Since the target variable that we must classify for this assignment is “Exited”, we first removed the features that we thought were useless in determining whether a customer will exit or not. The features that we removed from the dataset were “RowNumber”, “CustomerID”, and “Surname”. After removing these features and analyzing the dataset, we also thought it was a good idea to encode the labels “Gender” and “Geography” in order to avoid any problems with non-numerical data later.

Table

Description automatically generated

After removing useless features and encoding labels, we split the data into a training set and a testing set using the common 80/20 split ratio.



Table

Description automatically generated

In logistic regression our accuracy score is 0.8005 or 80.05%. In the confusion matrix below, we can see that logistic regression accurately classified 1601 out of the 2000 customers in the testing set, giving us the before mentioned accuracy score.

Graphical user interface, text, application, email

Description automatically generated

Chart, treemap chart

Description automatically generated

|  |  |
| --- | --- |
| ML Model | Accuracy on Test Set (Provide accuracy in %) |
| Naïve Bayes |  |
| KNN |  |
| SVM |  |
| DT |  |
| Logit |  |