**Bank Customer Churn Prediction**

Project submitted to the

SRM University – AP, Andhra Pradesh

for the partial fulfillment of the requirements to award the degree of

**Bachelor of Technology**

In

**Computer Science and Engineering**

**School of Engineering and Sciences**

Submitted by

**Sreeram Kandukuri-AP21110010100**

**Mainak Das Gupta-AP21110010112**

**Vijaya Vyshnavi Muvvala-AP21110010119**

**A picture containing text

Description automatically generated**

Under the Guidance of

**Ratna Raju Mekala**

SRM University–AP

Neerukonda, Mangalagiri, Guntur

Andhra Pradesh – 522 240

May, 2023

**Table Of Contents**

1. Introduction…………………………………………………… 3
2. Problem Statement…………………………………………… 4
3. Dataset…………………………………………………………. 4
4. Proposed Approach…………………………………………… 5
5. Evaluation……………………………………………………… 7
6. Conclusion………….………………………………………….. 10

**Introduction**

In the ever-changing world of financial services, focusing on keeping customers is a crucial necessity for banks and financial institutions. Maintaining current customers not only guarantees a steady income but also nurtures long-lasting connections and strengthens brand loyalty. However, the ongoing issue of customer churn continues to be a challenge, with clients ending their banking relationships for reasons such as dissatisfaction with service, competition from other banks, or personal life changes.  
Effectively dealing with customer turnover requires a proactive approach based on data-driven insights and predictive analytics. Banks can use historical data on customer interactions, transactional behaviour, demographic profiles, and nuanced behavioural patterns to create predictive models that can identify potential churners early on. These models help to efficiently allocate resources, allowing for personalized retention efforts that connect with different customer groups and ultimately reducing churn rates.  
This report thoroughly investigates the complexities of predicting bank customer churn, exploring the methods, algorithms, and best practices in the industry used to accurately forecast churn. By clarifying the factors that cause customers to leave and utilizing the forecasting skills of analytics, banks can strengthen their efforts to keep customers, improve marketing plans, and cultivate lasting relationships with clients. By conducting thorough analysis including churn prediction strategies and relevant case studies, this report aims to provide useful insights to banks dealing with customer churn in a highly competitive market environment.

**Problem Statement**

Bank customer churn prediction is essentially formulated as a classification problem, with the goal of constructing a classifier capable of discerning whether a customer is likely to churn or remain with the bank. The target variable in this classification paradigm distinguishes between churn and non-churn scenarios.

The aim of the problem is to utilize various machine learning algorithms to predict bank customer churn and subsequently determine the most effective algorithm for this task.

**Dataset:**

The dataset comprises information on bank customers, encompassing a range of features such as credit score, geographic location, gender, age, tenure, balance, number of products held, credit card possession, active membership status, estimated salary, and an indicator of whether the customer has exited the bank (denoted as the target variable).The target variable, Exited, is binary, with a value of 1 indicating that the customer has exited the bank and a value of 0 indicating that the customer is still active.

# 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

**Proposed Approach**

Using machine learning methods like Random Forests, Gradient Boosting, Decision Trees, Logistic Regression, k-Nearest Neighbours (k-NN), Support Vector Machines (SVM) with SVC, and Logistic Regression can be very helpful in this problem. A brief summary of each algorithm's contribution to churn prediction is provided below:

**Support Vector Machine (SVM):**

SVM, known for its efficacy in high-dimensional spaces, can effectively delineate classes.

In churn prediction, SVC is trained on labelled historical data, with features representing customer behaviours, demographics, and usage patterns.

Subsequently, the trained SVC model predicts the likelihood of churn for new customers based on their feature vectors.

**k-Nearest Neighbors (kNN):**

kNN is a straightforward algorithm that classifies data points based on the majority class of their k nearest neighbors.

In churn prediction, kNN identifies similar customers by assessing feature similarity and predicts churn based on the churn status of their nearest neighbors.

**Logistic Regression:**

Logistic Regression models the probability of a binary outcome (churn or retention) based on predictor variables.

In churn prediction, Logistic Regression utilizes historical data to estimate the probability of churn for new customers, leveraging their feature values.

**Decision Trees:**

Decision Trees recursively partition the feature space into subsets, guided by the most informative features at each node.

In churn prediction, Decision Trees identify significant predictors of churn and make predictions based on these features.

**Random Forests:**

Random Forests, an ensemble learning method, construct multiple decision trees and aggregate their predictions.

In churn prediction, Random Forests enhance prediction accuracy and mitigate overfitting by leveraging multiple Decision Trees.

**Gradient Boosting:**

Gradient Boosting sequentially builds an ensemble of weak learners, with each new learner rectifying errors made by its predecessors.

In churn prediction, Gradient Boosting algorithms such as XGBoost or LightGBM iteratively refine predictive models, leading to high accuracy.

**Evaluation**

**Confusion Matrix and Formulas:**

A Confusion Matrix provides a snapshot of a classification model's performance, categorizing predictions into four groups:

* True Positive (TP)
* True Negative (TN)
* False Positive (FP)
* False Negative (FN)

|  |  |  |
| --- | --- | --- |
|  | Negative | Positive |
| Negative | True Negative | False Negative |
| Positive | False Positive | True Positive |

**Formulas:**

Accuracy (all correct / all) = TP + TN / TP + TN + FP + FN

Precision (true positives / predicted positives) = TP / TP + FP

Recall (true positives / all actual positives) = TP / TP + FN

F1 Score=2\*(precision+recall)/precision\*recall

sklearn.metrics is a module in scikit-learn, a popular machine learning library in Python, that provides various functions for evaluating the performance of machine learning models.

**Accuracy:**

**Models** **Accuracy**

1 LR 0.791274

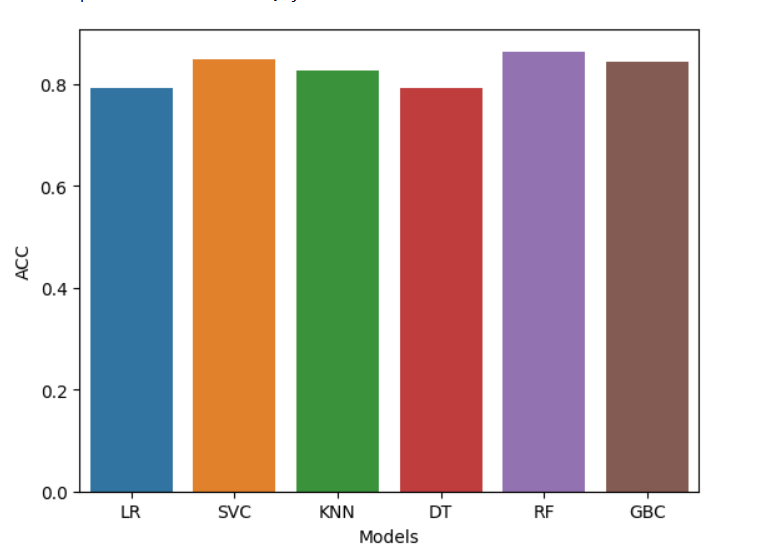
2 SVC 0.847772

3 KNN 0.826428

4 DT 0.791274

5 RF 0.863465

6 GBC 0.843691



**Precision:**

**Models Precision**

0 LR 0.776119

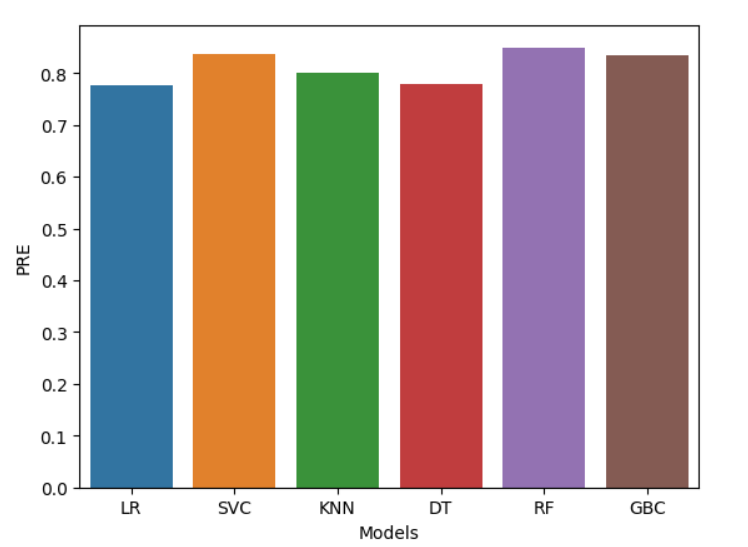
1 SVC 0.835849

2 KNN 0.799760

3 DT 0.777847

4 RF 0.849375

5 GBC 0.834921



**Conclusion**

After thorough experimentation and evaluation, the Random Forest model emerged as the optimal choice for predicting bank customer churn based on both accuracy and precision metrics. This indicates that the Random Forest algorithm achieved the highest level of overall correctness in predicting churn, while also demonstrating superior accuracy in identifying customers who actually churned. As a result, the Random Forest model is recommended for deployment in predicting bank customer churn, providing reliable insights for proactive customer retention strategies.