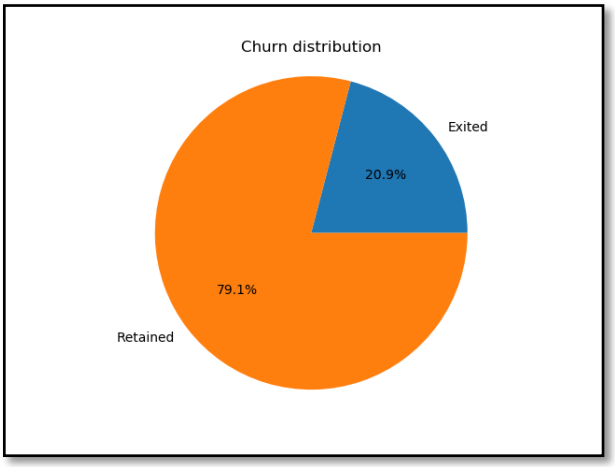
Report

One of the common business problems in many companies is to find out which customers will potentially be using their products or services continuously in a long-term perspective, becoming loyal clients, and which ones will possibly churn. It is very important to identify these types of customers to manage the time and resources more efficiently, also trying to take precautionary measures to keep those who have tendency to leave. The churn estimation is based on the machine learning algorithms of binary classification predicting whether a client will stop doing business with the company or not, given multiple parameters. The target variable, the churn of customers, usually represents a sparse vector due to the fact that only some customers tend to leave.

The data are obtained from a mobile company, anonymously shared on Kaggle. The dataset consists of 66469 entries depicting the customers from few months of 2013. There are 66 variables related to different measurements of calls, messages, durations and other information extracted via data mining techniques. The variables include the id, which was dropped before modeling since different for each customer and, therefore, useless in predictions, and the churn itself, which is the target variable in the models. The churn distribution in the data is as follows:

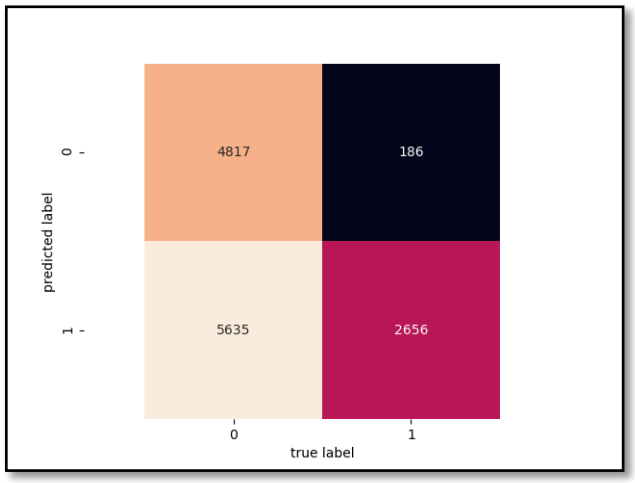


Such distribution implies that guessing that all customers are retained in the business regardless of any variable would result in almost 80% of accuracy score, which is definitely a wrong solution. The underrepresented group of exited clients has to be dealt more carefully. For the same reason different measurements along with the accuracy score, such as the AUROC or, in other words, the area under the Receiver Operating Characteristic (ROC) curve, which constitutes a tradeoff between the true positive (TP) and the false positive (FP) rates, have to be used since the accuracy score can be misleading.

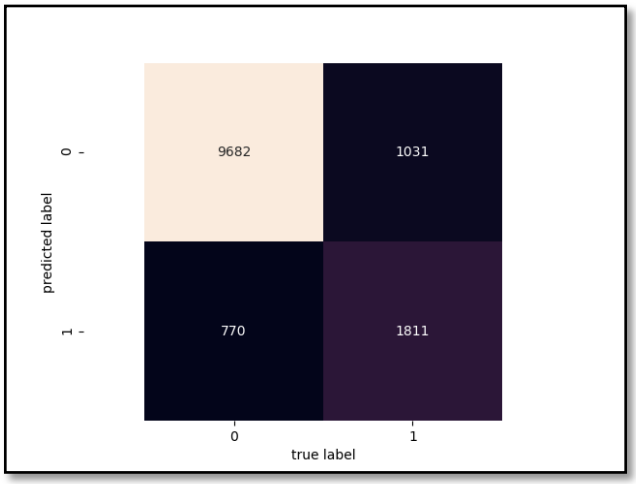
Before the application of various machine learning models, the data was normalized using the min-max scaler, which normalizes by only keeping the relative distances between the

data points. It was done to use the algorithms more efficiently and faster, because normalization allows some algorithms to approach to the solution faster.

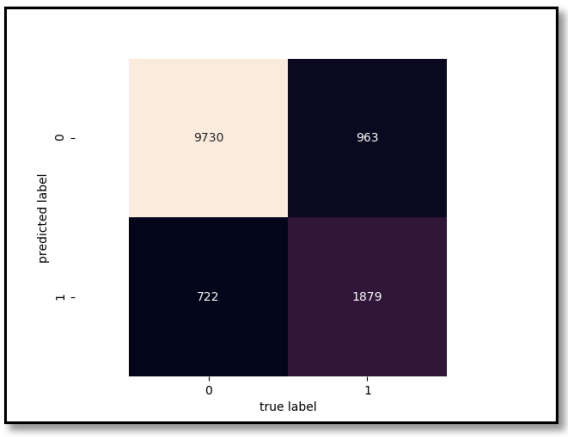
The first model is the Naïve Bayes classifier. It is among the simplest and fastest models for classification, using a simple formula with a naïve assumption of no covariance between the variables. It sometimes even outperforms the more complicated models. Here, the calculated accuracy score is 56.21%, which is quite low but better than random, and the area under ROC (AUROC) is 69.77%, which is higher due to better accuracy for the underrepresented group, as seen in the confusion matrix:



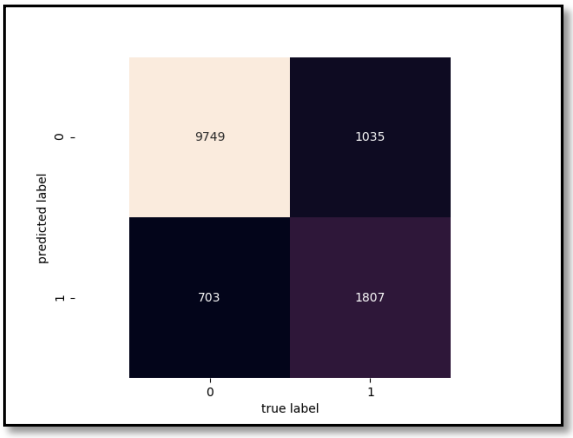
The second model is the Logistic Regression, which is classical for this type of problems. It uses a sigmoid-shaped activation function to produce binary output, either zero or one. The accuracy score is 86.45% and the AUROC score is 78.18%. Its confusion matrix is summarized below:



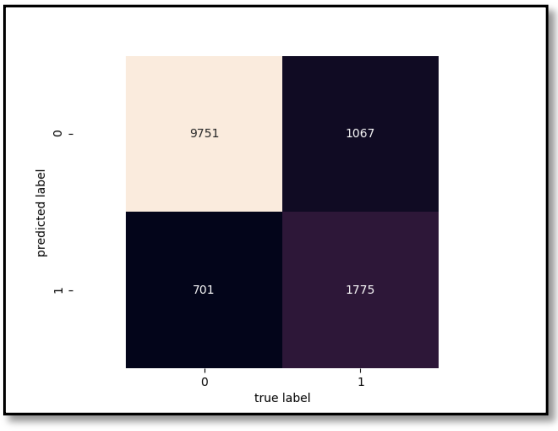
The next model is K Nearest Neighbors classifier. It is a nonparametric method, using only the distances and closest neighbors to decide on prediction. The number of neighbors was chosen to be 50, although more neighbors would improve the result; however the change is not so significant. The observed accuracy score is 87.33% and AUROC is 79.6%. Here is the confusion matrix:



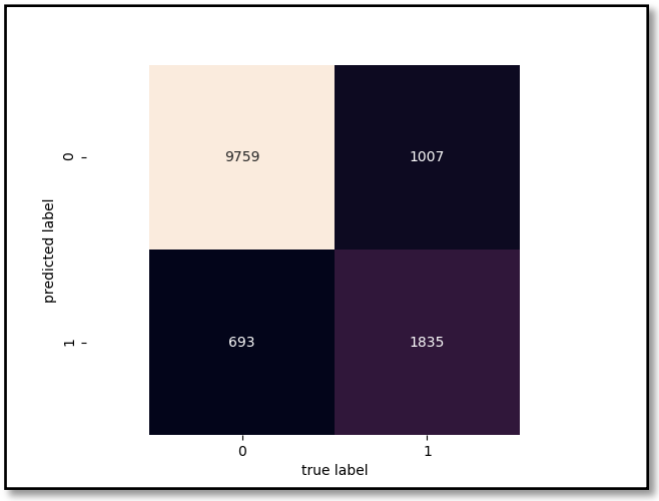
The fourth model is Random Forest classifier, which is an ensemble of decision trees. A decision left alone is a very weak nonparametric model; however as a combination of many trees using the boosting aggregation, it is a powerful tool. Similarly, the number of estimators was chosen as 50 and not more because the improvement was insignificant and time consuming. The accuracy and AUROC scores are 86.93% and 78.43%, respectively. The confusion matrix is given below:



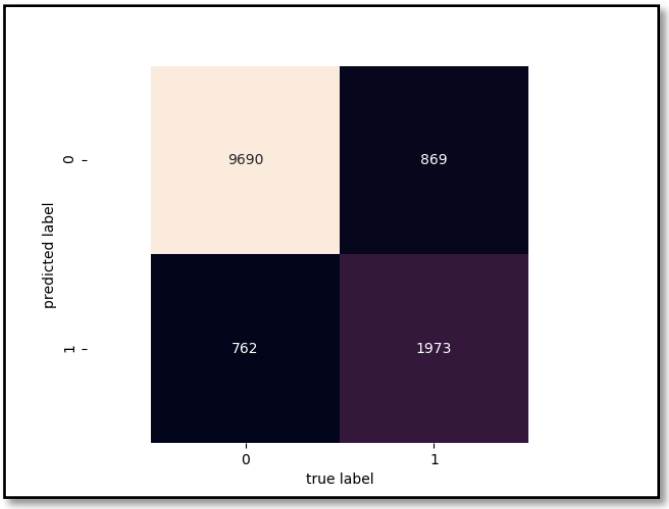
The fifth model is Support Vector Machines classifier. It is a strong model, based on several optimizations, which require some time for computations. The value for C was chosen through available grid search cross validation function. The resultant accuracy score is 86.70% and the AUROC is 77.87%. Its confusion matrix is summarized below:



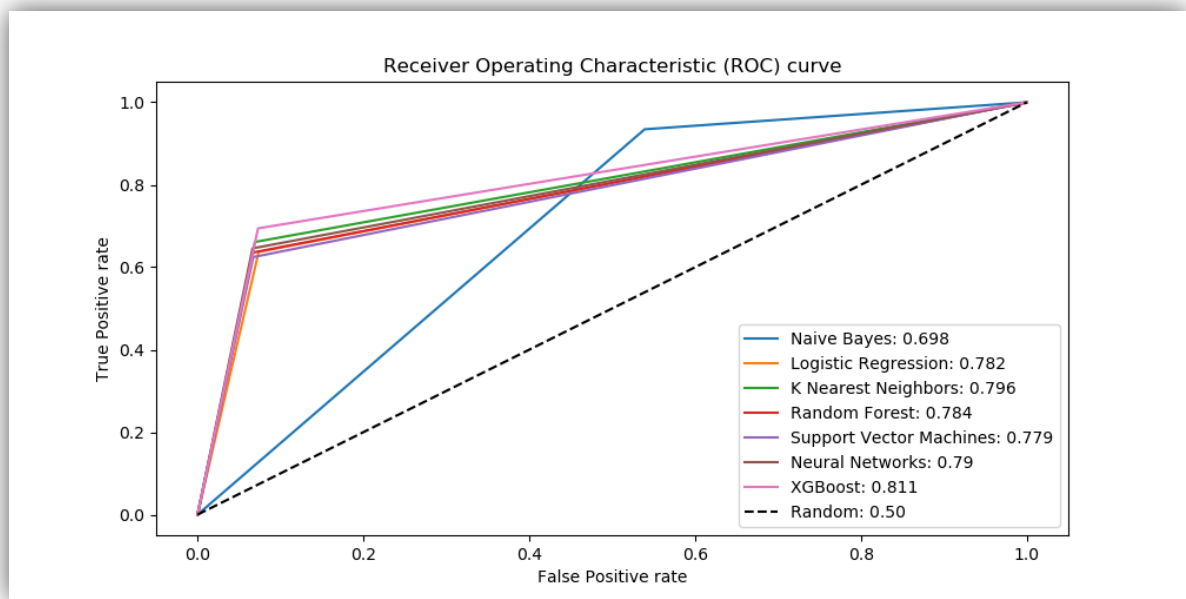
The following model is Neural Networks classifier. It is mysterious model which works as black box but it also takes some time to compute the weights for each layer to produce the result. The parameters were tunes manually, choosing such two layer representation. The accuracy score and AUROC are 87.21% and 78.97%, accordingly. This is the confusion matrix:



The last model is Extreme Gradient Booster (XGBoost) classifier. It is also an ensemble learning model. It has been the winner in many programming contests. The number of estimators was tuned with the grid search cross validation function. Its accuracy score is 87.73% and AUROC score is 81.07%. Its confusion matrix is given below:



To conclude, among all the used models, both the highest accuracy and the highest area under the ROC curve (AUROC) belong to XGBoost classifier, although it is slightly better than its alternatives. To summarize the comparison even better, the ROC curve is demonstrated below:



As seen in the graph, there are two peaks where ROC curves concentrate. On the left, the dominant model is XGBoost, followed by the K Nearest Neighbors, which is followed by the Neural Networks model. The peak on the right represents the Naïve Bayes classifier, which is best in predicting the underrepresented group of exited clients, although its accuracy score in total is not high enough. All other accuracy scores are greater than 85%, which is far better than guessing all as retained clients (79.1% of accuracy) and this is fascinating.