – A DEEP DIVE UTILISING SUPERVISED LEARNING ALGORITHMS WITH THE CRISP-DM METHODOLOGY**

By: Onyiriuba Leonard & Onwumbiko Chidera

TechBros :**TEAM NAME**

Applied Techniques of Data Mining and Machine Learning :**COURSEWORK**

[Kaggle](https://www.kaggle.com/datasets/radheshyamkollipara/bank-customer-churn) :**DATASET**

**TABLE OF CONTENT**

1 - [**Business Background Understanding**](#Business_Background_Understanding)

* 1. – Introduction
  2. – What is Churn?
  3. – Why Churn Analysis?
  4. – Project Context and Objective
  5. – Understanding the CRISP-DM Methodology
  6. – Literature Review: Data Mining in the Application Domain

2 - [**Data Understanding and Exploration**](#Data_Understanding_and_Exploration)

2.1 – Data Set Overview  
2.2 – Data Characteristics and Quality Evaluation  
2.3 – Initial Insights and Patterns (Python & Pandas Profiling)  
2.4 – Interactive Data Analysis (Power BI)

3 - [**Data Preparation and Pre-Processing**](#Data_Preparation_and_PreProcessing)

3.1 – Data Cleaning and Transformation  
3.2 – Handling Outliers  
3.3 – Feature Engineering   
3.4 – Feature Selection  
3.5 – Summary of Further Steps

4 - [**Modeling and Model Evaluation**](#Modelling_and_Model_Evaluation)

4.1 – Test Classifier Evaluation – With Complain Variable and Class Imbalance  
4.2 – Test Classifier Evaluation – Without the Complain Variable and Class Imbalance Fixed   
4.3 – Hyperparameter Tuning  
4.4 – Interpretation and Analysis of Results

5 - [**Project Evaluation and Summary**](#Project_Evaluation_and_Summary)

5.1 – Summary of Findings and Insights  
5.2 – Recommendations and Future Work  
5.3 – Impact on Banks, Business, and Society

6 - [**References**](#References)

7 - [**Appendix**](#Appendix)

1 - **Business Background Understanding**

1.1 – Introduction

Every bank wishes a scenario where customers who patronize them forever continue operations with the bank. For this, banks set customer retention as part of their primary goals while setting targets towards achieving this. It is impossible to imagine a scenario where a bank successfully achieves 100% customer retention. Several factors influence customers decisions to leave the bank, and this can be based off region, diversity, inclusion, competition from other banks, product offerings, etc. With customer acquisition being a major objective of the bank, along with customer retention, banks invest heavily towards minimizing loss of customers in business operations – Churn.

1.2 – What is Churn?

Churn is a situation where customers of a business switch to competitors or refuse to continue business proceedings with a business due to inherent factors created by the business or external processes outside the business. It is the decision of customers to leave a business after some time because of better alternatives or dissatisfied with the products and offerings of the business (Amplitude).

1.3 – Why Churn Analysis?

Competition pushes companies, businesses, and for our case-study banks towards pushing to achieve the best possible overall performance for product and service offerings to attract new customers and retain old ones. No business can exist without a customer; hence, churn analysis is paramount for any enterprise to succeed. The need for churn analysis cannot be neglected at any level of business operations as customers remain the backbone upon which businesses survive. Data has provided a way for companies and businesses to bridge the gap between efforts put forward by businesses and statistics to measure output which aids further analysis and decision-making processes.

1.4 – Project Context and Objectives

The banking industry faces intense competition, requiring banks to continually strive for better customer retention to maintain profitability and market position. Despite efforts to provide superior products and services, customer churn remains a significant challenge. Customer churn occurs when clients stop using a bank's services, either due to dissatisfaction or the allure of competitors' offerings. High churn rates can drastically impact a bank's revenue and growth.

In this context, leveraging data analytics and machine learning to predict and understand customer churn is essential. By identifying patterns and factors that lead to churn, banks can implement targeted strategies to retain at-risk customers, thereby improving customer satisfaction and loyalty.

The primary objective of this project is to build a predictive model that accurately identifies customers who are likely to churn. This model will enable the bank to take measures to retain these customers, thereby reducing churn rates and enhancing overall customer satisfaction. This project aims to provide insights and recommendations based on the model’s predictions to help the bank develop targeted customer retention strategies, as well as identify key indicators for likelihood of churn and suggest interventions to mitigate these factors.

1.5 – Understanding the CRISP-DM Methodology

With the ever-growing need for customer churn analysis, this project analyses customer churn in banks. For this, we make use of a bank dataset from [Kaggle](https://www.kaggle.com/datasets/radheshyamkollipara/bank-customer-churn) and arrive at our conclusions following the CRISP-DM methodology. The CRISP-DM methodology stands for Cross-Industry Standard Process for Data Mining, and it defines the standard procedure data scientists should follow in offering deliverables from any data mining project. It defines a step-by-step process that allows a deep-dive, thorough understanding of the project goals, and how to achieve them. The steps followed in our project as defined by the CRISP-DM methodology are:

* Business Background Understanding
* Data Exploration and Understanding
* Data Preparation and Pre-Processing
* Modelling and Evaluation
* Project Evaluation and Summary

The CRISP-DM will be employed as a solution for mapping our project workflow from Business Understanding to Project Evaluation.

1.6 – Literature Review: Data Mining in the Application Domain

Our literature is focused on six studies focused on predicting bank customer churn, offering insights into methodologies, data handling, and model performance.

Li and Chen (2021) explored the use of logistic regression, decision trees, random forests, and support vector machines (SVM) to predict customer churn. By analyzing a dataset that included customer demographics, account information, and transaction history, they found that random forests and SVM provided superior accuracy. The study highlighted the significance of transaction frequency and customer service interactions as critical predictors of churn.

In another study by Banki and Shajari (2020), they proposed a hybrid model combining Kernel Support Vector Machines (KSVM) and Artificial Neural Networks (ANN) to enhance churn prediction accuracy. Their approach, tested on a dataset with various customer attributes, demonstrated that the hybrid model outperformed individual models like SVM, KSVM, and ANN. The combination of KSVM’s feature mapping and ANN’s classification capabilities resulted in higher predictive performance, suggesting that hybrid models can effectively capture complex patterns in customer behaviour and improve retention strategies.

In their study, Liu et al. (2021) proposed an improved Whale Optimization Algorithm (WOA) combined with SVM to predict customer churn. This model was tested on VIP customer data from a domestic commercial bank and showed superior accuracy and hit rate compared to standard SVM, GA-SVM, multilayer perceptron, and logistic regression models. The improved WOA-SVM model’s ability to handle high-dimensional, nonlinear data effectively highlights the importance of using optimization algorithms to refine machine learning models for better churn prediction.

Ghassemi et al. (2022) examined the use of a Bi-LSTM model, a type of deep neural network, to predict customer churn in the Iranian banking sector. Utilizing daily transactional data from Pasargad Bank, their study revealed that the Bi-LSTM model significantly outperformed traditional machine learning techniques like decision trees, random forests, and logistic regression.

Wang and Zhang (2020) focused on improving the Fuzzy C-Means (FCM) algorithm to enhance churn prediction. By introducing a new method for calculating the effectiveness function, their approach demonstrated higher accuracy in predicting customer churn compared to traditional FCM methods.

Negash et al. (2022) investigated customer churn prediction at the Commercial Bank of Ethiopia using various machine learning algorithms, including logistic regression, random forest, SVM, K-nearest neighbours, and deep neural networks (DNN). Their study, analyzing a dataset with 204,161 records, found that the DNN model achieved the highest accuracy, precision, and recall. This research highlights the effectiveness of using advanced techniques like SMOTE for class imbalance and the critical role of feature selection in optimizing model performance, providing valuable insights for improving customer retention strategies in banks.

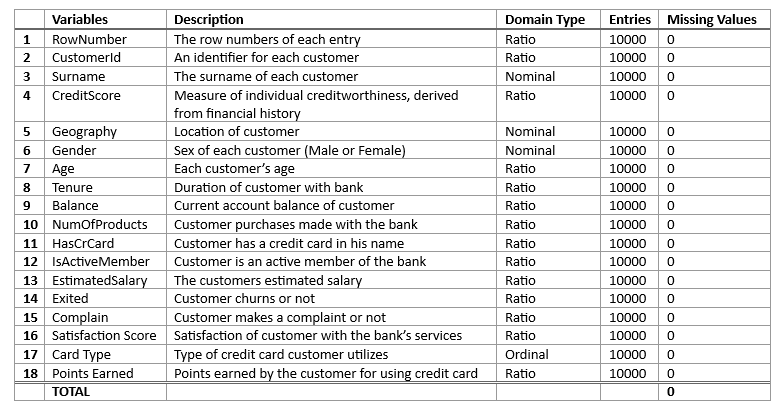
The findings from these studies emphasize the critical role of advanced machine learning and deep learning models in predicting customer churn. By leveraging these models, banks can more accurately identify at-risk customers and implement targeted interventions to reduce churn rates. The insights gained from these studies will guide the application of the CRISP-DM methodology in this research, with the aim of creating a predictive model for customer churn in the banking sector.

2 - **Data Understanding and Exploration**

2.1 – Data Set Overview

The dataset is gotten from [Kaggle](https://www.kaggle.com/datasets/radheshyamkollipara/bank-customer-churn) which makes available, open-source datasets, and allows for collaborations among data enthusiasts. Our dataset comprises 10,000 rows and 18 columns. The domain types of each column are specified below:

**Table 1**: Variables, description, domain type, number of valid entries, and missing values in the dataset.

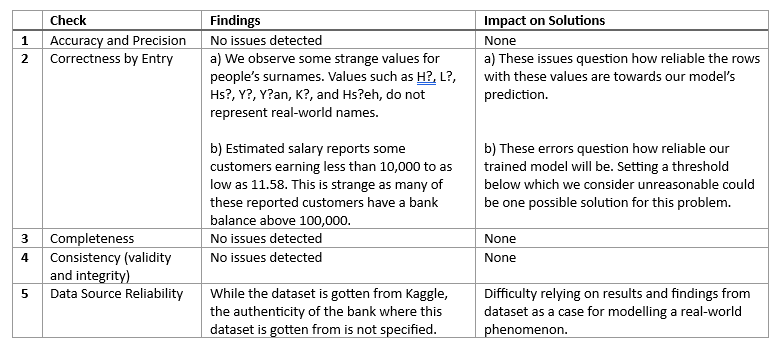


From table 1, we observe that the dataset gotten from [Kaggle](https://www.kaggle.com/datasets/radheshyamkollipara/bank-customer-churn) without any data cleaning or pre-processing conducted has no missing values.

2.2 – Data Characteristics and Quality Evaluation

The quality of our dataset will be judged by the following - Accuracy and Precision, Correctness by Entry, Completeness, Consistency (validity and integrity), and Data Source Reliability Checks.

**Table 2**: Summary on data quality checks



Many of the issues highlighted in table 2 are further discussed in later sections during the iterative model building phase.

2.3 – Initial Insights and Patterns (Python & Pandas Profiling)

Exploratory data analysis (EDA) helps us derive insights about our dataset, understand the structure of the variables, get a glance at the level of data cleaning and preparation we need to undertake before modelling. For our analysis, we make use of Pandas Profiling – A library for EDA and use python scripts for our initial EDA.

As part of our EDA, we considered the following:

* The count of categories among categorical variables.
* The number of duplicate entries in the rows.
* The number of missing values.
* Descriptive statistics of variables.
* Correlation among variables.
* The distribution of each column.

**Table 3**: Count of class categories among the variables – Gender, Geography, Card Type, Tenure, Satisfaction Score, NumOfProducts, IsActiveMember, HasCrCard, Exited, and Complain.

Here, we want to look at the distribution of the Gender, Geography, and Exited variables. All other variables mentioned above are provided in [Appendix A](#Appendix_A).

The colour scheme for highest and lowest values in each table are defined below:  
- Highest  
- Lowest

Figure 3.1: Geography



Figure 3.2: Gender

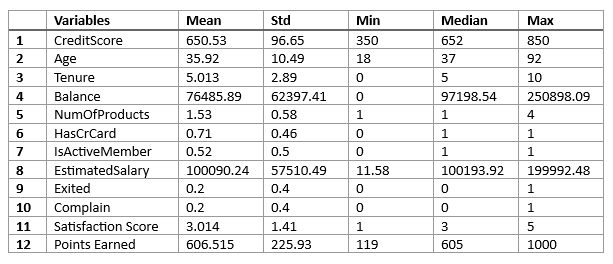


Figure 3.3: Exited



Our target variable is Exited, and from figure 3.3 in Table 3, we observe an unequal distribution for Exited, indicating a binary classification problem with class imbalance. Although the difference of about a thousand customers between gender classes is negligible as both are well represented, the location of our customers is unevenly distributed, with half being French. This class imbalance suggests the model might be better suited to a specific demographic of bank customers. In our final model analysis, attention needs to be drawn to this problem.

**Table 4:** Descriptive statistics of variables – CustomerId, RowNumber, Geography, Gender, Card Type, and Surname ignored as either categorical data or irrelevant for our analysis.



As seen in table 1, the dataset has no missing value. From our dataset, we also have no duplicate entries recorded in the data as each row represent a new customer record for the bank.

The distribution of the variables as seen in [Appendix B](#Appendix_B), provides some useful insight. The age distribution is skewed to the left, highlighting extremes for elderly people in our data. The balance variable's distribution shows a high peak at zero balance followed by a normal distribution for all non-zero balances, indicating a zero-inflated distribution. This needs to be properly handled to avoid the model underestimating zeros in customers' account balances, which could lead to biased parameter estimates and predictions. Steps to handle this issue are outlined in our data preparation section. The credit score variable is closely normally distributed. [Appendix C](#Appendix_C) shows the scatter matrix of the interaction among all variables in our dataset. From our feature space in the scatter matrix, all interactions show no clear distinction or pattern between the people who leave the bank, and those who stay.

A comprehensive exploratory data analysis is conducted including correlation analysis across variables using Pandas Profiling. From Pandas Profiling, the following problems have been flagged in the dataset and need to be resolved.

* Surname has a high cardinality: 2932 distinct values.
* Exited is highly overall correlated with Complain.
* Complain is highly overall correlated with Exited.
* RowNumber is uniformly distributed.
* RowNumber has unique values.
* CustomerId has unique values.
* Tenure has 413 (4.1%) zeros.
* Balance has 3617 (36.2%) zeros.

A snapshot of the pandas profile report is seen in [Appendix E](#Appendix_E). The full report and analysis can be accessed from the provided link. ([Pandas Profiling Report – EDA](https://leonardleo.github.io/Analysing-Key-Performance-Indicators-Leading-to-Customer-Churn-in-Banks/pandas_profiling_report/initial_EDA.html))

2.4 – Interactive Data Analysis (Power BI)

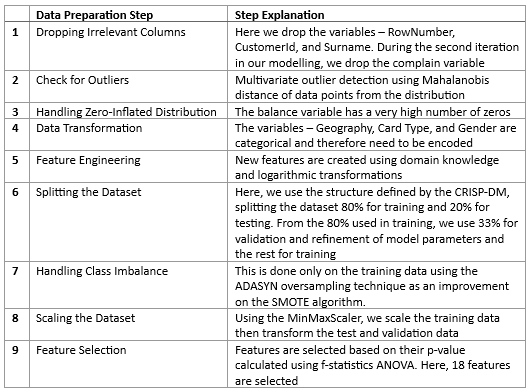
Data analysis is conducted on the dataset using Power BI. From our dashboard, we observe that France has the highest number of customers, which is twice the size of Germany and Spain. However, France has the least customer churn ratio, with only 811 customers leaving, while Germany has the highest churn ratio. We observe from the dashboard, every customer who has purchased high number of products with the bank churns. [Appendix D](#Appendix_D) shows a snapshot of the dashboard created with Power BI while link to the interactive dashboard is provided here.

([Power BI – Data Analysis](https://leonardleo.github.io/Analysing-Key-Performance-Indicators-Leading-to-Customer-Churn-in-Banks/interactive%20exploratory%20data%20analysis/index.html))

3 - **Data Preparation and Pre-Processing**

This section covers the data preparation, data cleaning, and data pre-processing steps carried out before creating a model from the data. Our data preparation step-by-step is tabled below.

**Table 5:** Steps for data preparation



Analysis of customer churn is a Binary Classification problem that predicts the likelihood of someone leaving the bank or staying with the bank. As seen from table 3 in section 2.3, the count of categories for the exited variable shows a class imbalance between the two categories we are trying to predict. Hence, we are faced with an Imbalanced Binary Classification problem.

3.1 – Data Cleaning and Transformation

We remove irrelevant columns from the data. The following columns are dropped for our analysis:

**Table 6:** Variables considered irrelevant and removed from our analysis.

A close-up of a text

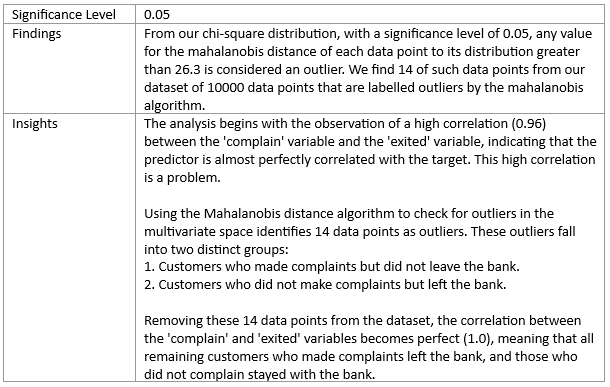
Description automatically generated

As part of our data transformation step, we convert the variables—Geography, Gender, and Card Type—into numerical data considering their respective domain types. Geography and Gender are nominal data types, so their conversion to numerical data is done by creating dummy variables. However, the Card Type variable is ordinal, given the inherent hierarchy in banks among different credit card types based on an internal metric that ranks customer performance. The hierarchy goes from Diamond, Platinum, Gold, and finally Silver, and is represented as such in our conversion.

3.2 – Handling Outliers

For our analysis, a multivariate outlier detection process was conducted using the mahalanobis algorithm, the chi-square distribution, and a specified significance level. The decision to proceed with a multivariate outlier technique as opposed to univariate process is due to the mahalanobis distance algorithms robustness in identifying outliers that occur jointly across multiple dimensions. Also, multivariate methods can handle more complex scenarios we find in analysis.

**Table 7:** Significance level, Handling outliers, and Insights.



3.3 – Feature Engineering

Feature engineering is the process of using domain knowledge to create new features or modify existing ones to improve the performance of machine learning models. The techniques used for data binning the derived features is determined by the distribution of the variable used to create the derived feature.

**Table 8:** Derived features table

**A screenshot of a paper

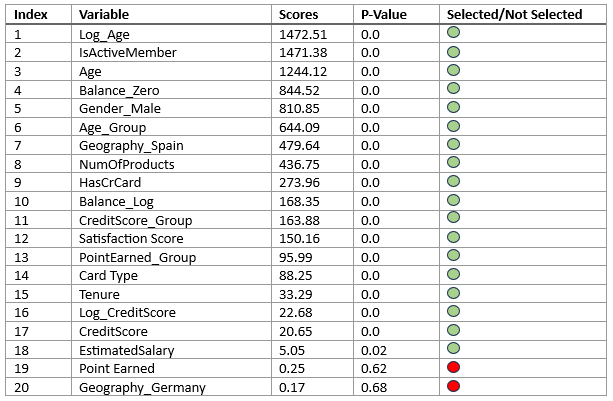
Description automatically generated**

3.4 – Feature Selection

As part of our data preparation step, feature selection is conducted using statistical p-values and f-statistics ANOVA. We use the Select False Positive Rate function from sci-kit learn, which selects features that reduce the model's false positives in predicting churn. This function identifies 18 relevant variables, assigns p-values, and provides corresponding ANOVA test scores. All 18 variables have p-values below the 0.05 significance level and are employed in our model.

The colour scheme for the features selected in each table are defined below. Variables with a p-value below 0.05 are not selected and vice-versa.  
- Not Selected.  
- Selected.

**Table 9:** Analysis of Features, P-Values, and Corresponding Scores from ANOVA Test



3.5 – Summary of Further Steps

Exploratory Data Analysis (EDA) is done a second time after data cleaning to verify the process. ADASYN is used to address class imbalance by creating more synthetic data for harder-to-classify minority class samples, improving on SMOTE. Resampling is applied only to the training data to prevent data leakage. Similarly, data scaling is done by fitting on the training data and then transforming the test and validation data.

4 - **Modeling and Model Evaluation**

Our model training, evaluation, and validation phase is divided into 2 sections:

* Modeling with the complain variable and class imbalance.
* Modeling without the complain variable and class imbalance fixed.

We compare 7 classifier algorithms built from the dataset. These algorithms are:

A white lined paper with black lines

Description automatically generated

These algorithms are evaluated based on time complexity, average score time, accuracy, model precision, precision for churners (PC1), precision for non-churners (PC0), model recall, average f1-score, true positive rate (TPR), true negative rate (TNR), cross-validation mean, cross-validation standard deviation. In our analysis, the prediction of customers leaving the bank is given higher priority across all metrics we compare.

We would also be drawing conclusions from how the classifier algorithms above perform in training, testing, and validation. All analysis results for training and validation sets including classification reports are seen in [Appendix F](#Appendix_F).

A math equations with black text

Description automatically generated with medium confidence

4.1 Test Classifier Evaluation – With Complain Variable and Class Imbalance.

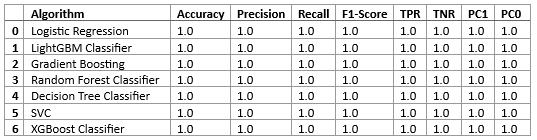
**Table 10:** Test results for classifiers with the complain variable and class imbalance.

Figure 10.1: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) for Algorithms as seen from Confusion Matrix

A table with numbers and letters

Description automatically generated

Figure 10.2: Evaluation Metrics for Algorithm Test Data

  
PC1 – Precision Class 1  
PC0 – Precision Class 0  
All algorithms have a cross validation mean of 100% and a standard deviation of 0.

4.2 Test Classifier Evaluation – Without the Complain Variable and Class Imbalance Fixed.

**Table 11:** Test results for classifiers without the complain variable and class imbalance fixed.

Figure 11.1: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) for Algorithms as seen from Confusion Matrix

A table with numbers and letters

Description automatically generated

Figure 11.2: Evaluation Metrics for Algorithm Test Data

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Algorithm | Accuracy | Model Precision | Model Recall | F1-Score | TPR | TNR | Test Precision (C\_1) | Test Precision (C\_0) |
| 1 | Gradient Boosting | 0.81 | 0.82 | 0.81 | 0.82 | 0.62 | 0.86 | 0.54 | 0.90 |
| 2 | Random Forest | 0.80 | 0.82 | 0.80 | 0.81 | 0.64 | 0.84 | 0.51 | 0.90 |
| 3 | Decision Tree | 0.78 | 0.80 | 0.78 | 0.79 | 0.60 | 0.83 | 0.47 | 0.89 |
| 4 | LightGBM | 0.84 | 0.83 | 0.84 | 0.83 | 0.50 | 0.93 | 0.64 | 0.88 |
| 5 | XGBoost | 0.85 | 0.84 | 0.85 | 0.84 | 0.54 | 0.93 | 0.65 | 0.89 |
| 6 | Logistic Regression | 0.77 | 0.77 | 0.77 | 0.77 | 0.44 | 0.85 | 0.44 | 0.86 |
| 7 | SVC | 0.81 | 0.80 | 0.81 | 0.80 | 0.48 | 0.89 | 0.52 | 0.87 |

Figure 11.3: Mean - Cross Validation (Train and Test), Fit Time, and Score Time

A table with numbers and a few black text

Description automatically generated with medium confidence

Figure 11.4: Standard Deviation - Cross Validation (Train and Test), Fit Time, and Score Time

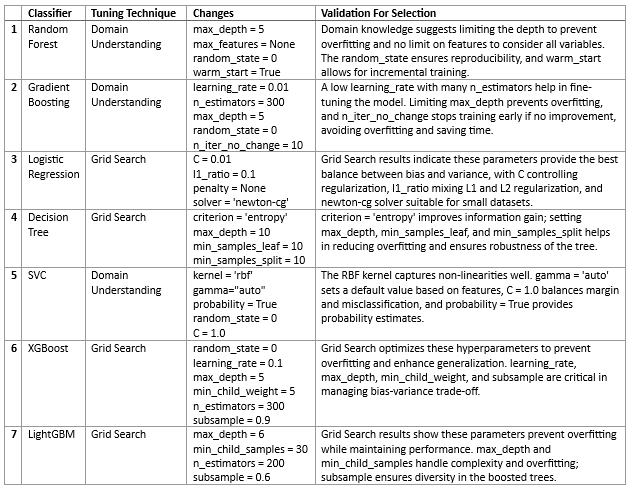
A table with numbers and a few digits

Description automatically generated with medium confidence

4.3 – Hyperparameter Tuning

For selecting the best parameters, we use a combination of GridSearch and domain understanding for parameter selection, which is a systematic way of testing multiple combinations of hyperparameters to find the optimal set for a given model. GridSearch thoroughly searches through a specified parameter grid, evaluating model performance for each combination. This approach helps in identifying the hyperparameters that yield the best performance for the model.

**Table 12:** Tuned model parameters from grid-search and domain understanding



4.4 – Interpretation and Analysis of Results

Our results show no model does an excellent job in predicting bank customers who churn. Not much difference can be found among our best performing models – LightGBM, Random Forest, Gradient Boosting, and XGBoost. However, the better performing model is judged with emphasis on its ability to predict class 1, fit time, and score time. From this, the LightGBM Classifier is selected as the best model with a TPR of 0.5 and precision for churners at 0.64, outperforming the Random Forest, Gradient Boosting, and XGBoost in fit time and score time. While we don’t see exceptional results, the two models which performed best are the XGBoost and LightGBM algorithms. When measured against time efficiency in producing results and retraining, the LightGBM algorithm performs best.

The results gotten from our chosen classifiers are convincing and are like results achieved from previous researchers. During our second iteration, we remove the complain variable as it introduces bias in our predictions given the complain variable and the target variable are almost perfectly correlated.

While the decision tree has the lowest fit time and score time at 0.006 seconds and 0 seconds respectively, our selected model LightGBM equally performs well with a fit time and score time of 0.31 seconds and 0.006 seconds respectively. The Gradient Boosting classifier has the worst fit time at 18.18 seconds. As seen from the tables in [Appendix F](#Appendix_F), observing the training performances of our classifiers, the classifiers do a brilliant job training the model with all classifiers having above 80% across all metrics.

When we compare these results with the test results in table 11 and the validation data also in [Appendix F](#Appendix_F), we can see that these algorithms are faced with the problem of Overfitting, therefore finding it difficult to generalize towards unseen data. With our classifiers overfit to the training data, the reliability of the results from our data mining process to be considered reliable, and deployed as a business solution is questioned.

Finally, from our chosen model (LightGBM), we recommend banks to pay close attention to customers – Estimated Salary, Account Balance, Age, Credit Score, Number of Products Purchased, and Duration with the Bank. These features provide the most information from data mining differentiating customers who leave the bank and those who stay.

5 - **Project Evaluation and Summary**

5.1 – Summary of Findings and Insights

In our study, we evaluated seven distinct algorithms to identify the most effective model for predicting customer churn at banks. Our findings indicate that no single model demonstrates exceptional predictive accuracy. However, the Light Gradient Boosting Machine (LGBM) Classifier emerged as the superior model, particularly in terms of its efficiency in fit time and score time. This model’s performance aligns with the outcomes by previous research in this domain.

Our analysis with Power BI provides useful insights pertaining to customer churn:

* Customers with a high number of products purchased invariably churn.
* Customers with a moderately high number of products and a non-zero account balance, who are not based in Germany, tend to churn.
* Germany exhibits the highest churn rate among bank customers.

These insights are crucial for banks to understand the patterns and behaviours associated with customer churn. It is recommended that banks closely monitor key customer attributes such as Estimated Salary, Account Balance, Age, Credit Score, Number of Products Purchased, and Duration with the Bank. These factors are instrumental in distinguishing between customers who are likely to churn and those who will remain loyal.

In conclusion, while the LGBM Classifier stands out for its efficiency and predictive capabilities, banks must exercise caution and further refine their predictive models to ensure they are robust and generalizable before implementation.

5.2 – Recommendations and Future Work

This research not only gives practical advice but also sets the stage for future studies to make bank predictions about customer churn even better. A key recommendation for improving our modeling process is to incorporate time as a variable. This will offer deeper understanding of the temporal aspects of customer churn, such as identifying critical periods when customers are more likely to leave the bank.

It’s also important to address potential biases that may arise from underrepresented data samples. Such biases can lead to models that disproportionately reflect the majority classes within certain demographics. To mitigate this, we recommend a more stratified approach to data sampling that ensures a balanced representation across all customer segments in each region.

For future work, we suggest dividing customers into two distinct groups based on their account balance—zero and non-zero—and developing separate models for each. This allows for better analysis of customer behaviour within these two fundamentally different segments. This approach could be expanded towards regional analysis. By creating region-specific models, we can capture unique patterns and factors influencing customer churn in different geographical areas.

Additionally, a hybrid approach that combines association rule mining with classification data mining could be explored. This method would enable the discovery of intricate rules and patterns that govern the relationship between customers who churn and those who remain. Such a hybrid model could potentially offer a more comprehensive view of customer behaviour.

Moreover, considering the rapid evolution of technology and customer expectations, it is vital to continuously update and validate our models with fresh data. This ensures that the models stay relevant and accurate in predicting churn.

Lastly, we recommend that banks consider the implementation of predictive intervention strategies. By leveraging the insights gained from data mining, banks can proactively engage with customers who are at risk of churning, offering personalized solutions and incentives to retain them.

Through these recommendations and future research directions, we aim to not only predict customer churn more accurately but also to empower banks with the tools to effectively prevent it.

5.3 – Impact on Banks, Business, and Society

The implications of accurately predicting customer churn extend far beyond the confines of banking institutions; they resonate across the business landscape and society at large. For banks, the ability to foresee and address churn translates into sustained customer relationships and financial stability. It enables banks to allocate resources more effectively, tailor customer experiences, and ultimately, retain a loyal customer base.

Businesses performing operations with banks, such as insurance companies and retail services, also benefit from these insights. A stable customer base in banks often means a more reliable market for their products and services. Moreover, the methodologies developed through this research can be adapted to predict behaviours in various customer-centric industries, fostering a culture of proactive engagement and customer satisfaction.

On a societal level, the reduction in churn can lead to a more stable economic environment. Customers who maintain long-term relationships with their banks contribute to the economic health of their communities by participating in lending and investment activities that stimulate growth.

Furthermore, the ethical considerations of data mining practices call for a balanced approach that respects customer privacy while striving for business innovation. As we refine our predictive models, it is imperative to uphold the highest standards of data integrity and ethical use of information.

6 – **References**

Amplitude. (n.d.). *What is Churn? Definition and Tracking Churn Rates*. [online] Available at: <https://amplitude.com/glossary/terms/churn> [Accessed 21 May 2024].

Cui, S. (n.d.). *Customer Churn Prediction Using Improved FCM Algorithm*.

Haddadi, S.J., Mohammadi, M.O., Bahrami, M., Khoeini, E., Beygi, M. and Khoshkar, M.H. (2022). Customer Churn Prediction in the Iranian Banking Sector. *2022 International Conference on Applied Artificial Intelligence (ICAPAI)*. doi:<https://doi.org/10.1109/icapai55158.2022.9801574>.

He, H. (2022). Bank Customer Churn Prediction Analysis Based on Improved WOA-SVM. *2022 3rd International Conference on Electronic Communication and Artificial Intelligence (IWECAI)*. doi:<https://doi.org/10.1109/iwecai55315.2022.00093>.

Hemalatha, P. and Amalanathan, G. (n.d.). *A Hybrid Classification Approach for Customer Churn Prediction using Supervised Learning Methods: Banking Sector*. ViTECoN.

Kaggle (n.d.). *Bank Customer Churn*. [online] www.kaggle.com. Available at: <https://www.kaggle.com/datasets/radheshyamkollipara/bank-customer-churn>.

Li, X. and Chen, Z. (2022). Customer Churn Prediction in Bank Based on Different Machine Learning Models. *2022 2nd International Signal Processing, Communications and Engineering Management Conference (ISPCEM)*. doi:<https://doi.org/10.1109/ispcem57418.2022.00061>.

7 - **Appendix**

**Appendix A**

*COUNT OF CLASS CATEGORIES AMONG VARIABLES*

Here, we count the categories among the variables – Card Type, Tenure, Satisfaction Score, NumOfProducts, IsActiveMember, HasCrCard, and Complain.

The colour scheme for highest and lowest values in each table are defined below:  
- Highest  
- Lowest

Card Type



Tenure



Satisfaction Score



NumOfProducts



IsActiveMember



HasCrCard



Complain



**Appendix B**

*DISTRIBUTION OF VALUES AMONG VARIABLES*

A group of blue graphs

Description automatically generated

**Appendix C**

*SCATTER MATRIX OF DISTRIBUTION AMONG VARIABLES*

A screenshot of a graph

Description automatically generated

**Appendix D**

*DATA ANALYSIS USING POWER-BI (DASHBOARDS)*

Main Dashboard

A screenshot of a computer

Description automatically generated

Customer Analysis Dashboard

A map of the world

Description automatically generated

**Appendix E**

*PANDAS PROFILING REPORT*

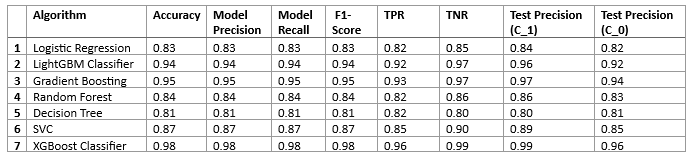
A screenshot of a computer

Description automatically generated

**Appendix F**

*TRAINING AND VALIDATION SETS RESULTS*

* **Training Data Results without Complain Variable and Classes Balanced**



* **Validation Data Results without Complain Variable and Classes Balanced**

A table with numbers and a number of objects

Description automatically generated with medium confidence