Customer Default Identification Report

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.

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

Female are more likely to take loans according to this data set, female will have more issues to pay the loan; however, given that the amount of women in this data set is significantly higher than the male amount, and that the percentage of default is not as high from the female side, I would say that maybe men are not as good payers either, we would need a more even data set in terms of female and male representation to observe that behavior in a better way.

There's no correlation seen between people being single, married or divorced and paying or not the loan.

The age group that presents more defaults in the loan payments is the group between 21 and 31 years of age, followed by the group between 31 and 41.

People with university degrees have likely taken loans to pay for college, interestingly enough, they are also the ones who represent a higher percentage of loan default.

So, we could say that the worst type of client to give a loan would be a woman (taking into account the observation above) independently of her marital status, between 21 and 31 years old with a college degree.

In order to ensure people will pay we need to identify some other variables, like a credit score, we would want to know history of payments with other financial entities, we would like to know whether they have had loans before and if they have been able to pay them or if they have defaulted as well. But based barely in this information, we would say that the company would need to pay attention to young clients since they are starting their professional life, in the country we are analyzing, financial education isn't as spread as in other countries, so it is probable that these young persons are taking loans to appear more opulent that they really are, they would not be able to pay these loans since they represent mere expenses and not investments so they won't be generating enough to pay and to gain something from it. Also, people with university degrees likely had to take a loan to pay for education, and if we analyze the socio-economic background of the country, we could see that Asian countries like Taiwan are increasing their unemployment rates, affecting mostly young professionals with little to none professional experience looking for a job, and without a job the probability of loan payment decreases significantly.

2. Can we approve customers with high certainty?

Using our model, we could generate a level of accuracy of 98% using the Support Vector Machine model, so we could say that we could predict the customer behavior with a high level of accuracy; however, to ensure these results we would like to analyze the data using other variables like credit scores and history of payments.

Some lessons the company learned from addressing a similar problem last year:

1. We cannot control customer spending habits

Customers tend to change behavior through time, since loans typically comprehend several years, it is understandable that they will change spending habits; however, through credit history we could have a basic idea of the behavior even if it changes in the future.

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

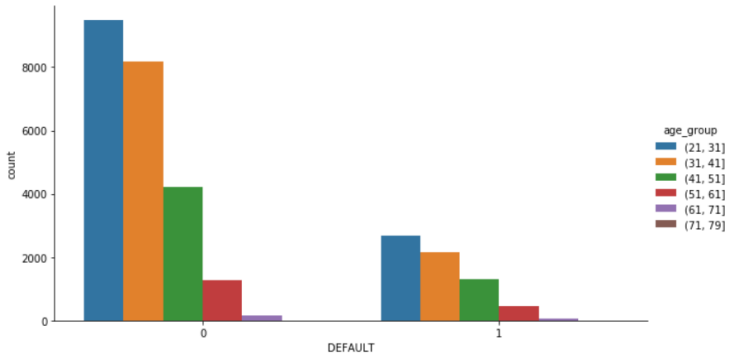
The data shows us information in a way we can analyze and get insight from; however, sometimes the data isn't showing what we need to solve the problem or understanding the background of it, sometimes the reasons are more complex, and we need other types or analysis in order to get to the bottom of it.

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?

Some attributes prove to be statistically significant from our analysis, for instance Age, shows an interesting behavior since mostly young people tend to default in the loan payment; and they are also the ones who have the most loans overall, so that would be expected given the amount of data.

```
Out[103]: <seaborn.axisgrid.FacetGrid at 0x1e426c456a0>
```




```
Out[105]: age_group
(21, 31]    12163
(31, 41]    10320
(41, 51]     5521
(51, 61]     1713
(61, 71]      204
(71, 79]       12
Name: age_group, dtype: int64
```

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

As mentioned before, it is difficult to derive some concrete information as the data is very polarized, there's a lot of women for example, is this the case in the whole credit portfolio of the company? Is it also the rule in the company that young people have the most loans? Do they have access to the credit scores of those customers?

We would need to access some more information in order to be able to gather a more concrete conclusion from this data. As for the current information, I would say no, we can't derive a concrete conclusion from this data set. We were able to assess very accurately the behavior of default or not default using SVM; however, to understand why we need more insight.

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

It is proven that credit scores are world wide used to determine if a person is a good creditor or not, it is very popular in the United States for example, where people have a score and it is used for literally any loan they apply for, it is even used for job applications. I would recommend gather this information if available in the country or find something similar that can give us more insight in the historical payment behavior.