# PROJECT REPORT

## GROUP 12

Customer Churn Prediction for Bank in the U.S. Market

Naveen Ajay karasu

11646981

MS in Cybersecurity

UNT

Varun Simha Reddy

11554660

MS in Artificial Intelligence

UNT

Santhoshi Kareddy

11600006

MS in Computer Science

UNT

Venkata Sai Karthik Aleti

11604572

MS in Computer Science

UNT

Gnana Naga siva sankar Kuralla

11597341

MS in

UNT

**Methodology**

**Data Exploration**

Our data exploration process commenced with an in-depth analysis of the dataset's features to discern their distributions and interrelationships, especially with regard to customer churn. This phase was crucial for gaining preliminary insights into the dataset and forming hypotheses about potential predictive factors.

**Abstract**

In this project we will address the problem of classifying bank customers into two categories. Those who will churn or those who will remain loyal. As we are dividing the data into two different outputs, we can call it as a binary classification problem, making machine learning algorithms like decision trees, Random Forest classifier, and neural networks as suitable candidates. Our primary goal in doing this project is to build efficient model which will have high accuracy, precision, and recall in predicting churn, it allows the banks to identify at-risk customers proactively. Customer churns also known as the loss of customers to other competitors in this case it is to other banks, as it presents a significant challenge to the banks. Our project aim is to study different things, behavior from day-to-day activities of the user to understand more clearly that why the customer churn is happening and after that we will we try to feed the behavior of the customer as data to the to a machine learning model by employing various machine learning model techniques on a bank customer dataset. The analyzing done on different factors like demographics, account activity, product usage etc..,. We think this predictive capability can empower banks to implement targeted different strategies, to reducing churn and enhancing customer loyalty.

**Introduction**

Before knowing about our project let’s understand the problem we are trying to solve here which is customer churn, it can be defined as the losing clients to banks, which is a very important challenge to any bank as it can significantly affect the working of the banks as it will cause decline in the sustainability and growth of the bank. In today’s modern world there are more and more banks are present for a customer to choose from in the competitive market, keeping the customer retained as a existing customers becomes a big challenge for any bank. As having more customer very importantly essential for maintaining a steady revenue for the bank to opereate. So, Understanding the patterns and trying to predict the customer churn has a very important necessity for every bank.

So, In our project we want to harnesses the power of machine learning models to analyze and try to predict reason for the churn using a dataset containing data on customer demographics, financial details, account activity etc..,. We think this approach allows us to move beyond traditional techiniques and strategies to deploy proactive create measures that can mitigate or reduce the risk of churn before it really happens.

The predictive model will help the banks to separate or divide the it’s customer’s database into distinct different groups based on their level of risk of churning. This segmentation can then help inform targeted intervention strategies which can be tailored to specific particular customer needs and preferences which thereby trying to increase the customer satisfaction and loyalty to the bank. By integrating these insights given by model into the bank’s strategic planning, we think banks can not only reduce churn rates but also enhance the overall customer engagement and profitability in the bank.

**Data Description**

The important thing in our analysis is a non-other than the data itself. So, for our project we are using the dataset provided by non-other than the U.S. banks itself. The data contains a wide range of customer characteristics carefully labeled with their churn status. This dataset plays an important role in developing our machine learning models to predict customer churn accurately.

**Data Composition**

The dataset consists of different varity of variable that makes a customer profile which can be used in creating a model for predicting the customer churn in the bank. The data is created based on the U.S. banks data. So, we can assume that the model we are going to develop is suitable for banks which are located and operated inside the U.S. itself. Knowing where the data came from and the targeted users of our model can help us greatly create a good model. For example, even if we develop a good model with high accuracy and prediction it can still fail if it is used on data from regions as we solely developed our model using data only from U.S. region.

**Data Quality and Preprocessing**

The data consists of 10000 rows of customer data in the dataset we are using and have 14 columns like customer ID, Surname, credit score, gender, age, tenure, balance, salary, does he credit card and many other. Having 10000 rows of data doesn’t mean we have data on 10000 unique customers, because same person can have multiple accounts in the same bank or in a different banks.

A screenshot of a computer

Description automatically generated

Figure 1: Showing no of rows in the dataset.

We cannot use the raw data directly as it may have anomalies or missing data for some customers, we need to preprocess the data before using it to build the model. As it will increase the accuracy and prediction of our model. We can do the preprocessing of the data using data normalization and scaling process by using techniques like min-max scaling, z-score normalization which helps in ensuring that each column/ feature contributes equally to the prediction of the model. Which helps in preventing the model’s decision being solely depended on a single feature/column.

A screenshot of a computer

Description automatically generated

Figure 2: Showing different features in the dataset

We can further process the data by making the data simple and easier for the machine learning model to understand like converting the alphabetic word into numeric or binary representation as 1 or 0. For example, we can use this to represent gender in this way so for male as ‘1’ and female as ‘0’. We can do the same thing for marital status, has creditcard etc..,. So, it makes easier to feed the data to the model and reduces the need for computation.

## BACKGROUND:

## EXPERIMENT METHODOLOGY:

## RESULTS:

Heat map

A screenshot of a computer

Description automatically generated

A graph showing different colored shapes

Description automatically generated with medium confidence

A graph of a normal distribution

Description automatically generated with medium confidence

A graph of a curve

Description automatically generated with medium confidence

CODE: