Week 4 Make Data Model Ready

The dataset consisted of 10,000 rows and 14 columns. There was no missing values for this dataset, but there were outliers for the Exit 0 group. This meant elder people were more willing to stay at the same bank. I would keep outliers as it was likely that outliers would be predictive of churn. Also, I dropped three columns which were not needed for modeling. The removed columns were row number, customer Id and surname. Among the rest 10 features, two of them were categorical attributes. I converted geography and gender into numerical variables. I used one hot encoding to convert geography variable into numerical variable because there was no sequence or order in geography. Then, I used a different strategy, which was labelencoder to convert gender attribute to numerical variable.

Then, I used stratified sampling to split the dataset. I split the dataset into train, validation and test sets. The train set was 80%, the test was 20% and the validation set was 20% of the training set. With the added validation set, I could use it to compare the performance of different models once I got to that step. I chose these percentage because my dataset was relatively small and I wanted the training data to be a big portion of the whole dataset. I used stratified sampling to avoid extreme case. Also, I did data normalization and standardization. The purpose of normalization was to change the values of numeric columns in the dataset to use a common scale, without distorting differences in the ranges of values or losing information. The reason why I did data standardization as well was to transform the values in a dataset to have a mean of zero and a standard deviation of one. The data was basically ready for training and model implementation at this point.

In addition, my dataset had no data leakage problem. Data leakage happens when information outside of the training dataset is used to create model. However, I divided the dataset into training, validation and testing sets, and only used the training set for model training to avoid leakage of information from the test set into the model.