An Intelligent Approach to Predict the Customer Churn: A Case Study of a Bank

***Abstract*—The modern age of information and communication technology has played a catalytic role in many industries. This includes, but not limited to, banking, education, government, production industries, and services industries. Specifically, the banking sector is the major beneficiary of this advancement. Apparently, the scenario looks benefiting. However, the gigantic rise in the customer data has created a pressing need to device intelligent methods to handle this massive data and transform it into knowledge. Artificial intelligence is reported to the best candidate to transform the massive data into actionable knowledge. The customer churn prediction is one of the frequent issues in the banking sector. However, the exhaustive exploration is this the research gape into the literature. This study has presented an intelligent approach based on the logistic regression model for customer churn prediction for a banking sector. The recent benchmark dataset has been employed to train the machine learning algorithm. In addition, the performance of the algorithm is also measured as the function of standard indicators.**

Keywords: Machine Learning, logistic regression, customer churn prediction, unstructured data, Artificial Intelligence, information retrieval

I. INTRODUCTION

Our study expands to take into account the changing landscape of technology and data analytics as we dig deeper into the topic of an intelligent way to predict client turnover within the banking sector. The growing dependence on digital platforms and the volume of data generated by customer interactions with financial institutions present an unparalleled chance to leverage artificial intelligence. Our study considers the dynamic nature of consumer expectations, industry trends, and economic effects in addition to closely examining the nuances of predictive modeling. By taking a holistic approach, we hope to provide insightful information about the larger context in which these predictive models function in addition to creating a strong predictive framework. Through this comprehensive examination, we strive to contribute to the ongoing discourse on leveraging intelligent solutions for mitigating customer churn in the banking industry.

Analyzing customer churn prediction entails predicting which consumers are most likely to stop using a product or service by applying machine learning and data analysis approaches. The procedure includes gathering data, preprocessing, and feature engineering, with an emphasis on finding pertinent variables affecting churn. Important phases include choosing a model, training it, and evaluating it. Selected models are then deployed into real-world settings, and updates are continuously monitored. For strategic decision-making to be

successful, churn factors must be interpreted, and their integration with company objectives guarantees the successful execution of initiatives such as customized customer interaction or targeted marketing campaigns. The study's scope includes ethical factors such as responsible data utilization and client privacy.

Customer churn prediction represents a widely employed use case of machine learning and data science, bearing significant relevance across diverse industries. This entails applying statistical and data mining methods to build models capable of foreseeing or categorizing customer behavior, thereby empowering businesses to formulate strategies focused on managing churn and enhancing retention. The bank sector, in particular, has derived substantial value from churn prediction by using it to anticipate potential customer attrition and develop retention tactics. Moreover, companies of all sizes, including startups, have harnessed churn prediction to bolster customer retention. The implementation process for customer churn prediction encompasses stages such as data transformation, modeling, testing, deployment, and monitoring.

The search results do not readily yield research dimensions specifically focused on customer churn prediction within the baking sector. Nonetheless, there are studies on customer churn prediction in other sectors, such as retail banking and general business, that offer valuable insights into the techniques and models used for predicting customer churn [6] These studies employ machine learning, predictive modeling, and logistic regression techniques to forecast customer behavior, like when customers are likely to stop buying or leave.

These models rank customers based on churn scores derived from the latest data, providing a foundation for developing a refined churn model that incorporates all

relevant variables pertaining to a customer's account with the baking firm.

II. Literature Review

The review of literature on intelligent methods for predicting customer churn in the banking industry underscores the importance of customer retention for banks. It is increasingly crucial for businesses to focus on managing customer relationships, as acquiring new customers can be up to five times as expensive as retaining current ones. [5] Predicting customer churn is essential for developing effective strategies to mitigate revenue loss and boost profits However, identifying churn in the banking sector poses challenges due to the large customer base and the necessity for precise predictive models and strategies to prevent customer attrition.

In 2016, Keramati, Ghaneei, and Mirmohammadi formulated a forecast model for customer churn from electronic banking services through data mining, underscoring the significance of customer retention for financial institutions. They asserted that a bank could potentially boost its profits by as much as 85% by enhancing the retention rate by up to 5%.[8]

Additionally, the assessment criteria employed for forecasting customer churn, including accuracy, precision, recall, and F1-score, were deliberated within the framework of data mining tools commonly utilized by numerous banks for predicting customer churn, researchers have also devoted attention to creating prediction models for customer churn by employing data mining and machine learning methods. For example, a study investigated the application of a distinctive customer-level dataset from a major Brazilian bank to forecast customer churn. The study underscored the significance of customer retention for banks and the opportunity to enhance profitability by boosting the retention rate [7].

Researchers have also devoted attention to creating prediction models for customer churn by employing data mining and machine learning methods. For example, a study investigated the application of a distinctive customer-level dataset from a major Brazilian bank to forecast customer churn. The study underscored the significance of customer retention for banks and the opportunity to enhance profitability by boosting the retention rate.

Recent research on predicting customer churn using machine learning has focused on developing churn prediction models, employing machine learning for predicting churn rates, and utilizing advanced deep learning techniques for customer churn prediction. These studies underscore the importance of comprehending customer data, preparing and processing data, and leveraging a variety of machine learning algorithms to uncover patterns in customer behavior and forecast churn. Furthermore, the literature explores the use of deep learning and artificial neural networks to capture intricate relationships among variables that impact customer churn. In addition, the research emphasizes the critical nature of taking early action to retain customers and the potential of machine learning in foreseeing churn [9,10,11]

Customer churn prediction studies often use a variety of techniques, including logistic regression, machine learning algorithms, and data mining. These methods are applied to analyze customer behavior and predict churn, especially in industries like retail banking. Their strengths lie in their ability to effectively model and analyze complex customer data, identify behavior patterns, and offer insights for developing customer retention strategies. However, they also have limitations, such as the potential for overfitting in machine learning models, the need for high-quality and comprehensive data, and the challenge of interpreting certain machine learning algorithms' black-box nature.

Methods

1. Logistic Regression: This approach was employed in a retail banking case study to forecast customer churn by analyzing the features of churning and non-churning customers. Logistic regression, a statistical method suited for binary classification tasks, was found to be applicable for customer churn prediction.

2. Machine Learning Algorithms: Various machine learning methods, including decision trees, support vector machines, and artificial neural networks, have been suggested for predicting customer churn. These algorithms utilize historical customer data to recognize patterns and make forecasts about potential churn.

3. Data Mining: In the domain of electronic banking services, data mining techniques have been utilized to construct prediction models for customer churn.

Key Strengths

- Predictive Capability: These approaches excel in effectively forecasting customer churn using historical data and customer behavior patterns.

- Guiding Business Tactics: The insights obtained from these approaches can guide the formulation of customer retention strategies and significantly contribute to the overall financial stability of organizations, especially within the banking sector.

- Scalability: Machine learning algorithms and data mining techniques can be expanded to handle large volumes of customer data, enabling the analysis of extensive datasets to identify potential churn indicators.

.

Methodology

The methodology is designed to facilitate the prediction of customer churn in a banking environment. By adhering to the specified steps, the bank can proactively forecast and prevent customer churn, ultimately leading to enhanced customer retention and overall business success.

The initial phase of the methodology entails the identification of suitable data sources, which may encompass customer transactional data, demographic information, and historical churn data. It is crucial to assemble dependable and pertinent data to ensure precise churn prediction.

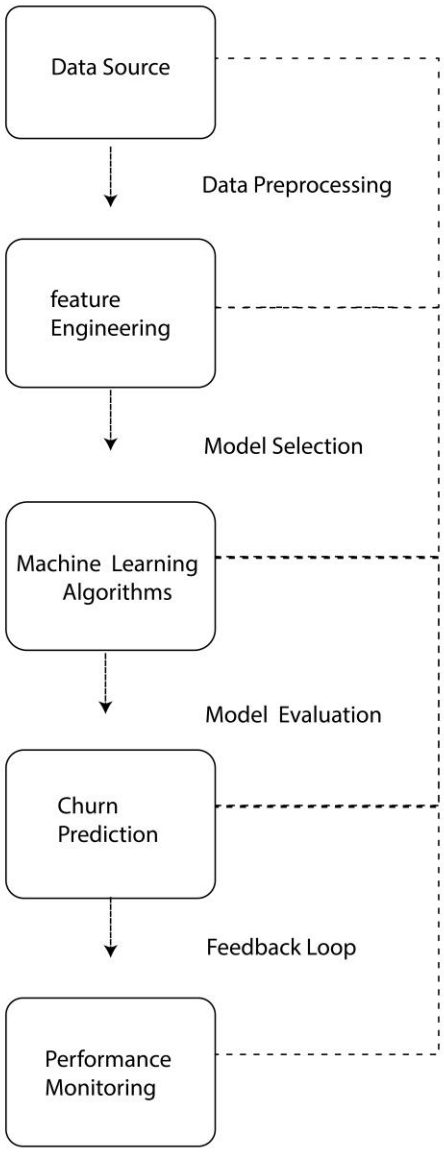
The thorough process of model selection entails comparing and assessing various algorithms. This process aids in identifying algorithms that demonstrate optimal performance and are aligned with the specific requirements and characteristics of the churn prediction problem within the banking domain.

One widely used method for model selection involves evaluating a broad spectrum of machine learning algorithms. This could encompass linear regression, logistic regression, decision trees, random forests, support vector machines (SVM), gradient boosting algorithms such as XGBoost or LightGBM, and neural networks. Each algorithm possesses distinct strengths, weaknesses, and underlying assumptions. Therefore, it is crucial to explore multiple options in order to identify the most suitable one for the task of predicting churn.

To gauge the performance of the chosen model(s), model evaluation is conducted. This crucial step involves using appropriate evaluation metrics such as accuracy, precision, recall, and F1 score. The evaluation offers insights into how effectively the model can forecast customer churn.

To achieve ongoing improvement, the methodology incorporates a feedback loop. This includes monitoring the model's performance in real-world situations and actively soliciting feedback on its accuracy. Any essential adjustments or improvements are implemented based on the insights obtained from this feedback loop.

During the process of selecting a model, various factors must be considered. One such factor is the interpretability of the algorithm. In certain instances, the ability to interpret and explain the predictions made by the model is vital, particularly in fields such as banking where regulatory compliance and transparency hold significance. In these scenarios, algorithms like logistic regression or decision trees may be favored due to their inherent interpretability.



1.1 *Dataset*

The dataset "Churn_Modelling.csv" comprises a comprehensive collection of customer information, enabling a thorough analysis of various factors related to customer behavior and churn. It encompasses a range of essential columns, including unique identifiers such as "Row Number," "Customer Id," and "Surname," allowing for individual customer tracking. Demographic aspects such as "Geography" and "Gender" provide insight into the distribution and diversity of customers across different countries and genders. Financial details like "Credit Score," "Balance," and "Estimated Salary" offer valuable indicators of customers' financial health and stability. These variables can be utilized to assess creditworthiness, identify patterns in customers' financial behavior, and potentially uncover correlations between financial status and churn. Additionally, the "Num Of Products" column indicates the number of products utilized by each customer, shedding light on their level of engagement with the company's offerings.

The dataset also includes specific variables related to customer engagement and loyalty. The "Tenure" column reveals the duration of the customer's relationship with the company, while "Is Active Member" and "Has Cr Card" provide insight into their level of activity and credit card usage. These factors can play a crucial role in understanding customer loyalty and identifying potential churn risks.

Furthermore, the dataset includes subjective measures such as the "Satisfaction Score." Analyzing customer satisfaction levels can help identify areas for improvement and assist in developing strategies to enhance overall customer experience and reduce churn rates. The "Card Type" column indicates the type of card held by customers, such as DIAMOND, GOLD, SILVER, or PLATINUM, providing insight into the customer segmentation based on card benefits and privileges. The "Point Earned" column represents the number of points earned by customers, which can serve as an indicator of their engagement and loyalty.

By delving into the various attributes within the dataset, businesses can gain a comprehensive understanding of customer behavior, identify potential churn risks, and develop targeted retention strategies. Analyzing demographic information, financial indicators, engagement levels, satisfaction scores, and card types can provide valuable insights for businesses to enhance their customer retention efforts, improve customer experience, and ultimately drive business growth.

1.2. Data preprocessing

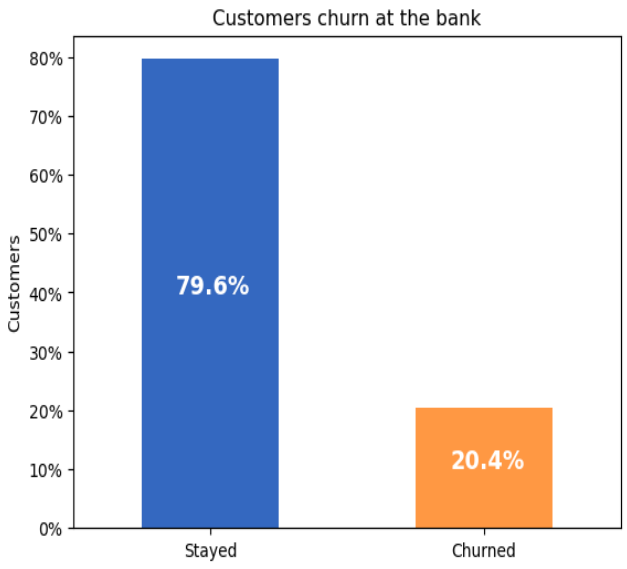
Data preprocessing plays a critical role in preparing the dataset for analysis and modeling. Initially, irrelevant columns like "Row Number," "Customer Id," and "Surname" are excluded as they do not add value to the analysis.

The "Row Number" serves as an arbitrary label for individual rows within the dataset, lacking meaningful information for analysis. Likewise, the "Customer Id" functions as a unique identifier for each customer but is not pertinent in this scenario. The "Surname" pertains to customers' last names, which may not significantly impact the analysis at hand.

Table 1: Variables Utilized in the Analysis for Predicting Bank Customer Churn

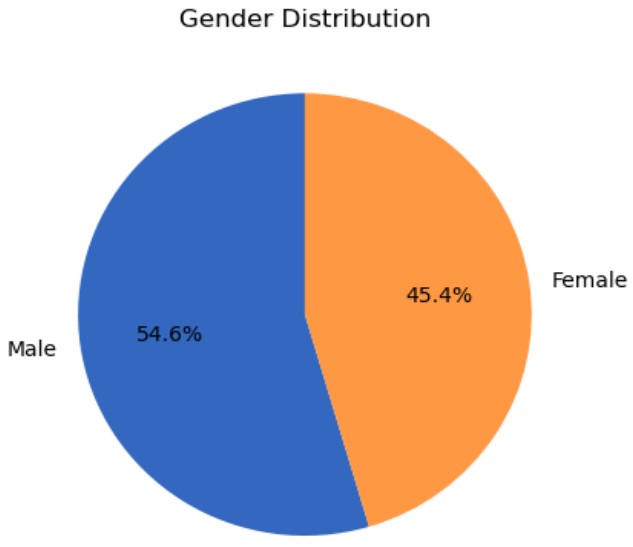
Column Name	Description
CreditScore	The credit score of the customer
Geography	The geographical location of the customer (France, Spain, Germany)
Gender	The gender of the customer (Female, Male)
Age	The age of the customer
Tenure	The number of years the customer has been with the company
Balance	The account balance of the customer
NumOfProducts	The number of products used by the customer
HasCrCard	Indicates whether the customer has a credit card (1 = Yes, 0 = No)
IsActiveMember	Indicates whether the customer is an active member (1 = Yes, 0 = No)
EstimatedSalary	The estimated salary of the customer
Exited	The churn status of the customer (1 = Exited, 0 = Active)
Complain	Indicates whether the customer has made a complaint (1 = Yes, 0 = No)
Satisfaction Score	The satisfaction score of the customer
Card Type	The type of card held by the customer (DIAMOND, GOLD, SILVER, PLATINUM)
Point Earned	The number of points earned by the customer

1.3 Comprehending the Data.



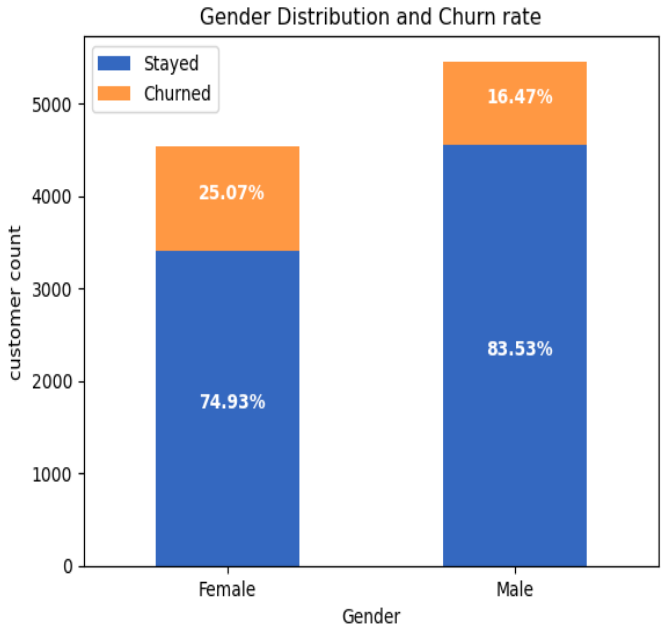
Graph 1: Distribution of customer Churn at bank

The chart illustrates customer retention at a bank, showing that 79.6% of customers stayed with the bank while 20.4% left. This indicates a relatively strong customer retention rate, suggesting high levels of loyalty and satisfaction among most customers. However, the churn rate reveals a significant number of customers decided to end their relationship with the bank, underscoring the need to identify the reasons behind churn and implement effective retention strategies to enhance customer satisfaction and loyalty.



Graph 2: Gender Distribution

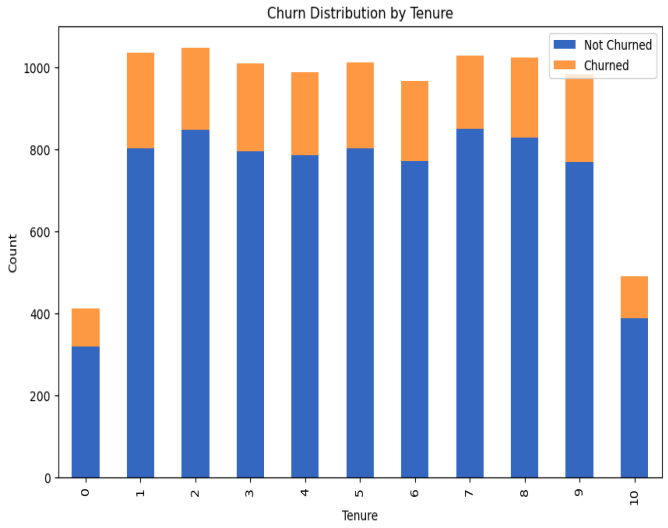
The pie chart visually displays the distribution of genders, with two distinct categories represented by separate slices. The blue slice, accounting for 54.6% of the total, represents the percentage of men in the population. Conversely, the orange slice, making up 45.4% of the distribution, symbolizes the percentage of women.



Graph 3: Gender Distribution and Churn Rate

Based on the chart, it is evident that most of the customer base is male (54.6%) rather than female (45.4%). Additionally, the churn rate is higher for male customers (25.07%) than for female customers (16.47%), indicating that a larger percentage of male customers have opted to end their relationship with the bank. Concerning customer retention, the graph demonstrates that a higher proportion of male customers (74.93%) chose to remain, in contrast to female customers (83.53%).

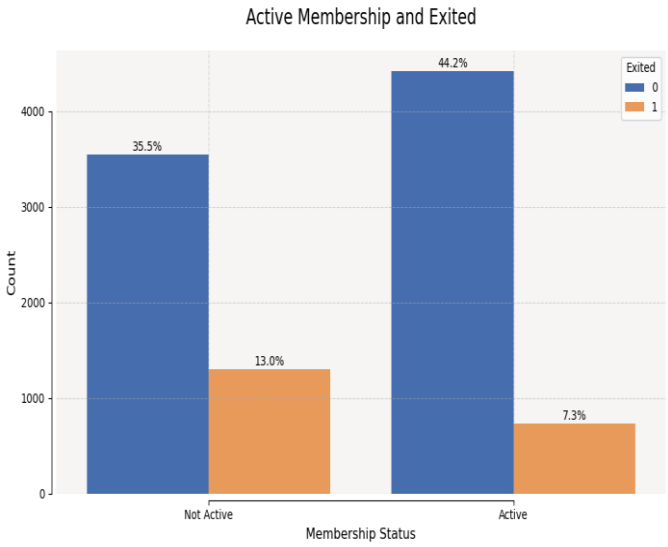
The chart illustrates the gender distribution of the customer base, indicating a greater presence of male customers. It also indicates a higher rate of customer churn among males compared to females, suggesting that a larger proportion of male customers have opted to end their association with the bank. Nonetheless, a higher percentage of male customers still opt to remain with the bank compared to females. These findings underscore the significance of gender-specific strategies to enhance customer retention and address the factors influencing churn, thereby improving overall customer satisfaction and loyalty.



Graph 4: Churn Distribution by Tenure

The plot aims to visually display the correlation between customer churn and their length of time with the bank. Examining the distribution of customer churn over various tenure periods allows for a deeper insight into how the duration of customer relationships with the bank influences their probability of churning.

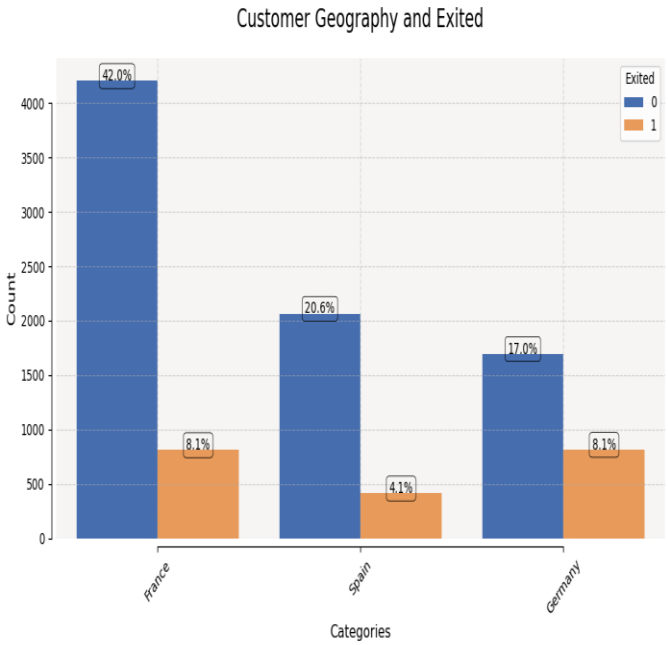
Analyzing churn is essential for customer retention strategies, enabling businesses to pinpoint potential churn risks and implement proactive measures to reduce customer attrition.



. Graph 5: Active Membership and Exited

The chart offers insights into customer retention based on their level of activity. It divides customers into "Active" (regular users) and "Not Active" (infrequent users). By analyzing retention rates in each group, clear patterns in customer behavior emerge.

For "Not Active" customers, approximately 35.5% continue using the service, while 13.0% opt to discontinue. In contrast, among active customers, a higher percentage—44.2%—remain, with only 7.3% choosing to leave. This indicates that engaged customers are more likely to stay, while less involved customers are more prone to churn.



.

. Graph 5: Active Membership and Exited

The analysis of customer churn prediction at a bank focused on data from various countries, providing valuable insights like observing who remained at a party and who left early. Across France, Spain, and Germany, distinct patterns emerged.

France: The analysis notably revealed that the highest retention rate was observed in France, with 42.0% of customers choosing to stay. This indicates a positive reception of the bank's offerings among its French clientele. However, it's important to note that despite the majority staying, approximately 8.1% still chose to leave, suggesting varying levels of satisfaction or potential competition.

Spain and Germany: Unlike Spain, Germany also showed lower retention rates, with approximately 20.6% and 17.0% of customers staying, respectively. However, despite these differences, the churn rate, which indicates the proportion of customers leaving, was almost equal in both countries at 4.1%. This similarity implies that similar factors may play a role in customer attrition in Spain and Germany, despite their varying overall retention rates.

Model building.

In the field of research, logistic regression is a fundamental tool for analyzing binary outcomes, such as predicting customer churn in the banking industry. This statistical approach works by estimating the probability of a binary outcome (e.g., churn or no churn) based on a set of independent variables. Using the logistic regression model, coefficients are calculated to indicate the impact of each independent variable on the log-odds of the dependent variable, helping to interpret the factors that influence customer behavior. By training the model with historical data, researchers can uncover patterns and relationships between customer characteristics and churn, thereby empowering banks to proactively identify at-risk customers and implement targeted retention strategies.

In logistic regression, the binary nature of the dependent variable, denoted as Y, distinguishes between instances of churn (Y=1) and non-churn (Y=0). Through a mathematical model, logistic regression estimates the probability ($p(Y=1)$) of churn transpiring based on a set of independent variables ($X1$, $X2$,...,Xn). This estimation is achieved by applying coefficients ($\beta 0$, $\beta 1$, $\beta 2$,..., βn) to a logistic function, yielding a predictive equation that evaluates the likelihood of churn given the values of the independent variables. By uncovering the impact of each independent variable on the log-odds of churn occurrence, logistic regression offers a powerful tool for understanding and predicting customer behavior, thereby enabling businesses to implement.

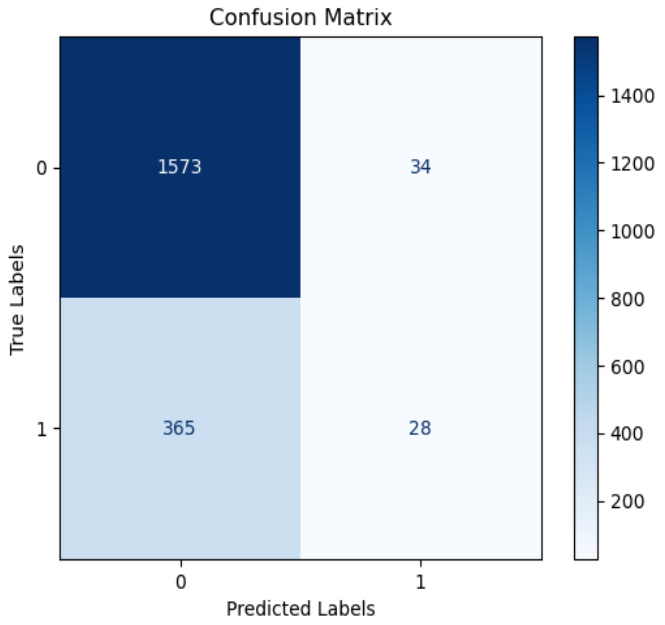
$$p(Y=1) = \frac{1}{1 + e^{(\beta 0 + \beta 1 X1 + \beta 2 X2 + \ldots + \beta n Xn)}}$$

.

In general, logistic regression is a powerful statistical method designed to analyze binary outcomes, and it provides a strong framework for forecasting customer churn in the banking sector. By understanding its basic principles and real-world uses, individuals can uncover important insights into customer behavior, enabling them to develop data-driven strategies to reduce churn and enhance customer retention efforts.

Model building.

The evaluation of the logistic regression model's performance in predicting churn presents both encouraging insights and opportunities for enhancement. With an accuracy of about 80%, the model displays a laudable capability to accurately categorize instances into churn and non-churn groups. However, a more detailed analysis of the confusion matrix highlights subtle challenges. Notably, while the model correctly identifies a significant number of true positives (1573 instances accurately predicted as churn), it also shows a worrisome number of false negatives (365 instances inaccurately classified as non-churn). This indicates a potential weakness in the model's ability to detect churn patterns, which could impede its efficacy in proactive intervention strategies aimed at retaining customers.



.

Furthermore, the occurrence of false positives (34 instances incorrectly identified as churn) emphasizes the model's inclination to misclassify non-churn cases, potentially resulting in unnecessary resource allocation or customer dissatisfaction. These findings underscore the necessity of enhancing the model's predictive capabilities, particularly in improving sensitivity to churn indicators while minimizing false alarms. Addressing these challenges could not only strengthen the model's overall accuracy but also enhance its practical usefulness in guiding targeted retention efforts and optimizing business outcomes. Therefore, additional research and iterative model refinement are needed to realize the full potential of churn prediction in driving sustainable customer engagement and business growth.