Relax Challenge_klagios

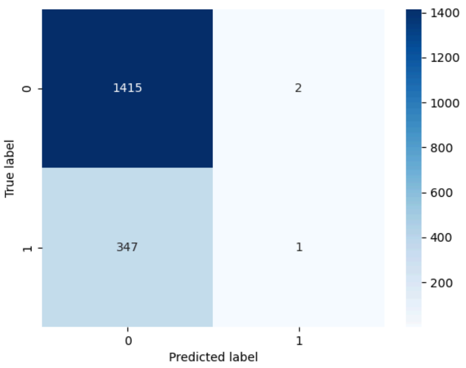
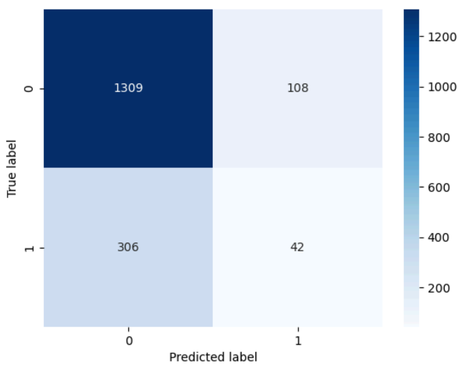
This problem presented is to look at the data about user profiles and login dates and develop a model that predicts who will adopt the software and what factors contribute to adoption. This is a classification problem and the inverse of a churn analysis.

Two CSV files were provided: one with user profile information and the other the login times per user. The data did not require much cleaning other than converting the date features to datetime objects and eventually imputing some of the data to fill NaNs.

In order to prepare the data for the adoption analysis, the adoption feature had to be created based on the provided criteria which stated that a user was considered adopted if they had, “logged into the product on three separate days in at least one seven day period.” Code was written to determine which users met this condition. Once this was created as a separate dataframe it was combined with the user profile dataframe.

Several algorithms were tested including Random Forest Classifier, Logistic Regression, Support Vector Machines and Gradient Boosting. In this case, the objective was to look for those who adopted, and it turned out the class was somewhat imbalanced, with only 18.7% of the users in the ‘adopted’ category. Because of this, when it came to evaluating the models, although the Gradient Boost had the highest accuracy (80.2%), the Logistic Regression model is preferred because it had the highest recall at 0.12 and correctly identified the most users in the True class.

Because the data is imbalanced, additional work could be done to help improve the performance of the models by resampling (over or under) the data prior to modeling. In addition, the Decision Trees algorithm could also be tried.

Confusion Matrix for Gradient Boosting Classifier	Confusion Matrix for Logistic Regression
 1415 2 347 1 0 1 True label 0 1 Predicted label 1400 1200 1000 800 600 400 200	 1309 108 306 42 0 1 True label 0 1 Predicted label 1200 1000 800 600 400 200