ALMA MATER STUDIORUM – UNIVERSITÀ DI BOLOGNA

Second Cycle Degree in Digital Transformation Management

Customer Churn Prediction Using Random Forest Classifier

Team Members:

Mustafa Kemal Kaya
Ezginur Cendik

Academic Year: 2022-2023

Contents

1 Introduction

In this Machine Learning and Data Mining project, we have developed a customer churn prediction model for a bank dataset. The model aims to predict whether new customers are likely to stay or leave the bank, addressing the critical challenge of customer churn. By accurately predicting churn, the bank can take proactive measures to retain customers and improve service quality. The project follows a well-structured approach, starting with data loading and exploration, followed by data preprocessing to ensure data quality. We analyze churn counts and visualize customer behavior to identify potential patterns related to churn. Using machine learning techniques, including the Random Forest Classifier, we evaluate the model's performance and present key metrics such as ROC curve, AUC score, and confusion matrix, providing valuable insights for customer retention strategies.

The project aims to build a customer churn prediction model for a bank's new customers, forecasting whether they will stay or leave. This model is crucial as it enables the bank to take proactive actions to retain customers, reduce churn rates, and optimize marketing and operational strategies. By leveraging data-driven insights, the bank can enhance customer satisfaction and strengthen its competitive position in the market, leading to improved business performance and long-term customer loyalty.

1.1 Related work

Guo-en and Wei-dong (2008) focused on building a customer churn prediction model using SVM in the telecommunication industry. They compared this method with other techniques such as DT, artificial neural networks, naïve Bayesian (NB) and logistic regression. The results proved SVM to be a simple classification method of high capability yet good precision. (4)

Nie et al. (2011) built a customer churn prediction model by using logistic regression and DT-based techniques within the context of the banking industry. (2)

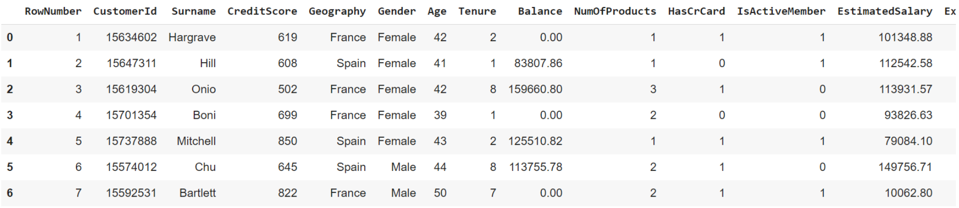
In their study, Lin et al.(2012) presented new-features-based logistic regression (LR), linear classifier (LC), NB, DT, MLP neural networks, and SVM. In their experiments, each technique produced a different output. Data mining by evolutionary learning (DMEL) could show the reason or probability of a churning phenomenon; DT, however, could only show the reason. LR, NB, and MLP could provide probabilities of different customer behaviors. LC and SVM could distinguish between a churner and a non-churner. (1)

Keramati et al. (2014) not only presented different approaches to data mining and classification methods such as DT, neural networks, SVM, and k-nearest neighbors, but also had the performances of these approaches compared. They analyzed, as a case study, data from an Iranian mobile company. (3)

2 Data Preparation

Data Preparation is a crucial stage in the customer churn prediction project, where we carefully clean, preprocess, and organize the raw bank dataset to make it suitable for further analysis and model building. This phase involves several key steps:

2.1 Loading and Exploration



	RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Ex
0	1	15634602	Hargrave	619	France	Female	42	2	0.00	1	1	1	101348.88	
1	2	15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	
2	3	15619304	Onio	502	France	Female	42	8	159660.80	3	1	0	113931.57	
3	4	15701354	Boni	699	France	Female	39	1	0.00	2	0	0	93826.63	
4	5	15737888	Mitchell	850	Spain	Female	43	2	125510.82	1	1	1	79084.10	
5	6	15574012	Chu	645	Spain	Male	44	8	113755.78	2	1	0	149756.71	
6	7	15592531	Bartlett	822	France	Male	50	7	0.00	2	1	1	10062.80	

Figure 1: Dataset

We start by loading the dataset into our data analysis environment. This allows us to understand the structure of the data, including the number of rows and columns, and the types of variables present. Exploring the dataset helps us gain initial insights into the data's characteristics and potential challenges we may encounter during data preparation.

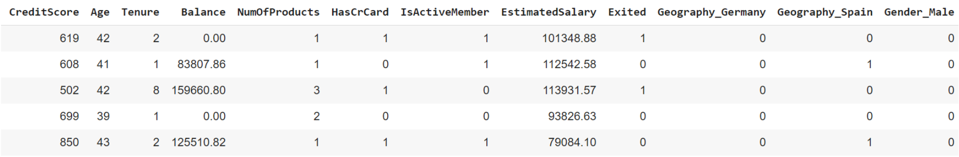
2.2 Data Preprocessing

Data preprocessing is an essential step in data preparation that addresses various data quality issues. This includes:

1. Handling Missing or Null Values: We identify any missing or null values in the dataset and decide on appropriate strategies to handle them. Depending on the extent of missing data and the nature of the variables, we may choose to impute missing values using methods like mean, median, mode, or advanced imputation techniques.

2. Dealing with Categorical Variables: Since many machine learning algorithms require numerical inputs, we convert categorical variables into numerical representations. To avoid the dummy variable trap, we may use techniques like one-hot encoding or label encoding.

3. Feature Scaling: To ensure fair treatment of different features during model training, we standardize the numerical features, bringing them to a similar scale.



CreditScore	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited	Geography_Germany	Geography_Spain	Gender_Male
619	42	2	0.00	1	1	1	101348.88	1	0	0	0
608	41	1	83807.86	1	0	1	112542.58	0	0	1	0
502	42	8	159660.80	3	1	0	113931.57	1	0	0	0
699	39	1	0.00	2	0	0	93826.63	0	0	0	0
850	43	2	125510.82	1	1	1	79084.10	0	0	1	0

Figure 2: Standardization

4. Data Splitting: Before model building, we split the dataset into dependent (target) and independent (predictor) variables. This separation allows us to train the model on one portion of the data and evaluate its performance on another unseen portion.

By effectively conducting data preparation, we can ensure that the data is ready for further analysis and model building. This will help us build an accurate and reliable customer churn prediction model, enabling the bank to proactively retain valuable customers and make data-driven decisions to improve customer satisfaction and loyalty.

3 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a critical phase in our customer churn prediction project that aims to gain deeper insights into the bank dataset and uncover meaningful patterns, relationships, and trends within the data. Through EDA, we explore and visualize the data to better understand its characteristics and potential factors that influence customer churn.

3.1 Churn Analysis

During the churn analysis, we focus on understanding the proportion of customers who have left the bank (churned) compared to those who have stayed. This analysis allows us to determine the churn rate, which is a fundamental metric for assessing customer retention performance. By identifying the churn rate, we can understand how many customers are leaving our business. This helps us measure the seriousness of the customer churn problem and its impact on our company's operations.

Additionally, our investigation into the relationship between customer churn and various independent variables (potential predictors) has enabled us to identify significant features that may influence customer churn. Understanding these predictors is vital for building an effective predictive model capable of accurately identifying customers at risk of churning. This model will empower the bank to take proactive measures and implement targeted retention strategies to improve overall customer satisfaction and reduce churn rates.

3.2 Visualizing Customer Behavior

Visualizations play a crucial role in our Exploratory Data Analysis (EDA), enabling us to understand complex relationships and patterns in the data. We have created various visual representations, such as count plots, pair plots, and hit map, to explore the distribution of customer churn across different categories. One aspect we focused on is the churn count in relation to gender, allowing us to determine if gender plays a role in customer churn. Additionally, we examined churn patterns across different countries, providing valuable insights into potential regional variations in customer behavior.

Moreover, our analysis extended to other relevant features, such as age, account balance, customer tenure, credit score, estimated salary, number of products, etc... Visualizing these relationships has helped us uncover hidden patterns and draw actionable conclusions.

The outcomes of our exploratory data analysis will guide us in selecting the most relevant features to be used as independent variables in our predictive model. With a comprehensive understanding of the data, we can build a robust customer churn prediction model that will assist the bank in implementing targeted strategies to retain customers, reduce churn, and enhance overall customer satisfaction.

4 Model Building

Model Building is a crucial phase in our customer churn prediction project, where we develop a predictive model using machine learning algorithms to identify potential churners and make accurate predictions based on the dataset's features.

4.1 Data Splitting and Standardization

Before building the model, we splited the dataset into two parts: the training set and the testing set. The training set is used to train the machine learning model, while the testing set is used to evaluate the model's performance on unseen data.

Additionally, to ensure fair and effective model training, we standardized the data. Standardization involves scaling numerical features to have a mean of 0 and a standard deviation of 1. This step is essential for certain machine learning algorithms, as it prevents features with larger scales from dominating the model during training.

4.2 Random Forest Classifier

For this project, we will employed the Random Forest Classifier, a powerful ensemble learning algorithm. The Random Forest builds multiple decision trees during training and combines their predictions to make more accurate and robust predictions. This algorithm is particularly well-suited for classification tasks like customer churn prediction, as it can handle both numerical and categorical data and can capture complex relationships between features.

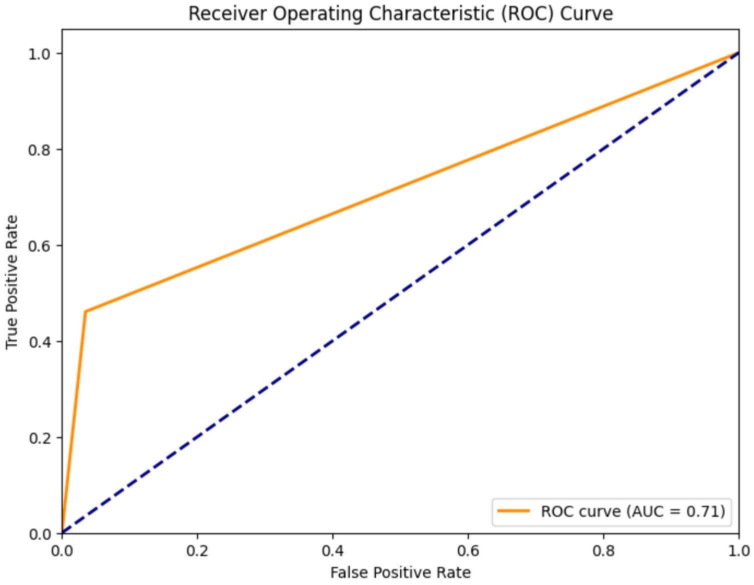
We trained the Random Forest Classifier on the training set, fine-tune its hyperparameters, and optimize its performance to achieve the best possible predictions.

4.3 Model Evaluation Metrics

After training the model, we evaluate its performance using various metrics to assess its predictive capabilities. Some of the key evaluation metrics we employ include:

1. Receiver Operating Characteristic (ROC) Curve: The ROC curve plots the true positive rate (sensitivity) against the false positive rate (1-specificity) at various classification thresholds. It provides a graphical representation of the model's ability to discriminate between churn and non-churn customers.

2. Area Under the Curve (AUC) Score: The AUC score quantifies the overall performance of the model. A higher AUC score indicates better discrimination and model performance.

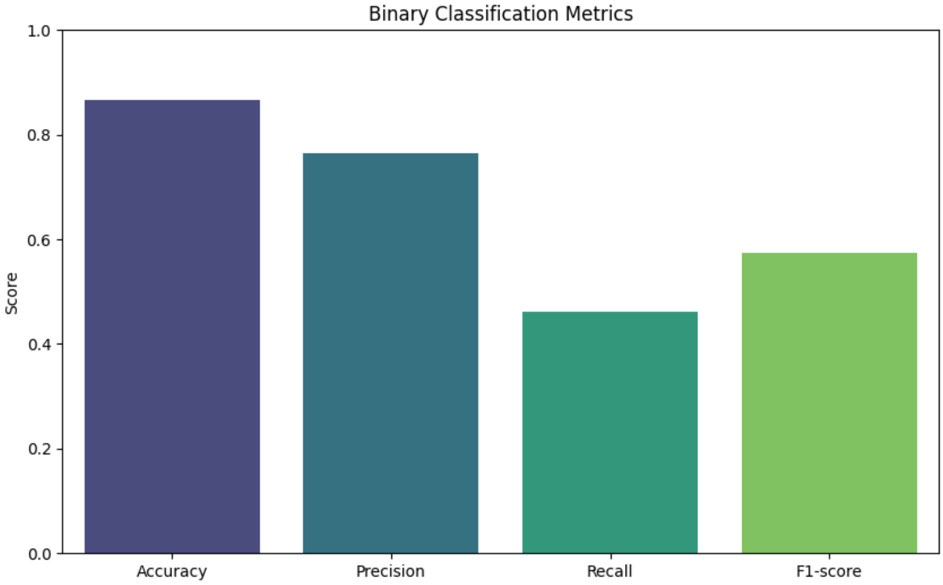


3. Accuracy: The accuracy metric measures the proportion of correct predictions made by the model.

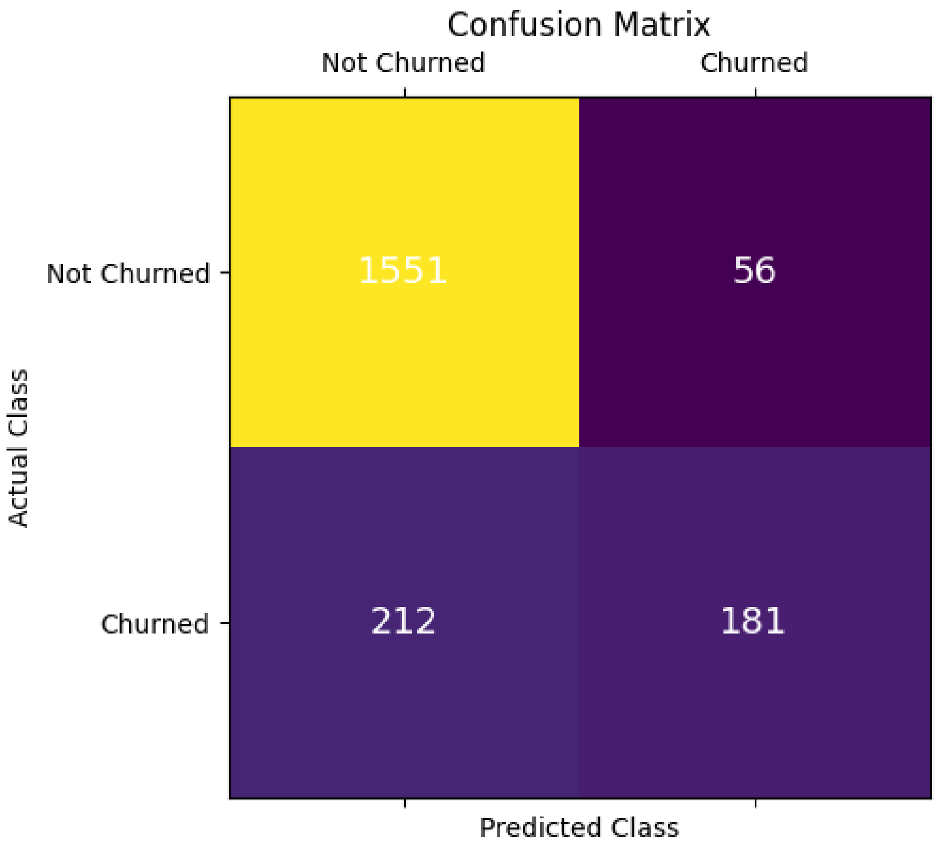
4. Precision: Precision measures the proportion of true churners among the total predicted churners. It indicates the model's ability to avoid false positives.

5. Recall (Sensitivity): Recall measures the proportion of actual churners correctly identified by the model. It indicates the model's ability to avoid false negatives.

6. F1-score: The F1-score is the harmonic mean of precision and recall and provides a balanced evaluation of the model's performance.



7. Confusion Matrix: The Confusion Matrix is a table that presents the model's predictions, showing true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). It provides a comprehensive view of how well the model correctly classifies churn and non-churn instances.



The model building phase was a critical step in our customer churn prediction project. By using the Random Forest Classifier and evaluating its performance with various metrics, we aimed to create a reliable predictive model that can help the bank proactively retain customers and improve customer satisfaction.

5 Conclusion

In conclusion, we successfully developed a customer churn prediction model for the bank dataset using the Random Forest Classifier. Through data preparation and exploratory data analysis, we gained valuable insights into customer behavior and identified potential factors influencing churn. The model's evaluation metrics, such as ROC curve, AUC score, accuracy, precision, recall, and F1-score, demonstrated its effectiveness in predicting customer churn. Armed with this predictive model, the bank can proactively retain customers, improve customer satisfaction, and optimize its services for better business outcomes. In our collaborative project, each team member played a vital role in contributing their effort to achieve our objectives.

Group Organization:
Google Colaboratory: Ezginur Cendik
LaTex Document Writer: Mustafa Kemal Kaya
GitHub: Ezginur Cendik, Mustafa Kemal Kaya

References

[1] Chiun-Sin Lin, Gwo-Hshiung Tzeng, and Yang-Chieh Chin. Combined rough set theory and flow network graph to predict customer churn in credit card accounts. *Expert Systems with Applications*, 38(1):8–15, 2011.

[2] Guangli Nie, Wei Rowe, Lingling Zhang, Yingjie Tian, and Yong Shi. Credit card churn forecasting by logistic regression and decision tree. *Expert Syst. Appl.*, 38(12):15273–15285, 2011.

[3] Guangli Nie, Wei Rowe, Lingling Zhang, Yingjie Tian, and Yong Shi. Credit card churn forecasting by logistic regression and decision tree. *Expert Systems with Applications*, 38(12):15273–15285, 2011.

[4] Guo-en XIA and Wei-dong JIN. Model of customer churn prediction on support vector machine. *Systems Engineering - Theory Practice*, 28:71–77, 01 2008.