# CUSTOMER CHURN PREDICTION

**A PROJECT REPORT**

***Submitted by***

**SOUMIYA.D.S (211701054)**

**K.A.PREETHI (211701038)**

**NEVETHITHA.B (211701037)**

***in partial fulfillment for the course***

**CS19643 - FOUNDATIONS OF MACHINE LEARNING**

***for the degree of***

**BACHELOR OF ENGINEERING**

**in**

**COMPUTER SCIENCE AND DESIGN**

**RAJALAKSHMI ENGINEERING COLLEGE**

**RAJALAKSHMI NAGAR**

**THANDALAM**

**CHENNAI – 602 105**

**NOVEMBER 2024**

# RAJALAKSHMI ENGINEERING COLLEGE

# CHENNAI - 602105

## BONAFIDE CERTIFICATE

Certified that this project report **"CUSTOMER CHURN PREDICTION"** is the bonafide work of **"SOUMIYA.D.S (211701054), K.A.PREETHI (211701038), NEVETHITHA.B (211701037)"** who carried out the project work for the subject **CS19643 - FOUNDATIONS OF MACHINE LEARNING** under my supervision.


**SIGNATURE**

Mr. Uma Maheshwara Rao
**HEAD OF THE DEPARTMENT**
Professor
Department of
Computer Science and Design
Rajalakshmi Engineering College
Rajalakshmi Nagar
Thandalam
Chennai – 602105

**SIGNATURE**

Ms.E.Preethi
**SUPERVISOR**
Assistant Professor
Department of
Computer Science and Design
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 _____.

**INTERNAL EXAMINER** **EXTERNAL EXAMINER**

# ABSTRACT

In the banking sector, customer retention is critical to maintaining stable revenue and competitive advantage. This project aims to create an end-to-end solution for predicting customer churn by leveraging a dataset that includes demographic, behavioral, and financial attributes of bank clients. Customer churn poses a significant challenge, impacting revenue and customer retention strategies. By analyzing a comprehensive dataset with features we aim to identify patterns and factors contributing to customer attrition. The model development process includes data preprocessing (handling missing values, encoding categorical features, and scaling), exploratory data analysis to uncover correlations, and feature engineering to enhance predictive capabilities. We employ various machine learning algorithms, such as logistic regression, decision trees, and gradient boosting, to classify customers as likely to churn or retain. This project not only demonstrates the application of machine learning to a real-world problem but also emphasizes the value of data-driven decision-making in customer relationship management. By reducing churn, the financial institution can strengthen customer loyalty, enhance its brand reputation, and achieve more consistent revenue growth in an increasingly competitive market.

## TABLE OF CONTENTS

# 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 **Mr.S.Meganathan, B.E, F.I.E.,** our Vice Chairman **Mr. Abhay Shankar Meganathan, B.E., M.S.,** and our respected Chairperson **Dr. (Mrs.) Thangam Meganathan, 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 our Head of the Department **Prof. Uma Maheshwara Rao,** Department of Computer Science and Design for his guidance and encouragement throughout the project work. We convey our sincere thanks to our internal guide and Project Coordinator, **Ms.E.Preethi**, Department of Computer Science and Design, Rajalakshmi Engineering College for her valuable guidance throughout the course of the project.

**SOUMIYA .D .S (211701054)**
**K. A. PREETHI (211701038)**
**NEVETHITHA.B (211701037)**

## LIST OF FIGURES

## LIST OF ABBREVIATIONS

| ABBREVIATION | ACRONYM |
|---|---|
| ML | Machine Learning |
| API | Application Programming Interface |
| SQL | Structured Query Language |
| EDA | Exploratory Data Analysis |
| DB | Database |

# CHAPTER 1

# INTRODUCTION

## 1.1. INTRODUCTION

In today's competitive banking and financial services landscape, customer retention has become a critical factor for profitability and sustained growth. One of the primary challenges financial institutions face is customer churn, where clients discontinue their services or switch to other providers. The ability to predict and prevent churn is essential for enhancing customer loyalty, reducing revenue losses, and building long-term relationships. With recent advances in machine learning and data science, it is now possible to predict which customers are likely to churn based on historical data and behavioral patterns.

This project focuses on creating a predictive model that identifies high-risk customers, allowing the bank to take proactive measures to improve retention. The dataset used in this project includes attributes related to customer demographics, financial behavior, and engagement, such as age, account balance, credit score, tenure, and usage of bank products. By leveraging machine learning, this project aims to provide insights into key factors driving churn and to develop a reliable model for predicting customer attrition.

## 1.2. OBJECTIVE

The primary objective of this project is to build a customer churn prediction model that accurately identifies customers likely to exit the bank's services. This involves several specific goals. First, analyzing the dataset helps to understand the characteristics and behaviors associated with churn. Next, the data undergoes preprocessing and transformation to optimize model performance. Following this, multiple machine learning models are developed to classify customers based on their risk of churning, either as high-risk or low-risk. Each model is carefully evaluated to determine which one demonstrates the highest predictive accuracy. Additionally, interpretability techniques are implemented to provide insights into which factors most significantly influence churn, offering valuable information for targeted retention efforts. Finally, the project includes deploying the model in a production-ready environment with a user-friendly interface, allowing for real-time churn prediction and enabling staff to take proactive action based on these insights.

## 1.3. EXISTING SYSTEM

In most banks, customer churn analysis is either absent or carried out in a limited way, relying heavily on historical trends or simple rule-based systems. These methods are often inadequate as they cannot capture complex patterns in customer data. Churn prediction relies on basic statistics and manually identified rules, which lack precision and scalability. In Traditional systems, Marketing and retention efforts are broad and not personalized, as there is no predictive model to segment customers based on churn risk.

## 1.4. PROPOSED SYSTEM

The proposed system introduces a machine learning-based approach to churn prediction that allows the bank to identify at-risk customers proactively. The key components of the proposed system include gathering customer data, including demographics, engagement metrics, financial attributes, and product usage and preprocess the data to handle missing values, scale numerical features, and encode categorical variables (e.g., Geography, Gender) for model compatibility. The model performs an analysis to uncover patterns and relationships in the data, identifying which features are most indicative of churn. Performs visualization of the distributions of features and their relationships to churn to gain a deeper understanding of the data. These insights enable the bank to better understand customer behavior and implement targeted retention strategies.

# CHAPTER -2

# LITERATURE REVIEW

Customer churn prediction is an essential task in customer relationship management across various industries, and the Random Forest algorithm has emerged as one of the most effective tools for this purpose. The Random Forest model excels due to its ability to handle both categorical and continuous variables, which are commonly found in datasets used for churn prediction. This flexibility, coupled with its ensemble approach of building multiple decision trees, enhances the accuracy and robustness of predictions while mitigating overfitting.

In the telecom sector, where churn prediction is particularly critical, Random Forest has been extensively used to deal with imbalanced datasets, where the number of customers who churn is often much lower than those who stay. By averaging the predictions of multiple trees, Random Forest significantly reduces bias, improving prediction reliability even in the face of this imbalance. Studies have shown that features such as usage frequency, transaction amounts, and customer service interactions play pivotal roles in determining churn probability. Additionally, Random Forest's ability to provide feature importance rankings allows businesses to prioritize variables that most influence churn, making it a powerful tool for not just prediction but also strategic decision-making.

Comparative studies between Random Forest and other machine learning algorithms, such as Decision Trees, highlight its superior performance in terms of both predictive accuracy and interpretability. Random Forest's ensemble learning mechanism helps it perform better by averaging results across several trees, thus preventing overfitting and enhancing its ability to generalize across

different customer profiles. Furthermore, by visualizing the importance of various features, organizations can gain deeper insights into the factors contributing to churn, enabling more targeted retention efforts.

In summary, research consistently supports Random Forest as a highly effective tool for churn prediction, particularly in sectors like telecom and banking where datasets tend to be large and imbalanced. Its adaptability, feature importance ranking, and ability to handle various data types make it an invaluable asset for predicting customer behavior and improving customer retention strategies.

# CHAPTER 3

# SYSTEM DESIGN

## 3.1 SYSTEM DESIGN

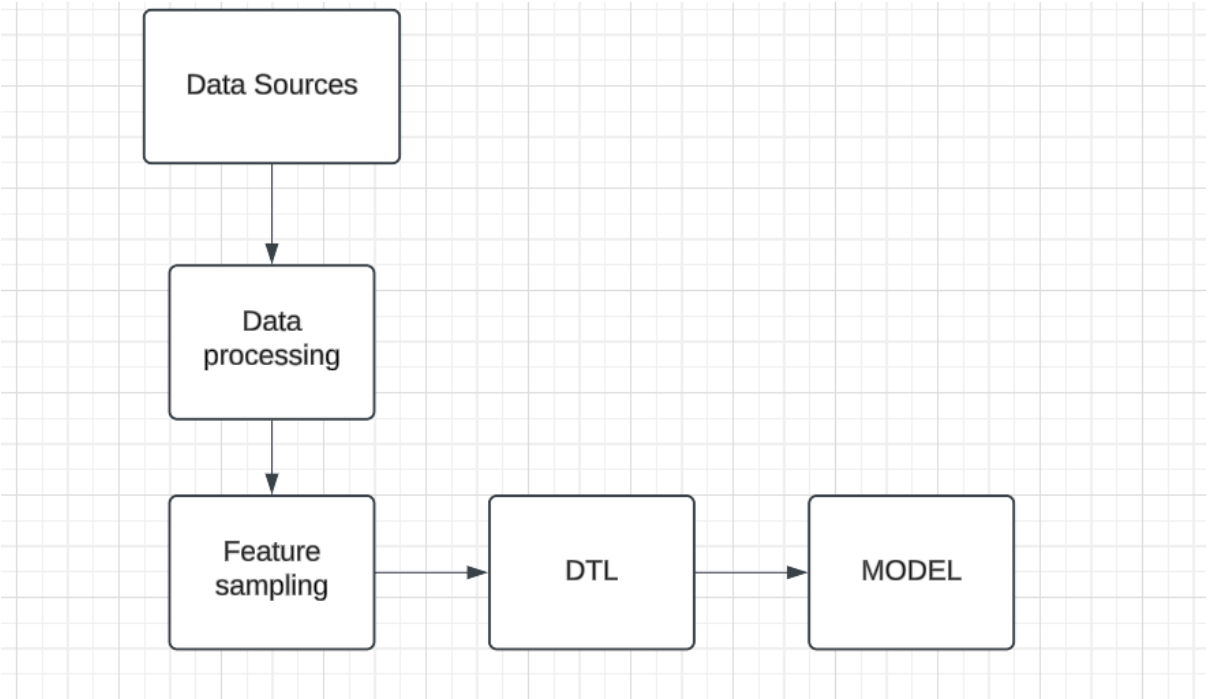


Figure 3.1 System Flow

The customer churn prediction system starts by collecting data from sources like CRM systems and transaction logs. This data is stored in a centralized database or data warehouse. It is then processed to clean, transform, and generate meaningful features, such as customer activity patterns. A Random Forest model is trained on historical data to identify churn patterns and predict which customers are likely to leave.

# CHAPTER 4

# PROJECT DESCRIPTION

## 4.1 MODULES

### 4.1.1. DATA COLLECTION AND PREPROCESSING

This module is focused on gathering relevant customer data and preparing it for analysis. Data is sourced from various customer interactions, demographic information, and financial attributes, including credit scores, account balances, tenure, and product usage. Data preprocessing includes handling missing values to ensure completeness, encoding categorical variables (like gender and geographic location) into numerical formats for model compatibility, and scaling numerical values to avoid model bias. Outlier detection and treatment are also performed to enhance data quality. Additionally, feature engineering techniques may be used to create new variables or transform existing ones, aiming to capture underlying patterns that influence customer churn. This preprocessing stage is essential as it ensures the data is in an optimal format, facilitating accurate model predictions and consistent results across various datasets. This module provides a clean, well-structured dataset, laying the foundation for the subsequent modules in the churn prediction pipeline.

### 4.1.2. EXPLORATORY DATA ANALYSIS

The EDA module is responsible for uncovering insights and patterns in the dataset, enhancing understanding of which features are associated with customer churn. This step involves visualizing the data through plots (e.g., histograms, box plots, and correlation matrices) to identify trends, outliers, and relationships between features. For example, analyzing the distribution of churn across different age groups, account balances, and credit scores provides clarity

on risk factors. EDA also includes summary statistics and bivariate analysis to compare churning and non-churning customers across multiple attributes, helping to prioritize features for the predictive model. By examining these relationships, the bank gains insight into customer behavior, allowing data scientists to make informed decisions regarding feature selection and engineering. The findings from EDA not only guide model development but also reveal business insights, which may be used to refine customer engagement strategies.

### 4.1.3. MODEL DEVELOPMENT AND TRAINING

This module involves developing and training machine learning models to predict churn based on the preprocessed dataset. Various algorithms are explored, including logistic regression for simplicity and interpretability, decision trees and random forests for handling non-linear relationships, and gradient boosting techniques like XGBoost for improved accuracy. Each model is trained on the training data and tested using a separate test set to evaluate its performance. The Random Forest model is well-suited for churn prediction due to its ensemble approach, which combines multiple decision trees to improve prediction accuracy and reduce overfitting. Evaluation metrics, including accuracy, precision, recall, F1-score, and ROC-AUC, are used to assess each model's effectiveness in predicting churn. The best-performing model is selected based on these metrics. This module ensures that the model captures relevant patterns and generalizes well to new data, forming the core of the predictive capability for identifying high-risk customers.

### 4.1.4. INSIGHTS AND DEPLOYMENT

This module aims to make the model's predictions understandable to stakeholders by identifying key factors that drive customer churn.

Interpretability tools are applied to highlight the contribution of each feature in the model's decision-making process. These tools allow the bank to understand which attributes (e.g., low account balance, short tenure, or low engagement) most significantly influence the likelihood of churn, offering transparency in predictions. The insights generated here are valuable for targeted retention strategies, as they reveal specific risk factors that can be addressed to reduce churn. By explaining the model's decisions, this module enables data-driven decision-making for the bank, allowing personalized outreach efforts tailored to each customer's profile.

# CHAPTER 5

# OUTPUT SCREENSHOTS

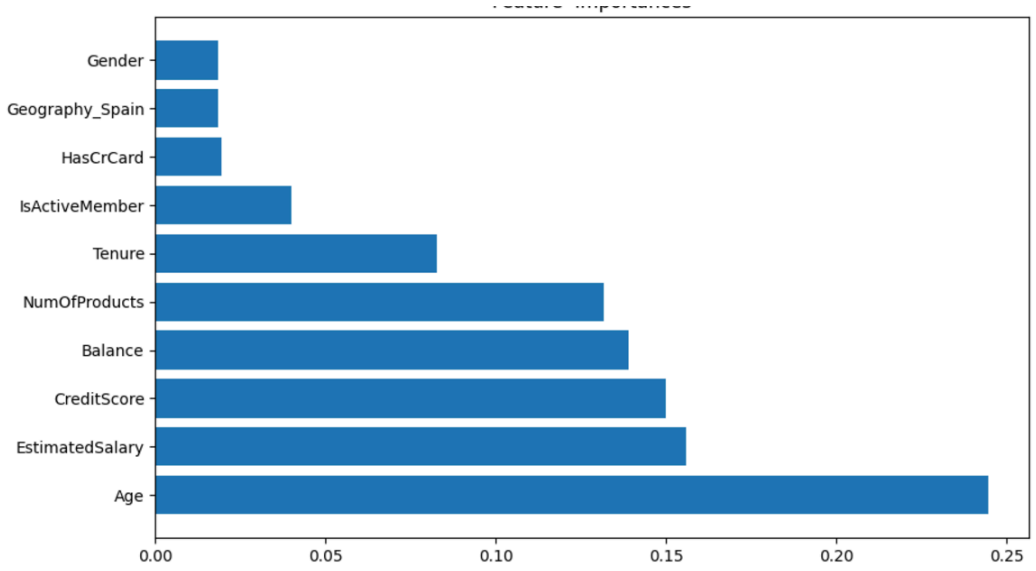


Figure5.1 Feature Importances

# CHAPTER 6

# CONCLUSION

This project successfully demonstrates the power of machine learning in predicting customer churn for financial institutions. By leveraging a comprehensive dataset with customer demographics, financial behaviors, and engagement levels, we developed a robust churn prediction model that accurately identifies high-risk customers. The data preprocessing, exploratory analysis, and feature engineering steps were instrumental in transforming raw data into valuable inputs for our models, enabling us to capture important patterns related to customer attrition.

The deployment of multiple machine learning models, including logistic regression, decision trees, and gradient boosting, allowed us to assess and select the best-performing model based on precision, recall, F1-score. Overall, this solution not only improves customer retention but also empowers the bank with data-driven decision-making capabilities. By reducing churn, the bank can enhance customer loyalty, drive sustainable growth, and gain a competitive edge in the financial services industry. In addition to improving customer retention, this project highlights the importance of leveraging advanced machine learning techniques to address real-world business challenges. By using data-driven insights, the bank can implement personalized strategies to mitigate churn, such as tailored marketing campaigns, targeted customer service interventions, and customized product offerings. These proactive measures, informed by the model's predictions, enable the bank to retain valuable customers and optimize its resources effectively.

# CHAPTER 6

# REFERENCES

[1]A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector, Irfan Ullah et al IEEE 2019

[2]Predictive Analysis of Customer Retention Using the Random Forest Algorithm. Yogasetya suhanadan et al Tiers information Technology Journal 2022.

[3]Customer churn prediction in telecom based on random forest algorithm, Aimin pal et al IEEE 2023

[4]Customer Churn Prediction Based on the Decision Tree and Random Forest Model, BCP business management publication. 2023

# APPENDIX

```
import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

from sklearn.model_selection import train_test_split

from sklearn.preprocessing import StandardScaler, LabelEncoder

from sklearn.ensemble import RandomForestClassifier

from sklearn.metrics import classification_report, confusion_matrix, accuracy_score

df=pd.read_csv('/content/Churn_Modelling.csv')

df.head()

df.isnull().sum()

df[df.duplicated()]

label_encoder = LabelEncoder()

df['Gender'] = label_encoder.fit_transform(df['Gender'])

df=pd.get_dummies(df,columns=['Geography'], drop_first=True)

features=['CreditScore','Age','Tenure','Balance','NumOfProducts','HasCrCard','IsActiveMember','EstimatedSalary','Gender','Geography_Spain']

x=df[features]

y=df['Exited']



x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=42)

scaler = StandardScaler()
```

```
x_train = scaler.fit_transform(x_train)

x_test = scaler.transform(x_test)

x_train[:5], x_test[:5]

model = RandomForestClassifier(n_estimators=100, random_state=42)

model.fit(x_train, y_train)

y_pred =model.predict(x_test)

conf_matrix=confusion_matrix(y_test,y_pred)

class_report=classification_report(y_test,y_pred)

accuracy=accuracy_score(y_test,y_pred)

print("Confusion Matrix:",conf_matrix)

print("Class report:",class_report)

print("Accuracy:",accuracy)

importances=model.feature_importances_

indices = np.argsort(importances)[::-1]

names=[features[i]for i in indices]



plt.figure(figsize=(10, 6))

plt.title("Feature  Importances")

plt.barh(range(x.shape[1]),importances[indices])

plt.yticks(range(x.shape[1]),names)

plt.show()

from sklearn.linear_model import LogisticRegression

log_reg=LogisticRegression(random_state=42)
```

```
log_reg.fit(x_train,y_train)


y_pred_log_reg=log_reg.predict(x_test)


conf_matrix_log_reg=confusion_matrix(y_test,y_pred_log_reg)
class_report_log_reg=classification_report(y_test,y_pred_log_reg)
accuracy_log_reg=accuracy_score(y_test,y_pred_log_reg)
print("Confusion Matrix:",conf_matrix_log_reg)
print("Class report:",class_report_log_reg)
print("Accuracy:",accuracy_log_reg)
from sklearn.svm import SVC
svm_model =SVC(kernel='linear', random_state=42)
svm_model.fit(x_train, y_train)


y_pred_svm = svm_model.predict(x_test)


conf_matrix_svm = confusion_matrix(y_test, y_pred_svm)
class_report_svm = classification_report(y_test, y_pred_svm)
accuracy_svm = accuracy_score(y_test, y_pred_svm)


print(conf_matrix_svm)
print(class_report_svm)
print(accuracy_svm)
```

```
from sklearn.neighbors import KNeighborsClassifier


knn_model = KNeighborsClassifier(n_neighbors=5)

knn_model.fit(x_train, y_train)


y_pred_knn = knn_model.predict(x_test)


conf_matrix_knn = confusion_matrix(y_test, y_pred_knn)

class_report_knn = classification_report(y_test, y_pred_knn)

accuracy_knn = accuracy_score(y_test, y_pred_knn)


print(conf_matrix_knn)

print(accuracy_knn)

print(class_report)

from sklearn.ensemble import GradientBoostingClassifier


gbm_model         =         GradientBoostingClassifier(n_estimators=100,
random_state=42)

gbm_model.fit(x_train, y_train)


y_pred_gbm = gbm_model.predict(x_test)


conf_matrix_gbm = confusion_matrix(y_test, y_pred_gbm)

class_report_gbm = classification_report(y_test, y_pred_gbm)
```

```
accuracy_gbm = accuracy_score(y_test, y_pred_gbm)



print(conf_matrix_gbm)

print(class_report_gbm)

print(accuracy_gbm)
```