Summary Report – Credit One – Default Predictability and Analysis

By Roberto Siegert – Nov 22nd 2019

This task requires that you prepare one deliverable and one Juypter Notebook:

1. Customer Default Identification Report that addresses:

Problem:
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.

Questions to Investigate:

1. How do you ensure that customers can/will pay their loans?

Answer: In order to ensure that the customer is able or will be able to pay his/her loan, we can look into the covariance matrix of the entire dataset. The covariance matrix will allow us to identify how the variable "Default Payment Next Month" relates with respect to the other variables.

But the covariance matrix also indicates that there is perhaps an even stronger direct correlation between the occurrence and the amount (value) of the BILL_AMTx variable.

2. Can we approve customers with high certainty?

Answer: The pay history correlates positively with a true default payment status. It seems that by looking into previous pay history, we (loan institution) can shield ourselves from extending loans to potential unreliable customers.

It was also demonstrated that the overall probability of default, based on the occurrence of the condition is 0.2212, or more than 1 in 5 customers will go into payment default.

As you progress through the task, begin thinking about how to solve the company's problem.

Here are some lessons the company learned from addressing a similar problem last year:

1. We cannot control customer spending habits

2. We cannot always go from what we find in our analysis to the underlying "why"

3. We must focus on the problems we can solve:

    1. Which attributes in the data can we deem to be statistically significant to the problem at hand?

Answer: From our data wrangler steps, graphical and covariance matrix analysis we identified the numerical PAY, BILL and PAY_ATM variables as being statistical significant. Showing a strong correlation with the occurrence of the default conditions.

But it was also shown that categorical variables such age are well suitable for the implementation of classification algorithms such as Random Forest.

    2. What concrete information can we derive from the data we have?

Answer: It can be derived that a set of identified variable and the corresponding historical data can be used to predict if a new client will or will NOT fault in default when applying for a loan.

    3. What proven methods can we use to uncover more information and why?

Answer: The use of robust and proven machine learning algorithm, commonly used in industry and with emphasis on classification – tree-based algorithms such as Random Forest.

With PAY_AMT6 as dependable variable we obtain

	Cross Val Score			Model Score	R Squared	RMSE
Random Forest	-1.59257	0.1073	0.112	0.92772	0.024	11088.343
Support Vector Regression	-0.085205	-0.0314	-0.0237979	-0.0239	-0.078	11654.238
Linear Regression	-10.5677	0.50328	0.352357	0.64615	-1.440	17535.702

Continuing with the model adjustment, we can evaluate using PAY_5 as dependable variable, obtaining

	Cross Val Score			Model Score	R Squared	RMSE
Random Forest	0.74359	0.678658	0.621435	0.93920	0.486	0.775
Support Vector Regression	0.00803	0.093577	0.04839	0.66384	0.069	1.043
Linear Regression	0.1084	0.01635	0.08182	0.13077	0.090	1.032

We can also evaluate using PAY_0 as dependable variable, obtaining

	Cross Val Score			Model Score	R Squared	RMSE
Random Forest	0.46261479	0.4429	0.234677	0.860176	0.042	0.968
Support Vector Regression	-0.0155584	-0.057687	-0.021255	0.56623	-0.028	1.003
Linear Regression	0.051957	0.038014	0.08311	0.09520	0.030	0.974

We can also evaluate using AGE as dependable variable, obtaining

	Cross Val Score			Model Score	R Squared	RMSE
Random Forest	-0.15877	-0.09306	-0.26619	0.771626	-0.052	9.590
Support Vector Regression	-0.025793	-0.0500009	-0.058576	0.116798	-0.053	9.598
Linear Regression	-0.07179	-0.01413	-0.013499	0.027679	0.019	9.265

Recommendations:

The dataset is robust and can be evaluated using common industry supported classification algorithms such as Random Forest. The dependency of every variable can be weighted using a platform such as Jupyter notebooks which allows us to import python libraries and package such as SciKit, Numpy and Pandas, and graphical tools such as Matplotlib and Seaborn.

Although regression algorithms were tested, the responds seems more robust (accuracy ~ 0.8) when using classification algorithms. In our case, we focused in the use of the Random Forest algorithms.

Some extracts of our analysis are presented in tabular form, as shown above.

It seems that the problem statement is structured such that different algorithms can be evaluated with ease and refined extensively.