Key Insights from Datasets and Model Training Results

Imbalanced Datasets

Do not train the model if you do not handle imbalanced datasets. Since we have only 388 data of "True" and 2,278 data of "False" in the Churn data, this imbalance can lead to the model being biased towards predicting "False" more often. Oversampling can be done by adding more copies to the minority class. This is done until the majority and minority class is balanced out nearly close.Oversampling can be a good choice when we don't have a ton of data.

LabelEncoder

For categorical variables that represent binary choices like "True/False," "Yes/No," in such cases label encoding can be used to convert these categories into numerical values.

Normalization(Min-Max Scaling):

Normalization is the process of rescaling data in a new range of **0 and 1.** Normalization is good to use when **our data does not follow a normal distribution(Gaussian distribution)**

Applying Principal Component Analysis (PCA) for Dimensionality Reduction

We know that PCA performs best with a normalized feature set, so we apply it above. After applying PCA, I reduced the features from 18 to 13 principal components, which capture 99% of the information from the dataset.

Model Performance Results

1. From Traditional Machine Learning Models **Random Forest** performs best with an accuracy score of 98% .
2. From the Deep Learning Neural Networks model **Seq2Seq with LSTM Classifier** performs best with an accuracy score of 98% .

My Recommendation:

Deep Learning Neural Networks generally excel at capturing complex relationships in data compared to traditional machine learning models. Therefore, the Seq2Seq with LSTM classifier performs best, achieving an accuracy score of 98%.