BANK CUSTOMER CHURN PREDICTION USING MACHINE LEARNING

Srinivas HB
Department of CSE
B.Tech in AI & ML
MS Ramaiah University of Applied Sciences
Bangalore, India
srinivashb12@gmail.com

Darshan Gowda B
Department of CSE
B.Tech in AI & ML
MS Ramaiah University of Applied Sciences
Bangalore, India
darshangowda554321@gmail.com

Rohith K
Department of CSE
B.Tech in AI & ML
MS Ramaiah University of Applied Sciences
Bangalore, India
rohithk9746@gmail.com

Gokul M
Department of CSE
B.Tech in AI & ML
MS Ramaiah University of Applied Sciences
Bangalore, India
gokulmohan710@gmail.com

***Abstract*— Predicting bank customer churn, a critical task for financial institutions aiming to retain their clientele. It contains a rich array of features including customer demographics, credit score, geographic location, account balances, product holdings, and activity indicators. Leveraging machine learning and predictive analytics techniques on this dataset enables the development of models capable of forecasting whether a customer is likely to churn or remain with the bank. By accurately identifying potential churners, banks can proactively implement targeted retention strategies, thereby mitigating customer attrition and fostering long-term customer relationships. This abstract outlines the dataset's significance in facilitating the development of effective churn prediction models crucial for the sustainable growth and competitiveness of banking institutions. the realm of banking and financial services, customer churn prediction stands as a cornerstone for proactive customer retention strategies. This dataset encapsulates a comprehensive array of attributes associated with bank customers, ranging from demographic information such as age, gender, and geography to financial metrics including credit scores, account balances, and product holdings. With the overarching objective of churn prediction, this dataset serves as a foundational resource for developing machine learning models capable of discerning patterns indicative of potential customer attrition. By leveraging advanced analytics techniques on this dataset, financial institutions can gain actionable insights into customer behaviour, enabling them to deploy targeted retention initiatives and mitigate the risk of losing valuable clientele. This abstract highlights the dataset's significance as a catalyst for fostering data-driven decision-making processes aimed at sustaining customer relationships and enhancing overall business performance within the banking sector.**

I. INTRODUCTION

In the dynamic landscape of banking and financial services, the ability to predict and mitigate customer churn has emerged as a critical imperative for institutions striving to maintain competitive advantage and foster sustainable growth. Customer churn, the phenomenon wherein customers cease their engagement with a bank by closing their accounts or ceasing to utilize its services, poses significant challenges to profitability, customer satisfaction, and market share retention. Recognizing the paramount importance of pre-emptive churn management, banks are increasingly turning to advanced analytics and machine learning techniques to anticipate and address customer attrition.

This dataset represents a rich and multifaceted repository of information, meticulously curated to facilitate the task of predicting bank customer churn. Comprising an extensive array of attributes spanning demographic details, financial metrics, and behavioural indicators, this dataset offers a holistic view of customer profiles and their interactions with the banking institution. From basic demographic information such as age, gender, and geographical location to more nuanced factors including credit scores, account balances, and product holdings, the dataset encapsulates a diverse spectrum of features integral to understanding customer behaviour and predicting churn propensity.

At its core, the dataset embodies the convergence of traditional banking domain expertise with the transformative power of data-driven insights. By leveraging the wealth of information contained within this dataset, financial institutions can embark on a journey towards developing robust churn prediction models capable of discerning subtle patterns and trends indicative of customer attrition. Through the application of sophisticated machine learning algorithms and predictive analytics techniques, banks can unlock actionable insights that empower them to proactively identify at-risk customers, tailor targeted retention strategies, and ultimately mitigate the adverse impacts of churn on their bottom line.

In light of the intensifying competition within the banking sector and the growing emphasis on customer-centricity, the significance of effective churn prediction cannot be overstated. As banks strive to differentiate themselves in an increasingly crowded market landscape, the ability to anticipate and mitigate customer churn represents a pivotal opportunity for fostering long-term customer relationships, enhancing operational efficiency, and driving sustainable business growth. Against this backdrop, the present dataset emerges as a potent catalyst for driving innovation, enabling financial institutions to harness the power of data and analytics to navigate the complex dynamics of customer churn with confidence and foresight

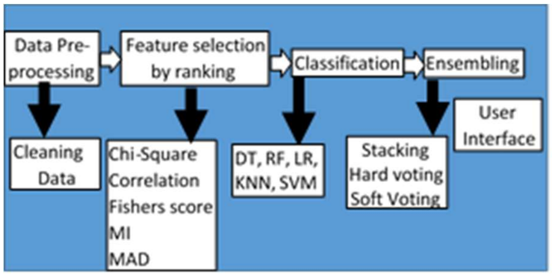


Figure.1 Implementation Block Diagram of BCC prediction

The project titled "Bank Customer Churn Prediction" assumes profound significance, offering a compelling avenue for leveraging advanced analytics and predictive modelling techniques to anticipate and mitigate the risks associated with customer churn.

The overarching goal of this project is to harness the power of data-driven insights to decipher the intricate nuances of customer behaviour and discern predictive patterns indicative of churn propensity. At its essence, the project encapsulates a multidimensional exploration of customer data, meticulously curated to encapsulate a diverse array of attributes encompassing demographic characteristics, financial metrics, and behavioural indicators. By delving deep into the intricate fabric of customer profiles and transactional histories, the project endeavours to unravel the underlying drivers and determinants of churn, thereby empowering banks to proactively identify at-risk customers and tailor targeted retention strategies that resonate with their evolving needs and preferences.

II. Feature Selection and Classification

A. Feature Selection

Input Data Extraction: Features are extracted from the dataset using iloc [:, 3:-1], which selects columns from the 4th column to the second-to-last column as features. This implies that the first three columns (presumably containing categorical data like geography and gender) are considered for inclusion in the model.

Categorical Data Encoding: Categorical variables are encoded using label encoding for "Gender" and one-hot encoding for "Geography". This step transforms categorical variables into a format that can be handled by the model.

Model Building: An Artificial Neural Network (ANN) is constructed using TensorFlow's Keras API. The network architecture consists of an input layer, two hidden layers with ReLU activation functions, and an output layer with a sigmoid activation function. Model Compilation: The ANN model is compiled with the Adam optimizer and binary cross-entropy loss function, suitable for binary classification tasks. Model Training: The ANN is trained on the training set using the fit method, with specified batch size and number of epochs. Prediction and Evaluation: Predictions are made on both a single observation and the test set. Accuracy metrics and confusion matrix are computed to evaluate the performance of the model.

Feature Selection Considerations:

Column Selection: Columns are chosen for feature inclusion based on their relevance to predicting customer churn. Typically, factors like demographics, account activity, and financial indicators are crucial in such predictions. Encoding Strategy: Categorical variables like "Geography" and "Gender" are appropriately encoded to ensure compatibility with the model. One-hot encoding is preferred for categorical variables with more than two categories to avoid ordinality bias

B. Classification

Classification Considerations:

Model Architecture: The choice of an ANN for classification suggests a need for a complex, nonlinear decision boundary to capture the relationships between features and the target variable.

Activation Functions: ReLU activation functions are used in hidden layers to introduce nonlinearity, while a sigmoid activation function is employed in the output layer for binary classification, providing probability-like outputs.

Loss Function and Optimizer: Binary cross-entropy loss is chosen as it is suitable for binary classification tasks, while the Adam optimizer is selected for efficient gradient descent optimization.

Data Pre-processing:

The dataset is imported, containing various features such as credit score, geography, gender, age, tenure, balance, etc. Features are extracted into the matrix X while the target variable (whether the customer exited or not) is stored in y.

Categorical variables are encoded using label encoding and one-hot encoding techniques to convert them into numerical format suitable for machine learning algorithms. The dataset is split into training and testing sets, and feature scaling is applied using StandardScaler to standardize the feature values.

Building the Artificial Neural Network (ANN):

An ANN model is initialized using TensorFlow's Sequential API. Input layer and two hidden layers with ReLU activation functions are added to the model. An output layer with a sigmoid activation function is added, as the problem is binary classification (predicting churn or not churn).

Training the ANN:

The model is compiled with the Adam optimizer and binary cross-entropy loss function. The model is trained on the training set using fit() method, specifying batch size and number of epochs.

Making Predictions and Evaluating the Model:

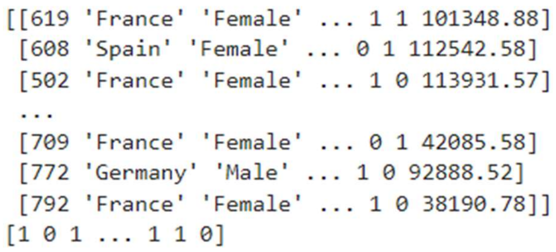
The model is used to predict whether a specific customer (with certain feature values) will churn or not. Predictions are made on the test set, and a confusion matrix is computed to evaluate the model's performance.

Accuracy score is calculated to quantify the model's predictive accuracy on the test set.

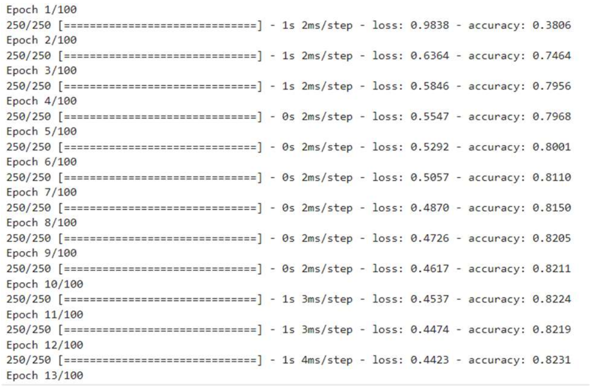
Final Prediction:

Finally, a single observation (customer) is passed to the trained model to predict whether they will stay or leave the bank.

III. Results



```
[[619 'France' 'Female' ... 1 1 101348.88]
 [608 'Spain' 'Female' ... 0 1 112542.58]
 [502 'France' 'Female' ... 1 0 113931.57]
 ...
 [709 'France' 'Female' ... 0 1 42085.58]
 [772 'Germany' 'Male' ... 1 0 92888.52]
 [792 'France' 'Female' ... 1 0 38190.78]]
[1 0 1 ... 1 1 0]
```



```
Epoch 1/100
250/250 [==============================] - 1s 2ms/step - loss: 0.9838 - accuracy: 0.3806
Epoch 2/100
250/250 [==============================] - 1s 2ms/step - loss: 0.6364 - accuracy: 0.7464
Epoch 3/100
250/250 [==============================] - 1s 2ms/step - loss: 0.5846 - accuracy: 0.7956
Epoch 4/100
250/250 [==============================] - 0s 2ms/step - loss: 0.5547 - accuracy: 0.7968
Epoch 5/100
250/250 [==============================] - 0s 2ms/step - loss: 0.5292 - accuracy: 0.8001
Epoch 6/100
250/250 [==============================] - 0s 2ms/step - loss: 0.5057 - accuracy: 0.8110
Epoch 7/100
250/250 [==============================] - 0s 2ms/step - loss: 0.4870 - accuracy: 0.8150
Epoch 8/100
250/250 [==============================] - 0s 2ms/step - loss: 0.4726 - accuracy: 0.8205
Epoch 9/100
250/250 [==============================] - 0s 2ms/step - loss: 0.4617 - accuracy: 0.8211
Epoch 10/100
250/250 [==============================] - 1s 3ms/step - loss: 0.4537 - accuracy: 0.8224
Epoch 11/100
250/250 [==============================] - 1s 3ms/step - loss: 0.4474 - accuracy: 0.8219
Epoch 12/100
250/250 [==============================] - 1s 4ms/step - loss: 0.4423 - accuracy: 0.8231
Epoch 13/100
```

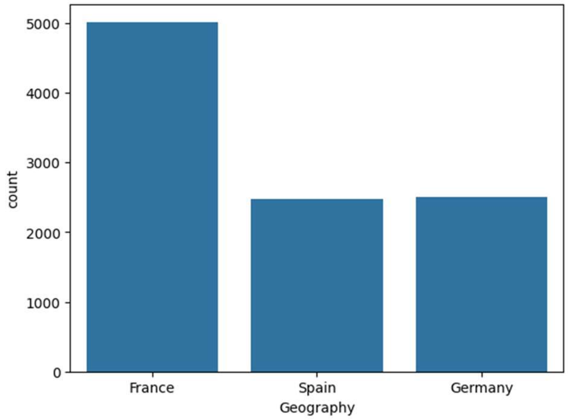
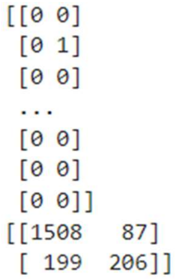


Figure.2 Count-plot of Geography



```
[[0 0]
 [0 1]
 [0 0]
 ...
 [0 0]
 [0 0]
 [0 0]]
[[1508   87]
 [ 199  206]]
```

Figure 3. Confusion Matrix

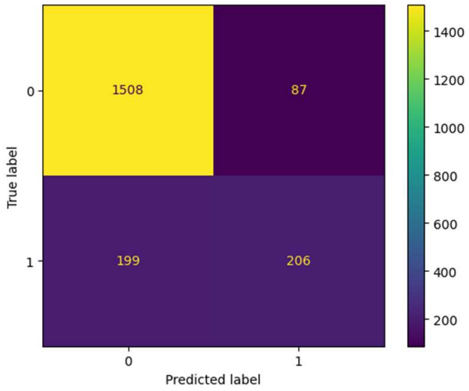
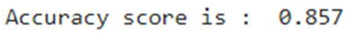


Figure 4. Confusion Matrix Display



```
Accuracy score is :  0.857
```

V. CONCLUSION

Artificial Neural Network (ANN) using TensorFlow and Keras to predict customer churn in a bank. Data Pre-processing: The dataset is imported and pre-processed.

Categorical variables are encoded using Label Encoding for "Gender" and One Hot Encoding for "Geography".

The dataset is split into training and test sets.

Feature scaling is performed to standardize the features.

Building the ANN: The ANN is initialized using the Sequential API. Layers are added: two hidden layers with ReLU activation and an output layer with sigmoid activation for binary classification. Training the ANN: The model is compiled with the Adam optimizer and binary cross-entropy loss. The ANN is trained on the training set for 100 epochs. Making Predictions and Evaluating the Model: A single observation is used to predict whether a customer will leave the bank. The prediction indicates that the customer will stay.Test set predictions are made and evaluated using a confusion matrix, displaying the model's accuracy score.