Gideon Blinick

Capstone 1 Milestone Report

Problem Statement

Goal

Credit card use is widespread in the United States. About 70% of all Americans use at least one credit card and there were 364 million open credit card accounts in the United States at the end of 2017, according to the American Banking Association (Gonzalez-Garcia, 2018). With use of credit cards comes the risk of default, which is defined as failing to make a payment for 180 days (Konsko, 2014).

The goal of this project is to effectively predict credit card payment default using demographic factors, credit data, repayment statuses, bill statements, and history of payments. Two sub-goals accompanying that goal are determining how the probability of default payment varies by categories of different demographic variables, and determining the strongest predictors of default payment among the variables (Default of Credit Card Clients Dataset, Kaggle.).

Client Interest

The client for this project is, naturally, credit card companies. Such companies have a significant interest in predicting which customers will default on their payments because such defaults cost them money, and thus, they would rather not extend money to individuals with a high probability of default. A good prediction model will enable them to lend to good customers.

A good prediction model will also enable credit card companies to make early and effective interventions with existing customers who are likely to default. Such interventions may take the form of debt and financial counseling, enabling the customer to maintain good credit and enabling the company to receive payment. In less fortunate circumstances, an intervention may allow credit card companies to be paid ahead of other creditors.

In summary, credit card companies are the primary stakeholder in this project, the decision being improved is the acceptance and rejection of credit applications as well as the decision of who to extend interventions to, and improvement of the decision results in greater profit for the company.

It is worth mentioning that this problem is not too much different from other problems such as bankruptcy prediction, and that a solution to one problem might generalize easily to the other. Hence, other stakeholders to this problem could include banks and other large creditors.

Dataset Description

Overview

From Kaggle: “This dataset contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005.”

There are 25 variables:

ID: ID of each client

LIMIT_BAL: Amount of given credit in NT dollars (includes individual and family/supplementary credit

SEX: Gender (1=male, 2=female)

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

MARRIAGE: Marital status (1=married, 2=single, 3=others)

AGE: Age in years

PAY_0: Repayment status in September, 2005 (-1=pay duly, 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)

PAY_2: Repayment status in August, 2005 (scale same as above)

PAY_3: Repayment status in July, 2005 (scale same as above)

PAY_4: Repayment status in June, 2005 (scale same as above)

PAY_5: Repayment status in May, 2005 (scale same as above)

PAY_6: Repayment status in April, 2005 (scale same as above)

BILL_AMT1: Amount of bill statement in September, 2005 (NT dollar)

BILL_AMT2: Amount of bill statement in August, 2005 (NT dollar)

BILL_AMT3: Amount of bill statement in July, 2005 (NT dollar)

BILL_AMT4: Amount of bill statement in June, 2005 (NT dollar)

BILL_AMT5: Amount of bill statement in May, 2005 (NT dollar)

BILL_AMT6: Amount of bill statement in April, 2005 (NT dollar)

PAY_AMT1: Amount of previous payment in September, 2005 (NT dollar)

PAY_AMT2: Amount of previous payment in August, 2005 (NT dollar)

PAY_AMT3: Amount of previous payment in July, 2005 (NT dollar)

PAY_AMT4: Amount of previous payment in June, 2005 (NT dollar)

PAY_AMT5: Amount of previous payment in May, 2005 (NT dollar)

PAY_AMT6: Amount of previous payment in April, 2005 (NT dollar)

default.payment.next.month: Default payment (1=yes, 0=no)

How the Data Was Obtained

This dataset was obtained from Kaggle (https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset/home) .

It was sourced from the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients) and was the subject of an academic paper (https://bradzzz.gitbooks.io/ga-dsi-

seattle/content/dsi/dsi_05_classification_databases/2.1-lesson/assets/datasets/DefaultCreditCardClients_yeh_2009.pdf).

Data Cleaning and Wrangling

Since the dataset was loaded from Kaggle, it was fairly clean. Nevertheless, inconsistencies were found in the data that needed to be corrected.

First, the data was examined for missing or Null values using the Pandas info method. No columns had Null values, and values of 0 were part of the domain in the columns in which they were found, so nothing needed to be done.

Next, outliers in the columns were examined. It was determined that no values should be excluded from the dataset because although there were plenty of values that met the statistical definition of being an outlier (greater than the 3rd quartile value + 1.5 times the interquartile range or less than the 1st quartile value minus 1.5 times the interquartile range) these values all made sense in context and there was no evidence that values had been entered incorrectly, or that the data was corrupt in any way.

And yet, some values did need to be changed because they violated the description of the dataset. There were three instances where this occurred.

In the first instance, it was discovered that the PAY_X columns (the columns determining how many months a customer was behind on credit) contained values of -2 and 0 when they should only have contained values of -1 (representing no months behind on repayment) and positive integers (representing the number of months behind payment for whom this applied). Given the definition of the column, it doesn't make sense to have negative values (because you cannot be ahead of bills, only behind or on time). Therefore, the decision as made to replace all negative values in these columns with 0.

In the second instance, it was discovered that the Marriage column contained values of 0, when it should have only contained values of 1 for "Married", 2 for "Single", or 3 for "Other". The logical decision was made to recode these values as 3, since the values of 0 and 3 effectively were 2 values for "Other".

Finally, in the third instance, there were values of 0 for Education which the dataset description did not account for. Furthermore, there was a column for "Other" and 2 values for "Unknown". Since all of these values represent essentially the same thing, it was decided to group all of these values under one coding. Therefore, values of 0 (unaccounted for in the dataset description), and 5 and 6 (both values for "unknown") were recoded to values of 4 representing "Other".

Additionally, it was also the case the PAY_0 column was awkwardly named, since the other PAY_X columns had values in range(2, 6) inclusive. Therefore, the column was renamed to PAY_1, which also made it conform to the convention used to name the BILL_AMTX and PAY_AMTX columns.

Exploratory Data Analysis

Data Storytelling

There were 2 main components to the Exploratory Data Analysis stage of this project. First, correlations between our variable of interest or dependent variable, 'default' were examined. And second, correlations and relationships between various independent variables were looked into.

Gideon Blinick

For the calculating of the correlation between default and other variables, a simple Pearson correlation was calculated for each of the variables and default, and their results were ordered from highest to lowest. It was discovered that none of the variables had significant or even moderate correlation (defined as having an absolute correlation values greater than 0.5), but that the LIMIT_BAL variable (measuring the credit extended to a customer) and the PAY_X variables (representing repayment status over 6 months) were most correlated with default. In the case of LIMIT_BAL, it was more negatively correlated with default than any other variable, having a correlation value of about -0.15, and in the case of the PAY_X variables it was found that they had correlations ranging from about 0.24 for PAY_6 all the way to about 0.4 for PAY_1 with the strength of the correlation increasing with later time (that is, lower X in PAY_X).

In the case of the PAY_X variables, it made sense that they would be more strongly correlated with default than any other variable or variable group, since they measure repayment status which can intuitively be understood as being very related to default. Furthermore it made sense that the correlation would increase with time since being behind on repayments in September should be more correlated with defaulting in October than being behind on repayments in April.

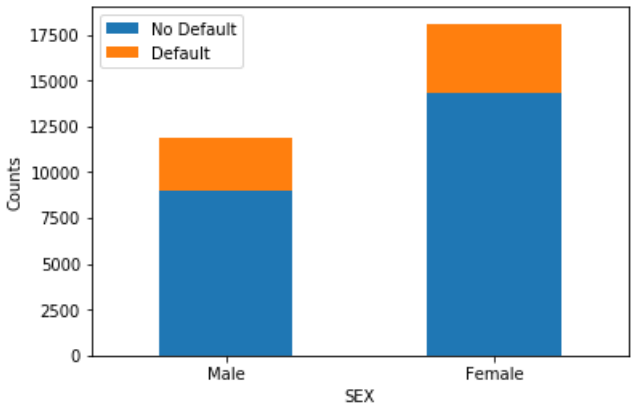
In the case of the LIMIT_BAL variable and the reason for the negative correlation with default, a working hypothesis is that the credit agency only extended more credit to individuals who they were more confident could pay back. Therefore, individuals with more credit should have had greater repayment abilities, and thus lower rates of default.

As for the second element of Exploratory Data Analysis, that of finding relationships among variables, a number of relationships were discovered, but only the top 4 will be mentioned for brevity.

They are:

1) Men default at a higher rate than women

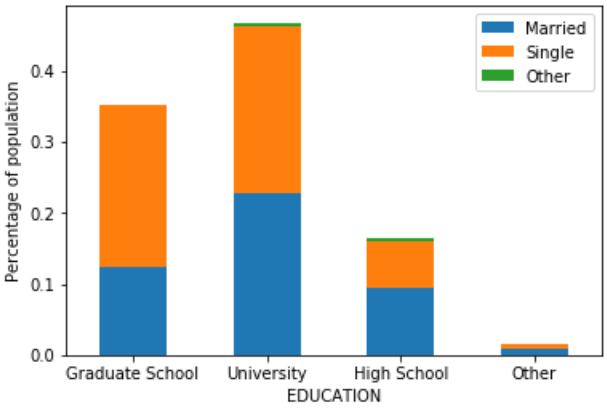
This can be seen in the following chart:



And table:

default	No	Yes	Percentage Default
SEX			
Male	9015	2873	24.2%
Female	14349	3763	20.8%

2) People with only a high school education have higher rates of marriage than people with graduate school education



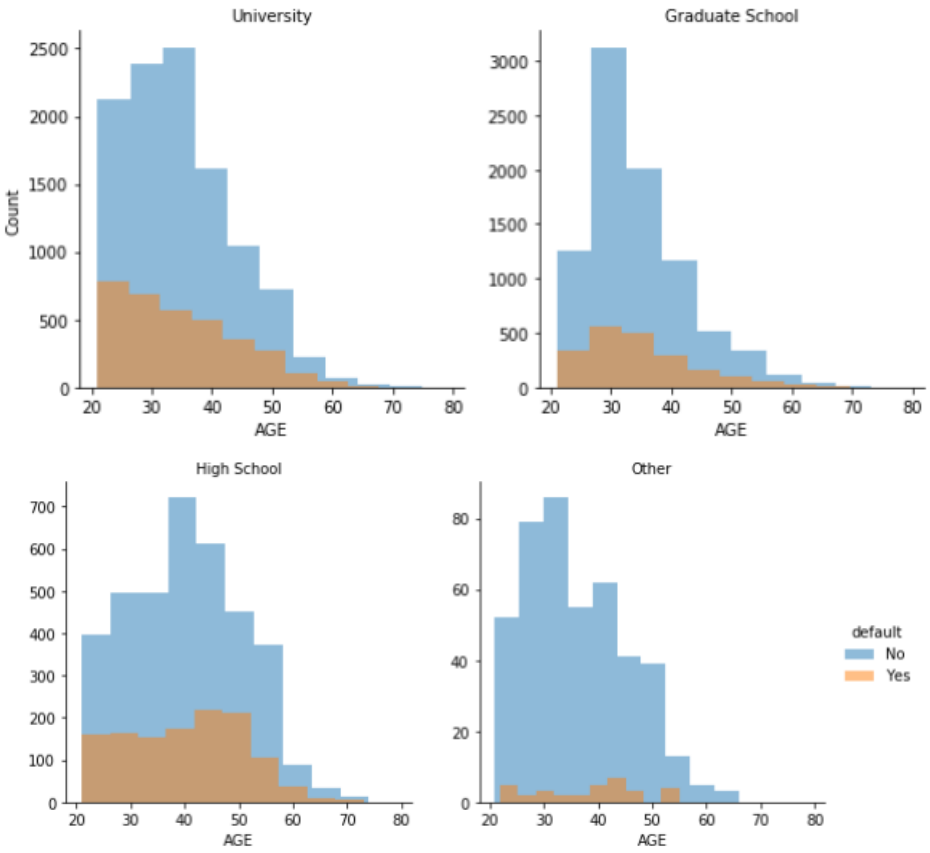
MARRIAGE	Married	Single	Other	Total
EDUCATION				
Graduate School	35.2%	64.3%	0.5%	100.0%
University	48.8%	50.0%	1.2%	100.0%
High School	58.2%	38.8%	3.0%	100.0%
Other	50.0%	48.3%	1.7%	100.0%

3) The mean age of people with only a high school education (about 40.3) is considerably higher than people with a graduate school education (about 34.2).

default	No	Yes
EDUCATION		
Graduate School	34.1	34.6
High School	40.3	40.2
Other	36.0	38.2
University	34.7	34.7

The histograms below also show also show that more education is associated with younger age. Additionally, from both the chart and the below histograms, we see that the ages of defaulters and non-defaulters is not significantly different.

Gideon Blinick



4) Married individuals are more likely to default than singles (23.5% vs. 20.9%)

default	No	Yes	Percentage Default
MARRIAGE			
Married	10453	3206	23.5%
Single	12623	3341	20.9%
Other	288	89	23.6%

After obtaining these insights, inferential Statistics were not applied to the dataset to determine if they were statistically significant, because the dataset constitutes a population of credit card holders and not a sample.