FINAL PROJECT

IST – 718

BIG DATA ANALYTICS

FALL 2023

BANK CUSTOMER CHURN PREDICTION

Group Members:

Collin Taylor
Emine Kakillioglu
Hendi Kushta
Sowmeya Maruthamuthu Kanagarathinam

1. Introduction

In the dynamic landscape of the financial industry, understanding and mitigating customer churn has become a pivotal challenge for banks worldwide. Customer Churn Prediction, the process of identifying customers likely to leave or unsubscribe from services, has emerged as a crucial aspect of customer relationship management. For financial institutions, the repercussions of high churn rates are far-reaching, impacting growth rates, sales, and overall profitability. Recognizing the significance of retaining existing customers over acquiring new ones, our project delves into the realm of predictive analytics using Apache Spark to develop a robust model for bank customer churn prediction.

1.1. The Impact of Customer Churn on Businesses

A high churn rate translates to a loss of subscribers, impeding growth and exerting a pronounced influence on sales and profits. In contrast, organizations adept at minimizing churn can cultivate customer loyalty, fostering prolonged and profitable relationships. The fundamental metric defining business success or failure, customer churn, underscores the imperative to retain existing clientele, considering the cost-effectiveness of selling to established customers over acquiring new ones.

1.2. Why Analyzing Customer Churn Prediction is Crucial

The essence of our project lies in acknowledging that customer retention significantly outstrips customer acquisition in terms of cost efficiency. Successful retention enhances the average lifetime value of a customer, rendering future sales more valuable and improving unit margins. The strategic reallocation of resources towards increasing revenue from recurring subscriptions and trusted repeat business, as opposed to constant customer acquisition endeavors, serves as the cornerstone for maximizing a company's long-term success and resilience.

1.3. Benefits of Analyzing Customer Churn Prediction

Increase Profits: The primary objective of churn analysis is to reduce customer churn, subsequently increasing profits. Prolonged customer relationships contribute to heightened revenue and improved profitability.
Improve Customer Experience: By understanding the reasons behind customer churn, businesses can identify weaknesses, rectify mistakes, and enhance the overall customer experience, thereby bolstering customer loyalty.
Optimize Products and Services: Insights gleaned from customer churn data provide an opportunity to refine products, services, or shipping methods, thus mitigating issues that contribute to churn and facilitating sustainable growth.
Customer Retention: Beyond mere churn reduction, the goal is customer retention. Building strong customer loyalty enables businesses to maximize the profitability of existing customers and extend their lifetime value.
How Customer Churn Prediction Works:

The project methodology involves a comprehensive process, commencing with Exploratory Data Analysis (EDA) on the dataset. Subsequently, we employ state-of-the-art Machine Learning Classification Algorithms to discern patterns and relationships within the data, selecting the most effective algorithm for our Bank Customer Churn Dataset. Our ultimate objective is to leverage the power of Apache Spark to develop a predictive model capable of identifying potential churners among bank customers. This predictive model will not only uncover the contributing factors to churn but also

explore the feasibility of assigning probabilities to churn predictions, enabling targeted and strategic customer retention efforts.
In essence, our project aims to empower financial institutions with the tools to proactively manage customer churn, fostering enduring customer relationships and, in turn, driving sustained business success.

2. Data Description

The dataset used for this project comprises information about bank customers, encompassing various attributes such as credit score, geography, gender, age, tenure, balance, number of products, credit card status, active member status, estimated salary, and an indicator denoting whether the customer has exited the bank.
The dataset consists of 10,000 rows and 14 columns.
Link to the dataset is https://www.kaggle.com/datasets/shantanudhakadd/bank-customer-churn-prediction

2.1. Variable Categories:

I. Demographic Information about Customers:

customer_id: Customer ID
vintage: Vintage of the customer with the bank in a number of days
age: Age of the customer
gender: Gender of the customer
dependents: Number of dependents
occupation: Occupation of the customer
city: City of the customer (anonymized)

II. Customer Bank Relationship:

customer_nw_category: Net worth of the customer (3: Low, 2: Medium, 1: High)
branch_code: Branch Code for a customer account
days_since_last_transaction: Number of days since the last credit in the last year

III. Transactional Information:

current_balance: Balance as of today
previous_month_end_balance: End of Month Balance of the previous month
average_monthly_balance_prevQ: Average monthly balances (AMB) in the previous quarter
average_monthly_balance_prevQ2: Average monthly balances (AMB) in the previous-to-previous quarter
current_month_credit: Total Credit Amount in the current month
previous_month_credit: Total Credit Amount in the previous month
current_month_debit: Total Debit Amount in the current month
previous_month_debit: Total Debit Amount in the previous month
current_month_balance: Average Balance of the current month
previous_month_balance: Average Balance of the previous month
churn: Indicator denoting whether the average balance of the customer falls below the minimum balance in the next quarter (1/0)

3. Data Preprocessing

Data preprocessing is an essential phase in the journey from raw data to actionable insights. In this project, we undertook various preprocessing steps using PySpark to enhance the quality and usability of the dataset for subsequent analysis and machine learning model development.

1. Standardizing Column Names: To ensure consistency and avoid potential issues related to case sensitivity, all column names were converted to lowercase. This standardization simplifies subsequent data manipulations and promotes uniformity in the dataset.

2. Column Removal: Certain columns, namely 'rownumber', 'customerid', and 'surname,' were identified as non-contributory to the churn prediction task. Consequently, these columns were dropped from the dataset, reducing unnecessary complexity and streamlining the data for focused analysis.

3. Handling Missing Values: An integral part of data preprocessing involves identifying and addressing missing values. A thorough examination of the dataset revealed that none of the columns contained missing values. The table below summarizes the count of missing values for each column.

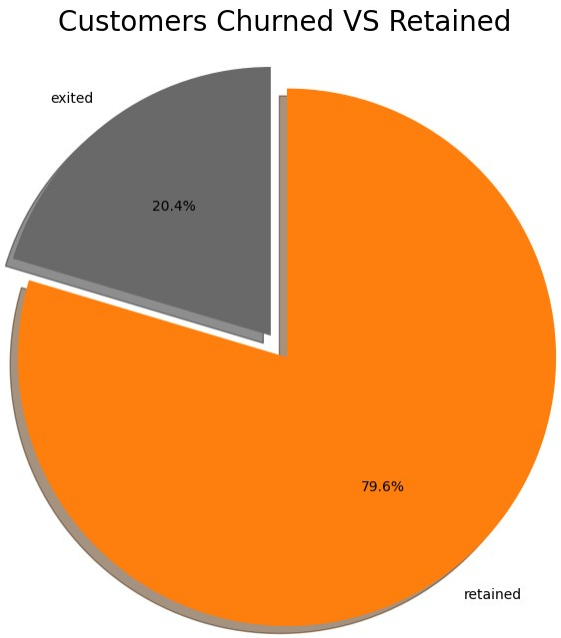
Column	Missing Values
creditscore	0
geography	0
gender	0
age	0
tenure	0
balance	0
numofproducts	0
hascrcard	0
isactivemember	0
estimatedsalary	0
exited	0

The absence of missing values underscores the completeness of the dataset, laying a solid foundation for subsequent analyses and predictive modeling. These preprocessing steps collectively contribute to the creation of a clean, standardized, and well-structured dataset for meaningful exploration of bank customer churn prediction.

4. Exploratory Data Analysis

Overall Customer Churn

Out of the total customers in the dataset, approximately 20% have exited, forming the baseline for anticipating customer churn.
Exited: 2037 customers
Not Exited: 7963 customers



Geography and Churn

France has the highest customer count, but the proportion of churned customers appears inversely related to the overall population of customers. This observation suggests potential challenges and areas for improvement in customer service resource allocation.

France:
- Exited: 810 customers
- Not Exited: 4204 customers

Germany:
- Exited: 814 customers
- Not Exited: 1695 customers

Spain:
- Exited: 413 customers
- Not Exited: 2064 customers

Gender and Churn

The proportion of female customers experiencing churn is higher than that of male customers, prompting further investigation into gender-specific factors influencing churn rates.

Female:
- Exited: 1139 customers
- Not Exited: 3404 customers

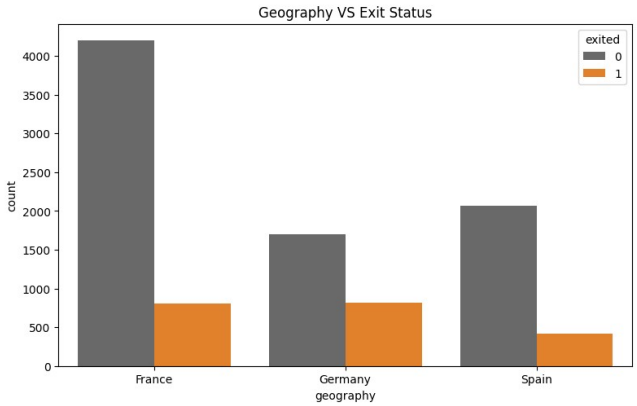
Male:
- Exited: 898 customers
- Not Exited: 4559 customers

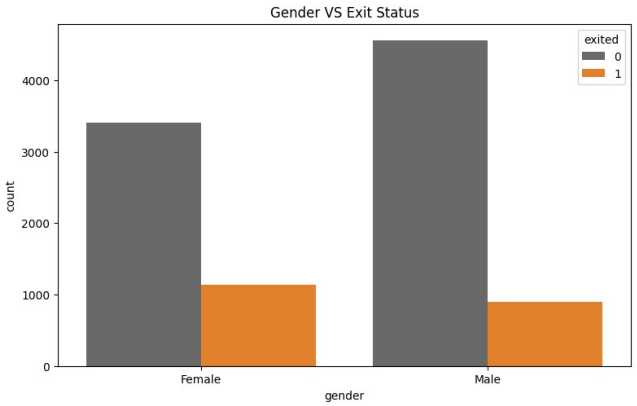
Credit Card Ownership and Churn

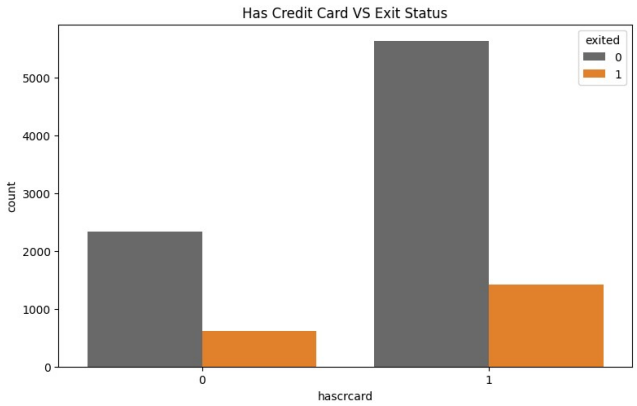
Despite the majority of customers having credit cards, the proportion of churned customers is notably higher among credit cardholders. This observation raises questions about factors influencing this trend.
No Credit Card:
    Exited: 613 customers
    Not Exited: 2332 customers
Has Credit Card:
    Exited: 1424 customers
    Not Exited: 5631 customers

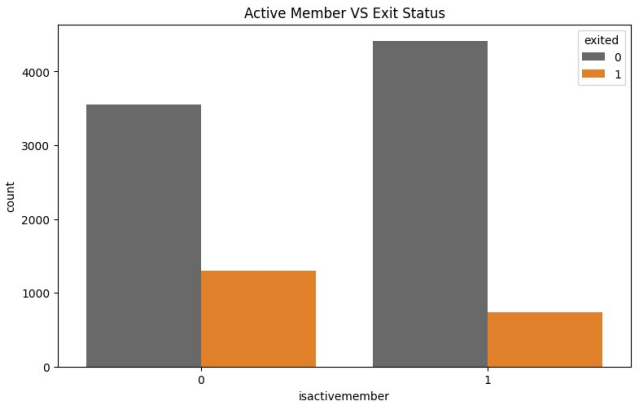
Active Membership and Churn

Inactive members exhibit a higher churn rate, emphasizing the need for initiatives to convert inactive members into active customers to mitigate churn.
Inactive:
    Exited: 1302 customers
    Not Exited: 3547 customers
Active:
    Exited: 735 customers
    Not Exited: 4416 customers









Credit Score and Age

The credit score distribution does not exhibit a significant variance between retained and churned customers. However, an interesting trend emerges concerning age: older customers demonstrate a higher likelihood of churning compared to their younger counterparts. This observation suggests a potential divergence in service preferences among different age groups. Consequently, the bank might consider reevaluating its target market or refining its retention strategies tailored to distinct age demographics.

Credit Score:

Churned customers:Minimum Credit Score: 350
Maximum Credit Score: 850
Mean Credit Score: 645.35
Retained customers:Minimum Credit Score: 405
Maximum Credit Score: 850
Mean Credit Score: 651.85

Age:

Churned customers:Minimum Age: 18
Maximum Age: 84
Mean Age: 44.84
Retained customers:Minimum Age: 18
Maximum Age: 92
Mean Age: 37.41

Tenure

In terms of tenure, customers at the extremes—those who have spent minimal time with the bank and those with a prolonged history—are more prone to churning compared to those with an average tenure. This highlights the importance of addressing the needs of both new and long-standing clients to enhance retention efforts.

Churned customers:Minimum Tenure: 0
Maximum Tenure: 10
Mean Tenure: 4.93
Retained customers:Minimum Tenure: 0
Maximum Tenure: 10
Mean Tenure: 5.03

Bank Balance

A concerning trend emerges regarding customers with substantial bank balances, as they show a higher propensity to churn. This could have implications for the bank's available capital for lending, urging a closer examination of strategies to retain customers with significant financial holdings.

Churned customers:Minimum Balance: 0.0
Maximum Balance: 250,898.09

Mean Balance: 91,108.54
Retained customers:Minimum Balance: 0.0
Maximum Balance: 221,532.8
Mean Balance: 72,745.30
Variety of Products and Estimated Salary

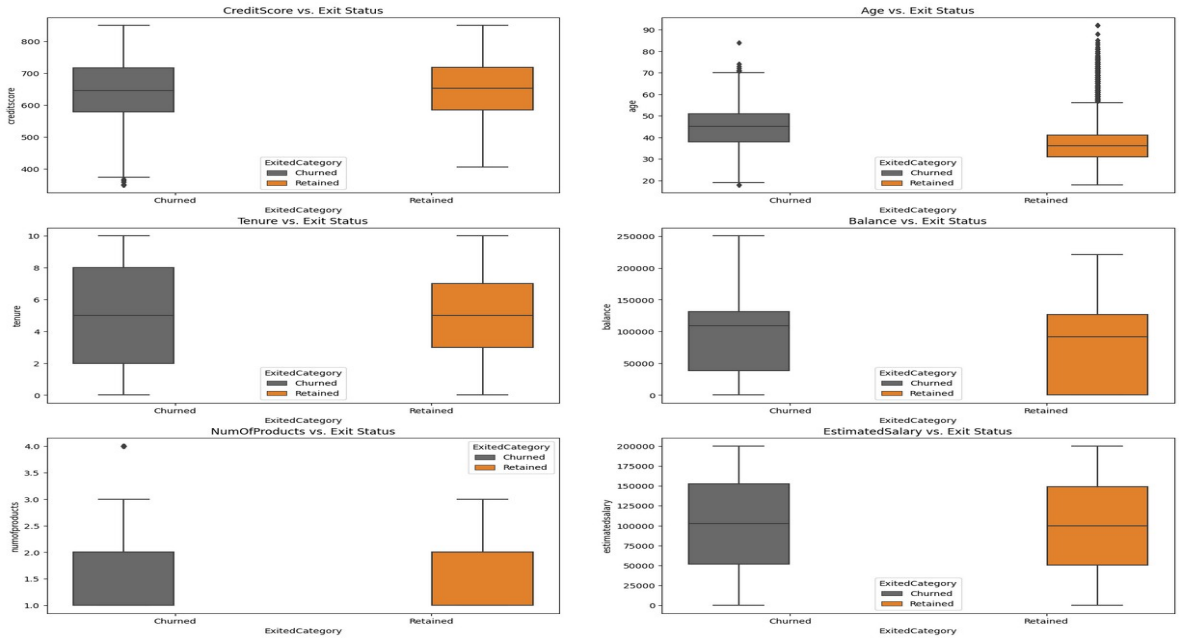
Surprisingly, neither the variety of products used nor the estimated salary demonstrates a substantial impact on the likelihood of churning. This insight suggests that other factors may play a more influential role in customer retention, prompting the need for a comprehensive analysis of additional variables to identify effective retention strategies.

Number of Products:

Churned customers:Minimum Number of Products: 1
Maximum Number of Products: 4
Mean Number of Products: 1.48
Retained customers:Minimum Number of Products: 1
Maximum Number of Products: 3
Mean Number of Products: 1.54

Estimated Salary:

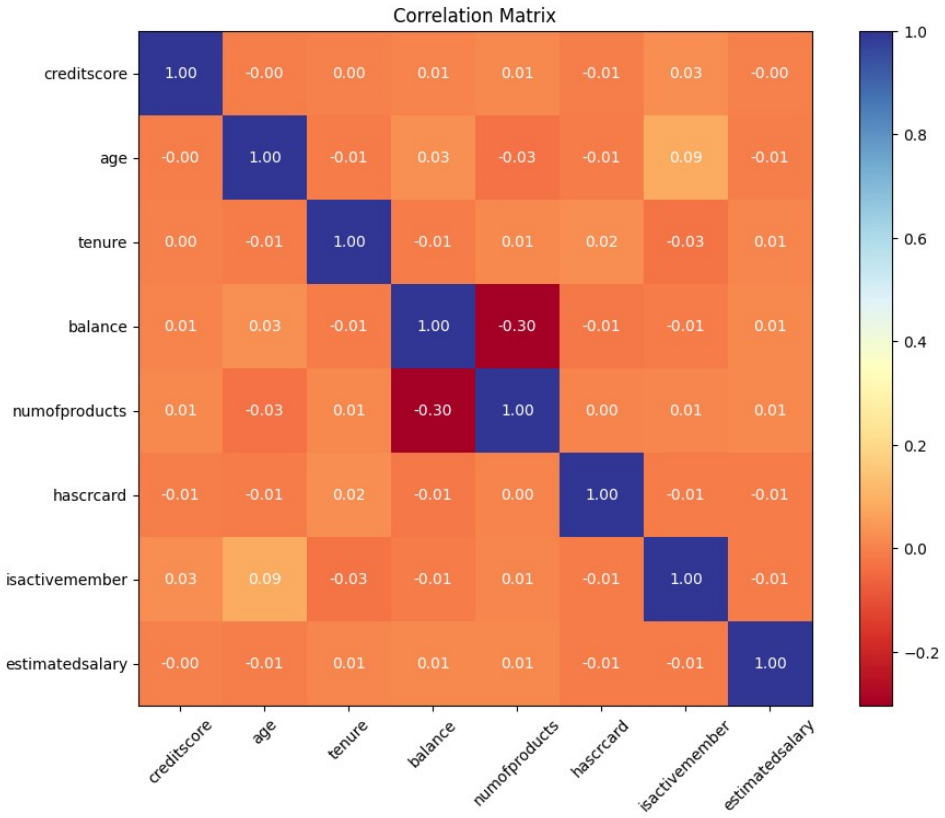
Churned customers:Minimum Estimated Salary: 11.58
Maximum Estimated Salary: 199,808.1
Mean Estimated Salary: 101,465.68
Retained customers:Minimum Estimated Salary: 90.07
Maximum Estimated Salary: 199,992.48
Mean Estimated Salary: 99,738.39



Correlation Matrix

The correlation matrix analysis reveals noteworthy insights. Notably, there exists a negative correlation between the 'balance' and the 'number of products.' This implies that as the quantity of products purchased increases, the balance tends to decrease.

Upon further examination, it's observed that the correlation values for the remaining features are relatively low. This suggests that no specific features exhibit a strong linear relationship with one another. Consequently, it seems there may not be singularly dominant features that significantly influence the overall correlation patterns within the dataset.



One-Hot Encoding

In our ongoing project, the subsequent task involves implementing one-hot encoding for categorical string values. This preprocessing step is crucial as we are working towards developing machine learning models. Within our dataset, there are two categorical attributes: 'geography' and 'gender.'

For 'geography,' which encompasses three distinct values—France, Germany, and Spain—we employ a numerical encoding strategy. Specifically, France is assigned the value 0, Germany is denoted by 1, and Spain is represented by 2. In the case of 'gender,' where the categories are 'female' and 'male,' a binary encoding scheme is applied, with 'female' mapped to 0 and 'male' to 1. This encoding process facilitates the translation of categorical attributes into a format that machine learning algorithms can effectively utilize for model training and analysis.

5. Choosing a Machine Learning Algorithm

Considering the paramount importance of addressing our current challenge, the selection of an optimal model is of utmost significance. In pursuit of this objective, we plan to delve into multiple machine learning algorithms to discern their respective performances. A key aspect of our evaluation involves conducting a meticulous comparative analysis of their accuracies.

Given the nature of our task, which involves classification, we have identified the following prominent algorithms for consideration:

K-Nearest Neighbor (KNN)

KNN is an intuitive algorithm that makes predictions based on the majority class of its nearest neighbors in the feature space. In the context of classification, it assigns a class to a data point by considering the most prevalent class among its k-nearest neighbors. For regression, it predicts a numerical value based on the average of the k-nearest neighbors' values. The choice of the parameter 'k' influences the algorithm's sensitivity to local patterns.

Logistic Regression (LR)

Contrary to its name, logistic regression is a classification algorithm suitable for binary outcomes. It models the probability of an instance belonging to a particular class using the logistic (sigmoid) function. The algorithm estimates coefficients for each feature, and the logistic function transforms the linear combination of these coefficients into a probability distribution. Its linear nature makes it interpretable, particularly when the decision boundary is assumed to be linear.

Gradient Boosting (GB)

Gradient Boosting is an ensemble learning technique that constructs a series of weak learners, typically decision trees, sequentially. Each new learner corrects the errors made by its predecessors, resulting in a robust predictive model. Decision trees are often used as weak learners, and each tree is fitted to the residuals of the previous one. While powerful, GB requires careful parameter tuning to achieve optimal performance.

Random Forest (RF)

Random Forest is another ensemble learning method that builds multiple decision trees and combines their predictions to enhance generalization and reduce overfitting. Each tree is trained on a random subset of the data and a random subset of features, introducing diversity. The final prediction is an aggregate of predictions from all the trees. RF is known for its robustness, handling of high-dimensional data, and providing insights into feature importance.

These algorithms offer diverse approaches to solving classification problems, each with its strengths and considerations. The selection of a specific algorithm depends on factors such as the dataset's characteristics, and the desired interpret ability of the model.

5.1. Simple - Model Building And Results

K-Nearest Neighbors (KNN)

The KNN model achieves an accuracy of 0.77, indicating that approximately 77% of predictions are correct. The Receiver Operating Characteristic Area Under the Curve (ROC AUC) stands at 0.54, providing a measure of the model's ability to distinguish between classes.

Logistic Regression Classifier

The Logistic Regression model exhibits an accuracy of 0.81, suggesting a high percentage of correct predictions. The ROC AUC, standing at 0.76, further emphasizes the model's capability to discriminate between classes.

Gradient Boosting Classifier

The Gradient Boosting model attains an accuracy of 0.87, indicating a strong predictive performance. The ROC AUC, registering at 0.87, underscores the model's effectiveness in classifying instances and distinguishing between classes.

Random Forest Classifier

The Random Forest model demonstrates an accuracy of 0.86, signifying a high level of correctness in its predictions. The ROC AUC, measured at 0.84, reinforces the model's efficacy in handling classification tasks and discerning between classes.

Model	Accuracy	ROC AUC
K-Nearest Neighbors	0.77	0.54
Logistic Regression	0.81	0.76
Gradient Boosting	0.87	0.87
Random Forest	0.86	0.84

5.2. Cross Validation and Hyperparameter Tunning

1st Iteration: Cross Validation

K-Nearest Neighbors (KNN)

In the first iteration, the K-Nearest Neighbors (KNN) model is subjected to cross-validation, a pivotal step in refining its performance. The hyperparameter n_neighbors is systematically varied across the values of 3, 5, and 7. Employing PySpark's sophisticated CrossValidator with 5 folds, this approach rigorously assesses the model's ability to generalize across different subsets of the data. Following this meticulous process, the KNN model demonstrates an accuracy of 0.77, indicating the proportion of correctly classified instances, and a ROC AUC of 0.54, reflecting its discriminative power.

Logistic Regression

Simultaneously, Logistic Regression is intricately tuned during cross-validation, with hyperparameters regParam and elasticNetParam exploring values of 0.01, 0.1, 1.0 and 0.0, 0.5, 1.0, respectively. This optimization process, facilitated by PySpark's CrossValidator with 5 folds, reveals that the Logistic Regression model attains an accuracy of 0.81 and a ROC AUC of 0.77. These metrics encapsulate the model's prowess in making accurate predictions and its effectiveness in distinguishing between classes.

Gradient Boosting

The Gradient Boosting model, a powerful ensemble technique, undergoes cross-validation with hyperparameters maxDepth and maxIter traversing values of 3, 5, 7 and 10, 20, 30, respectively. Employing PySpark's CrossValidator with 5 folds, this meticulous tuning process yields an accuracy of 0.87 and a ROC AUC of 0.87. These metrics signify the model's proficiency in capturing complex relationships within the data and its ability to discriminate between classes.

Random Forest

The Random Forest model, a robust ensemble of decision trees, is fine-tuned through cross-validation by varying hyperparameters numTrees and maxDepth across values of 10, 20, 30 and 3, 5, 7, respectively. Leveraging PySpark's CrossValidator with 5 folds, this iterative process culminates in an accuracy of 0.87 and a ROC AUC of 0.86. These metrics underscore the model's excellence in capturing the nuances of the dataset and its discriminative power.

Model	Accuracy	ROC AUC
KNN	0.77	0.54
Logistic Regression	0.81	0.77
Gradient Boosting	0.87	0.86
Random Forest	0.87	0.86

2nd Iteration: Hyperparameter Tuning

K-Nearest Neighbors (KNN)

Transitioning into the second iteration, the KNN model embarks on hyperparameter tuning, delving deeper into the refinement of its architecture. The parameter n_neighbors is scrutinized across values of 3, 5, and 7, employing GridSearchCV with 5 folds. This meticulous tuning process leads to an accuracy of 0.77 and a ROC AUC of 0.54, solidifying the model's configuration.

Logistic Regression

Simultaneously, Logistic Regression intensifies its optimization journey through hyperparameter tuning. The parameters C and penalty traverse values of 0.001, 0.01, 0.1, 1, 10, 100 and 'l1', 'l2', respectively, utilizing GridSearchCV with 5 folds. This comprehensive tuning process reveals an accuracy of 0.80 and a ROC AUC of 0.67, reflecting the model's adaptability to different regularization

strategies.

Gradient Boosting

The Gradient Boosting model extends its hyperparameter tuning, delving into the exploration of n_estimators and max_depth across values of 50, 100, 150 and 3, 5, 7. Employing GridSearchCV with 5 folds, this thorough tuning process leads to an accuracy of 0.87 and a ROC AUC of 0.86. These metrics affirm the model's adaptability and capacity to generalize across diverse scenarios.

Random Forest

Concluding the second iteration, the Random Forest model focuses on hyperparameter tuning, optimizing n_estimators and max_depth over values of 50, 100, 150 and 3, 5, 7. Leveraging GridSearchCV with 5 folds, this meticulous process results in an accuracy of 0.86 and a ROC AUC of 0.86. These metrics attest to the model's robustness and its ability to capture intricate relationships within the dataset.

In both iterations, the documentation provides detailed insights into the parameters explored, the approach adopted for cross-validation or hyperparameter tuning, and the resulting accuracy and ROC AUC for each model.

Model	Accuracy	ROC AUC
KNN	0.77	0.54
Logistic Regression	0.80	0.67
Gradient Boosting	0.87	0.86
Random Forest	0.86	0.86

6. Conclusion

In the dynamic landscape of the financial industry, our exploration into Bank Customer Churn Prediction has uncovered valuable insights crucial for strategic decision-making. Recognizing that customer retention is not only cost-effective but also a key driver of increased profits, our project aimed to empower financial institutions with tools for proactive churn management.

The significance of customer churn prediction became evident through our in-depth analysis of a dataset containing diverse customer attributes. We delved into factors such as geography, gender, credit card ownership, active membership, credit score, age, tenure, bank balance, and product usage. These insights provided a nuanced understanding of customer behavior and potential churn indicators.

In evaluating machine learning models for churn prediction, Gradient Boosting emerged as the most promising, boasting an accuracy of 0.87 and a ROC AUC of 0.87. This algorithm demonstrated a superior ability to capture complex relationships within the data, offering a robust predictive tool for financial institutions.

Moreover, our iterative approach involving cross-validation and hyperparameter tuning refined each model's performance. This meticulous process not only enhanced accuracy but also ensured the

adaptability and generalization capability of the models across diverse scenarios.

In essence, our project serves as a strategic guide for financial institutions navigating the challenges of customer retention. By leveraging advanced analytics and machine learning, banks can proactively identify potential churners, implement targeted retention strategies, and, most importantly, foster enduring and profitable customer relationships. In the ever-evolving financial landscape, the ability to predict and mitigate customer churn is not just a competitive advantage but a strategic imperative for sustained business success.

In wrapping up our study on predicting customer churn in banks, there are some exciting possibilities for future research. First off, it would be valuable to compare our results with those from different prediction methods. This way, we can see if there are other approaches that might perform even better. Also, finding a larger dataset with a more even distribution of customer churn and retention cases could help us create models that work well in a wider range of situations. Currently, our dataset might have some biases, and a more extensive and representative dataset would be ideal. Finally, we could do more testing with different features, trying out new ideas and variables to see if they improve our predictions. This way, we can uncover hidden patterns and create more accurate models that work effectively in the ever-changing world of banking and customer relationships.