April 27, 2020

To: Guido Rossum, Senior Data Scientist, CreditOne
From: Kevin Burr, Analyst

Subject: Response for Analysis of Credit One Data

Dear Guido,

Below is the results from the assessment of the problem to solve for CreditOne:

Is it possible to reliably predict a client's ability to repay a line of credit balance in order to better manage default rates, thereby retaining clients, ultimately protecting revenue?

Three models were developed to predict the possibility of credit default in the month following the latest bill period. Additionally, three different feature "configurations" were utilized to assess potential impact on outcome predictability.

The three models used were: KNN, RandomForestClassifier, and SVM. Each variant has a classification model that was applied to a subset of the data – the training data – and model performance was then measured. Finally, the best model based on the accuracy score for the training data was applied to the test data, from which accuracy – and therefore a measure of reliability – is derived.

Three different feature sets were run through each model.

1. FeatureSet1 (FS1) - All the features with the exception of BILL_AMT1, which is highly collinear with BILL_AMT2 (0.9515). All attributes except the categorical attributes were scaled. The models were then applied.
2. FeatureSet2 (FS2) – All features with collinear relationships calculated at .90 or above were removed. This resulted in all BILL_AMTX features removed EXCEPT BILL_AMT6. The collinear values are presented in the notebook for reference. All attributes except the categorical attributes were scaled. The models were then applied.
3. FeatureSet3 (FS3) - All features with collinear relationships calculated at .90 or above were removed. This resulted in all BILL_AMTX features removed EXCEPT BILL_AMT6. The collinear values are presented in the notebook for reference. . All attributes except the categorical attributes were scaled. Additionally, all categorical attributes were one-hot-encoded (dummified).The models were then applied.

Following are the model results for each feature set as described:

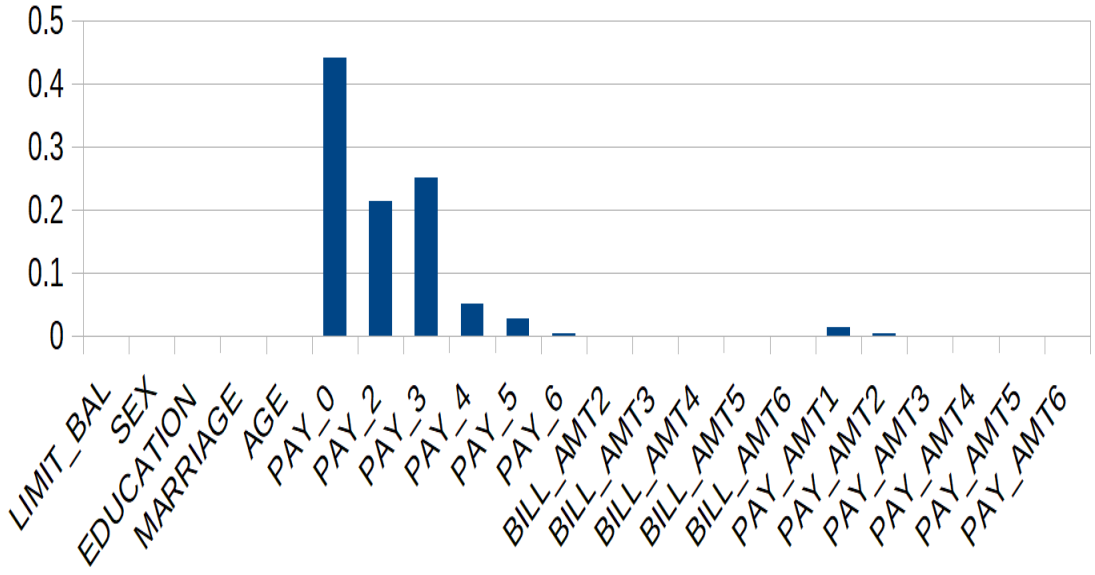
Model Cross Validation Accuracy Scores by Feature Set	FS1	FS2	FS3
KNN	.81	.809	.809
SVM	.801	.801	.801

RF	.819	.812	.801

The Random Forest Classifier model using the first feature set was the best performing model with an accuracy score of .819 – meaning it was accurate 81.9% of the time on the training data set.

This model was then applied to the test data set with an accuracy rate of 81.8% - meaning the model can predict the next month's account status almost 82% of the time accurately. So, while it may not be possible to ENSURE clients WILL pay their balances, it is possible to predict whether or not clients CAN pay their balances.

Furthermore, it is interesting to understand the features that carry the most weight in predicting a client's ability to pay, specifically, the most recent three month's Payment Status (PAY_0, PAY_2, PAY_3), with the last month's payment status being overwhelmingly important.



In summary, it is possible to be almost 82% confident in predicting the next month's account status based on the features in FS1 (provided), paying significant attention to the pay status for the three most recent past months.

We could potentially uncover more information by adding more features to the data set that may provide more insight into a client's propensity to pay such as a credit score, employment information (employed, income level, number of years at a specific company, etc.). There may also be a correlation between a client's ability to pay and current market conditions, which may be very interesting to investigate.