Summary Report – Credit One – Default Predictability and Analysis by Roberto Siegert – Dec 11th 2019

This task requires that you prepare one deliverable and one Jupyter 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 evaluate the customer's historic data as far as billing and payment data. We can also evaluate the performance of regression and classification algorithms applied not only to the payment and bill amount data, but also to key features such as age, sex and level of education.

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.

But in our analysis, we evaluated both, regression and classification algorithms that were conducive to low accuracy and kappa values. Hence, it seems that further analysis and the evaluation of alternatives models need to be explored, in order to generate more robust predictors.

Literature point towards the use of logistic regression algorithms to be more suitable for this type of task.

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_AMT variables as being statistically significant. Showing a strong correlation with the occurrence of the default conditions.

As shown in the notebook, we conducted a scoring analysis in order to determine what parameters have a higher contribution to the overall default prediction. It was found that the variables of higher relevance were billing and payment amount. It was also found that the customer "age" has some relevant weight as far determining the probability of default.

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.

Results from the regression analysis

	Cross Val Score			Model Score (X_test, y_test)	R Squared	RMSE
Random Forest Regression	0.115727	0.114513	0.111340	0.125311	0.125	0.388
Support Vector Regression	-0.01969	-0.022437	-0.01672	-0.022913	-0.023	0.420
Linear Regression	0.12312	0.11677973	0.11192	0.13705	0.137	0.386
Naïve Bayes	0.4083	0.36346	0.393652	0.38413	-2.572	0.785

Results from the Classification Analysis

	Cross Val Score			Model Score (X_test, y_test)	Kappa
Random Forest Classifier	0.806959	0.8104	0.80744	0.8104	0.282294
ADA Boost Classifier	0.814558	0.81573	0.813308	0.8288	0.408186
Gradient Boost Classifier	0.8157579	0.8152	0.81544	0.8289333	0.410312
K-Nearest Neighbor	0.77589	0.77706	0.77583	0.77586	0.065671

From the above tables, it can be derived that the classification algorithms exhibit a greater performance than their regression counterparts, for this particular dataset. But although the classification models are scoring high, a poor performance in predicting default is observed, due to the low Kappa values obtained.

Furthermore, we evaluated the number of True Positive and True Negative counts and compared them with the number of False Positive and False Negative counts. This was conducted for all classification algorithms and in all cases, the number of False Positives and False Negatives resulted into 5x to 10x the counts of True Positives and True Negative results.

This additional evaluation confirmed the inherent weakness associated to the use of classification algorithms for this particular dataset.

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

Answer:

At first glance, classification algorithms seem to show more robustness than linear regression algorithms, mainly due to a higher scoring. But low Kappa values and a significantly higher count of false positives and false negative render these algorithms as weak for this particular dataset.

From a brief literature review, it seems that this type of datasets are commonly are addressed with Logistic Regression algorithms and much more detailed and time intensive techniques.