Josh Martin
Data Analytics / Big Data
Course 2, Task 2
2021-05-18

Exploring Credit One Data

Project Overview

Credit One has recently seen an increase in the number of customers defaulting on their loans. The company is interested in understanding if it is possible to ensure customers will pay their loans and if possible, how to do so.

Dataset Details

Credit One provided a data set with 30,202 transactions and 25 features. After some cleaning efforts, the dataset was cut down to 30,000 transactions (202 duplicates or unintentional rows dropped) and 24 features (the ID feature was discarded). Loosely speaking, the features broke down into two types, customer demographics or loan information.

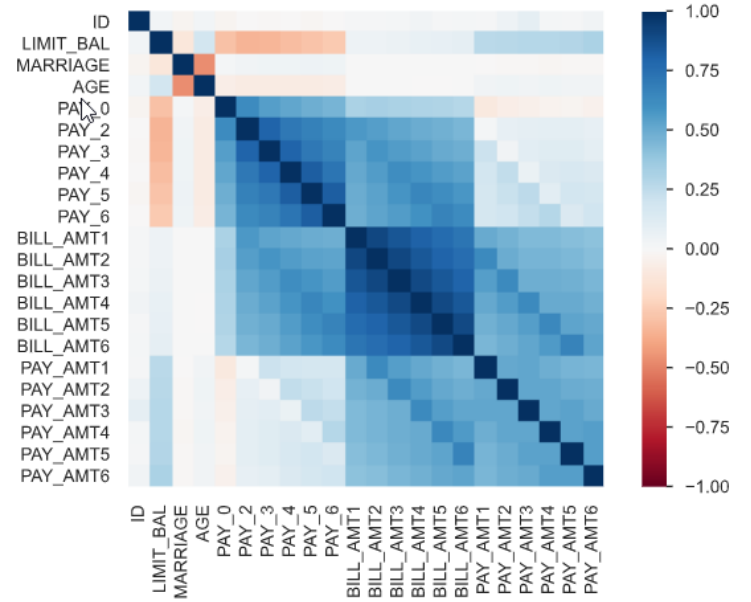
Demographic Features	SEX, EDUCATION, MARRIAGE, AGE
Loan Information Features	PAY_0, PAY_2,PAY_3, … ,PAY_6, BILL_AMT1, BILL_AMT2, … ,BILL_AMT6, PAY_AMT1,PAY_AMT2, … , PAY_AMT6, LIMIT_BAL, DEFAULT PAYMENT NEXT MONTH

A full description of these features can be found at the end of the report

Analysis

Correlations

After using Pandas profiling to find any obvious patterns in the data, it became apparent that many of the loan information features were highly correlated to one another.



From this Spearman correlation matrix, it is apparent that the groupings

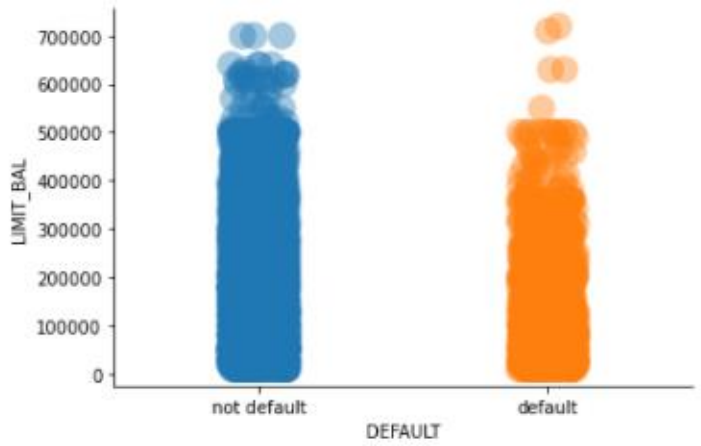
- *PAY_0, … ,PAY_6,*
- *BILL_AMT1, … , BILL_AMT6,*
- *and PAY_AMT1, … , PAY_AMT6*

are highly correlated to themselves and each other.

For future analysis, it could be helpful to remove or aggregate those strongly correlated features.

Larger loan limits

When comparing default status to the loan limit balance, the dataset seems to indicate that higher loan balances do not present a high risk of default.

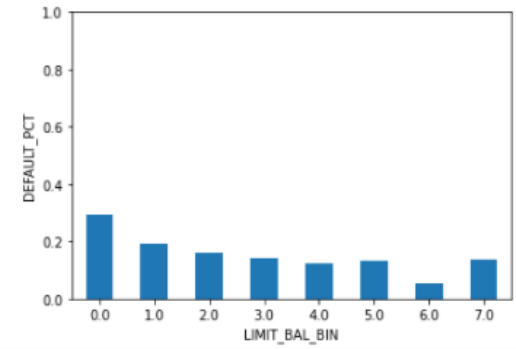


The scatter plot indicates that loan limits above about 500k tend to default less often then loans in the less-than-300k range. Perhaps focusing on the business processes behind approving these loans would be helpful.

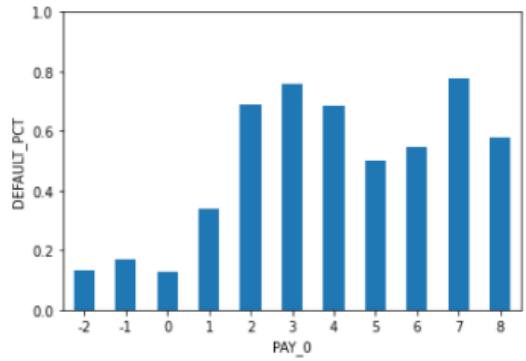
Default percentages by group

Finally, to help understand which groups are defaulting the most, a new variable was added to the dataset:

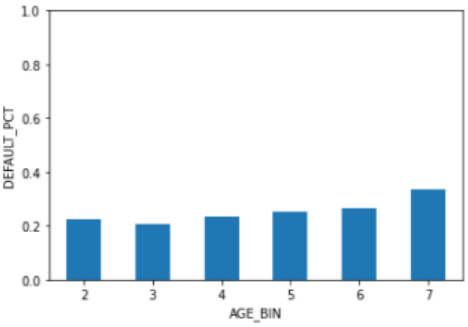
DEFAULT_PCT = sum of customers who defaulted in a particular group divided by total number of customers in that group



The bar chart indicates that lower loan limits seem to default at higher percentages.
(Note:
LIMIT_BAL_BIN: 0 = 0 – 100k, 1 = 100k – 200k, … 7 = 700k – 800k)



This bar chart shows that larger PAY_0 values result in drastically higher loan default rates. Essentially this means that the borrowers historic monthly payment statuses loosely correlate to whether the customer will default.



This last bar chart indicates that the higher the age of the customer, the more likely the customer will default.
(Note:
AGE_BIN: 2 = 20 – 30 years old, 3 = 30 – 40 years old, … 7 = 70 – 80 years old)

Conclusions

Through the exploratory data analysis, it became clear that many of the loan information variables were correlated and will need to be reduced before further modeling can occur. Also, it seems clear that loans over $500K default less frequently than loans under $300k. Finally, age, payment status, and loan limit seem to be linked to how likely the customer is to pay back their loan.

Feature Descriptions

LIMIT_BAL: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.

SEX: (1 = male; 2 = female).

EDUCATION: (1 = graduate school; 2 = university; 3 = high school; 0, 4, 5, 6 = others).

MARRIAGE: (1 = married; 2 = single; 3 = divorce; 0=others).

AGE: (year).

PAY_0 – PAY_6: History of past payments. (-2: No consumption; -1: Paid in full; 0: The use of revolving credit; 1 = payment delay for one month; 2 = payment delay for two months; . . .; 8 = payment delay for eight months; 9 = payment delay for nine months and above.)

BILL_AMT1 – BILL_AMT6: Amount of bill statement (NT dollar).

PAY_AMT1- -PAY_AMT6: Amount of previous payments (NT dollar).

DEFAULT PAYMENT NEXT MONTH: (default, not default)