12/07/2021

Guido Rossum

senior data scientist, credit one

Overview:

As per your request, I have further analyzed the customer data in order to provide insight into how we can ensure that customers pay their loans and how we can approve customers with high certainty.

DATA:

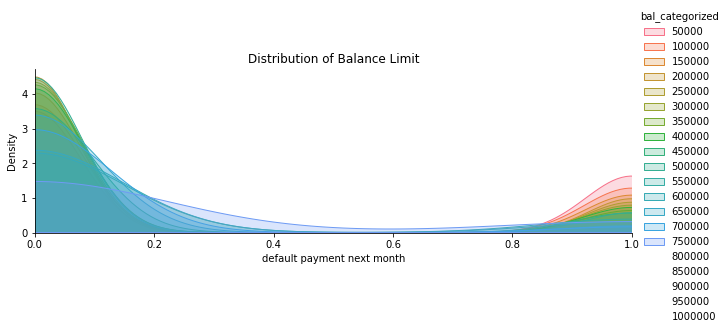
As this is an extension of the previous task, I am still utilizing the customer data present in the company’s SQL database. The data contains 30,000 customer records and 25 features. This includes age, balance limit, gender, education level, marital status, whether or not the account is in default, and half a year of billing and payment transactions. The data also needed to be cleaned.

Analysis:

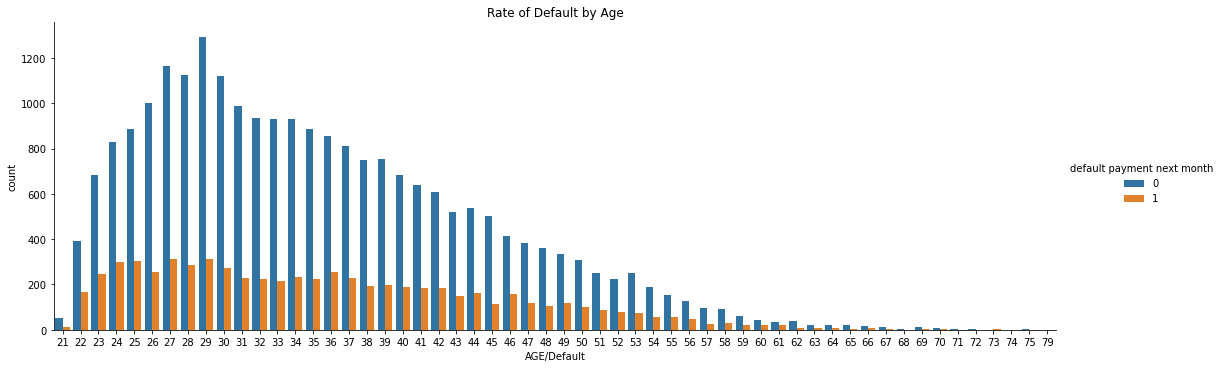
I utilized Random Forest Regressor and Linear Regression by using the balance limit as the dependent variable. This produced a similar value, however, Linear Regression was insignificantly higher. In order to make predictions, I calculated the r2 and RMSE scores, 0.469 and 93736.642, respectively. The RMSE tells us that the average deviation between balance limits is $93,736.64. The r2 value tells us how well the model can predict the value of the response variable in percentage terms. In this case, 46.9% , which is not a very good indicator.

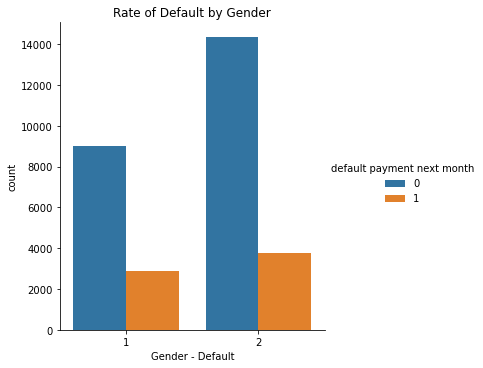
Therefore, I built two classification models utilizing Random Forest Classifier and Decision Tree Classifier. The first was to discretize the balance limit into 20 bins of 50,000 and set as the dependent variable. The second was to use the account default status as the dependent variable. The cross\_val\_scores for discretized balance limit did not go above .25, which is not a strong indicator of a good model. The cross\_val\_scores for default status bore the best results at .816, which indicates a reliable model.

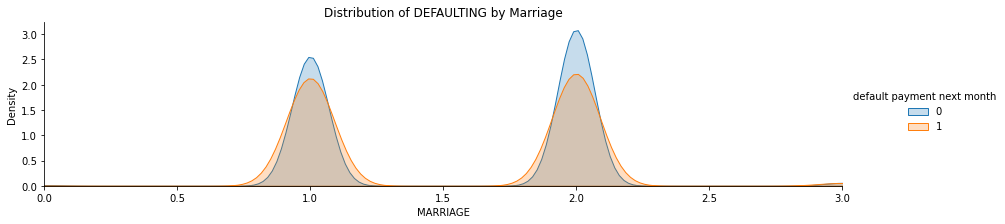
By using Seaborn and Histograms, I created visual models to help determine who is defaulted. The majority of defaulted accounts are those with the lowest balance limits. The rate of defaulting is lower the higher the balance limit is.

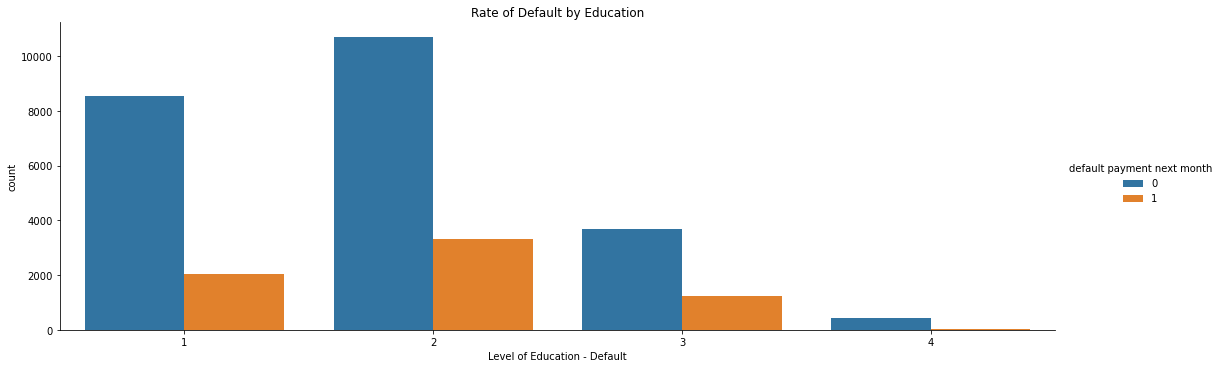


Conclusion:

Based on my findings, the following are not good indicators of whether a customer will default: age, marital status, gender, or education. However, the data shows that a customer is more likely to default (not default = 0, defaulted = 1) if they are younger, but most of the clients are younger. Men default at a slightly higher rate than women. In the graph, 1 signifies male and 2 signifies female.



Marital status also is a poor indicator of a customer who will default. Very few customers are divorced or in the other status. There is a slightly higher default rate for married customers, however it’s not significant. In the graph, 0 signifies other, 1 signifies married, 2 signifies single, and 3 signifies divorced. 

The majority of customers are college educated, closely followed by graduate school, then high school. There are very few customers who do not fall into any of those categories. High school educated customers have the highest rate of defaulting. In the graph, 1 signifies graduate school, 2 signifies university, 3 signifies high school, and 4 is other.

The greatest indicator of a customer who might default is education. However, in each feature of education, there are defaulted clients. Only the rate is slightly higher for those with only a high school education. The lower the limit balance the more defaulted accounts are present. It must be noted that the mean balance limit is $167,484.32. In the future, eliminating the highest balance limit customers with a balance limit over $800,000 should be considered, since there are very few at that level.

Moving forward, I suggest requiring current and potential customers to disclose their salary amount. I believe this will provide more insight and be the deciding factor on whether a client will default or not. The limit balance could then be set to a threshold which is easily met by customers without defaulting.

GRANT JOHNSON