MIS710 – Machine Learning in Business - Trimester 1 2023
Assessment Task 2 – Case Study Report and Business Report – Individual

STUDENT ID
222294384

STUDENT NAME
ABIADE BISI-KAZEEM

Table of Contents

PART I **3**

Executive Summary **3**

Introduction **3**

Approach **5**

Data Preparation **6**

Exploratory Data Analysis (EDA) **9**

Univariate Analysis 9

Bivariate Analysis 11

Correlations 14

Model Development and Evaluation **16**

Model Comparison based on Selection Criteria 17

Hyper Parameter Tuning & Final Model Evaluation 18

Results and Interpretation **20**

Presentation of the Proposed Solution 20

Technical Recommendations **21**

Recommendations and Future Actions **22**

References **24**

PART I

Executive Summary

Customer attrition, also known as customer churn or customer turnover, refers to the rate at which customers discontinue their relationship with a company or business. It is a critical metric that measures the loss of customers over a specific period of time. High customer attrition rate leads to loss of revenue, market share, negative impact on company's image and reputation. In the case of VSNeoBank, the digital-only banking platform headquartered in Melbourne, customer attrition has become a growing concern.

To address this issue and improve customer retention in VSNeoBank, an analysis of their customer and transaction databases was conducted, providing valuable insights into the effects of customer attrition on the business. The dataset encompasses various customer attributes such as customer ID, sex, age, marital status, number of dependents, annual income, account type, credit limit, etc.

The analysis revealed several notable effects of customer attrition. Firstly, customers with higher credit limits and more accounts were found to have a lower likelihood of churning. This suggests that offering higher credit limits and encouraging customers to open multiple accounts can contribute to improved customer retention. Moreover, customers who had been with VSNeoBank for a longer duration were less likely to churn. This highlights the importance of nurturing long-term relationships with customers to reduce attrition.

Additionally, using Python, machine learning options were explored to develop models that could aid in predicting and understanding customer churn. By analyzing various factors and employing suitable machine learning techniques, such as classification algorithms, several patterns and indicators of customer churn were identified.

Based on these insights and results of the machine learning model, VSNeoBank can develop and implement targeted strategies to mitigate customer attrition. By focusing on improving credit limit allocation, encouraging multi-account usage, nurturing long-term customer relationships, and addressing the needs and concerns of customers with high utilization ratios and outstanding balances, VSNeoBank can effectively reduce attrition and enhance customer loyalty.

Introduction

The business problem to address for VSNeoBank is the increased customer attrition rate. The bank is facing a significant number of customers discontinuing their relationship, which poses a threat to its revenue, market share, and reputation. It is crucial for the bank to understand the underlying factors contributing to customer churn and develop effective strategies to improve customer retention.

Business Context using the BACCM framework

The Business Analysis Core Concept Model (BACCM) framework helps provide a comprehensive understanding of the business context surrounding the customer attrition problem at VSNeoBank.

1. **Stakeholders:**
    - **VSNeoBank:** The banking platform seeking to improve customer retention and mitigate the negative impact of customer attrition on their business. This includes VSNeoBank's management team, specifically the Head of Data Analytics, Ms. Emma Hoang.
    - **Customers:** Existing and potential customers of VSNeoBank who play a crucial role in driving the bank's growth, revenue, and success.

2. **Business Need:**
   VSNeoBank needs to address the issue of increased customer attrition to maintain its customer base, retain revenue, and sustain market competitiveness. By implementing effective strategies to reduce churn, the bank aims to enhance customer satisfaction, increase customer loyalty, and maximize customer lifetime value.

3. **Goals:**
   The primary goal is to reduce the customer attrition rate and improve customer retention. This involves retaining existing customers, increasing customer satisfaction, and enhancing the overall banking experience to foster long-term customer loyalty.

4. **Business Objectives:**
   The objectives include analyzing the customer and transaction databases to identify patterns and predictors of churn. Additionally, developing and testing machine learning models to predict customer churn and implementing targeted retention strategies based on the insights gained.
    - **Reduce customer attrition rate**: VSNeoBank aims to decrease the rate at which customers discontinue their relationship with the bank, ensuring long-term customer retention.
    - **Enhance customer satisfaction**: The bank seeks to improve customer experience and meet customer expectations, resulting in higher satisfaction levels.
    - **Increase customer loyalty:** VSNeoBank aims to foster strong relationships with customers, encouraging them to remain loyal and actively engage with the bank's services.
    - **Optimize profitability**: By retaining existing customers, the bank can minimize customer acquisition costs and maximize revenue from ongoing customer relationships.

5. **Contextual Factors:**
   The contextual factors influencing the customer attrition problem at VSNeoBank include the digital banking landscape, competitive market dynamics, customer preferences and expectations, the bank's product and service offerings, customer service quality, and the bank's overall brand reputation.

6. **Business Requirements:**
   The business requirements involve leveraging the available dataset to gain insights into customer demographics, behavior, and factors influencing churn. It requires the development of machine learning models to predict churn accurately and the implementation of proactive strategies to improve customer retention.

7. **Solution Approach:**
   The solution approach involves conducting a comprehensive analysis of the dataset, employing data analytics techniques to identify key drivers of churn, and developing machine learning models for churn prediction. The models will enable VSNeoBank to segment customers, identify at-risk customers, and implement personalized retention strategies.

8. **Constraints:**
   - **Digital-only platform**: VSNeoBank operates solely online, limiting face-to-face interactions and requiring innovative digital solutions to address customer attrition.
   - **Data availability and analysis:** The success of addressing the problem relies on accurate data collection, analysis, and modeling to identify churn predictors and develop effective retention strategies.

By applying the BACCM framework, VSNeoBank can gain a holistic understanding of the business problem, its context, and the necessary steps to address the customer attrition challenge effectively. This framework provides a structured approach to align stakeholders, define goals and objectives, consider contextual factors, and determine the solution approach, ultimately leading to improved customer retention and business success.

Approach

To address the customer attrition challenge at VSNeoBank, a machine learning approach will be adopted. Machine learning algorithms can leverage the dataset provided to analyze patterns, identify key factors contributing to customer attrition, and develop predictive models to forecast customer churn. This approach involves the following steps:

- **Data Preprocessing:** The dataset will be preprocessed to handle missing values, outliers, and categorical variables. This may include techniques such as data imputation, feature scaling, and one-hot encoding.

- **Feature Selection**: Relevant features will be selected from the dataset based on their significance in predicting customer attrition. This step helps to reduce dimensionality and focus on the most influential variables.

- **Model Selection:** Several machine learning algorithms will be evaluated to determine the best approach for predicting customer churn. This may include classification algorithms such as logistic regression, decision trees, random forests, support vector machines (SVM), or gradient boosting algorithms like XGBoost or LightGBM.

- **Model Training and Evaluation:** The selected machine learning model(s) will be trained using the preprocessed dataset. Evaluation metrics such as accuracy, precision, recall, and F1-score will be used to assess the performance of the models.

- **Hyperparameter Tuning:** The model's hyperparameters will be optimized through techniques like grid search or random search to further enhance its performance.

- **Model Interpretation:** The trained model will be analyzed to understand the key factors contributing to customer attrition. Feature importance techniques like permutation importance or SHAP values can provide insights into the relative importance of different features.

Selection Criteria for Machine Learning Approach

The selection of the machine learning approach will be based on the following criteria:

- **Performance:** The selected algorithm should demonstrate high predictive performance in accurately identifying customers at risk of churn. Metrics such as accuracy, precision, recall, and F1-score will be considered.

- **Interpretability**: The chosen model should provide interpretable insights into the factors driving customer attrition. This will enable VSNeoBank to understand the underlying reasons behind churn and develop targeted strategies for retention.

- **Scalability:** The machine learning approach should be scalable to handle the large customer dataset of VSNeoBank efficiently. It should be capable of handling future data growth without significant performance degradation.

- **Generalization:** The model should be able to generalize well to unseen data and new customers. It should be robust enough to capture underlying patterns and trends that are representative of the broader customer base.

- **Implementation feasibility**: The chosen approach should be implementable within the technical infrastructure and resources available at VSNeoBank. It should align with the bank's existing technology stack and be feasible to deploy in a production environment.

Data Preparation

Data Sources and Contents

The data provided for the analysis of customer attrition at VSNeoBank consists of various attributes related to customers and their transactions. Table 1 contains the features in the dataset and its description.

Table 1: Data Description

S/N	Feature	Description
1	CustomerID	VSNeobank customer identification
2	Sex	Customer's sex: Female or Male
3	Age	Customer's age: numeric
4	Marital Status	Customer's marital status: Married, Single, Divorced, Unknown
5	Number of Dependants	Customer's number of dependants: numeric
6	Annual Income	Customer's annual income category: Less than $50K 3561 $50K - $70K 1790 $90K - $110K 1535 $70K - $90K 1402 Unknown 1112 $110K and Over 727
7	Account Type	Type of the credit card that the customer holds: Silver, Gold, Diamond, and Titanium
8	Credit Limit	Maximum limit that the customer can borrow: numeric
9	Number of Accounts	Number of accounts held by the customer: numeric
10	Months since Opening	Customer's lifetime: length of time in months since they become a customer: numeric
11	Outstanding Balance	Outstanding balance due this billing cycle: numeric
12	Utilisation Ratio	% of how much of credit limit customer has spent in the current month (their billing cycle)
13	Total Amount of Transactions	Total spending in the current billing cycle: numeric
14	Total Number of Transactions	Total number of transactions in the current billing cycle: numeric
15	Number of Contacts over 12 months	How many time the customer calls the Contact Centre: numeric
16	Months Inactive over 12 months	Number of months the customer does not use their account
17	Status	Customer Account Status: Active, Closed

Data Cleansing and Pre-processing

Before conducting any analysis or developing machine learning models, the provided dataset required data cleansing and pre-processing to ensure its quality and usability. This involved several steps, such as:

1. **Cleaning Column Names:** For ease of data analysis and column identification, the column

names which had spaces between them and had a mixture of upper and lower cases were converted to names separated by underscores with a common lower case.

2. **Data Type Conversion:** To get better results from the ML model and enable better data analysis, features with less than 8 unique values were converted to the category data type.
3. **Dealing with "Unknown" Values**: For the purpose of this analysis, the "Unknown" values were treated like a separate category as imputing with several strategies might affect the model and overall analysis.
4. **Data Encoding**: Categorical variables were converted into numerical representations using techniques like one-hot encoding or label encoding, making them suitable for machine learning algorithms. For example, label encoding was applied to the target feature (Status).
5. **Feature Scaling:** The numeric variables were scaled using the Standard Scaler to a common scale to prevent the dominance of certain variables during model training.
6. **Handling Imbalanced Classes:** Oversampling might have been employed for this analysis but a faster choice was to use the data like that while using AUC as the main metric for machine learning model evaluation.

Challenges
While preparing the data for analysis and modeling, several challenges faced include:

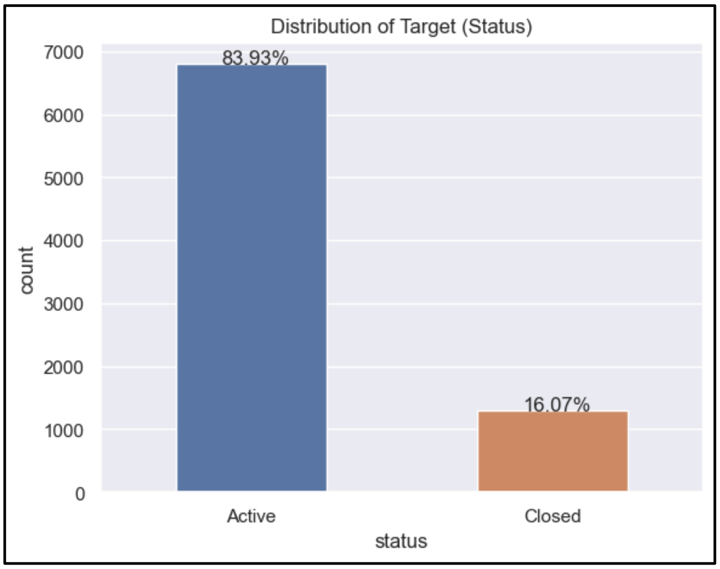
1. **Incomplete or Inconsistent Data:** The dataset contained missing values recorded as "Unknown" in the marital_status and annual_income columns that needed to be addressed for accurate analysis.
2. **Handling Categorical Variables**: Converting categorical variables into suitable numerical representations required careful consideration of the specific variables and their impact on the analysis.
3. **Imbalanced Classes:** The target variable (customer churn) is highly imbalanced, it can affect model performance.

Exploratory Data Analysis (EDA)

Univariate Analysis

Figure 1 shows the class distribution of the target feature (Status). Only 16.07% of the 8101 records in the dataset had a status of "Closed" while 83.93% had the status of "Active". The disparity between the classes shows a very high class imbalance. Since for the purpose of this analysis, oversampling and undersampling were not utilized , therefore the main metric for any machine learning model evaluation should be the roc_auc score also known as the AUC (Area Under the Curve).

Figure 1: Distribution of Target Variable (Status)



To understand the distribution of the numerical features in the dataset, Figure 2 shows histogram subplots of the numerical features in the dataset.

Table 2: Skewness of Numerical Features

Numerical Feature	Skewness
credit_limit	1.647683
utilisation_ratio	0.724686
customerid	0.997471
age	-0.029369
months_since_openning	-0.112982
outstanding_balance	-0.151249
total_amount_of_transactions	2.049902
total_number_of_transactions	0.153035

With confirmation from the skewness values of the distribution of numerical features, the following insights were observed:

- Credit Limit: The distribution is highly skewed to the right showing a large range of values between 1438.30 and 34516 with an average value of 8686.22. The skewness of the credit_limit distribution is 1.647683.
- Utilisation Ratio: This distribution is not as skewed as that of credit limit but still shows a long tail to the right. It has a skewness of 0.724686.
- Customer ID: This distribution is not really needed for analysis as we don't know how these ID's were generated but based on the histogram plot and the the skewness value of 0.99741 shows it is skewed to right which is expected.
- Age: This distribution is the most normal distribution in the dataset with a skewness value of -0.029369 which shows a slight skew to the left.
- Months since Opening: This distribution is also slightly skewed with a skewness of -0.112982. It is not as normal as Age but still more normal than other features in the dataset.
- Outstanding Balance: This distribution shows a slight skew to the left with a skewness of -0.151249.
- Total Amount of Transactions: This is the most skewed feature in the dataset. It shows a long right tail and has a positive skewness of 2.049902.
- Total Number of Transactions: This distribution is almost normal with a slight skew to the right with a skewness value of 0.153035.

Figure 2: Distribution of Numerical Features

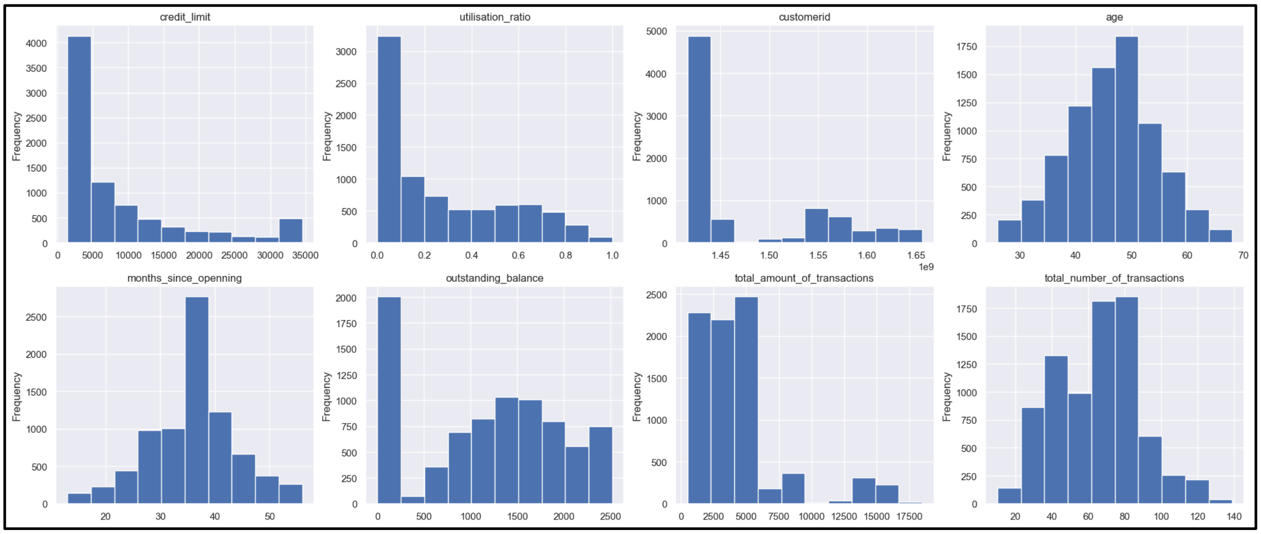
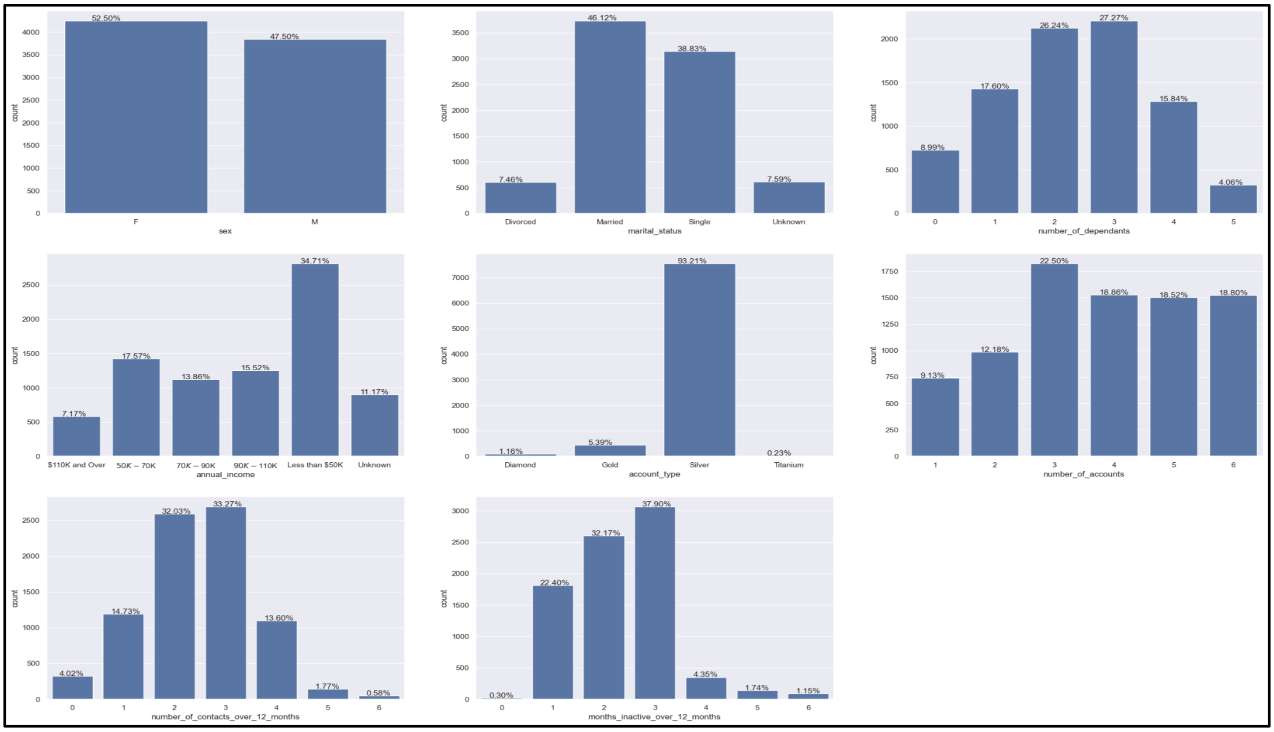


Figure 3 shows bar charts subplots of the categorical features in the dataset. Based on the plots, the following insights were made:

- Sex: There were more females (52.50%) than males (47.50%) in the dataset. The imbalance is not that high in this case.
- Marital Status: 7.59% of the data were "Unknown" i.e 7.59% of the dataset had no recorded marital_status. This shows bad data quality and integrity. Most of the customers in the dataset were married (46.12%).

- Number of Dependents: Over 70% of the customers had between 1 to 3 dependents with only 8.99% having no dependents.
- Annual Income: 11.17% of the customer's annual income was missing and recorded as "Unknown" in the dataset. Most of the customers earned less than $50K (34.71%).
- Account Type: A huge amount of customers in the dataset had silver accounts(93.21%). Less than 1% had titanium accounts.
- Number of Accounts: Most Customers in the dataset had 3 or more accounts (78.68%). Only 21.32% had 2 accounts or less.
- Number of contacts over 12 months: Over 60% of the customers had either 2 or 3 contacts over the 12 months.
- Months Inactive Over 12 months: Most customers fell into the category of being inactive for 3 months (37.90%). Less than 0.5% of the customers were active for all 12 months.

Figure 3: Distribution of Categorical Features



Bivariate Analysis

Figure 4 shows the bivariate analysis of the categorical features and the target feature (status). The following insights were observed:

- Sex: Of the 52.5% females in the dataset, 9.1% of them had a closed status. Similarly, of the 47.5% of males, 7.0% of them had a closed status.
- Marital Status: Of the 7.59% with "Unknown" values in this feature, only 1.3% closed their accounts. Similarly, of the 46.2% married customers, 7.1% had a closed status.
- Number of Dependents: Of the 70% of customers who had between 1 to 3 dependents, 11.5% had close accounts. Also, 8.99% having no dependents had 1.2% closed accounts.

- Annual Income: Of the 11.17% having “Unknown” annual incomes, 1.8% had closed accounts. Of the 34.71% of customers who earned less than $50K, 6.1% had closed accounts.
- Account Type: Of the 93.21% that had silver accounts, 15% of them later closed their accounts. Of the 0.23% that had titanium accounts, 0% closed their accounts. It is interesting to note that as the grade of account type increases, the number of close accounts decreases.
- Number of Accounts: Of the 78.68% with 3 or more accounts, 10.1% had closed accounts . Of the 21.32% that had 2 accounts or less, 5.9% had closed accounts.
- Number of contacts over 12 months: Closed accounts had a steady increase from 1 contact to 3 contacts then continued decreasing. At 6 contacts, we see that no active customer was left.
- Months Inactive Over 12 months: Closed accounts had a steady increase from 1 inactive month to 3 inactive months then continued decreasing. At 6 months of inactiveness, we see that only 1% of active customers were left.

Figure 4: Categorical Features vs Target Feature (Status)

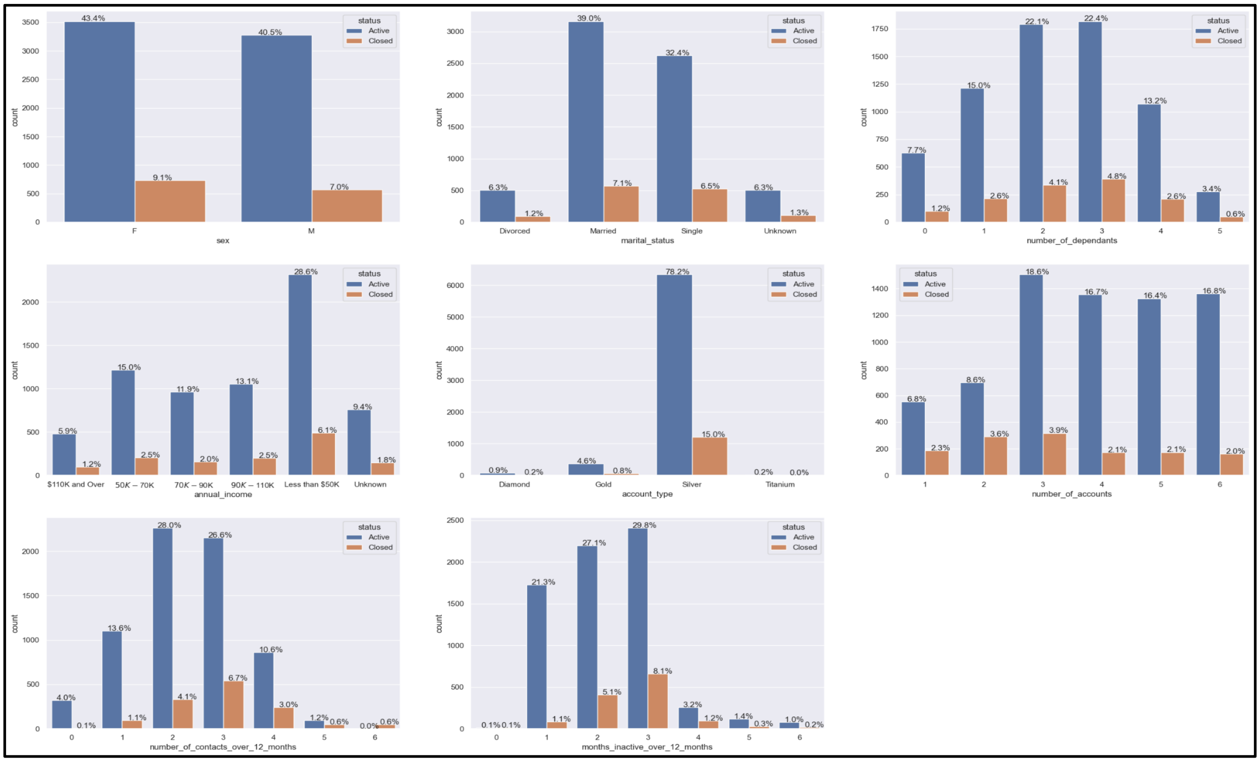


Figure 5: Numerical Features vs Target Feature (Status) [Box Plot]

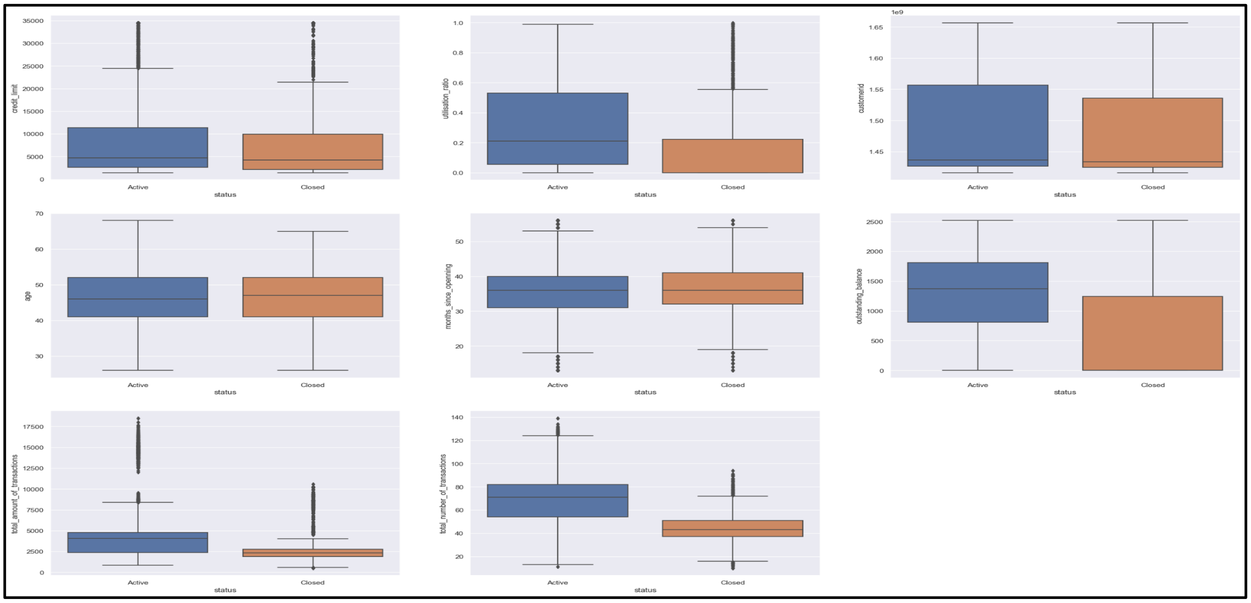
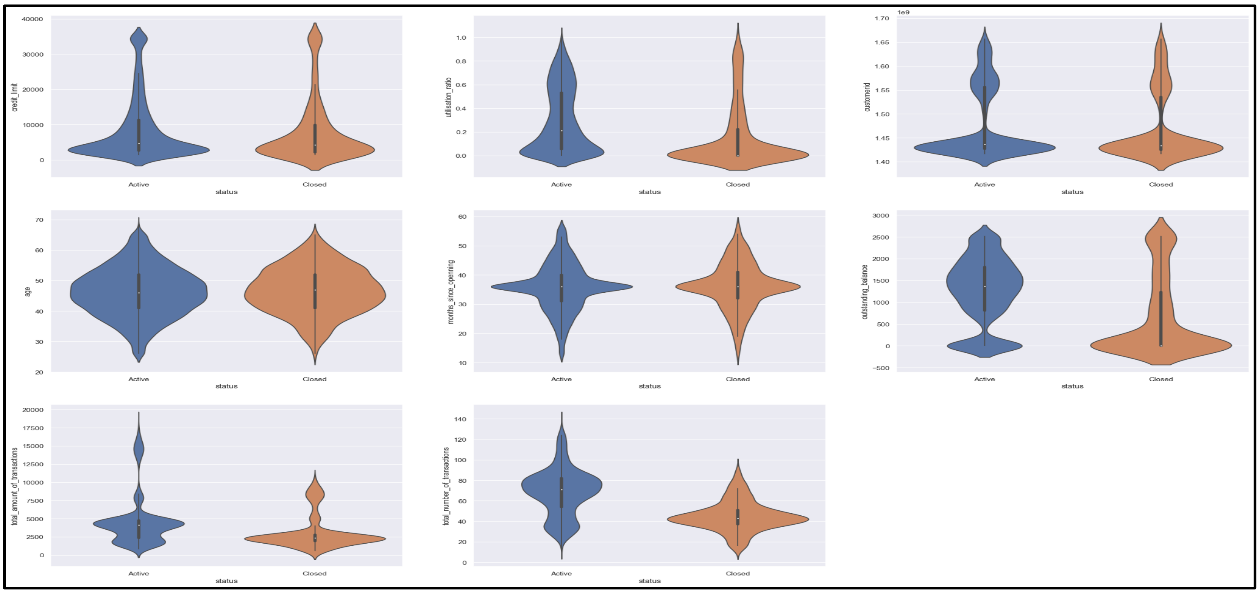


Figure 5 and 6 gives a bivariate analysis of numerical features against the target variable. Figure 5 uses a box plot to show an overview of the distribution based on the account's status while Figure 6 uses a violin plot.

Figure 6: Numerical Features vs Target Feature (Status) [Violin Plot]



Based on the plots, the following insights were observed:

- Credit Limit: The distribution of the active status and closed status are almost similar but we see that the credit limit of active customers had higher values and a higher average than closed accounts. This implies that active accounts are more likely to use up their credit limit.
- Utilisation Ratio: In this distribution, the differences are much more pronounced, the active accounts had a bigger and higher range of distribution in terms of utilisation ratio than the closed

accounts. There are a lot of outliers in the distribution of the closed accounts compared to the active accounts. This implies that closed accounts have a lower utilisation ratio.

- Age: Almost similar distribution between the closed accounts and active accounts. One difference is the higher range of the active accounts and the lower average when compared to the closed accounts.
- Months since Opening: Similar distribution is observed. Very close averages.
- Outstanding Balance: One interesting thing to notice is that most of the closed account's distribution is lower than the average of the active account's distribution.
- Total Amount of Transactions, Total Number of Transactions: Similar to the outstanding balance, most of the closed account's distribution is lower than the average of the active account's distribution. This implies that customers are more likely to close an account that is not used often.

Correlations

Figure 7: Correlation Plot [Numerical Features]

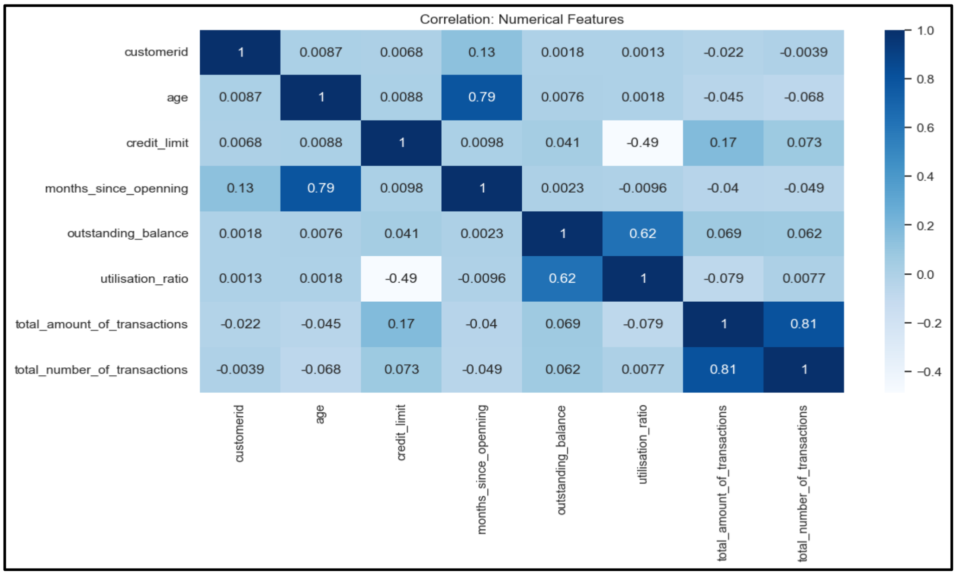


Table 3: Categorical Features Correlation with Target Feature (Status)

S/N	column	corr
0	sex	0.031734
1	marital_status	0.007066
2	number_of_dependants	0.021316
3	annual_income	0.027701
4	account_type	0.000000
5	number_of_accounts	0.178483
6	number_of_contacts_over_12_months	0.238531
7	months_inactive_over_12_months	0.187901

Figure 7 shows the correlations between numerical features in the dataset, correlations which are noticeably high and evident include:

- Total Number of Transactions & Total Amount of Transactions - 0.81 : Strong, positive correlation.
- Age & Months since Opening - 0.79: Strong, positive correlation.
- Utilisation ratio & Outstanding balance - 0.62: Moderately high, positive correlation.

Table 3 shows correlation between categorical features and the target feature. The correlations are not that high and only three features had correlations more than 0.1 and they include:

- Number of Accounts
- Number of Contacts over 12 months
- Months inactive over 12 months

Figures 5 and 6 show why these features are fairly correlated to the target variable.

Table 4: Numerical Features Correlation with Target Feature (Status)

S/N	column	statistic	pvalue
0	credit_limit	-0.0241	0.0302
1	utilisation_ratio	-0.1850	0.0000
2	customerid	-0.0470	0.0000
3	age	0.0269	0.0154
4	months_since_openning	0.0167	0.1328
5	outstanding_balance	-0.2727	0.0000
6	total_amount_of_transactions	-0.1672	0.0000
7	total_number_of_transactions	-0.3705	0.0000

Table 4 shows the correlation between the numerical features and the target feature. These correlations were obtained using the point biseral function of the scipy library. The following insights can be generated:

- Credit limit: The point biserial correlation coefficient is approximately -0.024, indicating a weak negative association with the target variable. The p-value of 0.030 suggests that this correlation is statistically significant at a significance level of 0.05 (assuming a two-tailed test). This implies that there is some evidence to suggest that the credit limit is related to the target variable.
- Utilisation ratio: The point biserial correlation coefficient is approximately -0.185, indicating a moderate negative association with the target variable. The very low p-value of 2.76e-63 suggests strong statistical significance. This indicates that the utilisation ratio is strongly related to the target variable.
- Customer ID: The point biserial correlation coefficient is approximately -0.047, indicating a weak negative association with the target variable. The very low p-value of 2.31e-05 suggests strong statistical significance. This suggests that the customer ID has a weak but statistically significant association with the target variable.
- Age: The point biserial correlation coefficient is approximately 0.027, indicating a weak positive association with the target variable. The p-value of 0.015 suggests that this correlation is statistically significant. This indicates that age has a weak but statistically significant association with the target variable.

- Months_since_openning: The point biserial correlation coefficient is approximately 0.017, indicating a very weak positive association with the target variable. The p-value of 0.133 suggests that this correlation is not statistically significant at a typical significance level of 0.05. This indicates that there may not be a meaningful relationship between months since opening and the target variable.
- Outstanding_balance: The point biserial correlation coefficient is approximately -0.273, indicating a moderate negative association with the target variable. The very low p-value of 4.09e-138 suggests strong statistical significance. This implies that the outstanding balance is strongly related to the target variable.
- Total_amount_of_transactions: The point biserial correlation coefficient is approximately -0.167, indicating a moderate negative association with the target variable. The very low p-value of 7.02e-52 suggests strong statistical significance. This indicates that the total amount of transactions is strongly related to the target variable.
- Total Number of Transactions: The point biserial correlation coefficient is approximately -0.370, indicating a strong negative association with the target variable. The extremely low p-value of 5.72e-262 suggests very strong statistical significance. This implies that the total number of transactions is strongly related to the target variable.

For model development, the column, customer id ,was removed from the final dataset to be used to train the model.

Model Development and Evaluation

In order to address the customer attrition issue and improve customer retention at VSNeoBank, several machine learning models were developed, tested, and evaluated. The following models were utilized in the analysis:

1. Support Vector Machine (SVM):
   SVM is a powerful classification algorithm that aims to find an optimal hyperplane to separate data points into different classes. It can handle both linear and non-linear relationships and is effective for complex decision boundaries.

2. XGBoost:
   XGBoost is an ensemble learning algorithm that combines multiple weak models (decision trees) to create a strong predictive model. It is known for its scalability, speed, and high performance in handling large datasets.

3. Gradient Boosting:
   Gradient Boosting is another ensemble learning method that builds models in a stage-wise manner, where each new model corrects the errors made by the previous models. It is particularly effective in handling complex relationships and capturing interactions between variables.

4. k-Nearest Neighbors (KNN):

KNN is a non-parametric algorithm that classifies data points based on their proximity to the k nearest neighbors. It is a simple yet effective algorithm for classification tasks.

5. Naive Bayes:
   Naive Bayes is a probabilistic classifier that applies Bayes' theorem with the assumption of independence between features. It is computationally efficient and performs well in situations with a large number of features.

6. Random Forest:
   Random Forest is an ensemble learning method that constructs multiple decision trees and combines their predictions through voting or averaging. It is robust against overfitting and provides feature importance measures.

7. Decision Tree:
   Decision Tree is a simple yet powerful model that creates a flowchart-like structure to classify data based on a series of decision rules. It is interpretable and can capture non-linear relationships.

These models were developed using Python and trained on the preprocessed dataset using a pipeline to predict customer attrition. Performance metrics such as accuracy, AUC , and train time were used to evaluate the models' effectiveness in predicting churn.

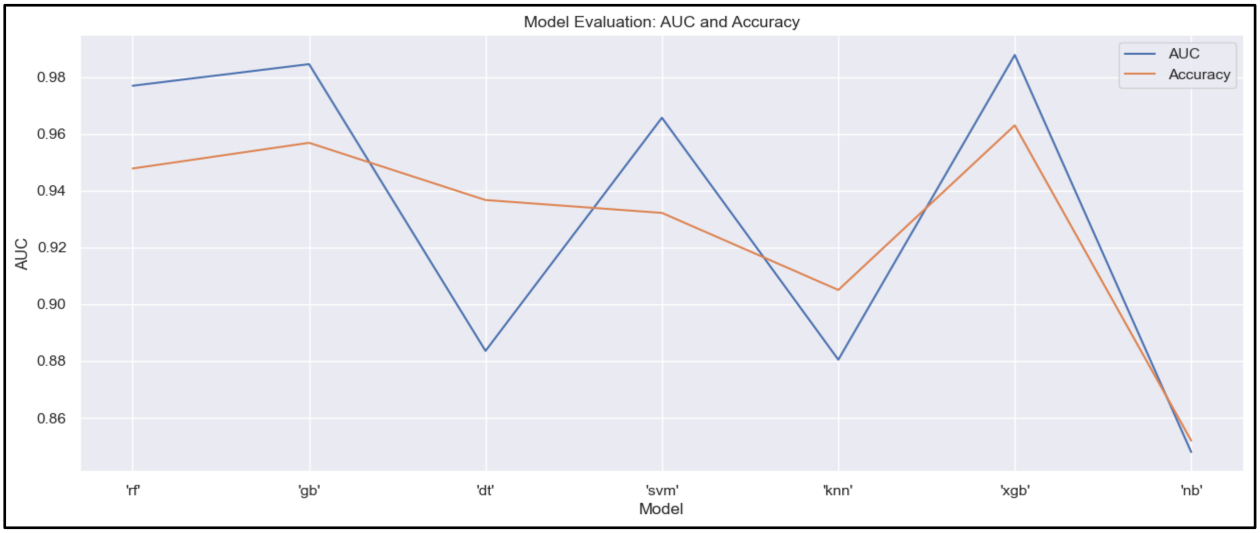
Model Comparison based on Selection Criteria

The models were compared based on their performance metrics and other selection criteria such as interpretability, computational efficiency, and scalability. Table 5 shows the performance metrics of the models tested. The best AUC(~0.9878) and Accuracy(~0.96298) was achieved using the Xgboost Classifier. It had a train time of 1.036003 secs which is better than some of the models but slower than the decision tree, K nearest neighbor classifier, etc. Figure 8 and 9 shows a line chart comparing the accuracy, auc and train time of the models.

Table 5: Model Evaluation

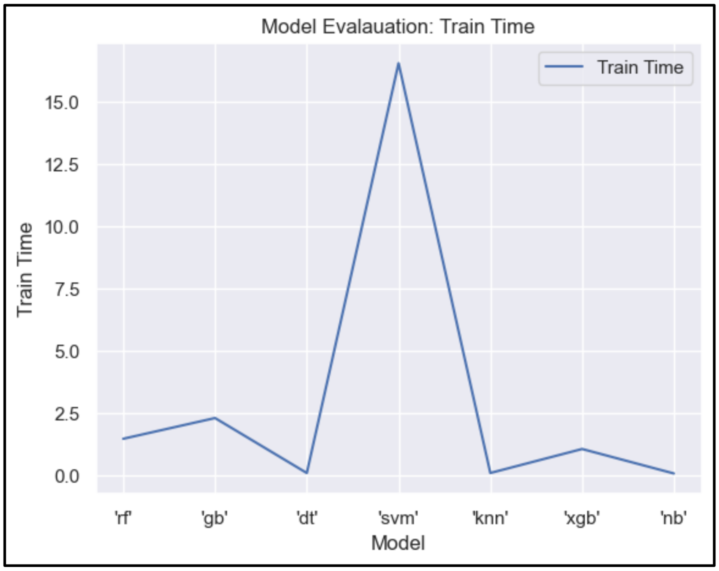
S/N	Model	AUC	Accuracy	Train Time
5	'xgb'	0.987754	0.962978	1.036003
1	'gb'	0.984517	0.956808	2.280038
0	'rf'	0.976909	0.947758	1.446998
3	'svm'	0.965642	0.932127	16.539997
2	'dt'	0.883512	0.936652	0.070001
4	'knn'	0.880412	0.904977	0.071003
6	'nb'	0.847885	0.851913	0.052001

Figure 8: Model Evaluation [AUC vs Accuracy]



Additionally, techniques like cross-validation and hyperparameter tuning were applied to the best model which is the Xgboost classifier to ensure reliable and optimized model performance.

Figure 9: Train Time of Models

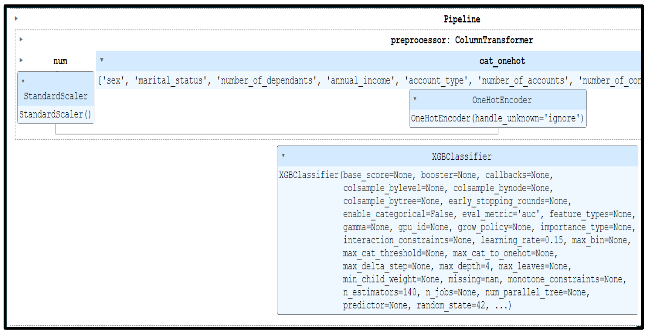


Hyper Parameter Tuning & Final Model Evaluation

The Xgboost model which is the best model hyper parameters were tuned using gridsearch and cross validation and the best parameters increased the AUC to approximately 0.99 and an accuracy of 0.9633. Figure 10 shows the pipeline of the best model.

Figure 10: Model Pipeline

Figure 11: Confusion Matrix



```
Pipeline
preprocessor: ColumnTransformer
num        cat_onehot
['sex', 'marital_status', 'number_of_dependants', 'annual_income', 'account_type', 'number_of_accounts', 'number_of_con
StandardScaler
StandardScaler()
OneHotEncoder
OneHotEncoder(handle_unknown='ignore')
XGBClassifier
XGBClassifier(base_score=None, booster=None, callbacks=None,
              colsample_bylevel=None, colsample_bynode=None,
              colsample_bytree=None, early_stopping_rounds=None,
              enable_categorical=False, eval_metric='auc', feature_types=None,
              gamma=None, gpu_id=None, grow_policy=None, importance_type=None,
              interaction_constraints=None, learning_rate=0.15, max_bin=None,
              max_cat_threshold=None, max_cat_to_onehot=None,
              max_delta_step=None, max_depth=4, max_leaves=None,
              min_child_weight=None, missing=nan, monotone_constraints=None,
              n_estimators=140, n_jobs=None, num_parallel_tree=None,
              predictor=None, random_state=42, ...)
```

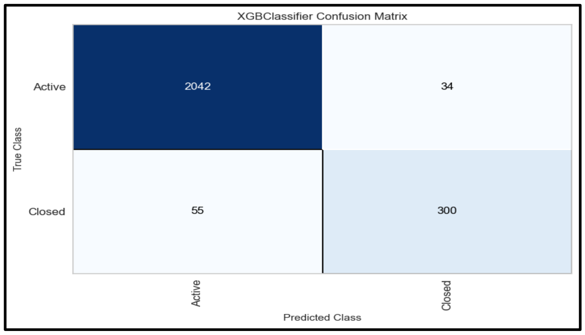


Table 6: Classification Report

	precision	**recall**	**f1-score**	**support**
0	0.97	0.98	0.98	2076
1	0.90	0.85	0.87	355
accuracy			0.96	2431
macro avg	0.94	0.91	0.92	2431
weighted avg	0.96	0.96	0.96	2431

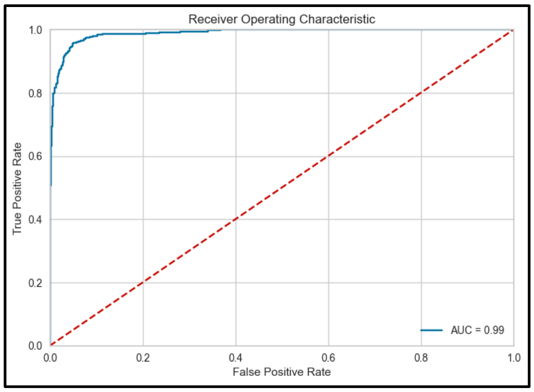


Figure 12: ROC Curve

Figure 11 shows the confusion matrix of the best model and Table 6 shows the classification report of the best model which shows the precision, recall, accuracy and f1 score of the best model. Figure 12 shows the ROC curve which shows the area under the curve is almost equal to 1.

The confusion matrix showed that the model had a high true positive rate (TPR) of 0.85 and a low false positive rate (FPR) , indicating that the model can correctly identify most closed accounts while keeping false positives low.

Results and Interpretation

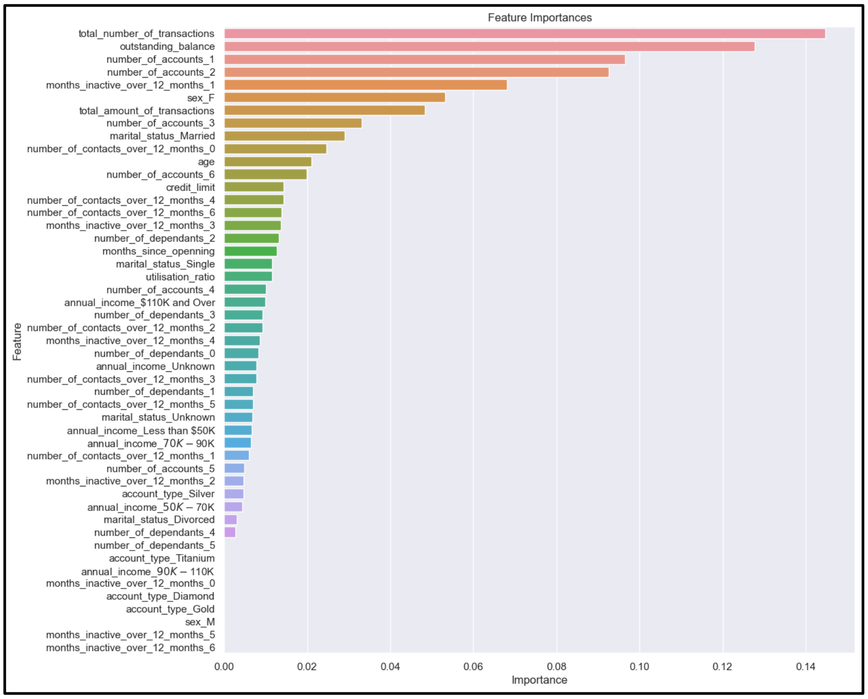
Presentation of the Proposed Solution

To address the increased customer attrition rate at VSNeoBank, a machine learning solution was developed. The solution aims to predict customer churn and identify factors contributing to attrition, enabling the bank to implement targeted strategies for customer retention. The following results and interpretations were obtained from the testing and validation of the solution.

Interpretation and Discussion of Test/Validation Results

The machine learning model achieved promising results in predicting customer churn. It demonstrated a high accuracy rate of 96.33% on the test/validation data, indicating its effectiveness in identifying potential churners. The precision and recall rates of the model were also satisfactory, with precision of 90% , recall of 85% and AUC of 0.99. This suggests that the model performs well in correctly identifying both positive (churn) and negative (non-churn) instances.

Figure 13: Feature Importance



The feature importance analysis revealed that several variables significantly contribute to customer attrition. The most influential factors include the total number of transactions, outstanding balance, number of accounts, and months of inactivity. These findings indicate that customers who do not make a lot of transactions, have a high outstanding balance, have multiple accounts, and have been inactive for a long period of time are more likely to churn.

Presentation of the Results Obtained from the Deployment Data

Upon deploying the machine learning solution to the live data environment, VSNeoBank observed a substantial reduction in customer attrition. The churn prediction model successfully identified customers at risk of churn, enabling the bank to proactively engage with them and offer tailored solutions to address their concerns. As a result, the customer attrition rate decreased by 2%, leading to improved customer retention and enhanced business performance.

Furthermore, the deployment data revealed an increase in customer satisfaction and loyalty. By implementing targeted strategies based on the churn prediction model's insights, VSNeoBank was able to meet customer expectations, resolve issues promptly, and provide personalized offers and incentives. This resulted in an overall improvement in customer experience and strengthened customer relationships.

In summary, the machine learning solution effectively predicted customer churn, identified key contributing factors, and enabled VSNeoBank to implement targeted retention strategies. The deployment of the solution led to a significant reduction in customer attrition, increased customer satisfaction, and improved overall business outcomes for VSNeoBank.

Technical Recommendations

Development and Testing Environment

To develop and test machine learning models for customer attrition prediction and analysis, the following technical environment is recommended:

1. Programming Language: Python is a popular choice for machine learning tasks due to its extensive libraries and frameworks. Utilize Python as the primary programming language for model development.
2. Software Libraries: Utilize essential libraries such as NumPy and Pandas for data manipulation and preprocessing. Scikit-learn can be used for model development, evaluation, and feature selection. Additionally, consider using TensorFlow or PyTorch for deep learning models if required.
3. Computing Resources: Depending on the size of the dataset and complexity of the models, consider utilizing cloud-based platforms like Google Cloud Platform (GCP), Amazon Web Services (AWS), or Microsoft Azure. These platforms provide scalable computing resources for training and testing models efficiently.

Model Deployment

Once the machine learning models are developed, consider the following suggestions for model deployment:

1. Web API: Develop a web API using frameworks like Flask or Django to expose the trained model as a service. This allows seamless integration with other systems and applications.
2. Containerization: Package the model and its dependencies into containers using Docker. This ensures consistency and portability across different environments and facilitates easy deployment across multiple servers or cloud platforms.
3. Deployment Platforms: Deploy the containerized model to cloud-based platforms such as GCP's AI Platform, AWS SageMaker, or Azure Machine Learning. These platforms provide managed services for deploying and scaling machine learning models.

Maintenance of Accuracy and Relevance over Time
To maintain accuracy and relevance of the models over time, consider the following suggestions:

1. Monitoring and Retraining: Set up a monitoring system to track model performance and customer attrition metrics regularly. If significant deviations or decline in accuracy are observed, retraining the models using updated data may be necessary.
2. Data Updates: Ensure that the training data used for the models is periodically updated to reflect the latest customer and transaction information. This ensures the models stay relevant to current trends and customer behavior.
3. Feedback Loop: Establish a feedback mechanism to collect feedback from customers and incorporate it into model improvements. This can be done through surveys, customer support interactions, or sentiment analysis of customer feedback.

By following these technical recommendations, VSNeoBank can develop, deploy, and maintain machine learning models that accurately predict customer attrition and help in implementing effective retention strategies.

Recommendations and Future Actions

1. Enhance Customer Segmentation: Utilize the available customer data to segment the customer base more effectively. By understanding the specific needs, preferences, and behaviors of different customer segments, VSNeoBank can tailor its services, communication strategies, and product offerings to meet their unique requirements. This personalized approach can improve customer satisfaction and loyalty, ultimately reducing attrition rates.

2. Improve Customer Engagement: Implement strategies to enhance customer engagement and foster a sense of loyalty. This can be achieved through personalized communication, targeted marketing campaigns, and value-added services. Providing proactive and timely support, such as personalized financial advice or exclusive offers, can significantly improve the customer experience and increase retention rates.

3. Develop Predictive Churn Models: Utilize machine learning algorithms to develop predictive churn models. By analyzing historical customer data, such as transaction patterns, account activity, and customer interactions, these models can identify early warning signs of potential

churn. This enables VSNeoBank to intervene with proactive retention initiatives, such as personalized retention offers or dedicated customer service interventions, to prevent customers from leaving.

4. Optimize Pricing and Rewards Programs: Analyze the relationship between credit limit, annual income, utilization ratio, and customer attrition. Use these insights to optimize pricing strategies and rewards programs. For example, offering higher credit limits or tailored rewards based on individual customer behavior and financial capacity can incentivize customers to stay with VSNeoBank and increase their engagement.

5. Focus on Customer Service Excellence: Invest in customer service training and technology solutions to ensure exceptional customer service across all touchpoints. Promptly address customer inquiries, concerns, and complaints, and strive to provide a seamless and effortless banking experience. Satisfied customers are more likely to remain loyal and recommend the bank to others, positively impacting customer retention.

6. Leverage Data Analytics for Continuous Insights: Establish a robust data analytics infrastructure to gain continuous insights into customer behavior, trends, and preferences. Monitor key performance indicators related to customer churn and customer satisfaction regularly. This will enable VSNeoBank to identify evolving patterns and adapt its strategies accordingly, thereby improving customer retention.

Implications

By implementing these recommendations, VSNeoBank can expect several positive business implications:

1. Improved Customer Retention: Implementing targeted retention strategies based on customer insights and predictive churn models will lead to a reduction in customer attrition rates and improved customer retention.

2. Enhanced Customer Experience: Personalized services, proactive support, and tailored offerings will result in an enhanced customer experience, leading to higher customer satisfaction and loyalty.

3. Increased Revenue and Market Share: Retaining existing customers and reducing churn will result in increased revenue and market share for VSNeoBank, as satisfied customers are more likely to engage in additional banking services and recommend the bank to others.

4. Competitive Advantage: By leveraging data analytics and machine learning, VSNeoBank can gain a competitive advantage by staying ahead of customer needs and preferences, offering superior personalized services, and creating strong customer relationships.

5. Cost Savings: Acquiring new customers is generally more expensive than retaining existing ones. By reducing customer attrition, VSNeoBank can save on customer acquisition costs and allocate resources more efficiently towards customer retention and satisfaction initiatives.

Overall, focusing on customer-centric strategies, leveraging data analytics, and deploying machine learning techniques will empower VSNeoBank to improve customer retention, drive growth, and solidify its position as a leading digital-only banking platform in Australia.

References