Project Report

Title: Customer Churn Analysis

Name: Ramashish

Internship at: Learntricks Edutech

Project Overview:

The project focuses on analyzing customer churn, which is a classification problem. The dataset contains various features related to customers such as credit score, geography, gender, age, tenure, balance, number of products, credit card status, active membership, estimated salary, and whether the customer has exited the bank or not.

Steps Undertaken:

1. **Data Exploration:**

- ❖ Checked the dataset for initial insights.
- ❖ Verified for any duplicate values, which were not found.
- ❖ And information about the dataset it is object, int or string etc. and if there are null values or not, no null values were found.

2. **Data Pre-processing:**

- ❖ Conducted value counts on 'Geography' and 'Exited' columns for understanding the distribution.
- ❖ Removed unnecessary columns 'RowNumber', 'Surname', 'CustomerId'.
- ❖ Applied one-hot encoding on 'Geography' and 'Gender' columns.
- ❖ Performed train-test split, where the 'Exited' column was designated as the output column (y), while all other features were considered input features (X).
- ❖ Performed scaling using StandardScaler for normalization.

3. **Model Building:**

- ❖ Used TensorFlow and Keras libraries for building a neural network model.
- ❖ Utilized a sequential model with two dense layers and activation function set to sigmoid.
- ❖ Chose 'Adam' optimizer and 'binary_crossentropy' loss function.
- ❖ Initially predicted accuracy of the dataset was 0.8125.

4. **Added Additional Dense Layers with ReLU Activation:**

❖ After the initial model building step, additional dense layers were introduced with the activation function set to 'ReLU' (Rectified Linear Unit). ReLU activation has been observed to alleviate the vanishing gradient problem and accelerate convergence in neural networks, potentially enhancing the model's performance.

5. **Model Summary:**

❖ Upon adding the additional dense layers with ReLU activation, the model summary revealed the following parameters:

- Total params: 40
- Trainable params: 40

❖ This indicates that the model comprises a total of 40 parameters, all of which are trainable. The model summary provides insight into the architecture and complexity of the neural network model, aiding in understanding its behavior and performance.

6. **Adjusting Threshold for Predictions:**

❖ To determine the predicted class labels, a threshold of 0.5 was initially set for the sigmoid activation output. However, to optimize model performance, the threshold was adjusted. The formula used for calculating y_pred is as follows:

y_pred = np.where(y_log > 0.5, 1, 0)

where y_log represents the output of the sigmoid activation function.

7. **Evaluation of Model Performance:**

❖ Following the addition of ReLU activation and three additional dense layers, the accuracy score of the model further improved. The revised accuracy score reached 86.25%, signifying the effectiveness of the model enhancements in achieving better classification results.

Recommendation:

- **Exploring Ensemble Methods:**

Implement ensemble learning techniques such as Random Forest, Gradient Boosting, or AdaBoost to improve model performance. Ensemble methods often provide better generalization and robustness compared to individual models.

- **Hyperparameter Tuning:**

Conduct a thorough hyperparameter tuning process using techniques like grid search or random search to optimize the neural network's architecture and parameters. This can help in achieving higher accuracy and better model convergence.

References:

- Kaggle notebook
- Scikit-learn Documentation
- Machine Learning Algorithms
- Keras Documentation

Conclusion:

The project successfully addressed the task of customer churn analysis by employing machine learning techniques. By preprocessing the dataset and building a neural network model, insights were gained into factors influencing customer churn in the banking sector. Further improvements can be made by experimenting with different architectures and hyperparameters to enhance model accuracy and performance.