Project Report: Customer Churn Prediction Model Using a Neural Network Model

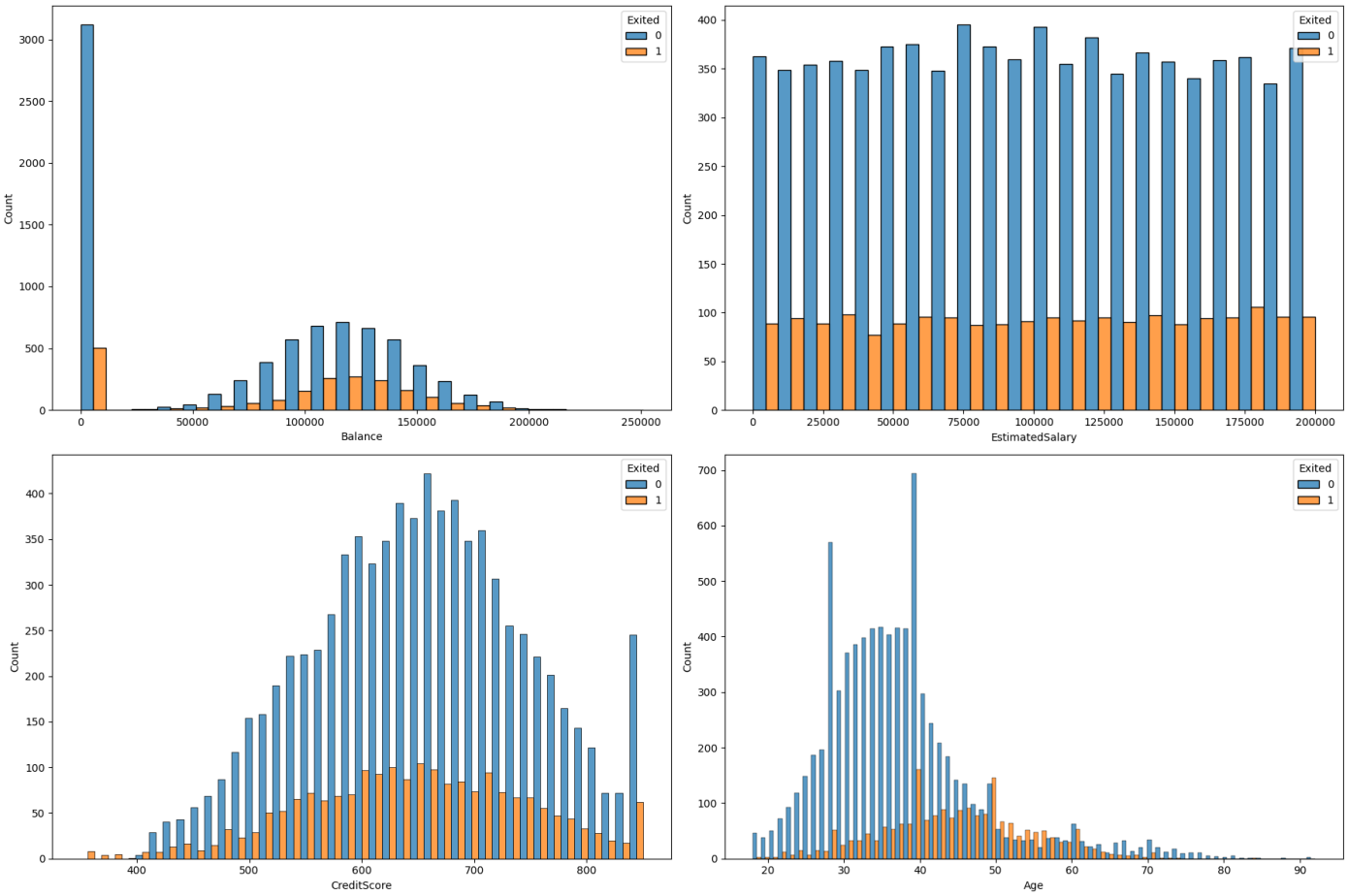
Rahman Mabano, MSC ECE-AD
Data Engineer

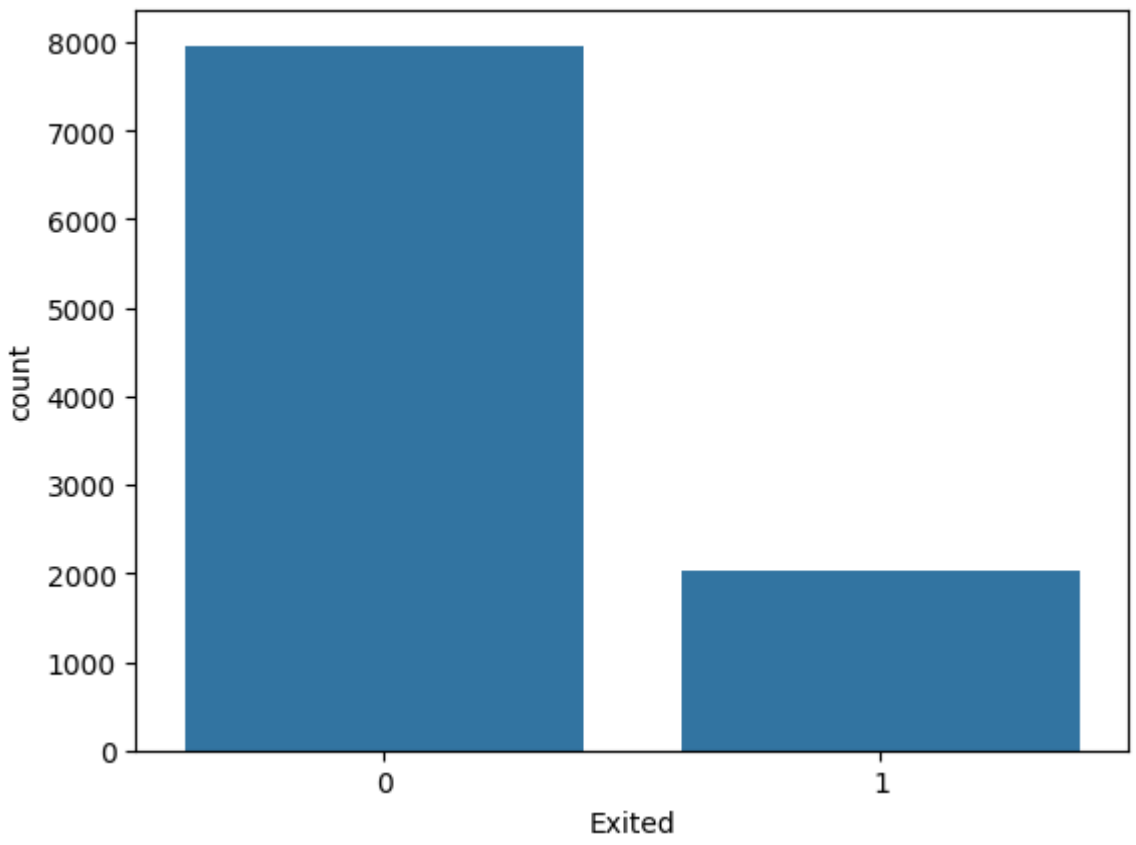
ABSTRACT
This project was conducted in collaboration with a bank in Cyprus which has also key branches in Germany, Spain and France, to predict customer churn. The analysis involved using key libraries such as TensorFlow and Keras for building machine learning models, as well as Pandas and Seaborn for data manipulation and visualization. The final model achieved an accuracy of 87%. Key steps included data preprocessing, exploratory data analysis (EDA), model building, training, dataset balancing, and hyperparameter tuning. The insights derived from this analysis aim to help the bank in identifying customers likely to churn, thereby informing strategies to improve customer retention.

INTRODUCTION
This study was undertaken for a bank in Cyprus aiming to introduce a new savings scheme. The provided dataset comprised well-curated customer data spanning two years, including high, medium, and low-income spenders. The primary objective was to predict customer churn and offer actionable insights to reduce it. Key libraries used in this project include TensorFlow for model building, Keras for neural networks, and Scikit-learn for data preprocessing and evaluation. The analysis steps included data cleaning, exploratory data analysis, feature encoding, model training, and validation.

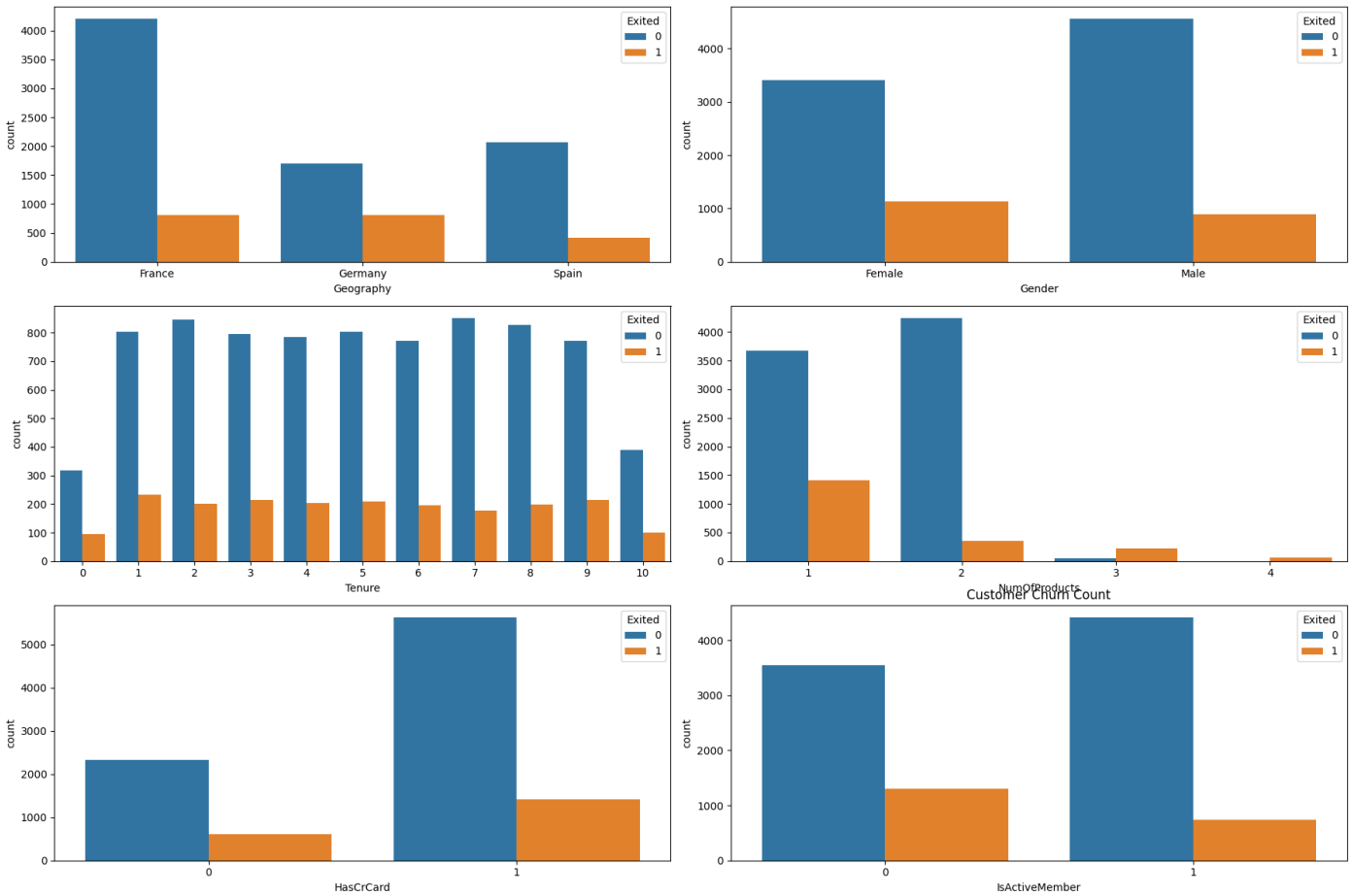
EDA STEPS
Exploratory Data Analysis (EDA) was conducted to understand the dataset's structure, identify patterns, and detect anomalies. Key steps included:
1. Checking data types and missing values.
2. Dropping unnecessary columns to reduce noise.
3. Analyzing the distribution of the target variable 'Exited'.
4. Visualizing categorical variables (Geography, Gender, etc.) and their relationship with the target variable.
5. Creating histograms for continuous variables (Balance, EstimatedSalary, CreditScore, Age) to observe their distributions and potential correlations with churn.

Data Count Plots & Histograms
The histogram plots show customer dropout occurrences by salary, account balance, credit score, and age. This is very essential to help get a full overall picture that cuts across each category as to which precise clients are leaving the bank.





Above we can see that only 20% of clients eventually churned, while 8000 are still active. Below we find the breakdown of these numbers by categories (Geography, Gender, Tenure, and so on). Analysing these numbers can give us a more clear picture as to why we are seeing such a significant drop in client base.



DATA PREPROCESSING

Data preprocessing involved several steps:

1. Encoding Categorical Variables: LabelEncoder was used to convert categorical variables like Geography and Gender into numerical values.

2. Feature Scaling: StandardScaler was applied to standardize features, which is crucial for ensuring that all variables contribute equally to the model training process.

3. Handling Class Imbalance: The dataset was initially imbalanced, so techniques like resampling were employed to create a balanced dataset for model training.

4. Train-Test Split: The data was split into training and test sets to evaluate model performance.

MODEL BUILDING AND TRAINING

The Keras library, a high-level API for building neural networks, was used for model building. Key components included:

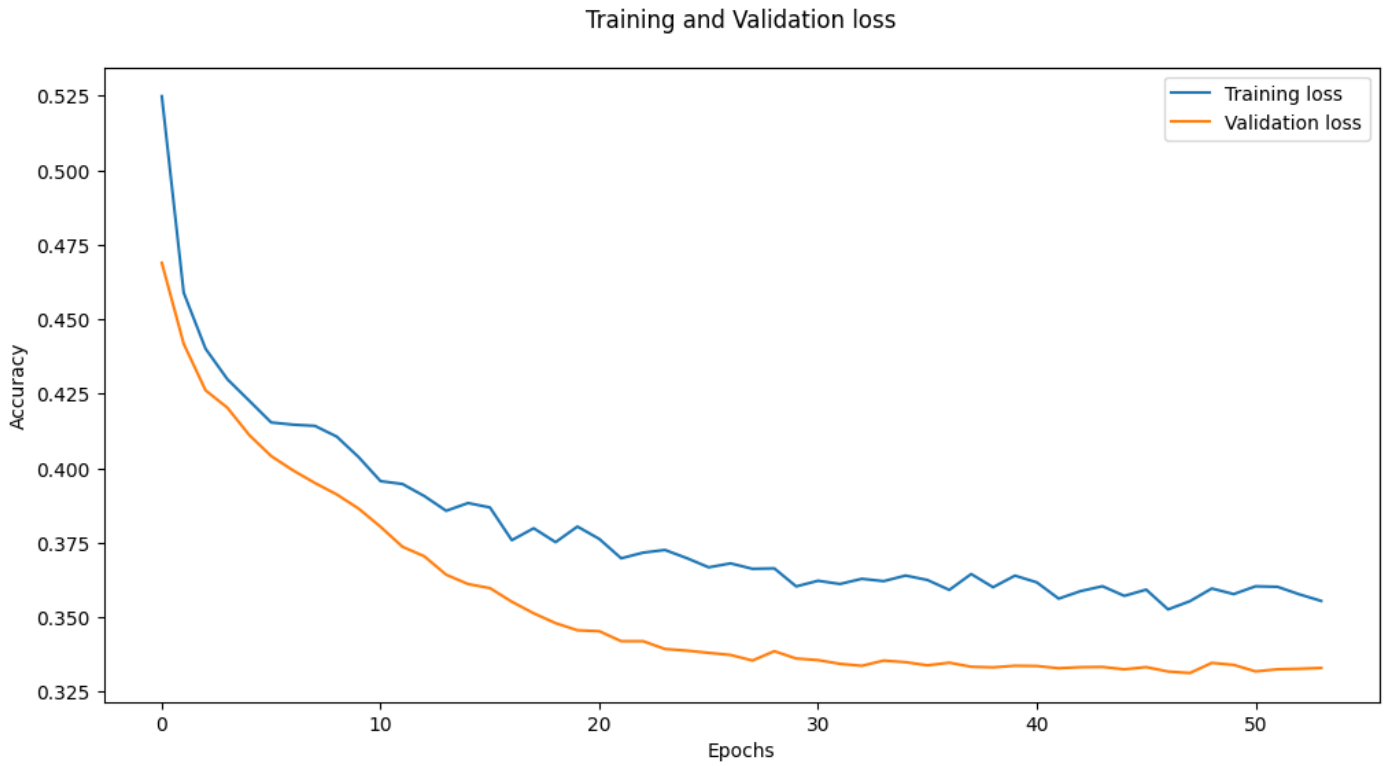
1. **Sequential Model**: A simple stack of layers where each layer has exactly one input tensor and one output tensor.

2. **Dense Layers:** Fully connected layers with ReLU activation functions were used.

3. Dropout Layers: To prevent overfitting, dropout layers were included.

4. Compilation: The model was compiled with the Adam optimizer and categorical cross-entropy loss function.

5. Training: The model was trained with early stopping to prevent overfitting and ensure optimal performance.



DATASET BALANCING

Dataset balancing is a technique used to address class imbalance, which can bias the model towards the majority class. In this project:

1. Resampling: The majority class (customers who did not exit) was downsampled to match the minority class (customers who exited).

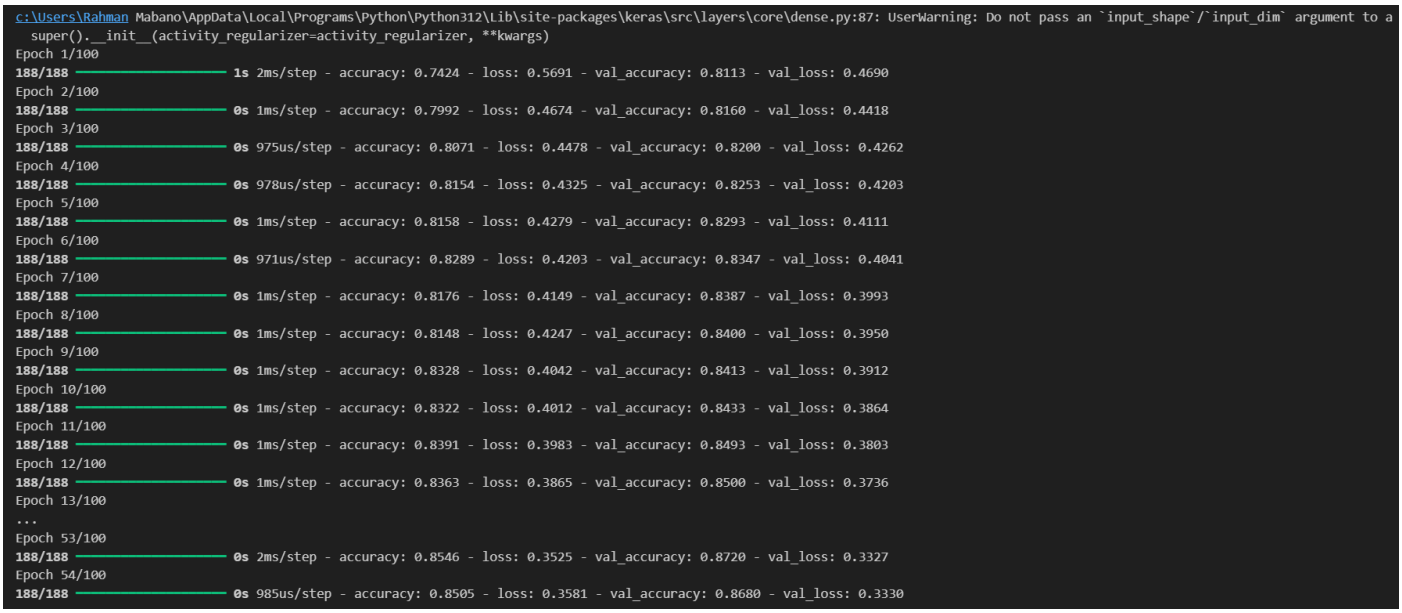
2. Shuffling: The dataset was shuffled to ensure randomness.

3. Evaluation: Despite the balancing attempt, using the entire dataset proved more beneficial, indicating the need for careful evaluation of balancing techniques.

HYPERPARAMETER TUNING

Hyperparameter tuning was performed using Keras Tuner:
1. Random Search: This method was used to find the optimal hyperparameters for the neural network.
2. Parameters Tuned: The number of units in each dense layer, activation functions, dropout rates, and learning rates.
3. Execution: Multiple trials were run, and the model's performance was evaluated based on validation accuracy.

The image below shows the final model found after exploring all parameters.



CONCLUSION
The neural network model achieved significant accuracy, demonstrating its effectiveness in predicting customer churn. However, exploring other models such as Random Forest, Logistic Regression, or SVM could provide a comprehensive understanding of model performance. The results highlight the importance of data preprocessing, proper model selection, and hyperparameter tuning in developing robust predictive models. Future work should include comparing different models and potentially integrating them to improve prediction accuracy further.