

CREDIT ONE FINAL REPORT

**Summary of Findings**

Alert Analytics

Ingrid Ortega Rojas

2020

# 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.

# Narrative of the data supported by the results

This report aims analyzing and developing models for predicting default of customers. This is an input for Credit One to design protocols and modify processes to decrease the number of customers who end in a default. A BADIR Data Science framework was used. First of all there were some questions to data which were brainstormed. Then an analysis plan was done making visual descriptive analysis then a predictive analysis. The latter is train, test, validate and compare performance metrics of different classification supervised algorithms. The data was collected, the plan was executed, and recommendations were made.

# Questions to investigate: Answers

The lessons previously learned:

* We cannot control customer spending habits
* We cannot always go from what we find in our analysis to the underlying "why"
  1. How do you ensure that customers can/will pay their loans?

Ensuring that customers pay their loans according to lessons learned from the past, analytics cannot ensure this behavior; but through data science we can provide insights that can be a base to take actions.

After doing some exploratory data analysis defaulters used more of their limit than non-defaulters. The most valuable insight from visualization is that the system allows customers to surpass the limit of credit; this should be revised.

Second of all, it is recommended to automatize the model developed, and apply it to every new customer. As the most important variable is the past credit behavior, this should be asked directly to other banks or if this is not legally allowed, then ask to the customer as requisites. If the model shows that the customer may be a default one, then it can be denied or either take some additional actions to enforce the money is recovered, such as penalties (negative) or financial education (positive). It is suggested to consult the Law Office for some ideas.

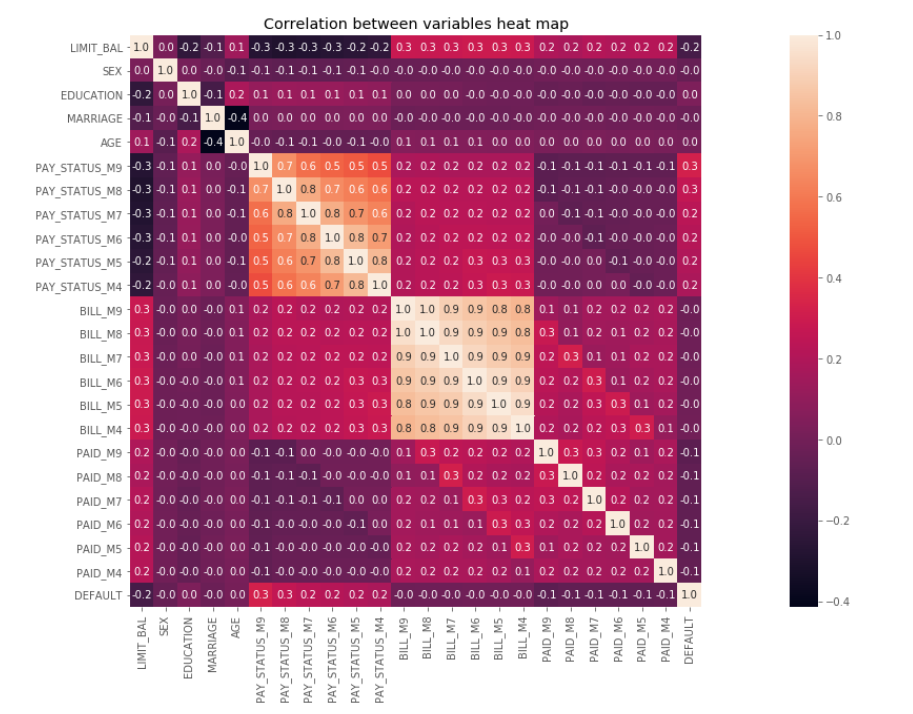
* 1. Can we approve customers with high certainty?

Yes, if high is greater than 70%. There is 77% (AUC) of chance that model will be able to distinguish between defaulters and non-defaulters.

# Problems to be solved

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

The main variable to solve the problem is default. The rest of the columns (attributes) were statistically evaluated among them and to default. The following heat map shows that the pay status from month 4 to 9 are the most significant variables that influence default.



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

The mean limit balance of the population is 130.109 (NT$) with a minimum of 10.000 (NT$) and a maximum of 740.000 (NT$). The mean age of customers in 35 years.

From 30.000 observations, 18.112 (60%) are female and 11.888 (40%) are males. The credit one default rate is 20.78% vs male default rate which is 24.17%.

The average bill in month 4 is 38.271(NT$) and increased to 48.509 (NT$) 5 months later.

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

The models can be extended to consider new calculated variables which can be a combination of the actual available ones. Actual trained models considered the raw variables; but as data miners, the data can be combined or treated in different creative ways.

The default can be left out and not considered as a dependent variable; instead it be evaluated and predicted the payment as there is some statistical relation between bill and payment. This can give interesting information to decision makers, so they can planned the money flow of Credit One for example.

Test of hypothesis is a proven statistical method that can be done.

# Relevant visualizations

|  |  |
| --- | --- |
|  |  |

|  |  |  |
| --- | --- | --- |
|  |  | |
|  | |  | |

# Rules that provide insights

* Customers tend to increase their bill through time.
* Customers are able to exceed customer limit balance.
* Logistic model is the best algorithm that represents reality.
* History of payment status and bill amount are critical to define default.

# Other relevant observations

Logistic model is fast and was the most reliable. The following are the scores based on time for fitting the estimator and time for scoring the estimator marked accordingly.

* Logistic: 1st 0.81809524 0.83238095, 2nd 0.82228571 0.82857143 3rd 0.82247619 0.82780952
* Decision tree: 1st 0.73309524 0.72190476, 2nd 0.81790476 0.82171429 3rd 0.80819048 0.81085714
* Random forest: 1st 0.81404762 0.82238095 2nd 0.81561905 0.81847619 3rd 0.81742857 0.81971429

Principal components analysis is a tool that allowed to consider only 5 components for numerical values instead of 14 attributes.

# Final recommendations

* Evaluate all the insights and take actions that improve actual performance.
* Automatize this model, including it in the software for client service.
* Update the model yearly according to the new behavior of data.