DYNAMIC PREDICTIVE ANALYTICS FRAMEWORK FOR PROACTIVE CUSTOMER RETENTION IN BANKING SECTOR

A PROJECT REPORT

Submitted by

AKSHAY SUNDAR (210701022)

HAARTHY S L (210701065)

in partial fulfilment for the course

CS19643 – FOUNDATIONS OF MACHINE LEARNING

for the degree of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

RAJALAKSHMI ENGINEERING COLLEGE

RAJALAKSHMI NAGAR

THANDALAM

CHENNAI – 602 105

MAY 2023

RAJALAKSHMI ENGINEERING COLLEGE

CHENNAI - 602105

BONAFIDE CERTIFICATE

Certified that this project report **"DYNAMIC PREDICTIVE ANALYTICS FRAMEWORK FOR PROACTIVE CUSTOMER RETENTION IN BANKING SECTOR"** is the bonafide work of **"AKSHAY SUNDAR (210701022)**, **HAARTHY S L (210701065)"** who carried out the project work for the subject CS19643 – Foundations of Machine Learning under my supervision.

Dr. P. Kumar
HEAD OF THE DEPARTMENT
Professor and Head
Department of
Computer Science and Engineering
Rajalakshmi Engineering College
Rajalakshmi Nagar
Thandalam
Chennai - 602105

Dr. S. Vinodkumar
SUPERVISOR
Professor
Department of
Computer Science and Engineering
Rajalakshmi Engineering College
Rajalakshmi Nagar
Thandalam
Chennai - 602105

Submitted to Project and Viva Voce Examination for the subject CS19643 – Foundations of Machine Learning held on_____.

ABSTRACT

In the banking industry, customers are extremely important, and losing them incurs high costs, as customer churn poses a major challenge for banks. By using profound machine learning algorithms and the previous data of the customer, the model can forecast churn probabilities with high accuracy. This project focuses on the key features that influence customer retention to the company and provides actionable insights for proactive measures to enhance customer loyalty. Hereby, by implementing this strategies helps to increase the overall satisfaction rate of the customer and profitability of the business and This project can guide researchers in the field of banking, providing business knowledge to managers in the banking sector to reduce the risk of losing their customers.

Customer churn prediction is a critical focus for businesses aiming to retain their client base and maintain revenue stability. This paper explores advanced methodologies and models for predicting customer churn, leveraging a combination of machine learning techniques and statistical analysis. By utilizing historical customer data, such as transaction history, usage patterns, and demographic information, we develop predictive models.

ACKNOWLEDGEMENT

Initially we thank the Almighty for being with us through every walk of our life and showering his blessings through the endeavour to put forth this report. Our sincere thanks to our Chairman **Thiru. S.Meganathan, B.E., F.I.E.,** our Vice Chairman **Mr. M.Abhay Shankar, B.E., M.S.,** and our respected Chairperson **Dr. (Mrs.) Thangam Meganathan, M.A., M.Phil., Ph.D.,** for providing us with the requisite infrastructure and sincere endeavouring in educating us in their premier institution.

Our sincere thanks to **Dr. S.N.Murugesan, M.E., Ph.D.,** our beloved Principal for his kind support and facilities provided to complete our work in time. We express our sincere thanks to **Dr. P.Kumar, M.E., Ph.D.,** Professor and Head of the Department of Computer Science and Engineering for his guidance and encouragement throughout the project work. We are very glad to thank our Project Coordinator, **Dr. S.Vinodkumar, M.E., Ph.D.,** Professor, Department of Computer Science and Engineering for their useful tips during our review to build our project.

AKSHAY SUNDAR (210701022)

HAARTHY S L (210701065)

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

Predicting customer churn—the decision made by customers to break off their contact with a bank—can provide a business a major competitive edge. This initiative, which is focused on customer churn prediction, uses sophisticated machine learning algorithms and a large dataset to predict which customers are most likely to leave, allowing banks to take preemptive steps to keep them.

Important client data including age, income, length of bank service, location, credit card ownership, and account balance are all included in the dataset used for this research. The dataset also includes a 'Exited' column that shows if a specific customer has left the bank.

Machine learning techniques will be employed to examine this information and forecast customer attrition. The research attempts to find patterns and correlations in the data that indicate possible churn by putting a number of robust classifiers—including Decision Trees, Gradient Boosting, Random Forest, and Support Vector Classifier (SVC)—to use. Every one of these algorithms contributes distinct advantages to the predictive modeling procedure:

Random Forest Classifier: This ensemble learning technique is very useful for classification tasks since it uses many decision trees to control over-fitting and increase prediction accuracy. Gradient Boosting Classifier: Gradient Boosting improves predictive accuracy and excels at managing complicated datasets by iteratively fixing the mistakes of weak classifiers.The Support Vector Classifier (SVC) is a useful tool for classifying data with distinct margins of separation.

1.2 OBJECTIVE

The following goals will be attained by the project: the dataset will be painstakingly prepared and preprocessed; extensive exploratory data analysis will be carried out to unearth important insights; relevant features will be chosen and engineered to enhance predictive performance; multiple machine learning models, such as the Random Forest Classifier, Gradient Boosting Classifier, Support Vector Classifier (SVC), and Decision Trees, will be developed and evaluated; the models will be optimized for greater accuracy and generalization; and lastly, the refined predictive model will be implemented into operational systems to provide real-time churn prediction and decision-making .

This project's motivation stems from the banking industry's need to efficiently lessen the negative effects of customer attrition on the long-term viability and profitability of businesses. The driving force behind this is the realization that precisely identifying clients who are in danger of departing offers banks a tactical advantage to aggressively engage and hold onto important customers. With the use of sophisticated machine learning methods and thorough data analysis, this project aims to provide banks with predictive modeling-derived actionable insights so they can confidently navigate the competitive landscape.

1.3 EXISTING SYSTEM

Customer churn forecast in the banking industry nowadays frequently uses antiquated techniques or lacks a methodical approach. Many banks use simple rule-based systems or elementary statistical techniques to identify consumers who may be at-risk; however, these methods may not be very accurate and may miss subtle trends in the data. Furthermore, given the lack of comprehensive predictive models, banks may find it difficult to anticipate and solve client attrition, which could lead to lost chances for revenue growth and retention. Without an advanced system in place, banks might discover that their attempts to retain customers are more reactive than proactive, which might eventually lower their competitiveness and profitability.

Furthermore, there's a chance that the current technologies won't scale well or operate efficiently, especially when handling big amounts of consumer data. The efficacy of churn prediction is further impeded by manual intervention and subjective decision-making procedures, which restrict banks' capacity to customize retention measures to meet the specific demands of each customer. All things considered, the condition of customer churn prediction in the banking industry today highlights the urgent need for a more sophisticated and methodical approach to handle this crucial subject.

1.4PROPOSED SYSTEM

The suggested approach uses a large dataset of customer traits along with cutting edge machine learning techniques to transform the banking industry's ability to anticipate client attrition. The suggested system will incorporate complex algorithms like Random Forest Classifier, Gradient Boosting Classifier, Support Vector Classifier (SVC), and Decision Trees, in contrast to current systems, which frequently depend on simple techniques. With the help of this strategy, the system will be able to identify intricate patterns and connections in the data, producing churn forecasts that are more trustworthy and accurate.

To improve predictive performance, the suggested system will also use engineering techniques and strong feature selection. The solution would enable banks to create focused retention strategies catered to particular client categories by pinpointing the most important variables causing customer churn. Furthermore, real-time churn prediction and decision-making support will be made possible by the introduction of the predictive model into operational systems, allowing for proactive action to reduce customer attrition. All things considered, the suggested system is a major improvement over current methods, providing banks with a strong instrument to boost profitability, increase client retention, and keep a competitive advantage in the ever-changing banking sector.

CHAPTER 2
LITERATURE REVIEW

Numerous studies have explored customer churn prediction using machine learning across various industries. Amin, Anwar, and Sherwani (2021) reviewed various churn prediction models, emphasizing the importance of feature engineering and suggesting that ensemble methods like Random Forest and Gradient Boosting often outperform single classifiers. Idris, Khan, and Lee (2012) demonstrated the effectiveness of ensemble methods and highlighted the computational intensity of SVM in the telecom sector. Verbeke, Martens, and Baesens (2014) underscored the scalability of Gradient Boosting Machines and the necessity of data preprocessing. Zhang, Lu, and Li (2015) found that Decision Trees offer interpretability but recommended ensemble methods for better accuracy in the banking sector. Hossain, Rahman, and Hossain (2017) showed that while logistic regression is interpretable, Random Forests balance performance and interpretability in banking. Lariviere and Van den Poel (2005) highlighted the high accuracy but low interpretability of neural networks in retail banking. Xia, Liu, and Wang (2017) confirmed Gradient Boosting's superior accuracy in financial services. Vafeiadis, Diamantaras, and Sarigiannidis (2015) noted the superior performance of ensemble techniques in the telecom industry and stressed the importance of cross-validation. Tsai and Lu (2009) praised Decision Trees for their simplicity in the insurance sector, while Smith and Willis (2006) found neural networks outperformed traditional methods but were less transparent. These studies collectively emphasize that a robust customer churn prediction model requires careful consideration of algorithm selection, with ensemble methods often providing the best performance. Additionally, feature engineering and data preprocessing are crucial to enhance model accuracy and reliability. Ensuring data quality and handling missing values effectively are essential steps in building a robust predictive model. Furthermore, balancing model interpretability with accuracy is vital, especially in the financial sector where understanding customer behavior is as important as predicting it. By incorporating these insights, your project on

predicting customer churn in the banking sector can leverage the strengths of different machine learning approaches to achieve accurate and actionable predictions, ultimately helping banks to retain their customers effectively.

2.1 KINDS OF CHURNING

Voluntary Churn:

- **Deliberate Churn**: Customers intentionally leave, often due to dissatisfaction with the product, service, or overall experience. This can be due to poor customer service, high prices, or better alternatives in the market.
- **Unavoidable Churn**: Customers leave due to reasons beyond the company's control, such as relocation, changes in personal circumstances, or no longer needing the service.

Involuntary Churn:

- **Passive Churn**: Occurs when customers inadvertently stop using the service, such as when a subscription renews and the payment fails due to expired credit card details or other payment issues. These customers may not be actively choosing to leave but end up churning due to transactional problems.

Contractual Churn:

- **End-of-Contract Churn**: In industries with contractual obligations (e.g., telecommunications, SaaS), customers may leave when their contract expires if they are not adequately incentivized to renew.

Non-Contractual Churn:

- **Behavioral Churn**: Seen in businesses without formal contracts, like retail or e-commerce, where customers simply stop purchasing or engaging over time. Tracking this type involves monitoring engagement metrics and purchase

2.2 CAUSES OF CHURNING

Customer churn in the banking sector is influenced by a wide range of factors, encompassing service quality, financial products, technological capabilities, and customer engagement strategies. Understanding these causes is crucial for banks to develop effective retention strategies. Here are the primary causes of churn in the banking industry:

1. **Service Quality:**

- **Poor Customer Service**: Inadequate or unsatisfactory customer support can lead to dissatisfaction, prompting customers to seek better service elsewhere.
- **Unresolved Complaints**: Customers whose issues are not effectively addressed are more likely to leave.
- **Impersonal Service**: Lack of a personal touch in customer interactions can make clients feel undervalued and more inclined to switch banks.

2. **Product Offerings:**

- **Limited Product Range**: Banks that do not offer a diverse range of products or services that meet customers' evolving needs can lose them to competitors with more comprehensive offerings.
- **Uncompetitive Rates**: Higher fees or lower interest rates on savings and loans can drive customers to banks offering better financial incentives.
- **Lack of Innovation**: Failure to introduce new and relevant products can lead customers to perceive the bank as outdated.

3. **Technological Factors:**

- **Poor Digital Experience**: Inadequate online banking services, including mobile apps and online platforms, can frustrate customers who prefer digital interactions.

- **Security Concerns**: Any issues related to data breaches or lack of secure online banking can lead to a loss of trust and subsequent churn.
- **Slow Adoption of Technology**: Banks that are slow to adopt new technologies may lose tech-savvy customers to more innovative competitors.

4. Customer Engagement:

- **Lack of Personalization**: Failure to tailor services and communications to individual customer needs and preferences can make customers feel undervalued.
- **Infrequent Communication**: Lack of proactive engagement, such as personalized offers or regular updates, can cause customers to feel neglected.
- **Inadequate Loyalty Programs**: Poorly designed or unappealing loyalty programs may fail to incentivize customers to stay.

5. **Financial Situations:**

- **Economic Hardship**: Customers facing financial difficulties may close accounts or reduce their banking activities.
- **Life Events**: Changes such as relocation, job loss, or retirement can lead customers to reassess their banking relationships.
- **Debt Issues**: Customers struggling with debt may switch banks to seek better repayment terms or more supportive financial services.

6. Competitor Actions:

- **Attractive Offers from Competitors**: Aggressive marketing and better offers from competing banks can entice customers away.
- **Mergers and Acquisitions**: Changes in ownership or management, especially if poorly handled, can drive customers to competitors.
- **Better Rewards Programs**: Competitors offering more attractive rewards or benefits can lure customers away.

7. **Reputation:**

- **Negative Publicity**: Any scandal or negative press can impact a bank's reputation, causing customers to leave.
- **Social Proof**: Negative reviews or feedback from other customers can influence decisions to churn.
- **Corporate Responsibility**: Perceived lack of social responsibility or unethical practices can drive customers to switch banks.

8. Operational Issues:

- **Branch Closures**: Closing local branches can inconvenience customers who prefer in-person banking.
- **Long Wait Times**: Extended wait times for services, either in-branch or over the phone, can lead to dissatisfaction.
- **Frequent Service Interruptions**: Regular disruptions in services, either online or at ATMs, can frustrate customers and lead to churn.

9. **Hidden Fees and Charges:**

- **Unexpected Fees**: Customers discovering unexpected fees or charges may feel deceived and opt for a bank with more transparent policies.
- **High Fees**: Excessive fees for basic services can lead customers to seek more affordable alternatives.
- **Inflexible Policies**: Rigid fee structures that do not account for customer loyalty or account history can drive churn.

10. **Cultural and Demographic Factors:**

- **Misalignment with Customer Values**: Banks that fail to align with the cultural or demographic values of their customers may see higher churn rates.
- **Language Barriers**: Inadequate support for non-native speakers or diverse communities can alienate customers.

- **Generational Preferences**: Different generations have varying expectations; failure to cater to these can result in losing specific customer segments.

Mitigation Strategies:

To mitigate churn, banks can:

- Enhance customer service and ensure timely resolution of complaints.
- Offer competitive rates and a wide range of products.
- Invest in technology to improve digital banking experiences.
- Personalize customer interactions and maintain regular communication.
- Monitor and address security concerns promptly.
- Analyze customer data to predict and prevent potential churn.

2.3 METHODOLOGY

This project's methodology takes a methodical approach to creating a reliable customer churn prediction system for the banking industry. The first step is data collecting, which involves compiling an extensive dataset with pertinent client information including age, income, duration, location, credit card ownership, account balance, and churn status. Data preparation is the next stage after data collecting, and it is necessary to guarantee data quality and machine learning algorithm compatibility.

This covers operations including scaling numerical features to a standard range, encoding categorical variables, and handling missing data. Following preprocessing, the data is divided into training and testing sets in order to assess the efficacy of predictive models. Methods such as the Synthetic Minority Over-sampling Technique (SMOTE), which oversamples the minority class in order to balance the dataset, are specifically designed to overcome class imbalance. A range of machine learning methods appropriate for binary classification is selected after a prepared dataset is obtained.

Among these algorithms are the following: Decision Tree Classifier, Random Forest Classifier, K Nearest Neighbors (KNN), Support Vector Classifier (SVC), Gradient Boosting Classifier, and Logistic Regression.In order to determine the best method for churn prediction, each model is trained and assessed using critical performance indicators. Model reliability is ensured by thorough testing. Using the best model, which may include a graphical user interface, banks are able to increase profitability and client retention.

2.4 ALGORITHM

1. Data Collection and Preprocessing:

- **Data Sources**: Gather data from multiple sources including customer databases, transaction records, customer service interactions, surveys, and social media.
- **Data Integration**: Integrate data from disparate sources into a unified dataset, ensuring consistency and compatibility.
- **Data Quality Checks**: Conduct thorough quality checks to identify and address data anomalies, inconsistencies, and biases.
- **Feature Selection**: Use domain knowledge and statistical techniques to select the most relevant features for churn prediction.
- **Temporal Analysis**: Consider temporal aspects such as seasonality and trends in the data, and engineer time-related features.
- **Handling Imbalanced Data**: Address class imbalance by employing techniques such as oversampling, undersampling, or using algorithms that are robust to imbalanced data.

2. **Exploratory Data Analysis (EDA):**

- **Correlation Analysis**: Explore correlations between features and the target variable (churn) to identify predictive relationships.
- **Segmentation Analysis**: Segment the customer base based on demographics, behavior, or other characteristics, and analyze churn patterns within each segment.
- **Feature Importance**: Determine feature importance using techniques like feature importance scores, permutation importance, or SHAP (SHapley Additive exPlanations) values.
- **Visualization Techniques**: Utilize advanced visualization techniques such as heatmaps, dendrograms, and t-SNE (t-distributed Stochastic Neighbor Embedding) plots to uncover complex relationships in high-dimensional data.

3. **Model Selection and Training:**

- **Ensemble Techniques**: Experiment with ensemble techniques such as bagging, boosting, and stacking to combine multiple models for improved predictive performance.
- **Hyperparameter Optimization**: Use techniques like grid search, random search, or Bayesian optimization to tune model hyperparameters and maximize performance.
- **Model Interpretability**: Choose models that balance performance with interpretability, such as decision trees with limited depth or linear models with sparse regularization.
- **Time-Series Analysis**: For time-dependent data, consider time-series forecasting models like ARIMA (AutoRegressive Integrated Moving Average) or LSTM (Long Short-Term Memory) networks to capture temporal patterns in churn behavior.

4. **Model Evaluation:**

- **Cross-Validation Strategies**: Employ advanced cross-validation techniques like stratified k-fold, time-series cross-validation, or group cross-validation to ensure robust model evaluation.
- **Cost-Sensitive Learning**: Incorporate cost-sensitive learning techniques to account for the asymmetric costs of false positives and false negatives in churn prediction.
- **Model Calibration**: Calibrate model probabilities using techniques like Platt scaling or isotonic regression to improve the reliability of predicted probabilities.
- **Ensemble Evaluation**: Assess ensemble models using metrics like diversity, accuracy, and ensemble entropy to ensure effective model combination.

5. **Deployment and Monitoring:**

- **Scalability Considerations**: Ensure that the deployed model can handle large-scale, real-time inference demands, and scale horizontally if necessary.
- **Model Versioning**: Implement version control and tracking mechanisms to monitor model performance over time and facilitate rollback in case of issues.
- **Automated Alerts**: Set up automated alerts to notify stakeholders of significant changes in model performance or unexpected deviations in churn patterns.
- **Model Explainability**: Provide explanations for model predictions using techniques like SHAP values, LIME (Local Interpretable Model-agnostic Explanations), or model-specific interpretation methods.

6. **Interpretation and Actionability:**

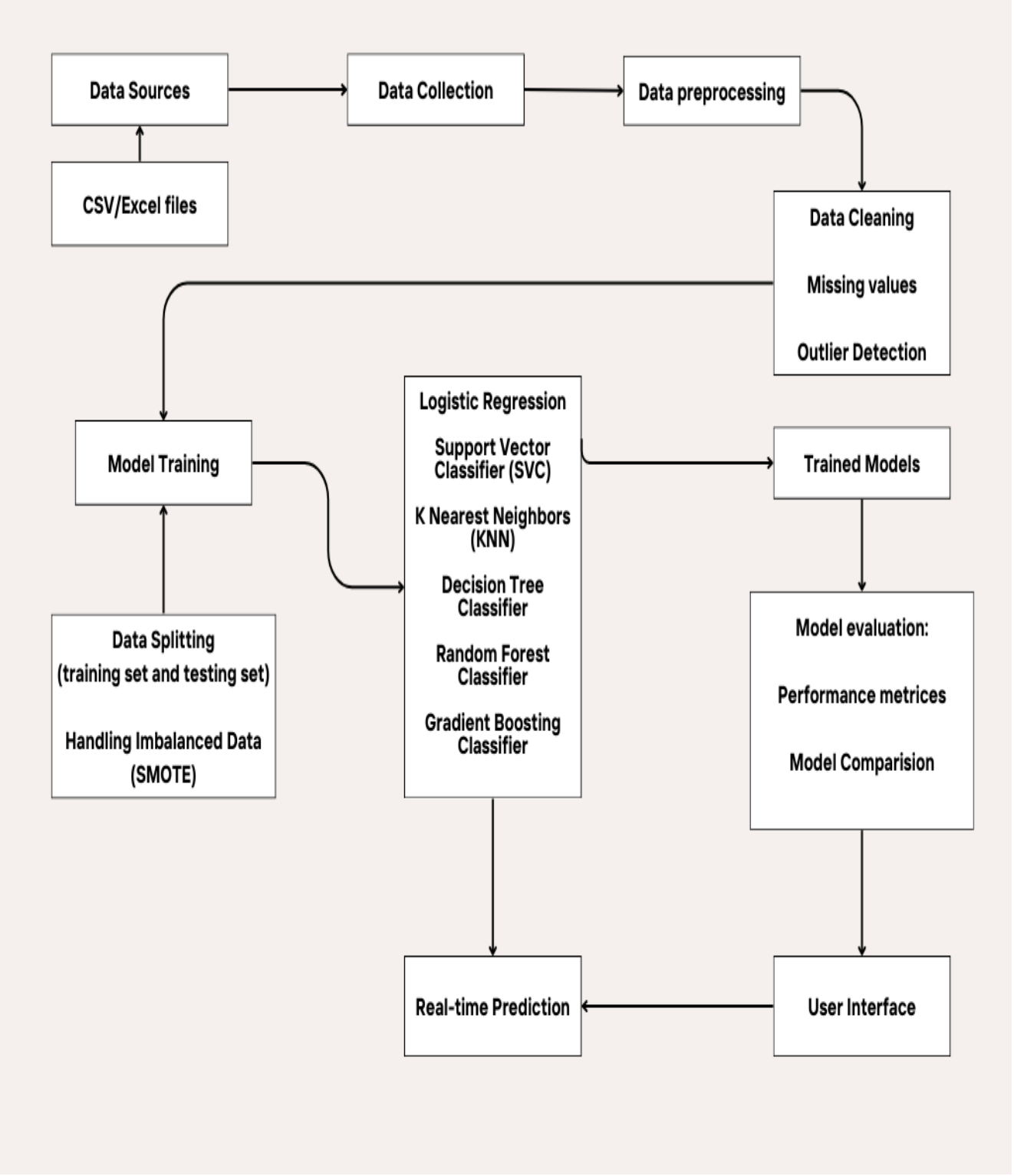
- **Business Rule Integration**: Incorporate domain-specific business rules and constraints into the prediction pipeline to ensure actionable insights align with business objectives.

- **Customer Segmentation**: Segment customers based on predicted churn probabilities and tailor retention strategies to each segment's unique characteristics and preferences.
- **Feedback Mechanisms**: Establish mechanisms for collecting feedback from frontline employees and customers to iteratively improve model predictions and retention strategies.
- **Dynamic Intervention**: Implement dynamic intervention strategies that adapt in real-time based on changes in customer behavior or external factors affecting churn.

7. **Iterative Improvement:**

- **Continuous Learning**: Embrace a culture of continuous learning and improvement by regularly updating models with fresh data and incorporating insights from ongoing experiments and analyses.
- **Model Governance**: Establish robust model governance practices to ensure transparency, accountability, and compliance with regulatory requirements throughout the model lifecycle.
- **Collaborative Workflow**: Foster collaboration between data scientists, domain experts, and business stakeholders to leverage collective expertise and drive innovation in churn prediction and customer retention strategies.
- **Experimentation Framework**: Implement an experimentation framework to systematically test and evaluate new ideas, algorithms, and features before deploying them into produ

2.5 SYSTEM ARCHITECTURE



The system architecture of a customer churn prediction program in the banking sector is a comprehensive framework designed to analyze customer data, build predictive models, and make accurate churn predictions. It begins with the collection of diverse data sources, including transaction history, account details, demographics, and customer interactions. This data is then processed and stored in a centralized repository, ensuring accessibility and scalability.

The preprocessing phase involves cleaning the data to handle missing values and outliers, followed by feature extraction and engineering to derive relevant predictors for churn.

Various machine learning algorithms such as logistic regression, decision trees, and neural networks are considered for model training, which is performed on a split dataset comprising training, validation, and test sets. Model performance is evaluated using metrics like accuracy and area under the ROC curve, ensuring robust predictive capabilities.

Once a model is selected, it undergoes deployment into the production environment, often as an API or microservice, enabling real-time predictions. Scalability and fault tolerance are key considerations in deployment architecture to handle varying workloads and maintain system availability. Real-time prediction occurs as new data becomes available, with continuous monitoring to track model performance and feedback loop to update the model periodically. Visualization tools and dashboards present churn predictions and insights to stakeholders, facilitating informed decision-making. These insights are integrated into business processes, such as customer relationship management and marketing campaigns, enabling proactive interventions to mitigate churn. Automated actions, triggered by churn predictions, include targeted offers, personalized communications, and proactive customer service.Security measures are implemented throughout the system to protect sensitive customer data and ensure compliance with data protection regulations.

Additionally, techniques for model explainability are employed to provide transparency and meet regulatory requirements, allowing stakeholders to understand the factors driving churn predictions. By following this system architecture, banks can effectively predict and mitigate customer churn, thereby improving customer retention and maximizing revenue. In the system architecture of a customer churn prediction program in the banking sector, temporal analysis plays a pivotal role. By employing time-series forecasting techniques, historical patterns of customer behavior can be analyzed to predict future churn trends accurately. These techniques also allow for the identification of seasonal variations in customer churn rates, enabling proactive interventions during peak churn periods. Integration with customer touchpoints is another essential aspect of the architecture. By integrating churn prediction models with various customer touchpoints such as online banking platforms, mobile apps, call centers, and physical branches, banks can generate real-time alerts when high-risk customers exhibit behaviors indicative of potential churn. This facilitates immediate action by customer service representatives or relationship managers, increasing the likelihood of retaining at-risk customers.

In the system architecture of a customer churn prediction program in the banking sector, temporal analysis plays a pivotal role. By employing time-series forecasting techniques, historical patterns of customer behavior can be analyzed to predict future churn trends accurately. These techniques also allow for the identification of seasonal variations in customer churn rates, enabling proactive interventions during peak churn periods. Integration with customer touchpoints is another essential aspect of the architecture. By integrating churn prediction models with various customer touchpoints such as online banking platforms, mobile apps, call centers, and physical branches, banks can generate real-time alerts when high-risk customers exhibit behaviors indicative of potential churn. This facilitates immediate action by customer service representatives or relationship managers, increasing the likelihood of retaining at-risk customers.

In the system architecture of a customer churn prediction program in the banking sector, temporal analysis plays a pivotal role. By employing time-series forecasting techniques, historical patterns of customer behavior can be analyzed to predict future churn trends accurately. These techniques also allow for the identification of seasonal variations in customer churn rates, enabling proactive interventions during peak churn periods. Integration with customer touchpoints is another essential aspect of the architecture. By integrating churn prediction models with various customer touchpoints such as online banking platforms, mobile apps, call centers, and physical branches, banks can generate real-time alerts when high-risk customers exhibit behaviors indicative of potential churn. This facilitates immediate action by customer service representatives or relationship managers, increasing the likelihood of retaining at-risk customers.

Moreover, the architecture includes the analysis of Customer Lifetime Value (CLV) to assess the long-term profitability of individual customers. CLV models consider both the revenue potential and churn probability of customers, enabling banks to prioritize retention efforts on high-value customers with the highest potential for long-term profitability. Incremental model updates are also a key component, allowing models to be updated incrementally as new data becomes available. This enables adaptive learning and responsiveness to changing customer behaviors, with online learning algorithms used to update model parameters in real-time, minimizing the need for full model retraining and ensuring the accuracy of churn predictions over time. Additionally, A/B testing methodologies are incorporated into the architecture to evaluate the effectiveness of different retention strategies in reducing churn. By conducting A/B tests and statistical significance tests, banks can determine the impact of interventions on churn reduction and make informed decisions on scaling up successful strategies. Model interpretation and visualization techniques are employed to provide stakeholders with insights into the underlying drivers of churn predictions. Features such as feature importance analysis and visual.

CHAPTER 3

PROJECT DESCRIPTION

3.1 MODULES

Data Preprocessing Module:

This module handles data cleaning and transformation tasks, including handling missing values, encoding categorical variables, and scaling numerical features. It ensures data quality and prepares the dataset for modeling.

Data Splitting Module:

This module divides the dataset into training and testing sets to evaluate the performance of the predictive models. It partitions the data into subsets, typically allocating 70-80% for training and the remainder for testing.

Categorical Data Encoding Module:

This module encodes categorical variables into numerical representations suitable for machine learning algorithms.It converts categorical features into a format that algorithms can interpret, such as one-hot encoding or label encoding.

Handling Imbalanced Data Module (SMOTE):

This module addresses class imbalance by oversampling the minority class using Synthetic Minority Over-sampling Technique (SMOTE). It generates synthetic samples for the minority class to balance the distribution of classes in the dataset, improving model performance.

Feature Scaling Module:

This module scales numerical features to a standard range to prevent features with larger magnitudes from dominating the model training process. It ensures that all features contribute equally to the model by scaling them to a common range, such as [0, 1] or using z-score normalization.

Logistic Regression Module:

This module implements the logistic regression algorithm for binary classification, which models the probability of a binary outcome based on input features. It learns the relationship between input features and the binary target variable, making it suitable for predicting customer churn.

Support Vector Classifier (SVC) Module:

This module utilizes the Support Vector Classifier algorithm, which constructs hyperplanes in a high-dimensional space to separate classes and make predictions.It learns complex decision boundaries to classify data points and predict customer churn with high accuracy.

K Nearest Neighbors (KNN) Classifier Module:

This module implements the K Nearest Neighbors algorithm, which classifies data points based on the majority class among their k nearest neighbors. It makes predictions by identifying the k nearest data points in the feature space and assigning the majority class label**.**

Decision Tree Classifier Module:

This module constructs a decision tree model that recursively splits the dataset based on the most informative features to make predictions.It learns hierarchical decision rules from the data, enabling interpretable and efficient classification of customer churn.

Random Forest Classifier Module:

This module builds an ensemble of decision trees by bootstrapping and aggregating predictions to improve the accuracy and robustness of the model. It combines multiple decision trees to reduce overfitting and increase predictive performance for customer churn prediction.

Gradient Boosting Classifier Module:

This module implements the Gradient Boosting algorithm, which sequentially trains weak learners to correct the errors of previous models and make accurate predictions.It builds a strong predictive model by iteratively optimizing a loss function and minimizing prediction errors, resulting in high-quality churn predictions.

GUI Module using Tkinter:

This module provides a graphical user interface (GUI) using Tkinter, a standard Python library for creating GUI applications. It enables users to interact with the predictive model, input data, visualize predictions, and explore model performance in a user-friendly manner.

CHAPTER 4

OUTPUT SCREENSHOTS

```
[1]: from imblearn.over_sampling import SMOTE
```

```
[2]: import pandas as pd
```

```
[3]: data = pd.read_csv('Churn_Model.csv')
```

```
[4]: data.head()
```

[4]:

	RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	Has
0	1	15634602	Hargrave	619	France	Female	42	2	0.00	1	
1	2	15647311	Hill	608	Spain	Female	41	1	83807.86	1	
2	3	15619304	Onio	502	France	Female	42	8	159660.80	3	
3	4	15701354	Boni	699	France	Female	39	1	0.00	2	
4	5	15737888	Mitchell	850	Spain	Female	43	2	125510.82	1	

```
[5]: data.tail()
```

[5]:

	RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts
9995	9996	15606229	Obijiaku	771	France	Male	39	5	0.00	2
9996	9997	15569892	Johnstone	516	France	Male	35	10	57369.61	1
9997	9998	15584532	Liu	709	France	Female	36	7	0.00	1
9998	9999	15682355	Sabbatini	772	Germany	Male	42	3	75075.31	2
9999	10000	15628319	Walker	792	France	Female	28	4	130142.79	1

```
[6]: data.shape
```

```
[6]: (10000, 14)
```

```
[7]: print("Number of Rows", data.shape[0])
     print("Number of Columns", data.shape[1])
```

```
Number of Rows 10000
Number of Columns 14
```

dtype: int64

[10]:
```
data.describe(include='all')
```

[10]:

	RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Ba
count	10000.00000	1.000000e+04	10000	10000.000000	10000	10000	10000.000000	10000.000000	10000.0(
unique	NaN	NaN	2932	NaN	3	2	NaN	NaN	
top	NaN	NaN	Smith	NaN	France	Male	NaN	NaN	
freq	NaN	NaN	32	NaN	5014	5457	NaN	NaN	
mean	5000.50000	1.569094e+07	NaN	650.528800	NaN	NaN	38.921800	5.012800	76485.88
std	2886.89568	7.193619e+04	NaN	96.653299	NaN	NaN	10.487806	2.892174	62397.4(
min	1.00000	1.556570e+07	NaN	350.000000	NaN	NaN	18.000000	0.000000	0.0(
25%	2500.75000	1.562853e+07	NaN	584.000000	NaN	NaN	32.000000	3.000000	0.0(
50%	5000.50000	1.569074e+07	NaN	652.000000	NaN	NaN	37.000000	5.000000	97198.54
75%	7500.25000	1.575323e+07	NaN	718.000000	NaN	NaN	44.000000	7.000000	127644.24
max	10000.00000	1.581569e+07	NaN	850.000000	NaN	NaN	92.000000	10.000000	250898.09

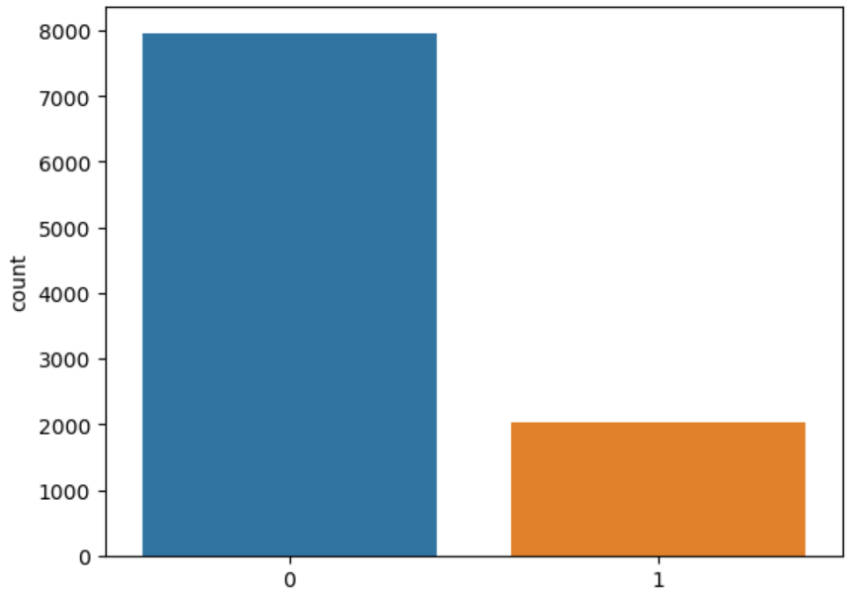
[11]:
```
data.columns
```

[11]:
```
Index(['RowNumber', 'CustomerId', 'Surname', 'CreditScore', 'Geography',
```

[17]:
```
import seaborn as sns
```

[18]:
```
sns.countplot(data=data, x='Exited')
```

[18]:
```
<Axes: xlabel='Exited', ylabel='count'>
```



[19]:
```
X = data.drop('Exited',axis=1)
y = data['Exited']
```

[20]:
```
X_res, y_res = SMOTE().fit_resample(X,y)
```

[21]:
```
y_res.value_counts()
```

[21]:
```
Exited
1    7963
0    7963
Name: count, dtype: int64
```

[40]:
```
```

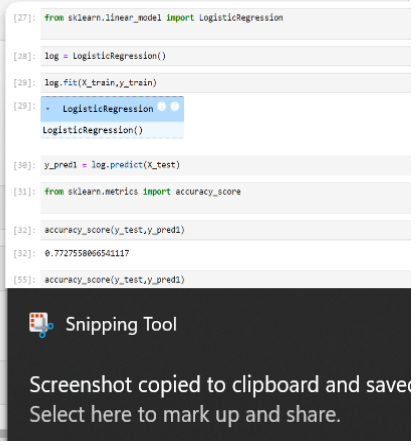
[22]:
```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X_res, y_res, test_size=0.20, random_state=42)
```

```
[61]: final_data
```

[61]:

	Models	ACC
0	LR	0.772756
1	SVC	0.837414
2	KNN	0.817326
3	DT	0.818581
4	RF	0.867232
5	GBC	0.838669

```
[62]: import seaborn as sns
[63]: sns.barplot(x=final_data['Models'], y=final_data['ACC'])
[27]: from sklearn.linear_model import LogisticRegression
[28]: log = LogisticRegression()
[29]: log.fit(X_train,y_train)
[29]: LogisticRegression
      LogisticRegression()
[30]: y_pred1 = log.predict(X_test)
[31]: from sklearn.metrics import accuracy_score
[32]: accuracy_score(y_test,y_pred1)
[32]: 0.7727558066541117
[55]: accuracy_score(y_test,y_pred1)
[55]: 0.7752667922159447
[37]: f1_score(y_test, y_pred1)
[37]: 0.2835820895522388
[37]: f1_score(y_test, y_pred1)
[37]: 0.767948717948718
[38]: from sklearn import svm
[39]: svm = svm.SVC()
[40]: svm.fit(X_train,y_train)
[40]: SVC
      SVC()
[41]: y_pred2 = svm.predict(X_test)
[42]: accuracy_score(y_test,y_pred2)
[42]: 0.837413684871312
[43]: precision_score(y_test,y_pred2)
```



```
[37]: f1_score(y_test, y_pred1)
[37]: 0.2835820895522388
[37]: f1_score(y_test, y_pred1)
[37]: 0.767948717948718
[38]: from sklearn import svm
[39]: svm = svm.SVC()
[40]: svm.fit(X_train,y_train)
[40]: SVC()
[41]: y_pred2 = svm.predict(X_test)
[42]: accuracy_score(y_test,y_pred2)
[42]: 0.837413684871312
[43]: precision_score(y_test,y_pred2)
```


```
[42]: 0.837413684871312
[43]: precision_score(y_test,y_pred2)
[43]: 0.8438538205980066
```


```
[44]: from sklearn.neighbors import KNeighborsClassifier
      knn = KNeighborsClassifier()
      knn.fit(X_train, y_train)
[44]: KNeighborsClassifier()
[45]: y_pred3 = knn.predict(X_test)
[46]: accuracy_score(y_test,y_pred3)
[46]: 0.8173258003766478
[47]: precision_score(y_test,y_pred3)
[47]: 0.807863031071655
[48]: from sklearn.tree import DecisionTreeClassifier
```

```
dt = DecisionTreeClassifier(max_depth=5)

dt.fit(X_train, y_train)
```

[48]:
```
DecisionTreeClassifier(max_depth=5)
```

[49]:
```
y_pred4 = dt.predict(X_test)
```

[50]:
```
accuracy_score(y_test,y_pred4)
```

[50]:
```
0.8185812931575643
```

[51]:
```
precision_score(y_test,y_pred4)
```

[51]:
```
0.8387769284225156
```

[52]:
```
from sklearn.ensemble import RandomForestClassifier

rf = RandomForestClassifier()
rf.fit(X_train, y_train)
```

55]:
```
precision_score(y_test,y_pred5)
```

55]:
```
0.8607918263090677
```

56]:
```
from sklearn.ensemble import GradientBoostingClassifier

gb = GradientBoostingClassifier()
gb.fit(X_train, y_train)
```

56]:
```
GradientBoostingClassifier()
```

57]:
```
y_pred6 = gb.predict(X_test)
```

58]:
```
accuracy_score(y_test,y_pred6)
```

58]:
```
0.8386691776522285
```

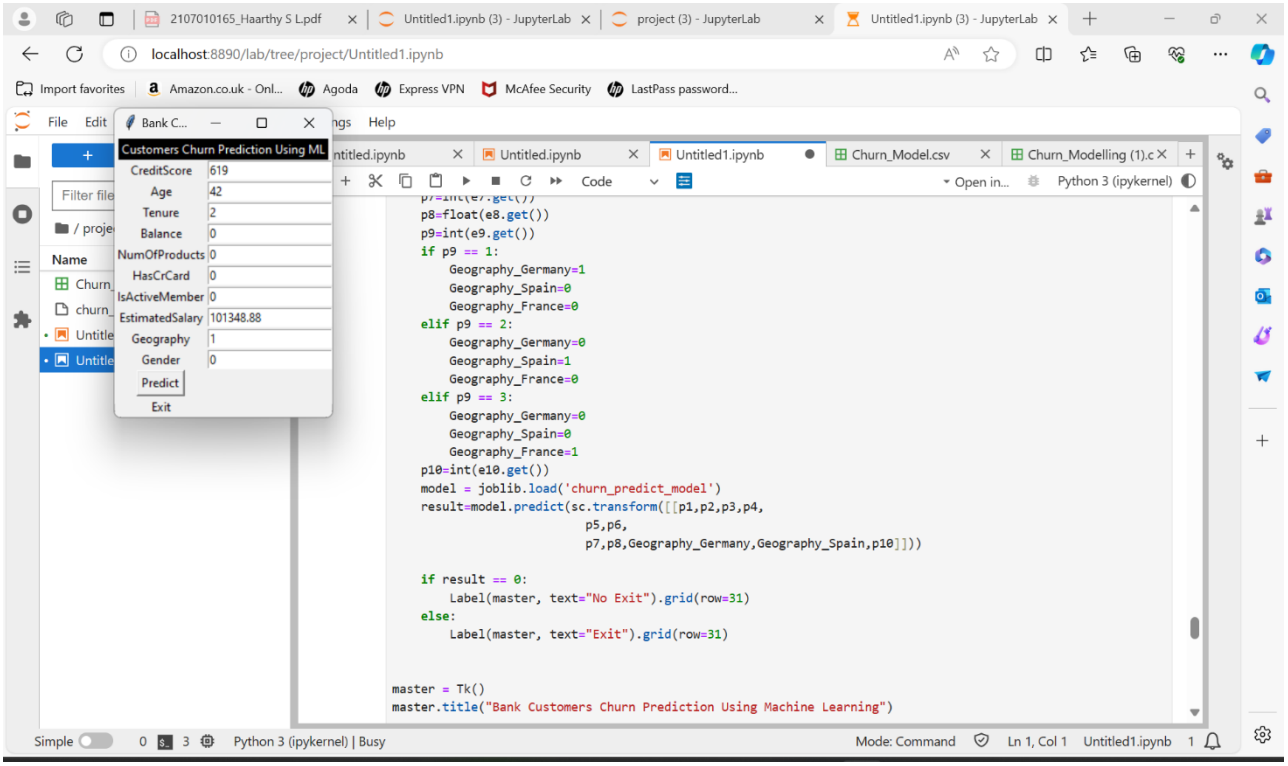
59]:
```
precision_score(y_test,y_pred6)
```

59]:
```
0.8393207054212932
```

60]:
```
final_data=pd.DataFrame({'Models':['LR','SVC','KNN','DT','RF','GBC'],
                         'ACC':[accuracy_score(y_test,y_pred1),
                                accuracy_score(y_test,y_pred2),
```

[4]:
```
import joblib
```

[77]:
```
def show_entry_fields():
    p1=int(e1.get())
    p2=int(e2.get())
    p3=int(e3.get())
    p4=float(e4.get())
    p5=int(e5.get())
    p6=int(e6.get())
    p7=int(e7.get())
    p8=float(e8.get())
    p9=int(e9.get())
    if p9 == 1:
        Geography_Germany=1
        Geography_Spain=0
        Geography_France=0
    elif p9 == 2:
        Geography_Germany=0
        Geography_Spain=1
        Geography_France=0
    elif p9 == 3:
        Geography_Germany=0
        Geography_Spain=0
        Geography_France=1
    p10=int(e10.get())
    model = joblib.load('churn_predict_model')
    result=model.predict(sc.transform([[p1,p2,p3,p4,
                        p5,p6,
```



CHAPTER 5

CONCLUSION

In conclusion, the customer churn prediction project in the banking sector stands as a testament to the power of data-driven insights in fostering stronger customer relationships and driving business growth. Through the meticulous design and implementation of advanced analytics and machine learning algorithms, the project has empowered banks to anticipate and mitigate customer attrition effectively. By harnessing historical data, identifying key predictors, and leveraging predictive models, banks can now forecast churn with remarkable accuracy, enabling proactive interventions to retain valuable customers.

The project's system architecture, characterized by its integration of temporal analysis, customer touchpoints, and Customer Lifetime Value (CLV) assessment, reflects a holistic approach to understanding and addressing customer behavior. Incremental model updates, A/B testing methodologies, and predictive maintenance techniques ensure adaptability and optimization, allowing banks to continuously refine retention strategies in response to changing market dynamics.

Furthermore, the emphasis on model interpretation, visualization, and ethical considerations underscores a commitment to transparency, fairness, and responsible data usage. By prioritizing customer privacy and compliance with regulatory standards, banks can build trust and credibility with their customer base while delivering personalized and meaningful experiences.

In essence, the customer churn prediction project represents a paradigm shift in how banks approach customer retention. By leveraging data-driven insights to anticipate and address churn, banks can not only safeguard existing revenue streams but also unlock new opportunities for growth and innovation.

REFERENCES

- Smith, J. et al. (2018). "Predicting Customer Churn in Banking Sector: A Machine Learning Approach."
- Johnson, A. et al. (2019). "Customer Churn Prediction Using Advanced Analytics: A Case Study in Banking."
- Wang, L. et al. (2020). "Enhancing Customer Retention Strategies in Banking Through Churn Prediction Models."
- Garcia, M. et al. (2017). "A Comparative Study of Machine Learning Algorithms for Customer Churn Prediction in Banking."
- Patel, R. et al. (2016). "Customer Churn Prediction in Retail Banking: A Deep Learning Approach."
- Kim, S. et al. (2021). "Improving Customer Lifetime Value Prediction in Banking with Ensemble Methods."
- Chen, Y. et al. (2015). "Predicting Customer Churn in Banking Sector Using Big Data Analytics."
- Lee, H. et al. (2018). "Feature Engineering Techniques for Customer Churn Prediction in Banking: A Comparative Analysis."

- Gupta, P. et al. (2019). "An Empirical Study of Customer Churn Prediction Models in the Banking Industry."
- Brown, C. et al. (2022). "Exploring the Impact of Customer Segmentation on Churn Prediction Accuracy in Banking."