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 2 – Assignment of the final project proposal. |
| *Assessment Due Date:* | 19/05/2024. |
| *Presentation Date* | 19/05/2024. |

**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 2](#_Toc166800354)

[1. Introduction 2](#_Toc166800355)

[2. Project Management and Planning 3](#_Toc166800356)

[Project Plan and Timeline 3](#_Toc166800357)

[Challenges and Solutions 4](#_Toc166800358)

[3. Data Analysis and Preprocessing 4](#_Toc166800359)

[4. Feature Engineering 6](#_Toc166800360)

[5. Machine Learning Application and Evaluation 7](#_Toc166800361)

[6. Results and Discussion 9](#_Toc166800362)

[Comparison of Model Performance 9](#_Toc166800363)

[Strengths and Weaknesses 9](#_Toc166800364)

[Business Implications 9](#_Toc166800365)

[Result 9](#_Toc166800366)

[7. Conclusion and Recommendations 10](#_Toc166800367)

[8. References 11](#_Toc166800368)

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

## Introduction

Estimating customer churn rates plays an important role in the customer relationship management and strategic planning processes of modern businesses. It has become necessary to develop effective strategies to minimize customer churn, increase revenues and strengthen customer loyalty. In this context, data analytics and machine learning techniques offer powerful tools to better understand customer behavior and predict customer abandonment. In this project, we aimed to predict the abandonment probability of a bank's customers. Being able to predict customer churn offers businesses the opportunity to take proactive measures and enable them to develop more targeted strategies to increase customer loyalty.

High customer abandonment rates in the banking industry can negatively impact profitability and weaken competitive advantage. Therefore, it is of great importance to understand the reasons for customer churn and take the necessary measures to prevent these losses. Data analytics and machine learning provide powerful tools to address these issues. In this project, we will use various data analysis and machine learning techniques to predict customer abandonment rates. This process will cover a wide range, starting from examining and cleaning the data set, to model development and evaluation stages.

The main purpose of our project is to determine the factors affecting customer abandonment rates and to predict the probability of customers leaving using these factors. For this purpose, we worked on a data set containing customer data of a bank. The data set includes customer demographic information, account information and data on customer behavior. Using this data, we sought to identify the main factors affecting customer abandonment and develop predictive models based on these factors.

During the data analysis process, we first examined the structure and general characteristics of the data set. We implemented data cleaning steps such as filling in missing data and handling outliers. By performing exploratory data analysis (EDA), we visualized the distribution and relationships of the data. At this stage, we examined the distribution of variables such as customers' age distribution, credit score and balance. We also analyzed the geographical distribution of customers and abandonment rates. Based on the EDA results, we identified notable patterns and anomalies in the dataset.

## Project Management and Planning

Project management and planning requires careful planning of specific steps and processes for the effective execution of the project.

### **Project Plan and Timeline**

| **Date** | **Work to be Done** |
| --- | --- |
| 6-7 May | Determining the project scope and objectives, introducing the data set, determining the methods and tools to be used. |
| 8-9 May | Loading the data set, examining it, filling in missing data, handling outliers. |
| 10-11 May | Calculating descriptive statistics, making data visualizations (histograms, scatter plots, etc.). |
| 12-14 May | Feature engineering is selecting the required features, creating new features, developing and training the machine learning model. |
| 15-16 May | Evaluation of the model with performance metrics, improvement and optimization of the model. |
| 17-18 May | Reporting EDA and model findings, evaluating results and business impacts. |
| 19 May | Uploading the Jupyter Notebook and Word report to GitHub, final checks of the project and submit. |

### **Challenges and Solutions**

Various difficulties were encountered during the project process. These difficulties arise from the characteristics of the dataset and the technical problems encountered while carrying out the project. Below are the main challenges and solutions to these challenges:

**Missing and Outliers**

Challenge: The presence of missing and outlier values ​​in the data set made data analysis difficult.

Solution: Missing values ​​were filled in using appropriate statistical methods. Outliers were detected according to specific criteria and cleaned or flagged as appropriate.

**Hyperparameter Optimization**

Challenge: The difficulty of performing hyperparameter optimization to improve the performance of machine learning models.

Solution: The best hyperparameters were determined using the Grid Search method. This process involved testing various parameter combinations to improve the model's performance.

**Data Visualization**

Challenge: Visualize data accurately and effectively.

Solution: Implemented various data visualization techniques using Python libraries such as Seaborn and Matplotlib. In this way, the distribution and relationships of the data could be seen more clearly.

**Evaluation of Model Performance**

Challenge: How to accurately compare the performance of different models.

Solution: Various metrics such as accuracy, precision, sensitivity and F1 score were used to evaluate the model performance. These metrics helped us identify the strengths and weaknesses of the models.

## Data Analysis and Preprocessing

Data analysis and preprocessing includes the processes of examining the data set, cleaning it, and exploratory data analysis (EDA).

**Examination of the Data Set**: The data set used within the scope of the project includes data on demographic information, account information and customer behavior of a bank's customers. As a first step, the data set was examined to understand the general structure and columns of the data set.

A screen shot of a black screen

Description automatically generated

**Handling Missing and Outlier Values**: One of the first challenges encountered in the data analysis process is the presence of missing and outlier values. Missing data can negatively impact the model's performance and lead to inaccurate results. Therefore, it is necessary to fill in missing values ​​with appropriate methods.

A screenshot of a computer

Description automatically generated

**Descriptive Statistics and Visualizations**: Exploratory data analysis (EDA) is an important step used to understand the overall structure and distribution of the data set. In this process, descriptive statistics of the data set were calculated and various visualizations were made.

A screenshot of a black screen

Description automatically generated

**Initial Findings and Patterns**: At the end of the EDA process, the initial findings and striking patterns obtained from the data set were analyzed. For example, the effects of variables such as age and credit score on the customer abandonment rate have been examined.

A graph of age distribution

Description automatically generatedA graph of credit score distribution

Description automatically generatedA diagram of a distribution

Description automatically generatedA blue and orange rectangular bars

Description automatically generated

## Feature Engineering

Feature engineering is a critical step that involves transforming existing features in the data set and creating new features to improve the performance of the model. This process helps the model make more accurate predictions by increasing its learning capacity. In this project, new features were created within the scope of feature engineering, existing features were transformed and categorical variables were converted into numerical data.

**Creation of New Features**: New features have been created that can better represent customer behavior. For example, a new feature called CreditAge has been created, which is customers' credit score multiplied by their age.



**Coding of Categorical Data**: Categorical variables such as Geography and Gender are converted to numerical values ​​using LabelEncoder.

**A screen shot of a computer program

Description automatically generated**

**Feature Selection**: The most meaningful features in the data set were selected and unnecessary features were removed. For example, properties such as RowNumber, CustomerId, and Surname are removed because they are not meaningful to the model.

**A black screen with red and blue text

Description automatically generated**

**Separation of the Data Set into Training and Testing**: 70% of the data set is reserved for training the model and 30% for testing the model.

****

**Standardization of Data**: Standardizing data ensures that the model processes each feature on the same scale.

A screen shot of a computer

Description automatically generated

## Machine Learning Application and Evaluation

In this section, models will be developed using different machine learning algorithms on the data set and the performances of these models will be evaluated. This process includes various steps, from algorithm selection to model development and hyperparameter optimization.

**Algorithm Selection and Application**: Logistic regression and random forest algorithms were chosen to predict customer abandonment rates. Logistic regression was chosen due to its simplicity and interpretability. Random forest was chosen due to its ability to capture more complex relationships and its high accuracy.

**Model Development and Validation**: Model development and validation processes were carried out for both algorithms. Training of the models and prediction processes were carried out in the following steps:

**Logistic regression model**: Used to predict customer abandonment rates. The model made predictions on the test data by performing learning on the trained data.

**Random Forest Model**: Random forest model has been used to predict customer abandonment rates. This model makes predictions using multiple decision trees and provides higher accuracy.

**Hyperparameter Optimization**: Hyperparameter optimization was performed to increase model performance. This process was carried out using the Grid Search method and the best hyperparameters were determined.

**Evaluation of Model Performance**: Metrics such as accuracy, precision, sensitivity and F1 score were used to evaluate the performance of the models. These metrics help measure the accuracy and reliability of the model's predictions.

**Logistic Regression Results:**

• Accuracy: 0.809

• Precision: 0.570

• Recall: 0.183

• F1 Score: 0.277

**Random Forest Results:**

• Accuracy: 0.863

• Precision: 0.770

• Recall: 0.452

• F1 Score: 0.569

**Optimized Random Forest Results:**

• Accuracy: 0.862

• Precision: 0.784

• Recall: 0.428

• F1 Score: 0.554

When we compare the performance of the models, it is seen that the random forest model provides higher accuracy and precision than the logistic regression model. However, both models have strengths and weaknesses in terms of sensitivity and F1 score.

## Results and Discussion

In this section, we will compare model performances, discuss strengths and weaknesses, and what the results mean from a business perspective. These analyzes will help us determine which approaches are more effective in predicting customer abandonment rates.

### **Comparison of Model Performance**

Two main machine learning algorithms were used in the project: logistic regression and random forest. The performance of both algorithms was evaluated using metrics such as accuracy, precision, sensitivity and F1 score. The table below summarizes the performance metrics for both models:

| **Metric** | **Logistic Regression** | **Random Forest** | **Optimized Random Forest** |
| --- | --- | --- | --- |
| Accuracy | 0.809 | 0.863 | 0.862 |
| Precision | 0.570 | 0.770 | 0.784 |
| Recall | 0.183 | 0.452 | 0.428 |
| F1 Score | 0.277 | 0.569 | 0.554 |

Looking at these performance metrics, it is seen that the random forest model provides higher accuracy and precision compared to the logistic regression model. The sensitivity and F1 score of the random forest model are also higher than the logistic regression model, indicating that the random forest model has a more balanced performance.

### **Strengths and Weaknesses**

**Logistic Regression**: Simple and interpretable, fast training time, low computational cost, but has difficulty capturing non-linear relationships and provides limited success with low sensitivity.

**Random Forest**: Ability to capture non-linear relationships, high accuracy and precision, determine the importance of features, but can be computationally costly and overly complex.

### **Business Implications**

Predicting customer abandonment rates provides significant strategic advantages for banks and other financial institutions. The results obtained can be used in customer relationship management and strategic planning processes.

### **Result**

Predicting customer abandonment rates using machine learning models can help banks improve customer relationship management and gain strategic advantage. In this project, customer abandonment rates were estimated using logistic regression and random forest algorithms. The random forest model has been found to be more effective in predicting customer abandonment rates because it provides higher accuracy and precision. In the future, it is possible to further improve model performance by using further feature engineering and alternative algorithms. Such projects can help businesses take proactive measures to increase customer loyalty and maintain profitability.

These findings show that the random forest model is more effective in predicting customer abandonment rates. However, both models have their own strengths and weaknesses. While the logistic regression model stands out with its simplicity and interpretability, the random forest model excels with its ability to capture more complex relationships.

## Conclusion and Recommendations

**Summary of Results (Summary of Findings)**

In this project, various data analysis and machine learning techniques were used to predict customer abandonment rates of a bank. The basic steps of the work are:

* Data Collection and Cleansing
* Exploratory Data Analysis (EDA)
* Feature Engineering
* Machine Learning Application
* Evaluation of Model Performance

These findings show that the random forest model is more effective in predicting customer abandonment rates. However, both models have their own strengths and weaknesses. While the logistic regression model stands out with its simplicity and interpretability, the random forest model excels with its ability to capture more complex relationships.

**Recommendations for Future Work**

The results obtained in this project provide important insights in predicting customer abandonment rates. However, there are some suggestions for future work to further improve model performance and more accurately predict customer abandonment rates:

* More Feature Engineering
* Using Alternative Algorithms
* Improving the Model
* Data Enrichment

## References

• Kaggle Dataset: Churn\_Modelling.csv

• Scikit-learn Documentation: https://scikit-learn.org/stable/documentation.html

• Seaborn Documentation: https://seaborn.pydata.org/

• Matplotlib Documentation: https://matplotlib.org/