1. Introduction

This document outlines the key steps involved in building a predictive model for customer churn. The project includes data cleaning and preprocessing, training and optimizing both traditional machine learning models and a neural network, and comparing their performance.

2. Data Cleaning and Preprocessing

Step 1: Data Loading and Column Removal

The dataset is first loaded, and unnecessary columns like RowNumber, CustomerId, and Surname are removed. These columns do not provide any useful information for predicting customer churn and are therefore excluded from the model.

Step 2: Encoding Categorical Variables

Categorical variables such as Geography and Gender are converted into numerical format using one-hot encoding. This method creates binary columns representing the different categories, which are necessary for the machine learning models to process these variables correctly.

Step 3: Feature Scaling

The numeric columns (such as CreditScore, Age, Balance, etc.) are standardized using feature scaling to ensure all features have the same scale. This step is crucial for algorithms like Logistic Regression and Support Vector Machines (SVM), which are sensitive to the magnitude of the input features.

Step 4: Splitting the Dataset

The cleaned and processed data is split into training and testing sets. Additionally, a smaller sampled subset of the training data is created for faster hyperparameter tuning during the model training process.

3. Training and Optimizing Classic Machine Learning Models

Step 1: Model Selection

Three classic machine learning models are chosen for this task:

Logistic Regression: A linear classifier that estimates the probability of a customer churning.

Random Forest: An ensemble learning method that builds multiple decision trees to improve accuracy.

Support Vector Machine (SVM): A classification model that finds the optimal boundary between classes.

Step 2: Hyperparameter Tuning with RandomizedSearchCV

Each model's performance is optimized by using a randomized search across a range of hyperparameters. RandomizedSearchCV helps find good parameter values quickly by testing a random subset of possible parameter combinations.

Step 3: Fine-tuning with GridSearchCV

After identifying the best hyperparameters from the randomized search, a more refined search is conducted using GridSearchCV. This method explores a smaller, more focused set of parameter values to further improve model performance.

Step 4: Model Evaluation

Once the best versions of the models are selected, they are evaluated on the test data. Classification metrics such as accuracy, precision, recall, and F1-score are used to assess how well the models predict customer churn.

4. Neural Network Construction and Training

Step 1: Label Encoding and Scaling

The same data preparation steps used for machine learning models (i.e., encoding categorical variables and scaling numeric features) are applied to prepare the dataset for the neural network.

Step 2: Neural Network Architecture

A simple feedforward neural network is constructed with two hidden layers. Each hidden layer uses ReLU activation to introduce non-linearity, while the output layer uses a sigmoid activation function to predict the probability of a customer churning.

Step 3: Model Compilation and Training

The neural network is compiled using the Adam optimizer and binary cross-entropy as the loss function. The model is trained with early stopping, which halts the training process when no further improvement is observed in validation loss. This helps avoid overfitting.

Step 4: Evaluation

The trained neural network is evaluated on the test set, and classification metrics such as accuracy, precision, recall, and F1-score are calculated. These metrics are used to determine the network's performance in predicting churn.

5. Comparison of Classic ML Models and Neural Network

Step 1: Accuracy Comparison

The test accuracy of the best-performing classic machine learning model is compared with the neural network. The model with the highest accuracy is noted as the better performer for this task.

Step 2: Performance Metrics

In addition to accuracy, other performance metrics such as precision, recall, and F1-score are used to give a comprehensive view of model performance, especially in handling imbalanced classes like churn.

6. Conclusion

This project demonstrates how both classic machine learning models and neural networks can be effectively used to predict customer churn. After cleaning and processing the data, hyperparameter optimization, and evaluation, the project concludes by comparing the performance of traditional machine learning models with a neural network.