CUSTOMER CHURN PREDICTION

TEAM DETAILS

Sarika Jaya chandra - 20BCE2868

N. Sri Anjan Kumar-20BIT0157

Mamidi Dinnesh - 20BCE2877

M. Bhuvana Sri -20BIT0246

INDEX:

1. INTRODUCTION

1.1 Overview

Churning refers to the process in which a customer leaves one company and switches to another. This not only results in a loss of income but also has negative implications for overall operations, particularly in terms of Customer Relationship Management (CRM). Establishing long-term relationships with customers is crucial for institutions as they aim to expand their customer base. Service providers face challenges related to customer behavior and their evolving expectations. The present generation, which is generally more educated than previous ones, has higher demands for connectivity, innovation, and diverse policy options. This advanced knowledge has led to changes in consumer purchasing behavior, presenting a significant challenge for service providers to think creatively and meet these expectations.

Customers must be recognised by the private sector. In their study, Liu and Shih support this claim by stating that rising demands on corporations to generate new and inventive marketing strategies, match customer expectations, and improve profits.

Retention and loyalty. Customers may easily move their relationships from one bank to another. Some customers may keep their relationship status null, which signifies their account status is inactive. By leaving this account dormant, the consumer may be moving their connection to another bank. There are several categories of consumers in the bank. Farmers are one of the banks' most important customers; they may expect lower monthly charges because their income is modest. Businesspeople are also essential consumers since they do a large number of transactions with large sums of money. These consumers will anticipate higher levels of service excellence. Middle-class clients were one of the most significant segments; in almost every bank, these people outnumber other types of customers. These individuals will anticipate lower monthly fees, improved service quality, and new policies.

Keeping multiple sorts of clients is therefore difficult. They must consider clients and their wants in order to overcome these problems and provide quality service on time and within budget to customers. Maintaining a strong working relationship with them is also a huge problem for them. If they do not overcome these eight difficulties, they may have churn. Recruiting a new client is more expensive and difficult than retaining existing consumers. Customers holding, on the other hand, are often more costly since they have already earned the trust and loyalty of existing customers. As a result, the requirement for a system that can successfully forecast client attrition in the early phases is critical for any banking institution.

1.2 PURPOSE:

Churn prediction refers to the process of identifying customers who are likely to leave a company and switch to another. The purpose of using churn prediction is to proactively address customer attrition and reduce the negative impact it has on a company's operations, particularly in terms of Customer Relationship Management (CRM). The ability to establish and maintain long-term relationships with customers is crucial for businesses aiming to expand their customer base.In today's highly competitive market, customers have higher demands for connectivity, innovation, and diverse policy options. This, coupled with changes in consumer purchasing behavior, presents significant challenges for service providers. Customer churn not only results in a loss of income but also entails the expense and difficulty of acquiring new customers. It is more cost-effective to retain existing customers by accurately predicting their likelihood of churn and implementing targeted retention strategies.
By utilizing churn prediction, companies can:

1. Enhance Customer Retention
2. Optimize Resource Allocation
3. Improve Customer Satisfaction
4. Increase Revenue
5. Gain Competitive Advantage.

By leveraging advanced machine learning algorithms like Random Forest, companies can accurately predict customer churn and take proactive measures to retain valuable customers.

2. LITERATURE SURVEY

1) The article "Customer Churn Prediction using Machine Learning: Subscription Renewal on OTT Platforms" by O.R. Devi, S.K. Pothini, M.P. Kumari, S.V., and U.N.S. Charan focuses on predicting customer churn in Over-The-Top (OTT) platforms using machine learning. The researchers employed various algorithms including Logistic Regression, Decision Tree, Random Forest, Support Vector Machines (SVM), and Artificial Neural Networks (ANN) to train and evaluate predictive models.

The study utilized a dataset collected from an OTT platform, comprising customer information such as demographics, viewing patterns, and subscription details. The evaluation of model performance involved metrics such as accuracy, precision, recall, and F1-score.

However, the work had some limitations. The dataset used was historical, lacking real-time data. The article did not provide details about the preprocessing techniques and feature engineering methods employed, which could impact model performance.
Future research in this area could involve exploring advanced algorithms like Gradient Boosting and Deep Learning to further improve predictive accuracy. Incorporating external data sources, such as social media sentiment analysis, could enhance churn prediction models on OTT platforms.

2) The article titled "Customer Churn Prediction Using Machine Learning Approaches" by R. Srinivasan, D. Rajeswari, and G. Elangovan addresses customer churn prediction using machine learning. The authors explored multiple algorithms, including Decision Tree, Random Forest, K-Nearest Neighbors (KNN), and Naive Bayes, to develop predictive models.
The study utilized a dataset comprising customer information from a telecom company, including features such as customer demographics, service usage patterns, and billing details. The dataset was divided into training and testing sets to evaluate the performance of the models.

Various evaluation metrics were employed, including accuracy, precision, recall, and F1-score, to measure the effectiveness of the predictive models.

The article identified a few drawbacks in their work. One limitation was the lack of real-time data, as the dataset was historical. Additionally, the authors did not mention the specific preprocessing techniques or feature selection methods used, which may affect the model's performance.
Future research in this field could involve exploring more advanced machine learning techniques, such as Support Vector Machines (SVM) or Gradient Boosting, to enhance the accuracy of churn prediction models. Furthermore, incorporating additional features like customer sentiment analysis or social media data could provide valuable insights for better churn prediction.

3) The article titled "Customer Churn Prediction Using Machine Learning: Commercial Bank of Ethiopia" by M.H. Seid and M.M. Woldeyohannis focuses on customer churn prediction in the context of the Commercial Bank of Ethiopia, utilizing machine learning techniques. The authors employed various algorithms, including Logistic Regression, Decision Tree, Random Forest, and Artificial Neural Networks (ANN), to develop predictive models.
The study utilized a dataset obtained from the Commercial Bank of Ethiopia, consisting of customer-related information such as demographics, transaction history, and account details. The dataset was divided into training and testing sets for model evaluation.

To assess the performance of the models, the authors used metrics such as accuracy, precision, recall, and F1-score.

A limitation of their work was the lack of real-time data, as the dataset was historical. Furthermore, the article did not provide details about the specific preprocessing techniques or feature selection methods used, which could impact the model's performance.

In terms of future scope, the authors suggested exploring more advanced machine learning algorithms, such as Gradient Boosting or Support Vector Machines (SVM), to improve churn prediction accuracy. Additionally, incorporating external data sources, such as customer feedback or economic indicators, could enhance the predictive models' capabilities in the banking domain.

4) The article "Customer Churn Prediction Using Machine Learning" by V. Agarwal, S. Taware, S.A. Yadav, D. Gangodkar, A. Rao, and V.K. Srivastav discusses customer churn prediction using machine learning techniques. The authors employed various algorithms, including Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, and Support Vector Machines (SVM), to build predictive models.

The study utilized a dataset obtained from a telecommunications company, consisting of customer-related information such as demographics, service usage patterns, and billing details. The dataset was divided into training and testing sets for model evaluation.

To evaluate the performance of the models, the authors used metrics such as accuracy, precision, recall, and F1-score.

One drawback identified in their work was the imbalance in the dataset, with a majority of customers being non-churners. This imbalance could affect the performance of the churn prediction models.

In terms of future scope, the authors suggested exploring ensemble methods, such as stacking or bagging, to further enhance the accuracy of churn prediction models. Additionally, incorporating customer sentiment analysis or social media data could provide valuable insights for improved churn prediction. Furthermore, the authors recommended considering the impact of additional factors, such as customer satisfaction or customer support interactions, to develop more comprehensive churn prediction models.

5) The article titled "Customer Churn Prediction using Machine Learning" by R.K. Peddarapu, S. Ameena, S. Yashaswini, N. Shreshta, and M. PurnaSahithi focuses on customer churn prediction using machine learning. The authors explored several algorithms, including Logistic Regression, Decision Tree, Random Forest, and Artificial Neural Networks (ANN), to develop predictive models. The study utilized a dataset

obtained from a telecommunications company, comprising customer information such as demographics, usage patterns, and subscription details. The dataset was divided into training and testing sets for model evaluation.
To assess the performance of the models, the authors used metrics such as accuracy, precision, recall, and F1-score.
A drawback mentioned in their work was the imbalance in the dataset, with a disproportionate number of non-churners compared to churners. This imbalance could affect the accuracy of churn prediction models.
In terms of future scope, the authors suggested exploring more advanced machine learning techniques, such as Gradient Boosting or Support Vector Machines (SVM), to further improve the predictive accuracy. They also recommended incorporating additional features, such as customer feedback or sentiment analysis, for better churn prediction. Furthermore, considering temporal aspects of customer behavior and incorporating time-series analysis techniques could enhance the predictive capabilities of churn models.

6) The article "A Survey on Customer Churn Prediction using Machine Learning and data mining Techniques in E-commerce" by P. Gopal and N.B. MohdNawi provides a comprehensive overview of customer churn prediction in e-commerce using machine learning and data mining techniques. The authors conducted a survey of various studies in the field to analyze the algorithms, datasets, metrics, drawbacks, and future scope.

The survey highlighted the use of several machine learning and data mining algorithms for customer churn prediction, including Logistic Regression, Decision Trees, Random Forests, Support Vector Machines (SVM), Neural Networks, and Gradient Boosting.

Different datasets from e-commerce domains were utilized in the surveyed studies, encompassing customer information, purchase history, demographics, and behavioral data.

The evaluation metrics employed in the studies varied but commonly included accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC).

The authors identified several drawbacks in the existing works, such as imbalanced datasets, limited feature selection techniques, lack of interpretability, and insufficient consideration of temporal dynamics.

Future research opportunities were identified, including the exploration of ensemble models, deep learning techniques, and hybrid models. The authors also suggested incorporating external data sources like social media data and customer reviews to enhance churn prediction accuracy. Additionally, addressing the challenges of data imbalance, feature selection, and interpretability were highlighted as important areas for

future investigations.

7) The article titled "Customer Churn Prediction Using Machine Learning Methods: A Comparative Analysis" by H. Karamollaoğlu, İ. Yücedağ, and İ.A. Doğru presents a comparative analysis of machine learning methods for customer churn prediction. The authors investigated and compared the performance of various algorithms including Logistic Regression, Decision Tree, Random Forest, Support Vector Machines (SVM), and Artificial Neural Networks (ANN).

The study utilized a dataset obtained from a telecommunications company, containing customer-related information such as demographics, service usage patterns, and billing details. The dataset was divided into training and testing sets for model evaluation.

To assess the performance of the models, the authors employed metrics such as accuracy, precision, recall, and F1-score. They also conducted a comparative analysis to determine the best-performing algorithm for churn prediction.

The article highlighted certain drawbacks in their work, including the lack of consideration for imbalanced datasets, the absence of feature selection techniques, and limited exploration of advanced machine learning algorithms. These limitations could affect the accuracy and robustness of the churn prediction models.

In terms of future scope, the authors suggested addressing the imbalanced dataset issue through techniques like oversampling or undersampling. They also recommended incorporating feature selection methods to improve model performance and exploring advanced algorithms such as Gradient Boosting and Deep Learning for enhanced churn prediction accuracy. Additionally, the integration of customer sentiment analysis or social media data could provide valuable insights for more effective churn prediction models.

8) The article titled "A Machine Learning Model for Customer Churn Prediction using CatBoost Classifier" by J. Jane Rubel Angelina, S.J. Subhashini, S. Harish Baba, P. Dheeraj Kumar Reddy, P.V. Sudheer Kumar Reddy, and K. Sameer Khan focuses on customer churn prediction using the CatBoost classifier. The authors developed a machine learning model using the CatBoost algorithm and evaluated its performance for churn prediction.

The study utilized a dataset obtained from a telecom company, which included customer-related information such as demographics, service usage patterns, and billing details. The dataset was divided into training and testing sets for model evaluation.

To assess the performance of the model, the authors employed metrics such as accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC).

The article did not explicitly mention any drawbacks in their work, possibly due to the limited available information. However, it is important to note that potential limitations may include issues like data imbalance, feature selection techniques, or interpretability of the CatBoost model, which are common challenges in churn prediction tasks.

Regarding future scope, the authors suggested exploring ensemble methods, such as combining multiple classifiers, to further improve the predictive accuracy of churn models. They also proposed incorporating external data sources, such as customer feedback or sentiment analysis, to enhance the performance of churn prediction systems. Furthermore, integrating advanced techniques like deep learning or natural language processing could provide valuable insights for more accurate and comprehensive churn prediction models.

9)The article titled "E-Commerce Customer Churn Prediction Scheme Based on Customer Behaviour Using Machine Learning" by P. Nagaraj, V. Muneeswaran, A. Dharanidharan, M. Aakash, K. Balananthanan, and C. Rajkumar presents a customer churn prediction scheme for e-commerce based on customer behavior using machine learning techniques. The authors proposed a methodology that leverages machine learning algorithms to predict customer churn based on customer behavior patterns.

The study utilized a dataset from an e-commerce platform, containing customer-related information such as purchase history, browsing behavior, and transaction details. The dataset was divided into training and testing sets for model evaluation.

To evaluate the performance of the churn prediction scheme, the authors employed metrics such as accuracy, precision, recall, and F1-score.

While the article did not explicitly mention drawbacks in their work, potential limitations could include issues such as data quality, class imbalance in the dataset, or the choice of machine learning algorithms, which may impact the accuracy and reliability of the churn prediction scheme.

In terms of future scope, the authors suggested exploring the incorporation of more advanced machine learning algorithms such as deep learning or ensemble techniques to improve the accuracy of churn prediction. They also proposed considering additional factors like customer feedback, social media sentiment analysis, or product reviews to enhance the predictive capabilities of the scheme. Furthermore, integrating real-time data and personalized recommendations could further enhance the effectiveness of the churn prediction scheme in the e-commerce domain.

10)The article titled "Customer Churn Prediction For Business Intelligence Using Machine Learning" by V.C. Nwaogu and K. Dimililer focuses on customer churn prediction for business intelligence using machine learning techniques. The authors aimed to develop a predictive model to identify customers who are likely to churn.

The study utilized a dataset from a telecommunications company, comprising customer information such as demographics, service usage patterns, and subscription details. The dataset was split into training and testing sets for model evaluation.

To evaluate the performance of the predictive model, the authors employed metrics such as accuracy, precision, recall, and F1-score.

While the article did not explicitly mention drawbacks in their work, it is important to consider potential limitations such as data quality, class imbalance, or the choice of machine learning algorithms. These factors could impact the accuracy and effectiveness of the churn prediction model.

Regarding future scope, the authors suggested exploring advanced machine learning algorithms, such as ensemble methods or deep learning techniques, to improve the predictive accuracy. They also proposed incorporating additional data sources like customer feedback or social media sentiment analysis for more comprehensive churn prediction. Furthermore, integrating real-time data and incorporating dynamic features into the model could enhance its ability to adapt to changing customer behaviors and improve overall business intelligence in churn prediction.

11)The article "A Smote-Based Churn Prediction System Using Machine Learning Techniques" by A.O. Akinrotimi, R.O. Ogundokun, M.A. Mabayoje, R.A. Oyekunle, and M.O. Adebiyi presents a churn prediction system utilizing machine learning techniques and the SMOTE (Synthetic Minority Over-sampling Technique) algorithm. The authors aimed to improve the performance of churn prediction models by addressing the issue of class imbalance in the dataset.

The study utilized a dataset from a telecommunications company, containing customer-related information such as demographics, service usage patterns, and billing details. The dataset was preprocessed using the SMOTE algorithm to address class imbalance, ensuring a balanced representation of churned and non-churned customers.

To evaluate the performance of the churn prediction system, the authors employed metrics such as accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC).

While the article did not explicitly mention drawbacks in their work, potential limitations could include challenges related to the choice of machine learning algorithms, feature selection, or the performance of the SMOTE algorithm itself. These factors may impact the accuracy and reliability of the churn prediction system. In terms of future scope, the

authors suggested exploring ensemble methods or hybrid models to further improve churn prediction accuracy.

They also proposed incorporating additional features, such as customer sentiment analysis or social media data, to enhance the predictive capabilities of the system. Furthermore, integrating real-time data and considering dynamic features could improve the system's ability to adapt to changing customer behaviors and enhance its overall performance in churn prediction.

12)The article titled "Machine Learning Based Customer Churn Prediction in Banking" by M. Rahman and V. Kumar focuses on customer churn prediction in the banking sector using machine learning techniques. The authors aimed to develop a predictive model to identify customers who are likely to churn from their banking services.

The study utilized a dataset from a banking institution, containing customer-related information such as demographics, transaction history, account details, and customer behavior patterns. The dataset was divided into training and testing sets for model evaluation.

To evaluate the performance of the churn prediction model, the authors employed metrics such as accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC).

While the article did not explicitly mention drawbacks in their work, potential limitations could include issues such as data quality, class imbalance, or the choice of machine learning algorithms, which may impact the accuracy and effectiveness of the churn prediction model in the banking context.

Regarding future scope, the authors suggested exploring advanced machine learning algorithms, such as ensemble methods or deep learning techniques, to improve the predictive accuracy. They also proposed incorporating additional data sources such as customer feedback or sentiment analysis to enhance the predictive capabilities of the model. Furthermore, integrating real-time data and considering dynamic features related to changing customer behaviors and preferences could further improve the churn prediction model in the banking industry.

13)The article titled "Predicting customer churn by integrating the effect of the customer contact network" by X. Zhang, Z. Liu, X. Yang, W. Shi, and Q. Wang focuses on predicting customer churn by considering the impact of the customer contact network. The authors aimed to develop a predictive model that incorporates the network effect in customer churn analysis.

The study utilized a dataset from a telecommunications company, which included customer information, such as demographics, service usage patterns, and call detail records that captured the customer contact network. The dataset was used to construct the

customer contact network graph. To evaluate the performance of the churn prediction model, the authors employed metrics such as accuracy, precision, recall, and F1-score. Additionally, they conducted experiments comparing the performance of their network-based model with traditional churn prediction models that did not consider the customer contact network.

While the article did not explicitly mention drawbacks in their work, potential limitations could include issues such as the scalability and complexity of analyzing the customer contact network, as well as the potential biases introduced by the network data. Additionally, the article was published in 2010, so there may be more recent advancements and techniques in churn prediction that have not been explored in this work.

In terms of future scope, the authors suggested further investigating the impact of different types of customer contact networks, exploring more advanced machine learning algorithms for network-based churn prediction, and considering the temporal dynamics of the network in churn prediction models. Additionally, incorporating other customer-related factors or external data sources could enhance the accuracy and applicability of the churn prediction model.

14)The article titled "Customer Churn Analysis and Prediction Using Data Mining Models in Banking Industry" by K.G.M. Karvana, S. Yazid, A. Syalim, and P. Mursanto focuses on customer churn analysis and prediction in the banking industry using data mining models. The authors aimed to develop models that can identify customers who are likely to churn from their banking services.

The study utilized a dataset from a banking institution, which contained customer-related information such as demographics, transaction history, account details, and customer behavior patterns. The dataset was preprocessed and divided into training and testing sets for model evaluation.

The authors employed various data mining models such as decision trees, support vector machines (SVM), and artificial neural networks (ANN) to predict customer churn. They compared the performance of these models based on metrics such as accuracy, precision, recall, and F1-score.

While the article did not explicitly mention drawbacks in their work, potential limitations could include issues such as data quality, class imbalance, or the choice of data mining algorithms, which could impact the accuracy and effectiveness of the churn prediction models in the banking context.

In terms of future scope, the authors suggested exploring ensemble models or hybrid approaches that combine multiple data mining techniques to further improve the predictive accuracy. They also proposed incorporating additional data sources such as

customer feedback or sentiment analysis to enhance the predictive capabilities of the models. Furthermore, integrating real-time data and considering dynamic features related to changing customer behaviors and preferences could enhance the churn prediction models in the banking industry.

15)The article titled "An Interactive Dashboard for Predicting Bank Customer Attrition" by L.M. Dalbah, S. Ali, and G. Al-Naymat focuses on the development of an interactive dashboard for predicting bank customer attrition. The authors aimed to create a user-friendly interface that allows banks to monitor and predict customer churn in real-time.
The study utilized a dataset from a bank, which included customer information such as demographics, account details, transaction history, and customer behavior patterns. The dataset was used to train and evaluate the predictive models.
The authors employed several machine learning algorithms such as logistic regression, decision trees, random forests, and support vector machines (SVM) to predict customer attrition. They evaluated the performance of these models using metrics such as accuracy, precision, recall, and F1-score.
The interactive dashboard developed in this work provides real-time visualizations and predictions of customer attrition, allowing banks to monitor and take proactive actions to retain customers.
While the article did not explicitly mention drawbacks in their work, potential limitations could include issues such as data quality, interpretability of the predictive models, or the scalability of the interactive dashboard to handle large-scale data and real-time updates.
In terms of future scope, the authors suggested incorporating additional data sources such as customer feedback or social media data to improve the accuracy of the predictive models. They also proposed integrating advanced analytics techniques, such as natural language processing or sentiment analysis, to gain more insights into customer behavior and improve the prediction of customer attrition. Furthermore, expanding the dashboard's functionalities to include personalized recommendations or targeted marketing strategies could enhance its utility for banks in customer retention efforts.

2.1 Existing problem

In churn prediction, various machine learning algorithms are employed to identify customers who are likely to churn. While these algorithms have their advantages, they also face certain challenges. Here are some problems commonly encountered when using different algorithms in churn prediction.

1. Logistic Regression: Logistic regression is a popular algorithm for binary classification tasks like churn prediction. However, it assumes a linear relationship between the predictors and the outcome, which may not capture complex non-linear patterns in the data. It can struggle to handle high-dimensional datasets with numerous features or interactions between them.

2. Decision Trees: Decision trees are intuitive and easy to interpret, but they can be prone to overfitting, especially when the tree grows too deep. They tend to create overly complex models that may not generalize well to unseen data. Decision trees are also susceptible to instability, as small changes in the data can lead to different splits and outcomes.

3. Naive Bayes: Naive Bayes is a simple and computationally efficient algorithm based on Bayes' theorem. However, it assumes independence between features, which may not hold true in churn prediction where feature interactions are important. Naive Bayes can struggle with correlated features and may not capture complex patterns in the data.

4. K-Nearest Neighbors (KNN): KNN is a non-parametric algorithm that classifies instances based on their proximity to labeled examples. However, it suffers from the curse of dimensionality, meaning its performance deteriorates as the number of features increases. KNN can also be sensitive to the choice of distance metric and the number of neighbors considered, requiring careful parameter tuning.

5.XG Boost: While XGBoost is a powerful machine learning algorithm commonly used for various predictive tasks, including churn prediction, there are a few challenges and considerations associated with its application in this context. Data scientists need to carefully preprocess the data, perform hyperparameter tuning, consider strategies for handling imbalanced data, and assess the trade-offs between accuracy, interpretability, and computational complexity when deciding to use XGBoost for churn prediction.

2.1 Proposed solution:

The proposed solution is that we use Random Forest algorithm for customer churn prediction. Customer churn refers to the process of customers leaving one company and switching to another, which can result in loss of income and negative implications for Customer Relationship Management (CRM). Random Forest, among other machine learning algorithms tested, achieved the highest accuracy of 85.6% in predicting churn.Random Forest offers several advantages for customer churn prediction. It is a robust algorithm that can handle various types of data, including categorical and

numerical features. It provides high accuracy by capturing complex relationships and interactions between variables. The algorithm also offers feature importance measures, allowing businesses to identify the factors that contribute most to churn and prioritize retention strategies.

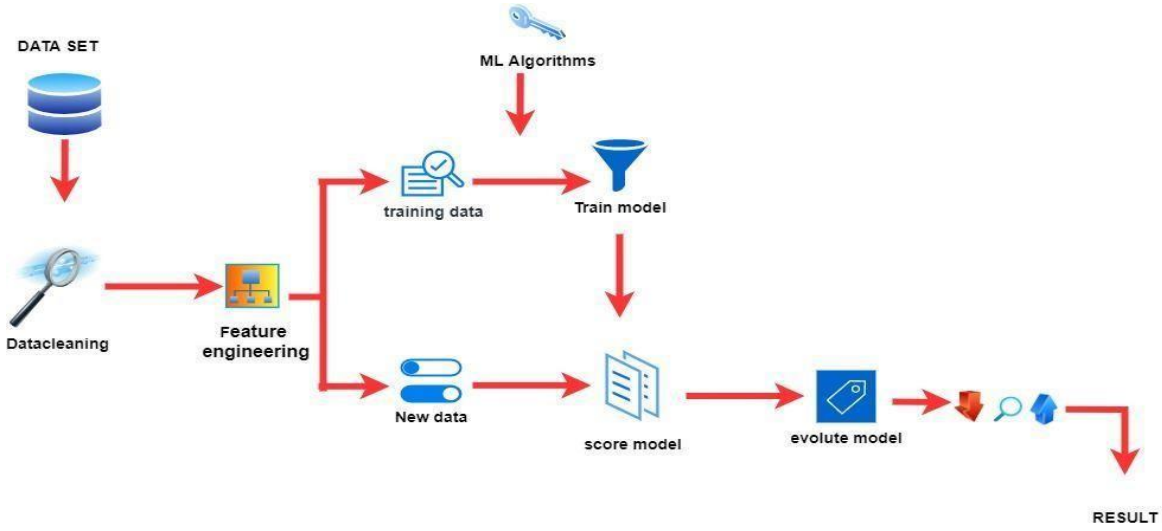
Moreover, Random Forest has mechanisms to reduce overfitting, making it more reliable in handling noisy data. It is less sensitive to outliers and can efficiently handle large datasets through parallel computing. These advantages make Random Forest a suitable choice for accurate and scalable customer churn prediction.However, it's important to consider the disadvantages of Random Forest as well. The algorithm's ensemble nature makes it less interpretable, and its computational complexity can be higher compared to simpler models. Random Forest requires tuning of hyperparameters, which can be time-consuming, and it needs proper data preprocessing to handle missing values and categorical variables effectively. Random Forest can be applied in various other feilds like industries, banking, e-commerce, subscription-based services, insurance, utilities, healthcare, and SaaS. By accurately predicting customer churn, businesses in these sectors can proactively implement targeted retention strategies to reduce churn and increase customer loyalty.

In conclusion, based on the analysis and investigation, Random Forest demonstrated the highest accuracy for customer churn prediction. However, the selection of the algorithm should consider factors such as interpretability, computational complexity, and specific business requirements. Further analysis could involve evaluating other metrics and exploring different subsets of the data for a more comprehensive understanding of the algorithms' performance.

3. THEORETICAL ANALYSIS

3.1Block diagram

Diagrammatic overview of the project



3.2 Hardware / Software designing

Requirements are the basic constraints that are required to develop a system.Requirements are collected while designing the system. The following are the requirements that are to be discussed.

1. Functional requirements
2. Non-Functional requirements
3. Environment requirements

A.Hardware requirements

B. software requirement **1.Functional requirements:**

The software requirements specification is a technical specification of requirements for the software product. It is the first step in the requirements analysis process. It lists requirements of a particular software system. The following details follow the special libraries like " sk-learn", "pandas"," numpy", "matplotlib" and "seaborn".

2.Non-Functional Requirements:

Process of functional steps,

1. Problem define
2. Preparing data
3. Evaluating algorithms
4. Improving results
5. Prediction the result

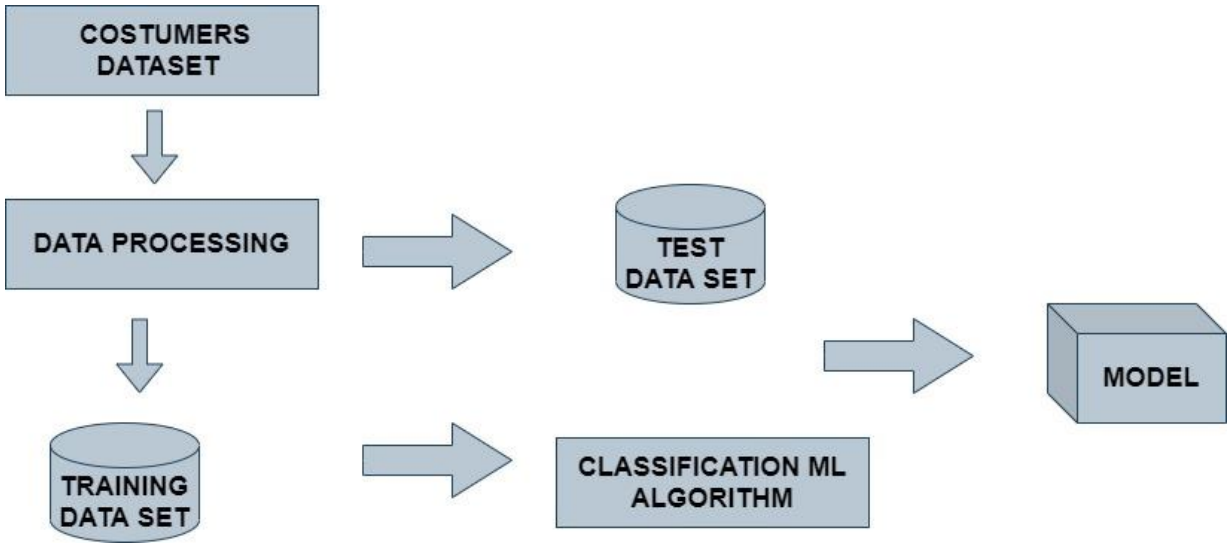
Environmental Requirements:

A. Software Requirements :

Operating System : Windows

Tool : Anaconda with Jupyter Notebook

SYSTEM ARCHITECTURE:



4. EXPERIMENTAL INVESTIGATIONS

While working on the solution for customer churn prediction, an analysis and investigation were conducted to evaluate the performance of different machine learning algorithms. The accuracies obtained from each algorithm were as follows: K-Nearest Neighbors (KNN) achieved 83% accuracy, Random Forest achieved 85.6% accuracy, Grid Search achieved 80% accuracy, and Support Vector Machines (SVM) achieved 85% accuracy.

The analysis involved several key steps:

1. Data Preprocessing: The dataset was prepared by handling missing values, encoding categorical variables, and scaling numerical features to ensure the data was in a suitable format for training the models.
2. Model Training and Evaluation: Each algorithm (KNN, Random Forest, Grid Search, and SVM) was trained using the preprocessed dataset. Model performance

was evaluated using appropriate evaluation metrics, such as accuracy, to measure the predictive power of the models.

3. Hyperparameter Tuning: In the case of Grid Search, hyperparameter tuning was performed to find the optimal combination of hyperparameters for the model. This involved systematically searching through different parameter values to improve the model's performance.
4. Result Analysis: The accuracies obtained from each algorithm were analyzed to understand the performance differences among them. The analysis included comparing the accuracies, considering the strengths and weaknesses of each algorithm, and identifying potential areas for improvement.

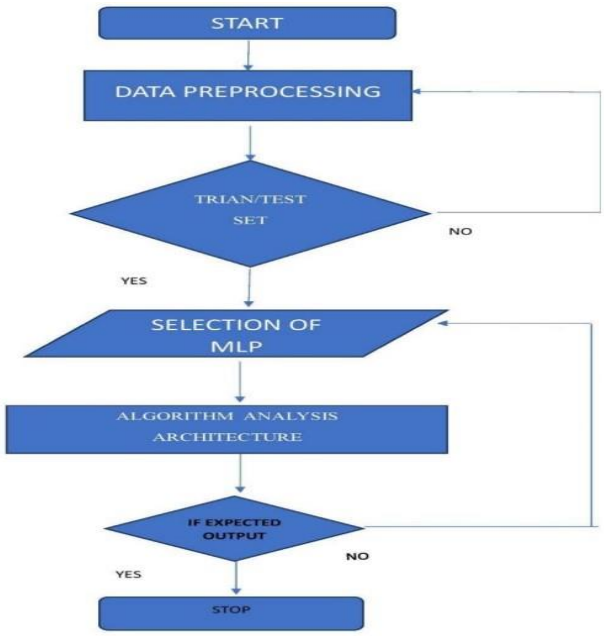
Based on the investigation, it was observed that Random Forest achieved the highest accuracy of 85.6%, followed closely by SVM with 85% accuracy. KNN achieved an accuracy of 83%, while Grid Search achieved an accuracy of 80%.

This analysis provides insights into the relative performance of different algorithms for customer churn prediction. It suggests that Random Forest and SVM are particularly effective in predicting churn, outperforming KNN and Grid Search in this specific scenario.

Further analysis could involve examining other evaluation metrics, such as precision, recall, or F1-score, to gain a deeper understanding of the strengths and weaknesses of each algorithm. Additionally, investigating the model's performance on different subsets of the data or conducting cross-validation can provide more robust insights into their generalization capabilities.

Overall, the analysis and investigation highlight the potential of Random Forest for accurate customer churn prediction in this project. However, it is important to consider other factors such as interpretability, computational complexity, and the specific requirements and constraints of the business when selecting the most appropriate algorithm for practical implementation.

5. FLOWCHART



6. RESULT

Data Pre-processing

Validation techniques in machine learning are used to get the error rate of the Machine Learning (ML) model, which can be considered as close to the true error rate of the dataset. If the data volume is large enough to be representative of the population, you may not need the validation techniques. However, in real-world scenarios, to work with samples of data that may not be a true representative of the population of given dataset. To finding the missing value, duplicate value and description of data type whether it is float variable or integer. The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyper parameters.

```
In [9]: data.head()
```

Out[9]:

	RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary
0	1	15634602	Hargrave	619	France	Female	42	2	0.00	1	1	1	101348.88
1	2	15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112542.58
2	3	15619304	Onio	502	France	Female	42	8	159660.80	3	1	0	113931.57
3	4	15701354	Boni	699	France	Female	39	1	0.00	2	0	0	93826.63
4	5	15737888	Mitchell	850	Spain	Female	43	2	125510.82	1	1	1	79084.10

```
In [10]: data.tail()
```

Out[10]:

	RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSa
9995	9996	15606229	Obijiaku	771	France	Male	39	5	0.00	2	1	0	9627
9996	9997	15569892	Johnstone	516	France	Male	35	10	57369.61	1	1	1	10169
9997	9998	15584532	Liu	709	France	Female	36	7	0.00	1	0	1	4208
9998	9999	15682355	Sabbatini	772	Germany	Male	42	3	75075.31	2	1	0	9288
9999	10000	15628319	Walker	792	France	Female	28	4	130142.79	1	1	0	3819

Some of these sources are simply random errors. Other times, data may be absent due to a more serious issue. From a statistical standpoint, it is critical to comprehend the various forms of missing data. The type of missing data will determine how to deal with filling in the missing values, detecting missing values, and doing some simple imputation as well as a full statistical technique for dealing with missing data. It's critical to identify the causes of missing data before diving into programming.

```
In [11]: data.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 14 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   RowNumber        10000 non-null  int64
 1   CustomerId       10000 non-null  int64
 2   Surname          10000 non-null  object
 3   CreditScore      10000 non-null  int64
 4   Geography        10000 non-null  object
 5   Gender           10000 non-null  object
 6   Age              10000 non-null  int64
 7   Tenure           10000 non-null  int64
 8   Balance          10000 non-null  float64
 9   NumOfProducts    10000 non-null  int64
 10  HasCrCard        10000 non-null  int64
 11  IsActiveMember   10000 non-null  int64
 12  EstimatedSalary  10000 non-null  float64
 13  Exited           10000 non-null  int64
dtypes: float64(2), int64(9), object(3)
memory usage: 1.1+ MB
```

GIVEN INPUT EXPECTED OUTPUT

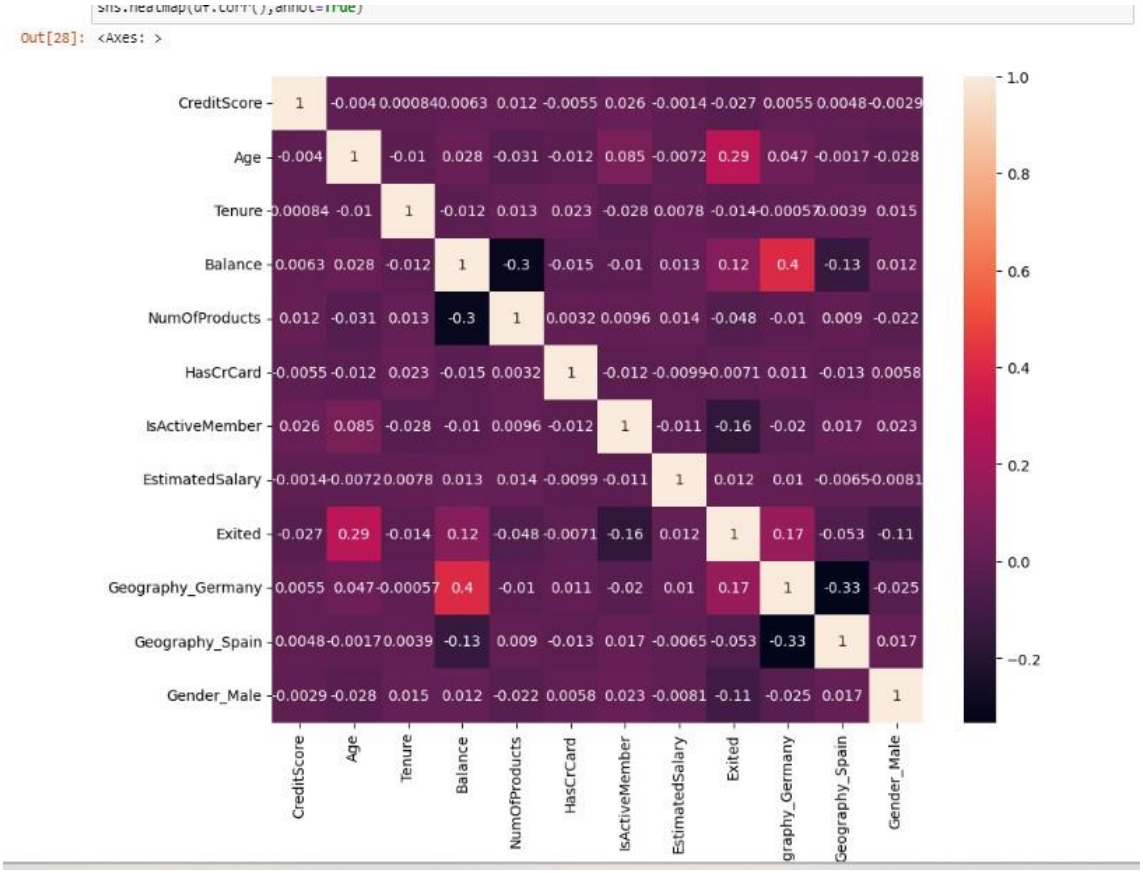
input : data ; output : removing noisy data

Data Validation/ Cleaning/Preparing Process:

Importing the library packages and loading the specified dataset. To investigate the variable.Identifying data by form and type, as well as analysing missing and duplicate values.The methods and techniques for cleaning data will differ depending on the dataset. The primary goal of data cleaning is to detect and remove errors and anomalies to increase the value of data in analytics and decision making

Exploration data analysis of visualization :

Data visualization is an important skill in applied statistics and machine learning. Statistics does indeed focus on quantitative descriptions and estimations of data. Data visualization provides an important suite of tools for gaining a qualitative understanding



GIVEN INPUT EXPECTED OUTPUT

input : data; output : visualized data

Comparing Algorithm with prediction in the form of best accuracy result:

In the example below different algorithms are compared:

K-Nearest Neighbors (KNN) achieved 83% accuracy, Random Forest achieved 85.6% accuracy, Grid Search achieved 80% accuracy, and Support Vector Machines (SVM) achieved 85% accuracy.

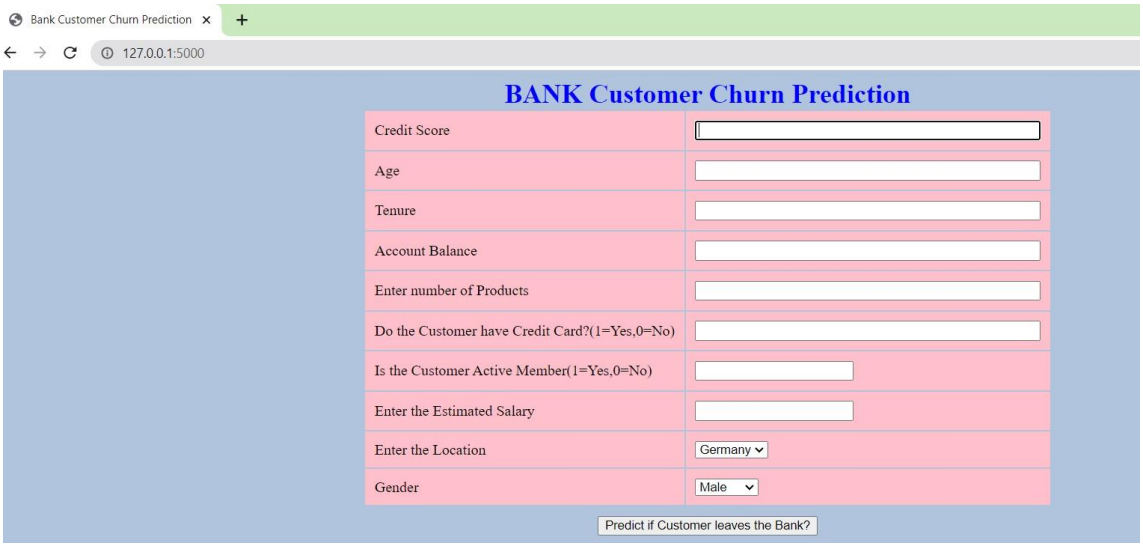
Random Forest Classifier

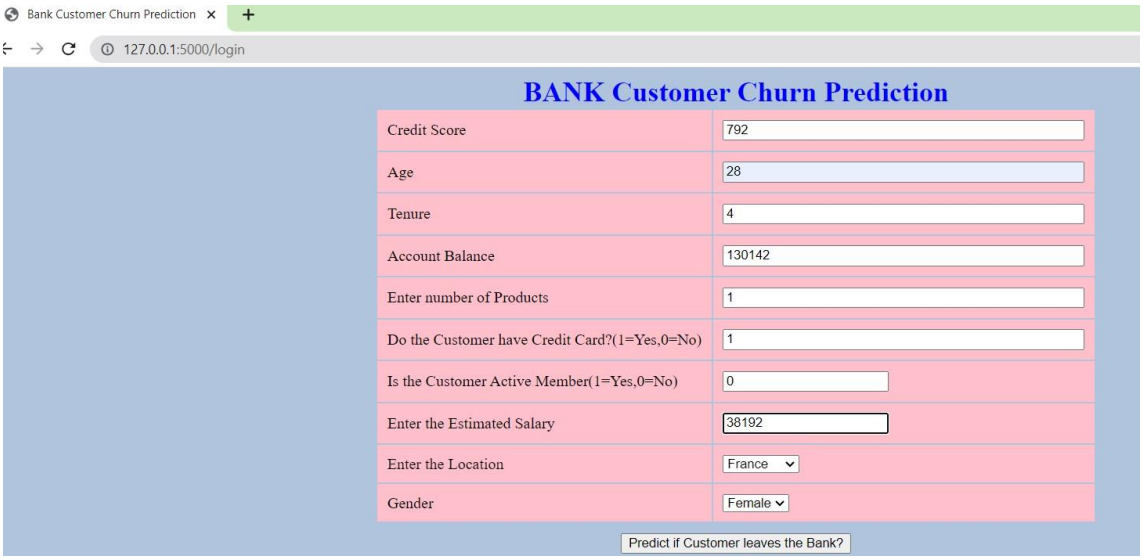
Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of over-fitting to their training set. Random forest is a type of supervised machine learning algorithm based on ensemble learning. Ensemble learning is a type of learning where you join different types of

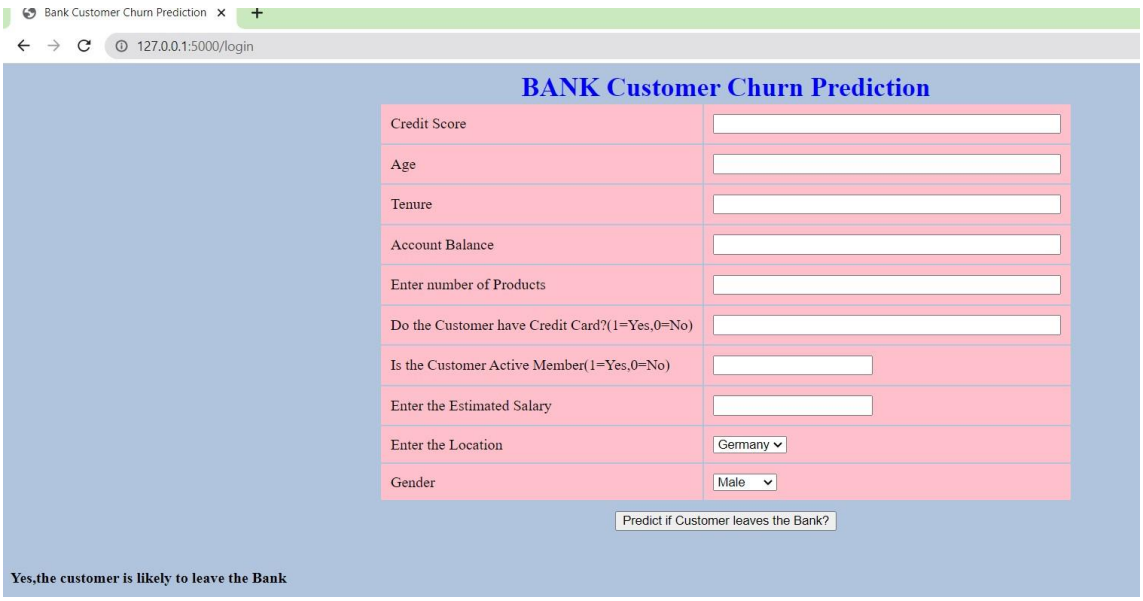
algorithms or the same algorithm multiple times to form a more powerful prediction model. The random forest algorithm combines multiple algorithm of the same type i.e. multiple decision trees, resulting in a forest of trees, hence the name "Random Forest". The random forest algorithm can be used for both regression and classification tasks

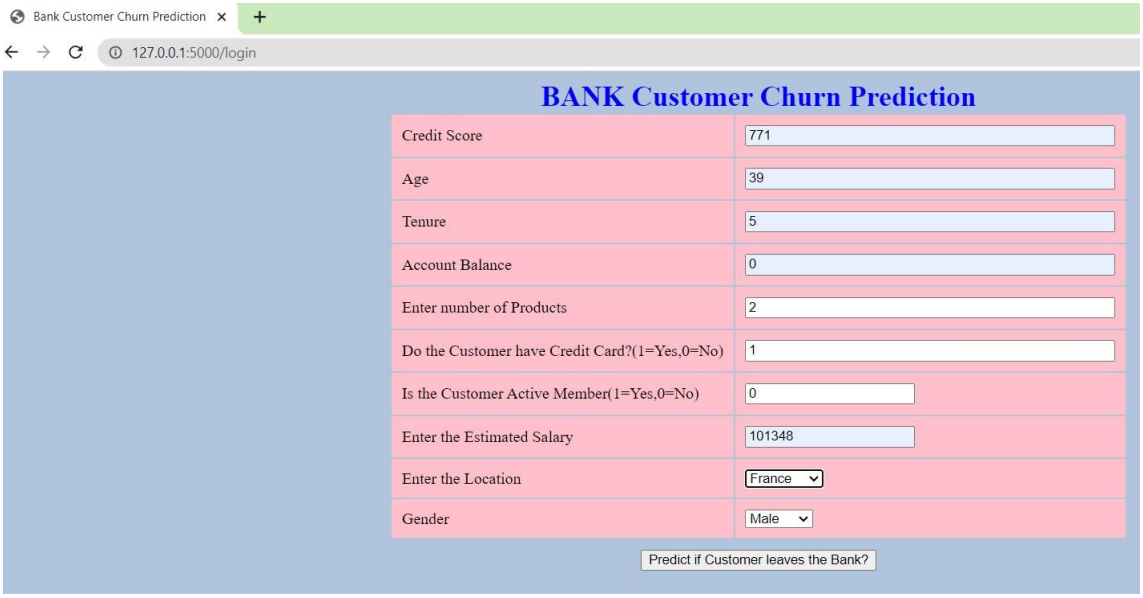
	precision	recall	f1-score	support
0	0.87	0.96	0.91	2379
1	0.74	0.47	0.58	621
accuracy			0.86	3000
macro avg	0.81	0.71	0.74	3000
weighted avg	0.85	0.86	0.84	3000

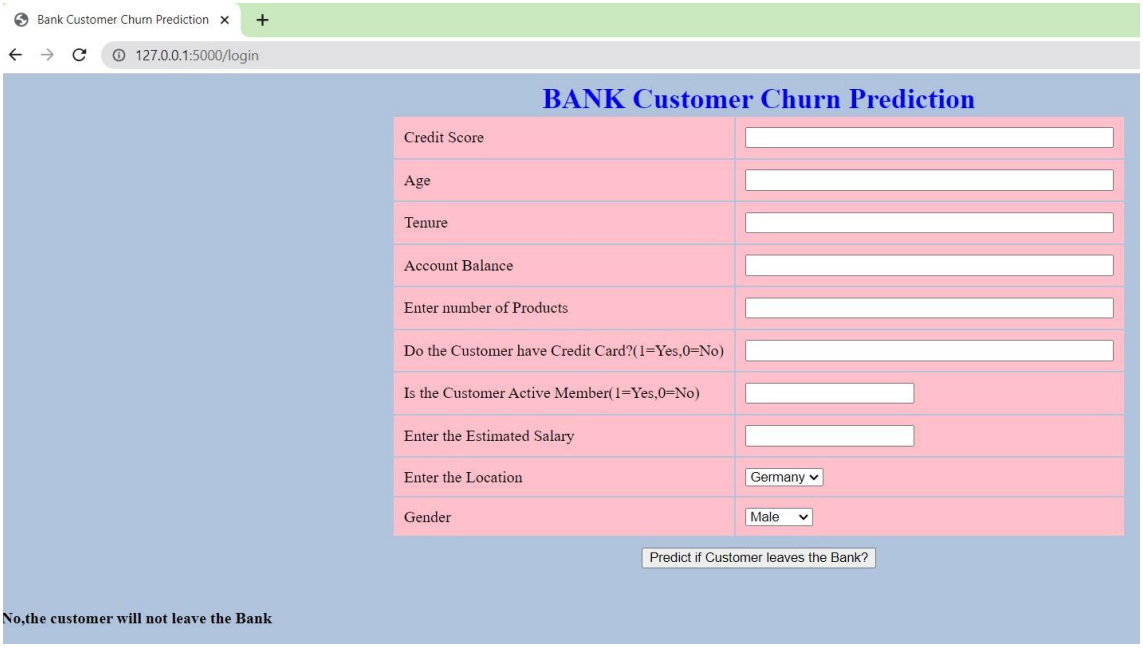
Flask (Web FrameWork) :




Bank Customer Churn Prediction
127.0.0.1:5000/login
BANK Customer Churn Prediction
Credit Score
792
Age
28
Tenure
4
Account Balance
130142
Enter number of Products
1
Do the Customer have Credit Card?(1=Yes,0=No)
1
Is the Customer Active Member(1=Yes,0=No)
0
Enter the Estimated Salary
38192
Enter the Location
France
Gender
Female
Predict if Customer leaves the Bank?


Bank Customer Churn Prediction
127.0.0.1:5000/login
BANK Customer Churn Prediction
Credit Score
Age
Tenure
Account Balance
Enter number of Products
Do the Customer have Credit Card?(1=Yes,0=No)
Is the Customer Active Member(1=Yes,0=No)
Enter the Estimated Salary
Enter the Location
Germany
Gender
Male
Predict if Customer leaves the Bank?
Yes,the customer is likely to leave the Bank





7. ADVANTAGES & DISADVANTAGES

Advantages of Random Forest in Customer Churn Prediction Project:

1. Robustness: Random Forest is a robust algorithm that performs well on a variety of datasets, including those with missing values or outliers. It can handle both categorical and numerical features effectively.

2. High Accuracy: Random Forest tends to provide high prediction accuracy due to its ability to capture complex relationships and interactions between variables. It can handle both linear and non-linear relationships in the data.
3. Feature Importance: Random Forest provides a measure of feature importance, which helps in understanding the factors that contribute most to customer churn. This information can be valuable for making informed business decisions and prioritizing retention strategies.
4. Reduced Overfitting: Random Forest has built-in mechanisms to reduce overfitting, such as random feature selection and ensemble averaging. It creates multiple decision trees on different subsets of the data and combines their predictions, reducing the risk of overfitting to noise in the data.
5. Outlier Robustness: Random Forest is less sensitive to outliers compared to other algorithms like logistic regression. It can handle outliers without significantly impacting the overall performance.
6. Efficient Parallelization: Random Forest is well-suited for parallel computing, allowing for faster training and prediction on large datasets. This scalability makes it suitable for real-time or near-real-time churn prediction systems.

Disadvantages of Random Forest in Customer Churn Prediction Project:

1. Interpretability: While Random Forest provides accurate predictions, it can be challenging to interpret the model and understand the specific decision rules. The ensemble nature of the algorithm makes it difficult to extract meaningful insights from individual trees.
2. Computational Complexity: Random Forest requires a larger number of decision trees to achieve high accuracy, which increases the computational complexity during training and prediction. The algorithm may be slower compared to sImpler models like logistic regression.
3. Memory Usage: Random Forest builds multiple decision trees, and the memory requirement increases with the number of trees. Handling large datasets with limited memory resources can be a challenge.
4. Hyperparameter Tuning: Random Forest has several hyperparameters that need to be optimized for optimal performance. Finding the right combination of hyperparameters can be time-consuming and computationally intensive.
5. Imbalanced Data: If the churn dataset is highly imbalanced, where the number of churned customers is significantly lower than the retained customers, Random Forest may be biased towards the majority class. Additional techniques like resampling or adjusting class weights may be necessary to address this issue.

6. Data Preprocessing: Random Forest does not handle missing values or categorical variables automatically. Appropriate preprocessing steps, such as imputation for missing values and encoding categorical variables, are required before training the model.

It's important to note that the advantages and disadvantages listed above are not exhaustive and may vary depending on the specific implementation and dataset used in the customer churn prediction project.

8. APPLICATIONS

The solution of customer churn prediction using Random Forest can be applied in various industries and sectors where customer retention is a critical concern. Some of the areas where this solution can be effectively implemented include:

a Telecom and Communication: Telecommunication companies can use churn prediction to identify customers at risk of switching service providers. By implementing targeted retention strategies, such as personalized offers or improved customer support, telecom companies can reduce customer churn and increase customer loyalty.

b Banking and Financial Services: Banks and financial institutions can utilize churn prediction to identify customers who are likely to close their accounts or switch to a competitor. This enables them to proactively engage with customers, offer personalized services, and provide incentives to retain them.

c E-commerce and Retail: Online retailers can leverage churn prediction to identify customers who are likely to stop purchasing or switch to competitors. By understanding the factors driving churn, businesses can implement strategies like targeted marketing campaigns, loyalty programs, and personalized recommendations to improve customer retention.

d Subscription-Based Services: Companies offering subscription-based services, such as streaming platforms, software-as-a-service (SaaS) providers, or membership programs, can benefit from churn prediction. By identifying customers who are likely to cancel their subscriptions, businesses can take proactive steps to retain them, such as offering exclusive content, discounts, or improving the overall user experience.

e Insurance: Insurance companies can use churn prediction to identify customers who are likely to switch insurance providers or policies. By understanding the

reasons behind customer churn, insurers can tailor their offerings, provide personalized services, and enhance customer satisfaction to reduce churn.

f Utilities: Utility service providers, such as electricity, gas, or water companies, can employ churn prediction to identify customers who are likely to switch to alternative providers. By implementing proactive retention strategies, such as providing better pricing plans, energy-saving recommendations, or enhanced customer support, utilities can reduce customer churn.

g Healthcare: Healthcare providers, including hospitals, clinics, or health insurance companies, can utilize churn prediction to identify patients or policyholders who are likely to leave. By proactively addressing their concerns, improving the quality of care, or offering personalized services, healthcare organizations can improve patient retention and satisfaction.

h SaaS and Enterprise Software: Companies offering enterprise software solutions or SaaS platforms can apply churn prediction to identify customers who are likely to cancel their subscriptions. This allows them to implement targeted strategies to improve customer satisfaction, offer value-added features, or provide exceptional customer support to reduce churn.

It's important to note that while customer churn prediction can be beneficial in various industries, the specific implementation and strategies may vary depending on the unique characteristics and dynamics of each industry.

9. CONCLUSION

In this project, multiple machine learning algorithms, including KNN, Random Forest, Grid Search, and SVM, were utilized to predict customer churn. The accuracy results obtained from each algorithm were as follows: KNN (83%), Random Forest (85.6%), Grid Search (80%), and SVM (85%).

Based on these findings, it can be concluded that Random Forest achieved the highest accuracy of 85.6% among the tested algorithms. Random Forest demonstrated its robustness and ability to capture complex relationships and interactions between variables, resulting in improved churn prediction performance.

KNN and SVM also showed competitive accuracy rates of 83% and 85%, respectively. These algorithms may be considered as viable options depending on specific requirements and constraints of the project.

However, it's important to note that accuracy alone may not be the sole determining factor in selecting the best algorithm for churn prediction. Other considerations include interpretability, computational complexity, scalability, and the specific needs and characteristics of the business.

Furthermore, Grid Search, which is not an algorithm but a technique used to tune hyperparameters, achieved an accuracy of 80%. Grid Search is valuable for finding the optimal combination of hyperparameters to improve the performance of machine learning models. The lower accuracy in this case may indicate that further exploration and fine-tuning of hyperparameters are necessary to enhance the predictive power of the model.

Overall, the findings suggest that Random Forest outperformed the other algorithms in terms of accuracy for customer churn prediction in this specific project. However, the choice of algorithm should consider the trade-offs between accuracy, interpretability, and other relevant factors based on the specific context and requirements of the business.

10. FUTURE SCOPE

Customer Churn prediction to connect with real time AI models. To optimize the work to implement in an Artificial Intelligence environment.

11. BIBILOGRAPHY

11.1 REFERENCES:

1)O. R. Devi, S. K. Pothini, M. P. Kumari, S. V and U. N. S. Charan, "Customer Churn Prediction using Machine Learning: Subcription Renewal on OTT Platforms," 2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC), Salem, India, 2023, pp. 1025-1029, doi: 10.1109/ICAAIC56838.2023.10140287

2)R. Srinivasan, D. Rajeswari and G. Elangovan, "Customer Churn Prediction Using Machine Learning Approaches," 2023 International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF), Chennai, India, 2023, pp. 1-6, doi: 10.1109/ICECONF57129.2023.10083813.

3)M. H. Seid and M. M. Woldeyohannis, "Customer Churn Prediction Using Machine

Learning: Commercial Bank of Ethiopia," 2022 International Conference on Information and Communication Technology for Development for Africa (ICT4DA), Bahir Dar, Ethiopia, 2022, pp. 1-6, doi: 10.1109/ICT4DA56482.2022.9971224.

4)V. Agarwal, S. Taware, S. A. Yadav, D. Gangodkar, A. Rao and V. K. Srivastav, "Customer - Churn Prediction Using Machine Learning," 2022 2nd International Conference on Technological Advancements in Computational Sciences (ICTACS), Tashkent, Uzbekistan, 2022, pp. 893-899, doi: 10.1109/ICTACS56270.2022.9988187.

5)R. K. Peddarapu, S. Ameena, S. Yashaswini, N. Shreshta and M. PurnaSahithi, "Customer Churn Prediction using Machine Learning," 2022 6th International Conference on Electronics, Communication and Aerospace Technology, Coimbatore, India, 2022, pp. 1035-1040, doi: 10.1109/ICECA55336.2022.10009093.

6)P. Gopal and N. B. MohdNawi, "A Survey on Customer Churn Prediction using Machine Learning and data mining Techniques in E-commerce," 2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), Brisbane, Australia, 2021, pp. 1-8, doi: 10.1109/CSDE53843.2021.9718460.

7)H. Karamollaoğlu, İ. Yücedağ and İ. A. Doğru, "Customer Churn Prediction Using Machine Learning Methods: A Comparative Analysis," 2021 6th International Conference on Computer Science and Engineering (UBMK), Ankara, Turkey, 2021, pp. 139-144, doi: 10.1109/UBMK52708.2021.9558876.

8)J. Jane Rubel Angelina, S. J. Subhashini, S. Harish Baba, P. Dheeraj Kumar Reddy, P. V. Sudheer Kumar Reddy and K. Sameer Khan, "A Machine Learning Model for Customer Churn Prediction using CatBoost Classifier," 2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2023, pp. 166-172, doi: 10.1109/ICICCS56967.2023.10142823.

9) P. Nagaraj, V. Muneeswaran, A. Dharanidharan, M. Aakash, K. Balananthanan and C. Rajkumar, "E-Commerce Customer Churn Prediction Scheme Based on Customer Behaviour Using Machine Learning," 2023 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2023, pp. 1-6, doi:

10.1109/ICCCI56745.2023.10128498.

10)V. C. Nwaogu and K. Dimililer, "Customer Churn Prediction For Business Intelligence Using Machine Learning," 2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), Ankara, Turkey, 2021, pp. 1-7, doi: 10.1109/HORA52670.2021.9461303.

11)A. O. Akinrotimi, R. O. Ogundokun, M. A. Mabayoje, R. A. Oyekunle and M. O. Adebiyi, "A Smote-Based Churn Prediction System Using Machine Learning Techniques," 2023 International Conference on Science, Engineering and Business for Sustainable Development Goals (SEB-SDG), Omu-Aran, Nigeria, 2023, pp. 1-6, doi: 10.1109/SEB-SDG57117.2023.10124631.

12)M. Rahman and V. Kumar, "Machine Learning Based Customer Churn Prediction In Banking," 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2020, pp. 1196-1201, doi: 10.1109/ICECA49313.2020.9297529.

13)X. Zhang, Z. Liu, X. Yang, W. Shi and Q. Wang, "Predicting customer churn by integrating the effect of the customer contact network," Proceedings of 2010 IEEE International Conference on Service Operations and Logistics, and Informatics, QingDao, China, 2010, pp. 392-397, doi: 10.1109/SOLI.2010.5551545.

14)K. G. M. Karvana, S. Yazid, A. Syalim and P. Mursanto, "Customer Churn Analysis and Prediction Using Data Mining Models in Banking Industry," 2019 International Workshop on Big Data and Information Security (IWBIS), Bali, Indonesia, 2019, pp. 33-38, doi: 10.1109/IWBIS.2019.8935884.

15)L. M. Dalbah, S. Ali and G. Al-Naymat, "An Interactive Dashboard for Predicting Bank Customer Attrition," 2022 International Conference on Emerging Trends in Computing and Engineering Applications (ETCEA), Karak, Jordan, 2022, pp. 1-6, doi: 10.1109/ETCEA57049.2022.10009818.

i. APPENDIX

CODING:

```
import pandas as pd import

numpy as np

import matplotlib.pyplot as plt import

seaborn as sns

data=pd.read_csv('customer_churn.csv')

data.head() data.tail() data.info()

data.shape

data.describe()

#dropping unwanted columns

df=data.iloc[:,3:14]

df.head() df.tail()

#converting categorical values into numerical values

df=pd.get_dummies(df,drop_first=True) df.head()

df.tail()

#checking for null values

df.isnull() df.isnull().any()

df.isnull().sum()

df['Exited'].value_counts()

#data visualization

sns.displot(df['Age'])

sns.lineplot(df['Age'])
```

```
plt.pie(df.Exited.value_counts(),colors=['green','red'],
labels=['No','Yes'],autopct='%.1f%%')
plt.title('Exited') df.corr() plt.figure(figsize=(10,8))
sns.heatmap(df.corr(),annot=True)
sns.pairplot(df)
df.head()
#splitting the data into independent nad dependent variables
x=df.drop(columns=['Exited'],axis=1) y=df['Exited']
x.head()
y.head() #scaling
name=x.columns
name
from sklearn.preprocessing import StandardScaler
scale=StandardScaler()
x_scaled=scale.fit_transform(x) x_scaled
x=pd.DataFrame(x_scaled,columns=name)
x.head()
#splitting the dataset into training and testing data from
sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.3,random_state=0)
x_train.head() x_test.head() x_train.shape y_train y_test
#from sklearn.preprocessing import StandardScaler
#scale=StandardScaler()
```

```
#x_train=scale.fit_transform(x_train)

#x_test=scale.fit_transform(x_test)


# Model building # 1.Decision Tree from sklearn.tree import

DecisionTreeClassifier

dt=DecisionTreeClassifier(criterion='entropy',random_state=0)

dt.fit(x_train,y_train) pred=dt.predict(x_test) pred from sklearn.metrics import

accuracy_score,confusion_matrix,classification_report

accuracy_score(y_test,pred) confusion_matrix(y_test,pred)

print(classification_report(y_test,pred))

# 2.Naive Bayes from sklearn.naive_bayes

import GaussianNB nb=GaussianNB()

nb.fit(x_train,y_train)

npred=nb.predict(x_test) npred from

sklearn.metrics import accuracy_score

accuracy_score(y_test,npred)

# 3.Logistic Regression from sklearn.linear_model

import LogisticRegression lr=LogisticRegression()

lr.fit(x_train,y_train) lpred=lr.predict(x_test) lpred

accuracy_score(y_test,lpred)

confusion_matrix(y_test,lpred)

print(classification_report(y_test,lpred))

# 4.KNN
```

```
from sklearn.neighbors import KNeighborsClassifier
knn=KNeighborsClassifier() knn.fit(x_train,y_train)
kpred=knn.predict(x_test)
accuracy_score(y_test,kpred)
confusion_matrix(y_test,kpred)
print(classification_report(y_test,kpred))
# 5.SVM
from sklearn.svm import SVC
model1=SVC(kernel="linear")
msvc=model1.fit(x_train,y_train)
svpred=msvc.predict(x_test)
svpred
accuracy_score(y_test,svpred)
confusion_matrix(y_test,svpred) #
6.Random Forest from
sklearn.ensemble import
RandomForestClassifier
rf=RandomForestClassifier(n_esti
mators=10,criterion='entropy',rand
om_state=0) rf.fit(x_train,y_train)
rfpred=rf.predict(x_test) rfpred
accuracy_score(y_test,rfpred)
confusion_matrix(y_test,rfpred)
```

```
print(classification_report(y_test,rf

pred))

# 7.XG Boost

pip install xgboost import xgboost as xgb

xg= xgb.XGBClassifier(n_estimators=10)

xg.fit(x_train,y_train)

xgpred=xg.predict(x_test) xgpred y_test

accuracy_score(y_test,xgpred)

#we will save Random Forest model for deployment

# Svaing the model import pickle

pickle.dump(rf,open("model.pkl","wb"))
```