A logo for university computing

Auto-generated description

**Evaluation cover page**

|  |  |
| --- | --- |
| *Student's Full Name:* | Taylan Ozgur Ozkan |
| *Student Number:* | 2024140 |
| *Module Title:* | Strategic Thinking |
| *Evaluation Title:* | CA 3 Final Submission |
| *Assessment Due Date:* | 10/11/2024. |
| *Presentation Date* | 10/11/2024. |

**Declaration**

By submitting this review, I confirm that I have read CCT's policy on academic misconduct and understand the implications of submitting work that is not mine or that does not appropriately reference material taken from a third party or other source.

I declare that it is my own work and that all third-party material has been properly referenced.

I further confirm that this work has not previously been submitted for evaluation by me or anyone else at CCT College Dublin or any other higher education institution.

Contents

[Using Machine Learning Techniques to Predict Customer Churn Rates : Report 1](#_Toc182081086)

[Introduction 1](#_Toc182081087)

[Project Purpose 1](#_Toc182081088)

[Problem Definition 1](#_Toc182081089)

[Goal and Success Criteria 2](#_Toc182081090)

[Dataset and Data Preprocessing 2](#_Toc182081091)

[Dataset Introduction 2](#_Toc182081092)

[Handling of Missing and Outlier Values 2](#_Toc182081093)

[Feature Encoding 2](#_Toc182081094)

[Standardization 3](#_Toc182081095)

[Exploratory Data Analysis - EDA 3](#_Toc182081096)

[Data Visualization and Distributions 3](#_Toc182081097)

[Customer Churn Analysis 3](#_Toc182081098)

[Initial Findings 4](#_Toc182081099)

[Feature Engineering 4](#_Toc182081100)

[Creating New Features 4](#_Toc182081101)

[Feature Selection 4](#_Toc182081102)

[Modeling and Evaluation 5](#_Toc182081103)

[Model Selection and Applied Algorithms 5](#_Toc182081104)

[Hyperparameter Tuning 5](#_Toc182081105)

[Model Performance Comparison 6](#_Toc182081106)

[Selection and Interpretation of the Best Model 6](#_Toc182081107)

[Results Interpretation 6](#_Toc182081108)

[Feature Importance and Business Implications 6](#_Toc182081109)

[Future Improvements and Recommendations 7](#_Toc182081110)

[Conclusion 8](#_Toc182081111)

# Using Machine Learning Techniques to Predict Customer Churn Rates : Report

## Introduction

### Project Purpose

This project aims to use data analytics and machine learning techniques to predict customer churn in the banking sector. Customer churn is a situation that directly affects customer loyalty in highly competitive sectors such as banking and can lead to major losses in business profitability. A customer leaving the bank can not only cause short-term revenue loss, but also damage brand loyalty in the long term.

Being able to predict customer churn in the banking sector provides strategic advantages to increase customer loyalty. For example, identifying customers at high risk of churn allows the bank to offer special offers, loyalty programs or personalized campaigns to these customers. In addition, predicting customer churn in advance helps banks use their resources more effectively. For example, customer service and marketing departments can optimize their costs by focusing on segments with high customer churn risk.

The aim of this project is to develop a powerful model that can predict churn risk based on customer characteristics. Customer data was analyzed, profiles of customers at risk of leaving the bank were determined and a prediction model was created in line with these characteristics. The modeling process took into account various features such as customer demographics (age, gender), financial status (credit score, balance, annual income) and customer interaction information (active membership status, number of products owned).

This project also provides banks with the opportunity to better understand the factors behind customer churn. The results of the model allow the bank to shape its customer loyalty strategies, provide important insights into which customer segments are at risk of churn, and thus make more informed decisions to increase customer satisfaction. For example, if the analysis determines that older customers and those with high balances are at high risk of churn, the bank can reduce the risk of churn by offering more attractive services to customers in this segment.

As a result, this project aims to proactively manage customer churn using data analytics and machine learning techniques, thus contributing to the bank's long-term sustainability by preserving its customer base.

### Problem Definition

Churn refers to the financial losses experienced by a business as a result of not being able to retain its existing customers. In the banking sector, customer churn occurs when customers turn to other banks, give up on the products and services offered by the bank, or end their relationship with the bank due to changes in their financial situation. Customer churn not only leads to short-term loss of income; it can also have long-term negative effects such as shrinking the customer base, damaging brand reputation, and reducing potential future profitability.

The negative effects of customer churn on a business result in decreased financial performance, less efficient use of resources, and increased marketing costs. For example, the cost of retaining an existing customer is usually lower than acquiring a new customer. Therefore, minimizing the loss of existing customers becomes a strategic priority for banks. In particular, the loss of loyal customers can weaken the competitiveness of the business and reduce customer trust. This increases the likelihood that customers will work with other banks.

Being able to predict customer churn in the banking sector is important in understanding why customers end their relationship with the bank and preventing this situation. Being able to detect customer churn tendencies in advance provides banks with opportunities to reduce churn risk and strengthen customer loyalty. For example, identifying customers at high churn risk can enable the bank to increase customer satisfaction by offering special services to these customers. This proactive approach allows for the development of targeted strategies to improve customer experience.

In this project, a model was developed to predict which customers are at higher churn risk by analyzing the bank’s existing customer data. This model aims to predict customer churn by taking into account customer demographics (e.g. age, gender), financial status (such as credit score, balance, annual income), and interactions with the bank (number of products, active membership status). This model, which predicts customer churn, can provide the bank with in-depth insights into which customer segments are at higher risk. These insights allow the bank to make more strategic decisions to reduce customer churn.

As a result, preventing customer churn is critical to the bank’s long-term profitability, sustainability of the customer base, and increasing customer satisfaction. This project provides an important tool to support data-driven decision-making mechanisms in the bank's customer churn management process and to develop proactive solutions by identifying customers at risk of churn in advance.

### Goal and Success Criteria

The main goal of this project is to develop a machine learning model that can predict customer churn by analyzing the bank's existing customer data. Predicting customer churn will help the bank improve customer relationship management, increase customer satisfaction, and strengthen customer loyalty. For banks, being able to predict customer churn in advance allows them to develop targeted strategies, especially focusing on customer segments with high churn risk. In this way, banks can increase customer satisfaction and minimize customer churn by offering special campaigns, advantages, and loyalty programs to customers with high churn risk.

The success criteria in developing the model are based on the model's ability to predict customer churn accurately and reliably. In this direction, a number of statistical metrics are used to evaluate the performance of the model:

Accuracy: Indicates the overall predictive success of the model. It refers to the rate at which the model correctly predicts customer churn. However, accuracy alone may not be a sufficient success criterion in unbalanced data sets.

Precision: Indicates whether the situations predicted by the model as customer churn are actually customer churn. That is, a high precision value indicates that the model keeps the false alarm rate low and correctly detects customer churn.

Recall (Sensitivity): Indicates how successfully the model can detect customer churn. Sensitivity is an indicator of the model's success in predicting the risk of customer churn and is a critical criterion, especially for predicting customer churn in advance.

F1 Score: Indicates the overall balance of the model by combining precision and sensitivity. The F1 score is used to minimize the effect of both false positive and false negative predictions in imbalanced data sets.

ROC-AUC (Receiver Operating Characteristic - Area Under Curve): It is one of the metrics that best shows the model's ability to predict customer churn. ROC-AUC shows the model's ability to distinguish between different classes. A high ROC-AUC value indicates that the model can correctly distinguish between customer churn and non-customer churn situations.

These performance metrics are critical for evaluating the model's accuracy in predicting customer churn. In particular, high ROC-AUC and Recall values ​​were determined as key success criteria indicating the model's ability to predict customer churn. Thanks to these metrics, not only the overall accuracy of the model but also its ability to accurately and effectively detect customer churn is measured. This is necessary for the bank to correctly identify customers at high risk of churn and take appropriate actions.

As a result, the success of this project depends on the ability of the model we developed to accurately predict customer churn. The high predictive power of the model will enable the bank to proactively manage customer churn and support strategic decision-making processes.

## Dataset and Data Preprocessing

### Dataset Introduction

The dataset used for this project includes customer information of a bank. The dataset includes demographic information of the customers (age, gender, country), financial status (credit score, balance, annual income), and interaction information with the bank (number of products, active membership status). The target variable is the Exited column, which indicates whether the customer has left the bank. This variable takes values ​​of 0 and 1, representing customer churn.

The features in the dataset include variables such as CreditScore, Geography, Gender, Age, Balance, NumOfProducts, HasCrCard, IsActiveMember, and EstimatedSalary. These features were analyzed and processed to improve the model's ability to predict customer churn.

### Handling of Missing and Outlier Values

The dataset contained missing values ​​in the Geography, Age, HasCrCard, and IsActiveMember columns. Since the missing data rate was as low as 0.01%, missing data were filled with methods such as mean or mode. In particular, the median value was used in the Age column, and the most frequent value was used in other categorical columns.

Features such as CreditScore and Balance were examined for outlier analysis. Outliers in these features were left as is instead of being removed from the data set, as they could potentially be important in understanding customer behavior. This strategy ensures that the model works without losing the true reflection of customer behavior.

### Feature Encoding

In order for the model to work correctly, categorical features were converted to numerical data. The Geography variable was coded with the one-hot encoding method, allowing each country (France, Germany, Spain) to be represented as a separate column. In this process, France was selected as the reference category and other countries (Geography\_Germany, Geography\_Spain) were added as new columns. The Gender variable was transformed by assigning 0 to the Female value and 1 to the Male value with the binary encoding method.

These coding processes were performed to make the categorical features in the dataset suitable for the model and to increase the predictive power of the model.

### Standardization

In order to increase the model performance and balance the effect of numerical data at different scales, some numerical features were scaled using StandardScaler. Numerical columns such as CreditScore, Age, Balance, and EstimatedSalary were transformed so that their mean was zero and their standard deviation was one. This process is especially important for distance-based algorithms and complex models, because differences in data scales can negatively affect the model's learning process.

Standardization positively affected the overall performance of the model by ensuring that all features were evaluated equally.

## Exploratory Data Analysis - EDA

### Data Visualization and Distributions

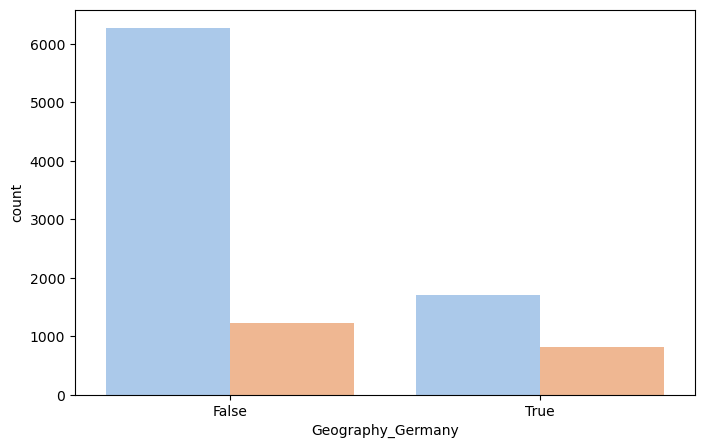
Various visualizations were made to understand the structure of the dataset and the distribution of features:

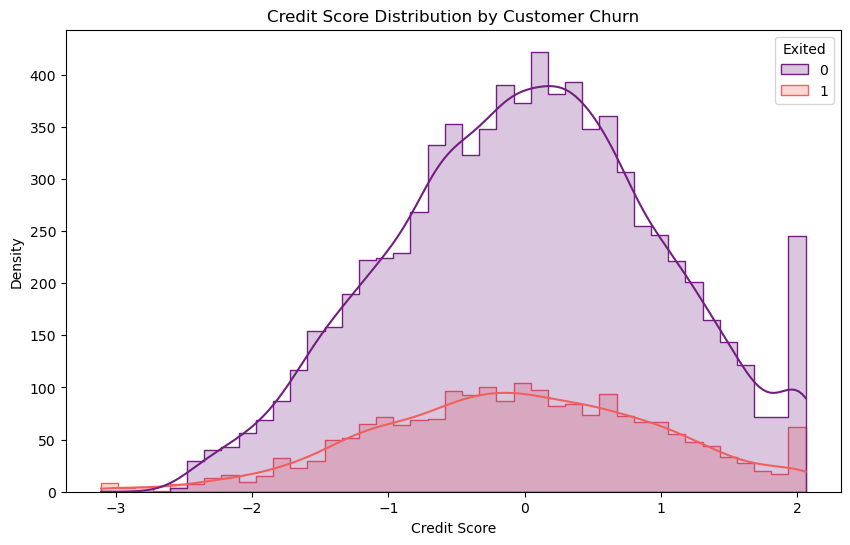
• Age Distribution: The age distribution of the customers was examined and it was observed that the majority were between the ages of 30-40. This information can be useful in understanding the demographic characteristics of the customers and determining the risky age groups. A graph of age distribution

Description automatically generated

• Customer Churn Rate (Exited): It was seen that approximately 20% of the customers in the dataset left the bank. This information understands the customer churn rate in the dataset and allows the model to work with unbalanced data. A chart with a green and blue bar

Description automatically generated

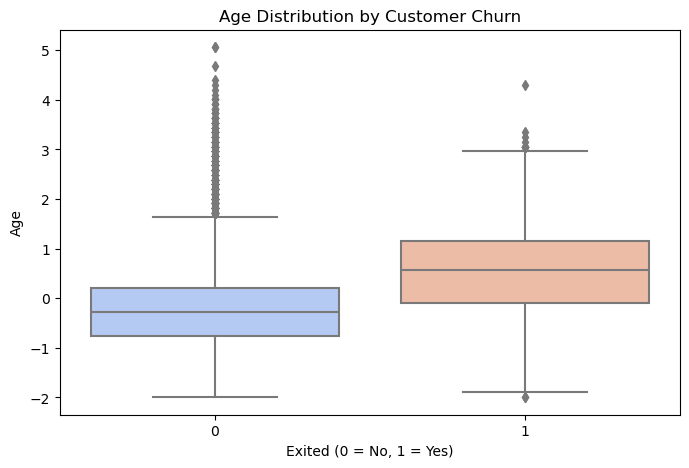
• Geography and Customer Churn: It was seen that the churn rate of customers in Germany was higher than in other countries. This information shows that geographic location can have an effect on customer churn and allows geography to be taken into account in the model.

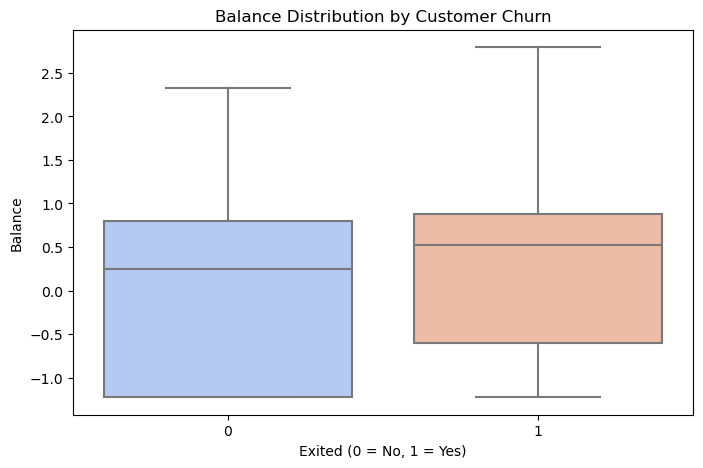
• Credit Score Distribution: The difference in credit scores between customers who lost and did not lose customers was visualized, and it was observed that customers who lost had lower credit scores.

These visualizations helped us understand the important variables related to customer churn and the general structure of the dataset.

### Customer Churn Analysis

A comparison of certain characteristics was made between the groups that experienced and did not experience customer churn:

• Age: The average age of the group that experienced customer churn was higher than the other group. This shows that age can be an effective factor on customer loyalty.

• Balance: It was observed that the balance rate of customers who left the bank was higher. This shows that customers with a high balance in the bank tend to leave.

• Active Membership Status (IsActiveMember): The churn rate of inactive customers was higher than active customers. This feature can be considered as a critical factor in terms of customer interaction and loyalty.

• Credit Score: Customers with low credit scores are more likely to leave. This shows that financial reliability is another factor affecting customer churn.

These analyses contributed to a clearer understanding of the characteristics associated with customer churn and to the prominence of these characteristics in the model.

### Initial Findings

According to the analysis, some basic features affecting customer churn have been identified:

• Age and Balance stand out among the most effective features in customer churn. These features play an important role in predicting customer loyalty.

• IsActiveMember has been observed as an effective feature in predicting customer churn by providing information about the customer's interaction with the bank.

• Credit Score and Geography show the impact of customers' financial reliability and demographic characteristics on customer churn.

These features stand out in the model's customer churn prediction process, allowing more accurate predictions to be made. These initial results also formed the basis for the selection of features to be used in the model.

## Feature Engineering

### Creating New Features

Two new features have been added to the dataset to better predict customer churn:

• CreditAge: This feature is created by multiplying the credit score by the age. The combination of credit score and age provides more detailed information about the customer's financial reliability and history. Especially for young customers, a high credit score can be considered a positive signal in terms of customer loyalty.

• BalanceTenure: This feature, obtained by comparing the customer balance to the loyalty period, provides information about the customer's financial mobility in the bank. People with a high balance who have been bank customers for a long time may have lower churn; this ratio allows for a deeper analysis of customer loyalty.

These new features have increased the model's ability to predict customer churn and provided a better understanding of customer behavior.

### Feature Selection

The SelectKBest method was used to determine the best features to optimize model performance and reduce the impact of unnecessary data. This method allowed us to select the features with the highest correlation with the target variable, customer churn (Exited).

As a result of the analysis, the following 10 features were determined to have the highest correlation:

• CreditScore, Gender, Age, Balance, NumOfProducts, IsActiveMember

• Geography\_Germany, Geography\_Spain, CreditAge, and BalanceTenure

These features increase the model's ability to predict customer churn, providing healthier results. These selected features were used in the modeling phase and formed an important basis for increasing the accuracy of the model.

## Modeling and Evaluation

### Model Selection and Applied Algorithms

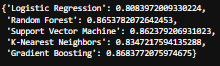
Various machine learning algorithms were used to predict customer churn. In this project, both simple and complex models such as Logistic Regression, Random Forest, and Gradient Boosting were tried. Each of these models was compared according to their predictive power of customer churn, and the model that gave the best results was selected.

• Logistic Regression: As a simple and interpretable model, it was used as a basic reference for predicting customer churn. However, its performance remained lower than other complex models.

• Random Forest: This model, which combines multiple decision trees to provide stronger predictions, has been successful in predicting customer churn and has achieved a high accuracy rate.

• Gradient Boosting: This model provides progressively more accurate results by improving erroneous predictions. As the model with the highest performance in the project, it has provided the best results in predicting customer churn.

Hyperparameter tuning has been performed to further optimize the performance of these models and the best parameters have been determined.



### Hyperparameter Tuning

Hyperparameter optimization has been applied to Gradient Boosting and Random Forest models to increase the performance of the model in predicting customer churn. Different parameter combinations have been tried with the GridSearchCV method and the best parameters have been determined.

• Gradient Boosting: The best results were achieved with the settings n\_estimators=200, learning\_rate=0.1, and max\_depth=3. These settings have increased the accuracy of the model and contributed to more accurate predictions.

• Random Forest: The best parameters were determined as n\_estimators=100, max\_depth=20, and min\_samples\_split=10. These settings increased the performance of the Random Forest model and allowed it to reach similar accuracy rates with Gradient Boosting.

This optimization process allowed both models to better predict customer churn, with Gradient Boosting standing out as the best model.

### Model Performance Comparison

The performance of three different models was compared using metrics such as accuracy, precision, recall, F1 score and ROC-AUC:

| **Model** | **Accuracy** | **Precision** | **Recall** | **F1 Score** | **ROC-AUC** |
| --- | --- | --- | --- | --- | --- |
| Logistic Regression | 0.808 | 0.556 | 0.208 | 0.303 | 0.763 |
| Random Forest | 0.868 | 0.760 | 0.495 | 0.599 | 0.852 |
| Gradient Boosting | 0.869 | 0.761 | 0.503 | 0.606 | 0.859 |

According to these results, Gradient Boosting provided the highest accuracy, F1 score and ROC-AUC value. It was selected as the model that showed the highest success in predicting customer churn. Random Forest was the second best model with very close results. Logistic Regression, on the other hand, fell behind these models because it showed lower performance.

### Selection and Interpretation of the Best Model

Gradient Boosting was selected as the best model because it provided the highest accuracy and ROC-AUC value in predicting customer churn. This model has been successful in predicting a complex structure such as customer churn, especially thanks to its ability to gradually learn the complexity of the data and correct errors.

The Gradient Boosting model allows us to understand the factors that affect customer churn by examining the importance of the features. The prominent features in the model are as follows:

• Age: As the feature with the highest importance, age is seen to play a critical role in predicting customer churn.

• NumOfProducts: The number of products the customer owns stands out as an important factor affecting customer loyalty.

• IsActiveMember: Whether the customer is an active member or not is an important factor in predicting customer churn.

The fact that these features have decisive effects on customer churn has contributed to the model being able to make accurate predictions. The bank can take strategic measures for high-risk customer segments based on these features.

## Results Interpretation

### Feature Importance and Business Implications

Feature importance analyses conducted on Gradient Boosting and Random Forest models revealed the main factors affecting customer churn:

• Age: Age stands out as the most important feature in predicting customer churn. It is seen that older customers have a higher churn risk. This information can help prevent customer churn by offering loyalty programs or special offers especially for young customers.

• NumOfProducts: The number of products purchased from the bank is an important factor affecting customer loyalty. Increasing the number of products reduces customer churn. Based on this information, it may be beneficial for the bank to increase loyalty by offering new products to its customers.

• IsActiveMember: The fact that the customer is an active member is another factor that reduces customer churn. Organizing interaction campaigns and feedback surveys for inactive customers can be effective in reducing customer churn.

• Balance: It has been observed that customers with high balances have a higher risk of loss. Loyalty can be increased by offering special financial advantages or privileged services for this customer group.

• Geography: It has been observed that customers in Germany in particular have a higher risk of loss compared to other countries. This situation indicates that geographically personalized marketing strategies should be developed.

In line with these characteristics, it is recommended that the bank take the following strategic steps to prevent customer loss:

1. Age-Specific Campaigns: Offer loyalty-increasing advantages to young customers.

2. Product Diversification: Increase product variety by offering new products to customers.

3. Interaction Programs for Inactive Customers: Increase loyalty by communicating more with customers with low interaction.

4. Geographically Focused Marketing: Develop special advantages and campaigns for customers in Germany.

These conclusions will make it easier for the bank to take strategic steps to reduce customer loss and increase customer loyalty.

## Future Improvements and Recommendations

Despite the successful results obtained in the project, some improvement suggestions have been developed to predict customer churn with higher accuracy and make the bank's strategic decisions more effective. The bank can make its efforts to reduce customer churn more sustainable and effective by implementing these suggestions.

Strengthening the Model with Additional Data Sources: Collecting more data on customer interactions with the bank and feeding the model with these additional data sources can strengthen the model's understanding of customer behavior. For example:

Customer Feedback: Feedback given by customers in call centers or surveys can provide information about customer satisfaction. Dissatisfied customers may be at risk of churn.

Call Center Records: Frequent customer calls or complaints about certain issues can be important for predicting future churn risk.

Purchase Frequency and Payment History: Customers who actively use bank services or make regular payments may have a lower churn risk. This data can increase the accuracy of the model and provide better predictions.

Time Series Analysis: Understanding how customer churn changes over time allows the bank to develop strategies to reduce customer churn in certain periods. For example:

Seasonal Churn Periods: Customer churn rates may increase during certain periods (such as the end of the year or during holiday periods). During these periods, the risk of churn can be reduced by offering special campaigns or incentives.

Monitoring Changes in Customer Relationships: If there are certain periods when customers' relationships with the bank weaken, proactive steps can be taken during these periods. These analyses can also help the bank plan campaign and promotion periods.

Developing Segment-Based Models: Customer churn dynamics may differ across different customer segments (such as age groups, income levels or geographic regions). Developing segment-based models allows for more precise estimates to be made for each segment. For example:

Geographic Segmentation: Separate models can be developed for customers in different countries or regions, and strategies suitable for the churn dynamics of each region can be applied.

Demographic Segmentation: Special models can be developed for young customers or customers in high-income groups, and more targeted measures can be taken for these groups.

Integration with Customer Lifetime Value (CLV) Analysis: Customer churn prevention strategies can be optimized by integrating with customer lifetime value (CLV). This approach allows strategic decisions to be made based on the potential gains customers bring to the bank:

Prioritizing High CLV Customers: The bank can develop stronger strategies to prevent churn by allocating more resources to high CLV customers.

Cost-Effective Strategies: More cost-effective strategies can be implemented for low CLV customers. For example, reaching these customers with automated messages or email campaigns can optimize resource usage.

Model Improvement by Adding New Features: Predictive power can be increased by adding other features or factors that may affect customer loyalty to the model. New features allow for more detailed analysis of the reasons behind customer churn:

Participation in Loyalty Programs: Participation in loyalty programs is a factor that strengthens customer loyalty. Adding this data to the model can help better understand the churn risk of loyalty program members.

Credit Application History: The customer's credit application history provides information about their financial needs and interactions with the bank. Customers who frequently apply for credit may be in a group with a higher risk of loss.

Usage of Mobile and Digital Banking: Loyalty may increase in customers who actively use the bank's digital services. More accurate predictions can be made by including this data in the model.

Implementing the Model in a Live Environment and Continuous Updates: The model should not only make predictions, but also be constantly updated with new data to maintain its performance. In this direction:

Periodic Retraining of the Model: Customer behaviors may change over time, so the model should be retrained at regular intervals.

Creating a Feedback Loop: The model's predictions should be compared with actual results, their accuracy should be analyzed, and necessary improvements should be made.

Developing Machine Learning Algorithms: The performance of the model can be further increased by improving the algorithms used or by trying more complex algorithms. For example, deep learning methods or ensemble models can be tried. More advanced algorithms can be useful, especially in predicting complex customer behaviors.

## Conclusion

This project aimed to develop a successful prediction model using data analytics and machine learning techniques to predict customer churn in the banking sector. Customer churn is a significant problem, especially in sectors with intense competition such as banking, and is a factor that directly affects the profitability and sustainability of businesses. During the project, a model that can predict churn risk based on customer characteristics was developed and business results were analyzed.

The data set included demographic, financial and interaction information of customers with the bank. Using this information, the main characteristics that could affect customer churn were determined and data pre-processing steps were performed to be used in the modeling process. Steps such as processing missing and outlier values, converting categorical features to numerical data and scaling numerical features were meticulously implemented to ensure that the model yielded more accurate results.

Some basic findings regarding customer churn were obtained through exploratory data analysis (EDA). It was observed that characteristics such as age, balance, number of products and active memberships were the determining factors in customer churn. It has been determined that customers who are older or have a large balance have a higher risk of loss, while customers who actively use the bank's products have a lower risk of loss. These analyses have helped us understand the demographic and financial factors that affect customer loss.

Various algorithms such as Logistic Regression, Random Forest, and Gradient Boosting were tested in the modeling process. As a result of the hyperparameter adjustments, the Gradient Boosting model was selected as the best model by achieving the highest accuracy and ROC-AUC score. In terms of model performance, Gradient Boosting proved to be a powerful tool in predicting customer loss; the Random Forest model stood out as an alternative option by giving very close results. Logistic Regression, on the other hand, fell behind other models in predictive performance as a simpler model.

In the analysis of the Gradient Boosting model that showed the best performance, it was seen that features such as Age, NumOfProducts, and IsActiveMember were determinants of customer loss. These features show the areas that the bank can focus on to increase customer loyalty and reduce the risk of loss. In particular, demographic factors such as active membership status and age play a critical role in predicting customer churn. Based on this data, the bank can identify customer segments with high churn risk in advance and develop strategies such as loyalty programs or special offers.

Based on the project results, strategies for preventing customer churn have been proposed. For example, strategic steps such as offering special benefits to young customers, increasing customer loyalty by increasing product variety, and increasing interaction with inactive customers are suggested. Personalized campaigns can be planned for specific regions, such as customers in Germany, considering geographical differences. In addition, it is suggested to further develop the model with additional data in the future and to examine the changes in customer churn over time with time series analysis.

In conclusion, this project demonstrates the power of data analytics in the process of predicting and preventing customer churn for banks. While in-depth analysis of customer data provides a better understanding of the factors affecting customer churn, machine learning models allow banks to identify risky customers in advance and develop targeted strategies. This project can be considered a valuable resource for increasing customer loyalty, reducing churn rates, and strengthening the bank's competitive advantage. Reducing customer churn not only provides financial gains, but also increases brand loyalty and contributes to the long-term success of the bank. Therefore, data analytics and machine learning-based churn prediction models have an important place in the strategic planning processes of banks.

## GITHUB LINK

https://github.com/TaylanOzgur96/Strategic-Thinking---Semester-Two---CA-3.git