CreditOne - Customer Default Identification Report

(Class 2, Task 3)

Summary

- The predictive model exercise produces a model with high accuracy to predict when a client would not default.
- However, the model has a low accuracy predicting when a client would default. No model produced better results.
- For the cases when the model predicts that a customer would default, further review would be necessary until this can be automated further.

Problem Statement

An increase in customer default rates is bad for Credit One since its business is approving customers for loans in the first place. This is likely to result in the loss of Credit One's business customers. CreditOne is asking our team for a model that can better predict if a client will be at default or not.

Analysis of the Data

We are using the data set provided by CreditOne via SQL database that includes 30K records (observations) each related to a CreditOne credit customer. Each record contains demographic information (education, age, marital status), credit limit, history of balances and payments for the last 6 months and the performance of the credit holder in terms of paying on time or not.

As a summary from our previous report, our Exploratory Data Analysis (EDA) shows the following:

Population View: Client's average age is 35.49 years-old with 75% of clients being 41 years-old or younger. Most of the clients (82%) have, at least, a university degree. Also, 99% are married or single, and we have a 60/40 split for female to male.
75% of the clients have a credit limit of $240K or below. The average credit limit is $167K with a large Std. deviation ($130K), the average monthly bill is $63K and average monthly payment is $5.2K.

Payment and Balances: Monthly balance distribution remains similar from month to month. Payments however, look more scatter as time elapses. This can be caused by people trying to do catch up payments. Some balances are negative which means a client paid more than necessary (creating a credit)

Delinquency: Delinquency has been increasing every month which is what CreditOne stated. September is the worst month with increased cases of one- and 3-month delays.

Correlation: There is a tendency (-0.32) where the lower the credit limit, the higher the incidence of delinquency. Also, there is a small tendency (0.28) where the higher the bill to pay, the higher delinquency cases. Also, divorce clients seem to have the worst time paying on-time. Other values were evaluated vs. delinquency but not much relation was found and they followed the overall composition of the data sample. Ex: Visualization shows that females have a higher incidence of delinquency than males (circa 3:1 ratio), but this is expected as 60% of the clients are females.

Modeling – Data Preparation

Based on the EDA, and thru an initial view at feature selection, we reduced the feature set for the models by removing Gender, Age and Education which correlation shows they have little to no relation with a client entering Default. Also, the initial supervised learning of the models showed an improvement of up to 3 percentage point in accuracy by removing these features from the training set. Another observation was that using the new columns created during the data wrangling phase and used for visualization, also reduce the accuracy of the models. Hence, the features set used was made out of the following:

- Limit Balance
- Bill amount to pay (6 months)
- Paid beahior (6months)
- Marital Status
- Amount paid (6 months)

Modeling – Choosing Algorithm

First, the question of how to ensure that customers can/will pay their loans is translated into a model that can predict if a customer will enter default or not (i.e., Will they pay the bill?). **Default** in this data set is a categorical value and so we need to use **classifier models** (instead of regressors).

After investigating our SciKit toolset, we chose the Supervised Learning algorithm models listed in the table below which we ran thru cross validation using a KFold technique of training the algorithms in 4 subsets of the data and then using a fifth subset to train the models. Then, we measure their accuracy (correct prediction/ total records), also listed below.

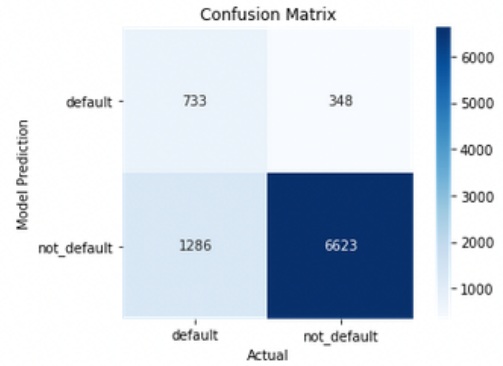
Classifier Model	Accuracy (%)
Decision Tree	**82.099**
Random Forest	81.312
Gradient Boosting	82.059
Logistic Regression	77.861
SVC (Support Vector Machines)	77.968
MLP Classifier (Neural Network)	72.688

The cross validation showed that the most accurate models is the **Decision Tree** which we had tested separately to improve its accuracy. The result is that the optimal depth for highest accuracy is 4 nodes per branch. For sake of completeness, it is important to note that we attempted some tuning to improve all models. This work is reflected in the pipeline that is part of this report stored at GitHub.

Modeling – Validation

As stated before, the **Decision Tree** (DT) model with a maximum node depth of 4 was found to be the most accurate (82.099%) in the cross-validation comparison.

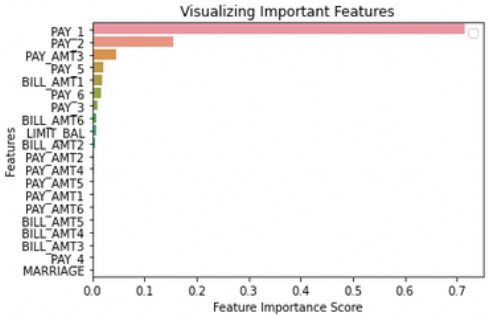
We focused on the DT model to evaluate its performance which can be summarized using the confusion matrix below.



	precision	recall	f1-score	support
default	0.68	0.36	0.47	2019
not default	0.84	0.95	0.89	6971
accuracy			0.82	8990
macro avg	0.76	0.66	0.68	8990
weighted avg	0.80	0.82	0.80	8990

The matrix shows that the model is more accurate predicting correctly when a customer will not default (95%) vs. predicting correctly when a customer will default (36%).

The features used by the model, shown in the figure below, are mostly the payment behavior over the last 6 month and it uses 60% of the features.



PAY_1	0.713989
PAY_2	0.153997
PAY_AMT3	0.045117
PAY_5	0.021521
BILL_AMT1	0.017283
PAY_6	0.016828
PAY_3	0.008455
BILL_AMT6	0.007797
LIMIT_BAL	0.007021
BILL_AMT2	0.004204
PAY_AMT2	0.001949
PAY_AMT4	0.001838
PAY_AMT5	0.000000
PAY_AMT1	0.000000
PAY_AMT6	0.000000
BILL_AMT5	0.000000
BILL_AMT4	0.000000
BILL_AMT3	0.000000
PAY_4	0.000000
MARRIAGE	0.000000

In addition, the ROC curve (left side on figure below) of the capability to predict accurately when a customer will not default shows an area under the curve total of 0.737 which is adequate and aligns with the 95% accuracy discussed before. The ROC Curve to the right shows the poor coverage of the model to correctly predict if a customer will default

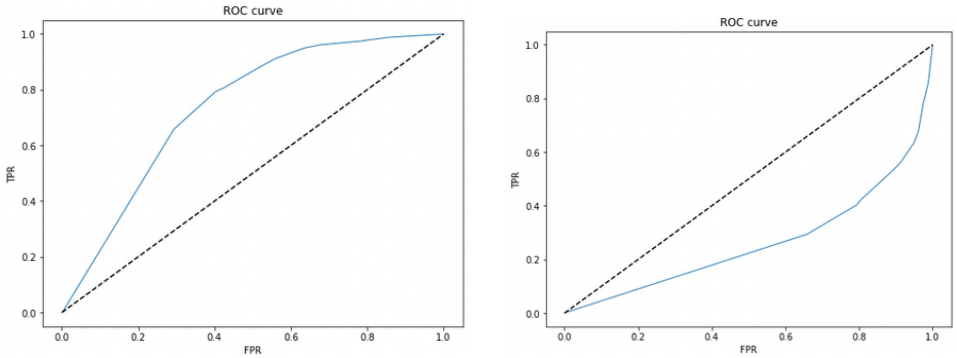


Figure – ROC Curve “Not Defult” vs. ROC Curve “Default”

Last, neither of the other models tested provided a better confidence of accurately predicting when the customer will default. Overall, the Decision Tree model is the best with 36% confidence value.

In addition, we ran three regression models to predict the credit limit for customers who would not default. The accuracy of the models was not higher than 32%. No extra tuning was employed as this is seen as a preview of future work. Once we have a more robust predictive model to screen customers for default, we can work further on tunning this regression model.

Recommendations

The resulting model has an overall accuracy of 82% which is acceptable, but it has a very low accuracy to predict defaults. Therefore, it is not recommended to go into production until it can be improved.

If the model predicts that a customer would default, then additional review would be necessary (in this case human intervention).
Else, the model predicts that a customer would not default with a high accuracy. It would still need further training to improve its 82% accuracy.

An option to improve the model is to add data on the state of the economy (Ex: Employment rate, inflation, interest rates) and more information on client’s income, credit score or finances aimed to improver the accuracy of the models.

Additional Technical Details

The provided data set has the following characteristics:

Number of observations	29,965 (each a single person credit behavior)
Attributes	24 [in-store, age, items, amount, region]
Duplicates	236. All dropped
Bad/Missing data	3. All dropped
Unused Attributes	Initial index in original data

Correlation result: