**Results**:  
  
In the competitive financial sector, retaining customers is critical for banks' profitability and market position. This study aims to develop an effective model to predict customer churn using machine learning techniques, balancing precision, and recall. By identifying key churn factors, banks can implement targeted holding strategies to enhance customer satisfaction and loyalty.

The main objective is to build a predictive model that accurately identifies at-risk customers. Various machine learning algorithms, including Random Forest, Logistic Regression, and XGBoost, are explored to determine the best performance for imbalanced data. Techniques like SMOTE and cost-sensitive learning are used to address class imbalance.

Feature importance analysis is crucial for understanding the factors influencing churn, helping banks focus holding efforts. Key features such as age, tenure, balance, and the number of products are examined for their impact on churn decisions. Additionally, the study investigates geographic and demographic influences on churn rates and the role of credit scores in enhancing predictive power.

**Data Collection and Preparation:**The dataset used in this study was sourced from Kaggle, containing 10,000 customer records and 14 attributes, including customer information such as credit score, geography, gender, age, tenure, balance, number of products, and whether the customer exited the bank. The initial steps involved loading the data, checking for null values, and performing basic exploratory data analysis (EDA).

A screenshot of a computer

Description automatically generated

**Step 1: Importing Libraries and Loading Data**

**Importing Libraries**

The first step in any data science project is to import the necessary libraries that provide the tools and functionalities required for data manipulation, model building, and evaluation. In this project, we use a variety of libraries, each serving specific purposes:

* Pandas: Used for data manipulation and analysis. It provides data structures like DataFrame, which are essential for handling tabular data.
* NumPy: A fundamental package for numerical computations in Python, providing support for arrays and matrices.
* Scikit-learn: This comprehensive library is utilized for preprocessing data, model selection, building machine learning models, and evaluating them.
* XGBoost: An efficient and scalable implementation of gradient boosting framework, widely used for regression and classification problems.
* Matplotlib and Seaborn: These libraries are used for data visualization, allowing us to create various plots to understand data distributions and model performance.
* Statsmodels: Used for statistical modeling and hypothesis testing.

**Loading the Dataset**

The dataset is loaded using the pd.read\_csv function from Pandas. This function reads the data from a CSV file and stores it in a DataFrame, which is a two-dimensional labeled data structure with columns of potentially different types.

After loading the data, the data.head() method is called to display the first five rows of the dataset. This preview helps in understanding the structure of the data, including the types of features (columns) and a quick overview of the data values.

The dataset includes the following columns:

* RowNumber: The row number of the data.
* CustomerId: Unique identifier for each customer.
* Surname: Customer's surname.
* CreditScore: Customer's credit score.
* Geography: The country where the customer resides.
* Gender: Gender of the customer.
* Age: Age of the customer.
* Tenure: Number of years the customer has been with the bank.
* Balance: The balance amount in the customer's account.
* NumOfProducts: Number of products the customer has with the bank.
* HasCrCard: Whether the customer has a credit card.
* IsActiveMember: Whether the customer is an active member.
* EstimatedSalary: Estimated salary of the customer.
* Exited: Whether the customer has exited the bank (churned).

Data Structure and Initial Insights

By examining the first few rows, we can gather some initial insights:

* The dataset includes both numerical and categorical features.
* The target variable (label) is Exited, indicating customer churn.
* Features such as Geography and Gender are categorical and will require encoding before model training.
* Numerical features like CreditScore, Age, Tenure, Balance, NumOfProducts, and EstimatedSalary will likely need scaling.

Understanding the data structure and its contents is critical for planning the following steps in the data preparation and modeling process. It allows us to identify potential issues such as missing values, outliers, or imbalanced classes that need to be addressed during data preprocessing. Additionally, it helps in selecting appropriate feature engineering and transformation techniques, ultimately improving model performance.

A screenshot of a computer

Description automatically generated

**Step 2: Initial Data Exploration**

In this step, we check the dataset's structure and look for any missing values. The data.info() function provides an overview of the data types and non-null counts, while data.isnull().sum() helps identify columns with missing values. Fortunately, the dataset does not contain any missing values, ensuring that all columns are complete and ready for processing.

A screenshot of a computer

Description automatically generated

**Step 3: Data Cleaning**

In the data cleaning step, unnecessary columns are removed to simplify the dataset and focus on the relevant features that can predict customer churn. The columns RowNumber, CustomerId, and Surname are dropped from the dataset. These columns serve as unique identifiers but do not provide any predictive power for the churn prediction model. The cleaned dataset now includes only the columns that contain meaningful information about the customers' demographics and account details.

By dropping these columns, the dataset becomes more manageable, and the later analysis and modeling steps are not influenced by irrelevant data. This step is crucial for improving the efficiency and accuracy of the machine learning models.

**Step 4: Defining Features and Target Variable**

Next, the features (X) and the target variable (y) are separated. The target variable Exited indicates whether a customer has churned (1) or not (0). Separating the features from the target variable is essential for supervised learning, where the model learns the relationship between input features and the target variable.

This step sets up the data for model training, where X contains all the features that will be used to predict customer churn, and y contains the labels indicating whether each customer has churned.

**Step 5: Train-Test Split**

To evaluate the model's performance, the dataset is split into training and test sets. The training set (80% of the data) is used to train the model, while the test set (20% of the data) is used to evaluate its performance. Stratified sampling is used to ensure that the class distribution in the target variable is preserved in both the training and test sets.

Stratified sampling is important in this context because the dataset is imbalanced, with fewer customers having churned compared to those who have not. Preserving the class distribution ensures that the model's performance metrics are reliable.

**Step 6: Preprocessing**

Preprocessing involves transforming the data to make it suitable for machine learning algorithms. Different preprocessing steps are applied to numerical and categorical features. Numerical features are scaled using StandardScaler, which standardizes the data to have a mean of 0 and a standard deviation of 1. This scaling helps in improving the performance of many machine learning algorithms.

Categorical features are encoded using OneHotEncoder, which converts categorical variables into a format that can be provided to machine learning algorithms to do a better job in prediction. ColumnTransformer is used to apply these transformations to the respective columns.

By applying these transformations, we ensure that the numerical features are scaled, and the categorical features are encoded consistently in both the training and test sets.

**Step 7: Handling Class Imbalance with SMOTE**

The dataset is imbalanced, with fewer customers who have churned compared to those who have not. To address this imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) is applied to the training data. SMOTE generates synthetic samples for the minority class (customers who have churned) to balance the class distribution in the training data.

Balancing the class distribution is crucial for training a robust machine learning model. Without addressing the imbalance, the model might be biased towards predicting the majority class, resulting in poor performance on the minority class. By using SMOTE, we ensure that the model has enough examples of both classes to learn from, improving its ability to predict customer churn accurately.

A screenshot of a computer program

Description automatically generated

**Step 8: Model Building and Evaluation**

In this step, we focus on building, optimizing, and evaluating the Random Forest classifier, which is one of the powerful ensembles learning methods. Random Forests are particularly effective for classification tasks because they combine the predictions of multiple decision trees to produce more accurate and robust results.

Pipeline Setup

First, a machine learning pipeline is created. This pipeline includes the Random Forest classifier. The pipeline facilitates reorganized preprocessing and model training steps, ensuring that the same preprocessing steps are consistently applied to all data.

Hyperparameter Optimization with GridSearchCV

GridSearchCV is used to find the optimal hyperparameters for the Random Forest model. The hyperparameters include the number of trees in the forest (n\_estimators), the maximum depth of each tree (max\_depth), and the minimum number of samples required to split an internal node (min\_samples\_split). These parameters are critical as they influence the complexity and performance of the model.

Best Model Selection

After the grid search, the best model is selected based on the highest ROC AUC score. The ROC AUC score is a comprehensive metric that evaluates the model's ability to distinguish between the classes (churned and non-churned customers).

Model Evaluation

The model is then evaluated on the test set. The predictions made by the best Random Forest model on the test data are compared to the actual labels to assess the model's performance.

Classification Report and ROC AUC Score

A classification report is generated to provide detailed performance metrics, including precision, recall, and f1-score for both classes. The overall accuracy of the model is also reported. The ROC AUC score is calculated to measure the area under the ROC curve, providing an total measure of the model's performance across all classification thresholds.

Results:

Accuracy: 85%

ROC AUC Score: 0.8494

The classification report for the Random Forest model shows that it performs well across various metrics. For the non-churned class (label 0), the model has a high precision (0.90) and recall (0.91), resulting in a high f1-score (0.90). For the churned class (label 1), the precision is lower (0.63) and recall is also lower (0.59), leading to a moderate f1-score (0.61). The weighted average metrics are strong, indicating that the model performs well overall.

The ROC AUC score of 0.8494 is particularly significant as it suggests that the model has a strong ability to differentiate between churned and non-churned customers. A ROC AUC score closer to 1 indicates better performance, and a score of 0.8494 is considered excellent.

A screen shot of a computer

Description automatically generated

Feature Importance of Random Forest

In this analysis, the feature importance of the Random Forest model was plotted to understand which features are most influential in predicting customer churn. Feature importance measures how valuable each feature is in the construction of the decision trees within the Random Forest model.

The plot reveals the following insights:

* Age: Age is identified as the most critical predictor of customer churn. This suggests that the likelihood of a customer leaving the bank varies significantly with age, making it a crucial factor for the bank to consider in its holding strategies.
* Number of Products: The number of products a customer holds is the second most important feature. Customers with more products are generally more engaged and less likely to churn, indicating that cross-selling and up-selling strategies could be effective in reducing churn rates.
* Balance: The balance in a customer's account also plays a significant role in predicting churn. Higher balances might be indicative of more satisfied and financially stable customers, who are less likely to leave.
* Estimated Salary: The estimated salary of a customer is another significant predictor. Customers with higher salaries may have different banking needs and expectations, affecting their likelihood to stay with or leave the bank.
* Credit Score: A customer's credit score, though not the top predictor, still holds considerable importance. It suggests that customers with lower credit scores might be at a higher risk of churn.
* Tenure: The length of time a customer has been with the bank (tenure) is also important, though to a lesser extent. Longer tenure generally indicates a stronger relationship with the bank, reducing the likelihood of churn.
* Geography: Geographic location, particularly being from Germany, also influences churn. This highlights potential regional differences in customer behavior and satisfaction.
* Activity Status: Whether a customer is an active member is another factor, indicating that actively engaged customers are less likely to leave.

A screen shot of a graph

Description automatically generated

The ROC (Receiver Operating Characteristic) curve visually represents the performance of a binary classifier as its threshold varies, plotting the true positive rate (sensitivity) against the false positive rate (1-specificity). The ROC curve for the Random Forest classifier shows a strong performance with an area under the curve (AUC) of 0.85, indicating good accuracy in distinguishing between customers who churn and those who do not.

The true positive rate (sensitivity) measures the proportion of actual positives correctly identified, while the false positive rate (1-specificity) measures the proportion of actual negatives incorrectly identified as positives. A higher true positive rate and a lower false positive rate indicate better model performance.

The ROC curve approaching the top-left corner signifies the Random Forest classifier's high sensitivity and low false positive rate. The diagonal line represents a random classifier with an AUC of 0.5. The Random Forest's AUC of 0.85 indicates a strong ability to differentiate between churned and non-churned customers, making it a reliable tool for predicting customer churn in the banking dataset.A screenshot of a computer program

Description automatically generated

Logistic Regression

In this step, the Logistic Regression model is developed and optimized. The logistic regression classifier is a linear model used for binary classification tasks. It estimates the probability that an instance belongs to a particular class by fitting a logistic function to the data.

Model Training and Hyperparameter Tuning

The model is trained using a pipeline that includes preprocessing steps and the classifier itself. GridSearchCV is utilized to perform hyperparameter tuning, which involves searching over a specified parameter grid to find the best combination of hyperparameters that maximize the model's performance.

The parameter grid for Logistic Regression includes:

C: This is the inverse of regularization strength. Smaller values specify stronger regularization.

The GridSearchCV performs cross-validation (cv=5) to evaluate the model's performance across different folds of the training data. The scoring metric used for evaluation is the ROC AUC score, which balances between true positive rate and false positive rate across various threshold settings.

Model Evaluation

After training, the best model parameters are selected, and the model is evaluated on the test set. The classification report generated provides detailed metrics, including precision, recall, f1-score, and support for both classes (0 for non-churned customers and 1 for churned customers).

The Logistic Regression model achieved:

Accuracy: 72%

ROC AUC Score: 0.7755

The classification report reveals that the model has high precision (0.90) for predicting non-churned customers but lower precision (0.39) for predicting churned customers. This discrepancy indicates that while the model is good at identifying non-churners, it struggles with accurately identifying churners.

A screenshot of a graph

Description automatically generated

Feature Coefficients for Logistic Regression: The visualization of the Logistic Regression model's coefficients provides valuable insights into how each feature influences the prediction of customer churn. This analysis helps in interpreting the model and understanding the relationship between different customer attributes and their likelihood of churning.

Positive Coefficients: Features with positive coefficients increase the likelihood of churn. In the graph, these features are:

* Age: Older customers are more likely to churn.
* Geography\_Germany: Customers from Germany are more likely to churn compared to those from other regions.
* Gender\_Female: Female customers have a higher likelihood of churning compared to male customers.
* Balance: Customers with higher account balances are more likely to churn.
* EstimatedSalary: Higher estimated salary is associated with a higher likelihood of churn.

Negative Coefficients: Features with negative coefficients decrease the likelihood of churn. In the graph, these features are:

* IsActiveMember: Active members are less likely to churn.
* Geography\_Spain and Geography\_France: Customers from Spain and France are less likely to churn compared to those from other regions.
* Gender\_Male: Male customers are less likely to churn compared to female customers.
* NumOfProducts: Having more products decreases the likelihood of churn.
* CreditScore: Higher credit scores are associated with a lower likelihood of churn.
* Tenure: Longer tenure slightly decreases the likelihood of churn.
* HasCrCard: Having a credit card has a very minimal effect on the likelihood of churn.

A screen shot of a graph

Description automatically generated

Logistic Regression ROC Curve

The ROC curve for the Logistic Regression model illustrates the trade-off between the true positive rate (sensitivity) and the false positive rate (1-specificity). With an area under the curve (AUC) of 0.78, the Logistic Regression model demonstrates decent performance in distinguishing between customers who churn and those who do not. However, this performance is notably lower than that of the Random Forest model, which achieved an AUC of 0.85. This indicates that while Logistic Regression is interpretable, it may not capture the complexity of the data as effectively as the Random Forest model.

A screenshot of a computer program

Description automatically generated

XGBoost

The XGBoost model implementation and optimization involve several key steps, as illustrated in the provided code. Below is an explanation of each segment of the code, focusing on the XGBoost classifier.

An XGBoost classifier is integrated into a pipeline, which ensures that the model's fitting and prediction processes are streamlined. The random\_state ensures reproducibility, use\_label\_encoder=False avoids deprecated warnings, and eval\_metric='logloss' specifies the evaluation metric during training.

A parameter grid is defined for hyperparameter tuning. It includes the number of trees (n\_estimators), the maximum depth of the trees (max\_depth), and the learning rate (learning\_rate). These parameters significantly influence the model's performance.

The GridSearchCV object is created to perform an exhaustive search over the specified parameter grid. The cross-validation (cv=5) ensures that the model is validated on different subsets of the data. The scoring metric used is roc\_auc, which evaluates the model's ability to discriminate between the classes.

After fitting the grid search, the best model (best\_xgb) is retrieved. This model is then used to make predictions on the pre-processed test data.

The classification report provides detailed performance metrics, including precision, recall, and f1-score for each class. The ROC AUC score is also calculated, showing the model's performance in distinguishing between churned and non-churned customers.

The output illustrates the model's effectiveness, with an accuracy of 84% and a ROC AUC score of 0.8366, indicating strong predictive performance. This thorough approach to model building and evaluation ensures that the XGBoost classifier is well-tuned and capable of providing valuable insights for customer churn prediction.

A screenshot of a computer

Description automatically generated

Feature Importances for XGBoost

In this step, the feature importances of the XGBoost model are visualized to understand which features contribute the most to predicting customer churn. The code used for this visualization and its explanation are as follows:

First, a list of feature names is created by combining numerical and categorical features. The feature importances are extracted from the trained XGBoost model, and the features are sorted based on their importance.

A horizontal bar plot is generated to visualize the feature importances. The y-axis represents the feature names, and the x-axis shows their relative importance.

Findings

* Number of Products: The number of products a customer holds is the most significant predictor of churn. This suggests that customers with more products are less likely to churn, possibly due to greater engagement or satisfaction with the bank's services.
* Customer Activity Status: The activity status of the customer (whether they are active members) is another crucial feature. Active customers are less likely to churn, highlighting the importance of customer engagement in holding strategies.
* Geography: Geographic location, particularly customers from Germany, shows high importance. This aligns with the earlier finding that churn rates vary significantly by region, indicating the need for region-specific holding strategies.
* Age: Age is also a significant predictor, with older customers showing different churn behaviours compared to younger ones. This could be due to different financial needs and satisfaction levels across age groups.
* Gender: Gender, specifically male customers, also shows importance, suggesting potential differences in churn behaviour between males and females.
* Balance: The account balance is another important feature, indicating that customers with higher balances are less likely to churn.
* Credit Score: Although less significant than the top features, the credit score still impacts churn prediction. Customers with lower credit scores may have higher churn rates.

A screen shot of a computer screen

Description automatically generated

XGBoost ROC Curve

The ROC curve for the XGBoost model, as shown in the figure, illustrates the trade-off between the true positive rate (sensitivity) and the false positive rate (1-specificity). The area under the curve (AUC) of 0.84 indicates that the XGBoost model has strong performance in distinguishing between churned and non-churned customers. This AUC value is comparable to the performance of the Random Forest model, demonstrating that XGBoost is effective in predicting customer churn. The ROC curve's shape and the AUC value reflect the model's ability to correctly classify customers while minimizing false positives and false negatives.

A screenshot of a computer

Description automatically generated

Statistical Analysis of Significant Features

A logistic regression model using Statsmodels analyzed the statistical significance of key features, including age, tenure, and the number of products. The output shows that the credit score, represented in this analysis, has a statistically significant negative impact on customer churn (p-value < 0.01). This means customers with higher credit scores are less likely to churn. The constant term is also significant, reinforcing the model's reliability. The results highlight the importance of considering credit scores when predicting customer churn.

A screenshot of a graph

Description automatically generated

Churn Rate Analysis by Geography

Churn rates were analyzed across different geographic regions, revealing significant disparities. Germany exhibited the highest churn rate at approximately 32.44%, followed by Spain at around 16.67%, and France with the lowest at about 16.15%. This variation suggests that geographic-specific factors, such as regional economic conditions, customer service quality, and competition, might influence customer holding rates. This insight indicates the need for tailored holding strategies for different regions to effectively address the factors contributing to higher churn rates.

A screenshot of a computer program

Description automatically generated

Impact of Credit Score

The impact of the credit score on customer churn is analyzed using logistic regression. The results indicate that the credit score has a statistically significant, though small, impact on the likelihood of churn (p-value = 0.007). The negative coefficient for the credit score (-0.0007) suggests that customers with lower credit scores are slightly more likely to churn, indicating a need for targeted holding strategies for customers with lower credit scores.

**Findings**

Model Performance

The evaluation of different models for predicting customer churn revealed notable insights into their performance metrics. Among the models assessed, the Random Forest and XGBoost classifiers appeared as the top performers. The Random Forest model achieved a ROC AUC score of 0.8494, indicating its strong capability in distinguishing between churned and non-churned customers. This performance metric suggests that the Random Forest model is highly effective in predicting customer churn with a good balance between sensitivity and specificity.

The XGBoost classifier also demonstrated robust performance, achieving a ROC AUC score of 0.8366. While slightly lower than that of the Random Forest, the XGBoost model still showed considerable predictive power. The high performance of these two models underscores their suitability for churn prediction tasks in the banking sector.

In compare, the Logistic Regression model, although interpretable and straightforward to implement, showed a lower ROC AUC score of 0.7755. This indicates that while Logistic Regression can provide valuable insights into the relationships between variables and churn, its predictive accuracy is not on average with the ensemble methods of Random Forest and XGBoost. The difference in performance highlights the importance of using more complex models when the objective is to maximize predictive accuracy in customer churn prediction.

Feature Importance

An analysis of feature importance across the models provided significant insights into the factors driving customer churn. The Random Forest model highlighted that age was the most critical predictor of churn, followed by the number of products held by the customer and their account balance. This suggests that older customers and those with fewer products are more likely to churn. The importance of the account balance indicates that customers with higher balances tend to be more loyal.

The XGBoost model, on the other hand, highlighted the number of products and customer activity status as the most crucial features. This model also identified age as a significant factor but gave more weight to the number of products and whether the customer is actively engaged with the bank's services. This finding aligns with the notion that active engagement and a spread portfolio of products are key to customer holding.

The consistency in the identification of significant features across different models strengthens the confidence in these findings. Age, number of products, and customer activity status were consistently recognized as vital predictors of churn. This consistency suggests that these factors should be the important points in strategies aimed at reducing churn rates.

Geographic Influence

The analysis of churn rates across different geographic regions uncovered substantial disparities. Notably, Germany exhibited the highest churn rate compared to other regions. This geographic-specific trend suggests that churn is influenced by regional factors that might include local economic conditions, cultural differences, or regional market dynamics.

The higher churn rate in Germany indicates that the bank may need to develop region-specific strategies to address the unique factors contributing to customer churn in this area. Tailoring holding strategies to the specific needs and preferences of customers in different regions can be more effective than a one-size-fits-all approach. For instance, customer service initiatives, marketing campaigns, and product offerings could be customized to better resonate with the local customer base in Germany.

Statistical Significance

The use of Logistic Regression for statistical analysis provided further insights into the significance of various predictors. The analysis confirmed that age and the number of products is statistically significant predictors of customer churn. This statistical validation reinforces the findings from the feature importance analysis, highlighting the critical role these factors play in predicting churn.

Interestingly, the tenure of the customer did not show a significant impact on churn. This finding suggests that the length of time a customer has been with the bank does not necessarily correlate with their likelihood of churning. This insight can help the bank focus its holding efforts on other more influential factors rather than relying on tenure as a key indicator.

Credit Score Impact

The impact of credit score on customer churn was also examined, and the findings revealed a minor yet statistically significant effect. Customers with lower credit scores were found to be slightly more likely to churn. Although the effect size is small, it is still noteworthy as it suggests that credit risk management practices could play a role in customer holding strategies.

Given this insight, the bank could consider integrating credit score analysis into its customer holding framework. By identifying customers with lower credit scores who are at a higher risk of churning, the bank can proactively offer support and tailored services to improve their satisfaction and loyalty. For example, financial counseling, personalized loan products, or flexible credit terms could be offered to these customers to enhance their overall experience and reduce the likelihood of churn.

Conclusion

The findings from this project provide a comprehensive understanding of the factors influencing customer churn in the banking sector. The robust performance of Random Forest and XGBoost models underscores their effectiveness in predicting churn. The consistency in identifying significant features such as age, number of products, and customer activity status across different models highlights the importance of these factors in churn prediction.

The geographic disparities in churn rates suggest the need for region-specific holding strategies, particularly in regions like Germany with higher churn rates. The statistical significance of age and the number of products as predictors, coupled with the minor impact of credit scores, provides actionable insights for the bank to refine its customer holding strategies.

By focusing on the identified key predictors and tailoring strategies to different geographic regions and customer profiles, the bank can enhance its efforts to reduce customer churn and improve overall customer satisfaction and loyalty. These findings not only contribute to the academic understanding of customer churn in the banking sector but also offer practical implications for the bank's operational strategies.