**Exploratory Data Analysis / Lessons Learned – Credit One**

Credit One has tasked us with examining current customer demographics to better understand what traits might relate to whether or not a customer is likely to default on their current credit obligations. Credit One has been collecting and storing the data with an external server which we were able to access and mine for analysis.

**Objective**

The main objective is to perform an Exploratory Data Analysis to determine if/why customer defaults have risen recently. We are asked overall how to evaluate/predict a customer’s creditworthiness and how to avoid future customer credit defaults. In other words, how do we ensure that customers can/will pay their loans? And can we even do so?

**Exploratory Data Analysis**

The available data used for this analysis consists of 30,000 individual customer entries, each comprised of these demographic variables:

* Age
* Gender
* Marital Status
* Education Level
* Credit Amount (aka limit balance)
* Bill Amount (past 6 months)
* Payment Amount (past 6 months)
* History of Past Payment (including measurement scale of 6 month repayment breakdown)
* Default Status (yes or no)

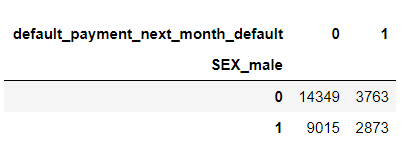
Of these 30,000 records, we find that 6636 have defaulted on payment – about 22% total. While that does not seem like a large percentage, we are talking about money lending and collection. So 22% in this instance can be seen as a fairly high rate of default and certainly not a rate that should be allowed to rise.

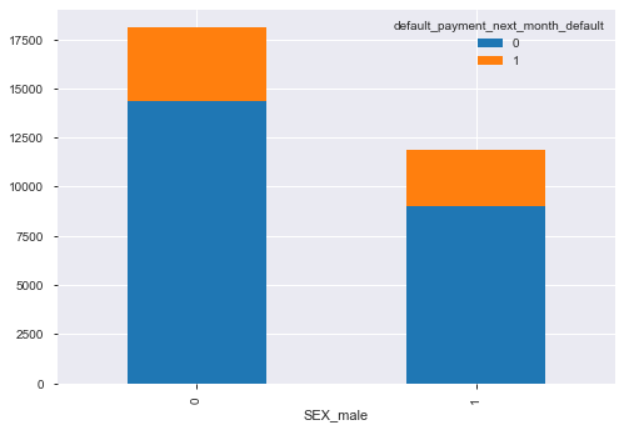
In order to handle and interpret the data offered, we inspect and prepare the information to produce a manageable data set. We examine the data provided as it is recorded and collected by the server, cleaning and validating to the best of our abilities. For this exploratory analysis we compare variables of the data to reach conclusions related to the pressing issue of increase in defaults on the rise – most notably region and amount spent, as well as number of items and amount spent.

**Results**

*Gender and Default Status*

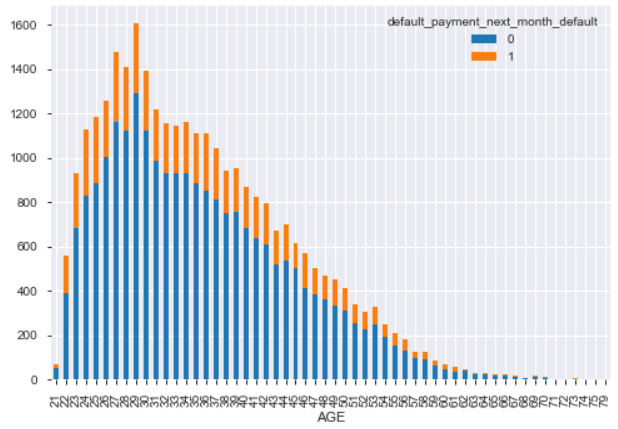
First, we compare gender to default status and we see women (indicated numerically as 0) have a higher number of defaults to men (indicated as 1). Females account for 3763 defaults while males account for 2873 defaults -- or 57% and 43% respectively **of all defaults recorded**. However, we also see more women represented as credit holders overall which may be creating an issue of ratio here. If we compare women to only other women represented in the data, the percentage of default becomes 26%, and male default represents 31% when compared only to other males in this dataset.





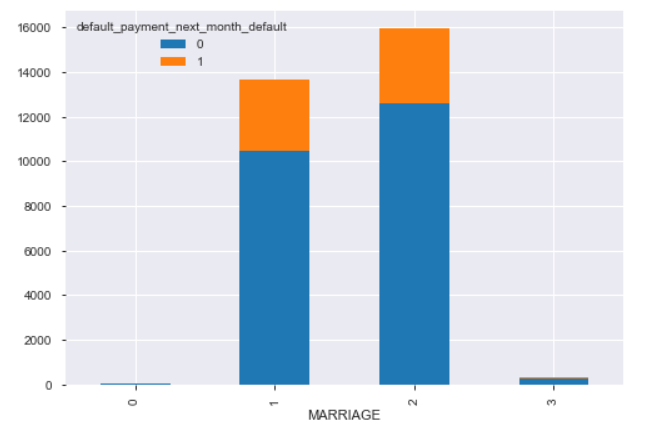
*Age and Default Status*

Ages across all 30,000 records range from 21 to 79. In that, we see the highest number of defaults between the ages of 23 to 37. The defaults in this age range are recorded at over 200 plus defaults per each age and combine to account for 3,897 of the 6,634 total defaults – meaning this age range accounts for 59% of all defaults recorded. Specifically, the ages of 27 and 29 have the highest numbers of defaults at 313 each. Again, this age range also contains much of the overall number of records as well.



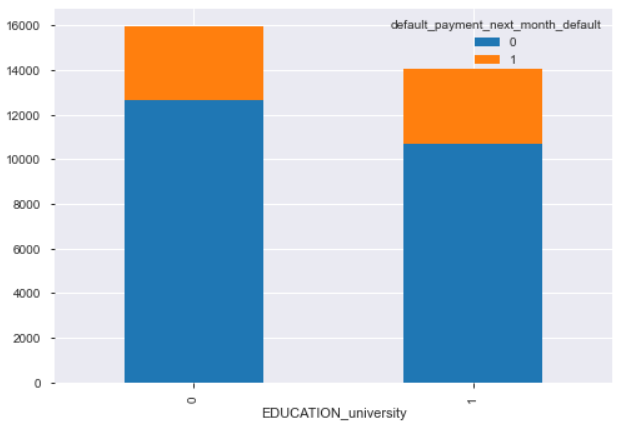
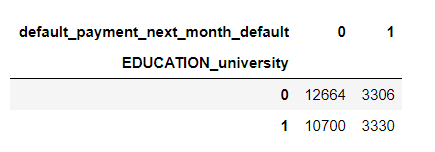
*Marital Status and Default*

Again, we see the largest number of records also contains the highest number of defaults. Married credit holders (represented as 1) account for 3206 defaults, singles (2) account for 3341 defaults, while divorced (3) and other (0) show the least number of defaults.



*Education and Default*

In keeping with what we have already discovered, again the categories with the highest number of records results in the highest number of defaults. Education in this data is separated by 4 options – high school, university, graduate school, and other. University educated represented the largest number of responses at 14030 with 3330 of those as defaults, and graduate school closely followed with 10585 responses and 2036 defaults. Overall, university educated credit holders (shown in below chart as 1) who defaulted (shown also as 1) accounted for 50% of all 6636 defaults recorded.

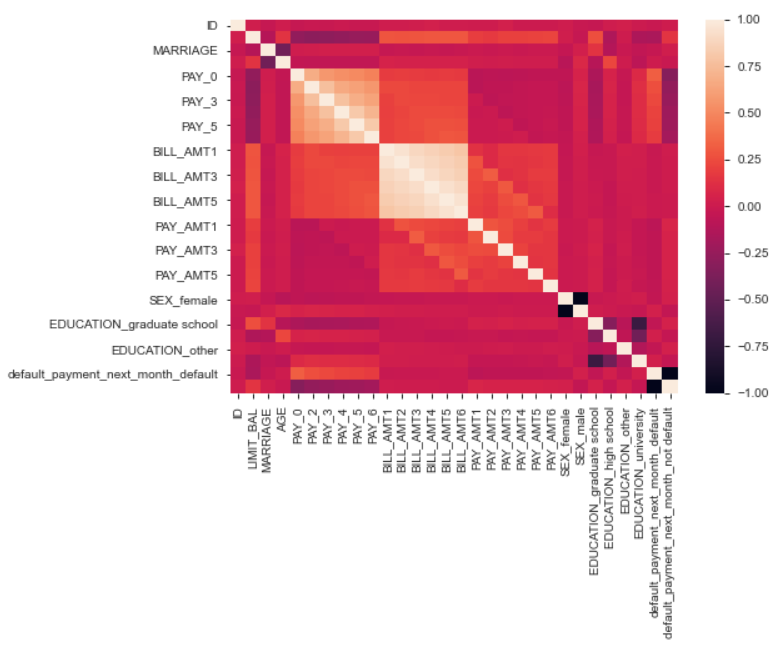
 

**Next Steps**

Moving forward I believe more in-depth analysis is needed. While none of these variables truly stood out, these demographic characteristics that ranked among the highest numbers of defaults and how they relate to one another would serve well to further investigate:

**females  
ages 23-37  
single (followed closely by married)  
university education (followed by graduate education)**

In addition to these demographics, **repayment history, bill amount, and payment amount** all demonstrate the strongest correlations. This is not surprising to see as they are all closely related anyway. Bill amount obviously relates to payment amount and payment amount obviously relates to repayment history, but they do not indicate default necessarily. We can also see that **limit balance, bill amount, and payment amount** are related as expected, but again do not show strong correlation with default.



All in all, customer spending habits are unpredictable and difficult to determine accurately with any confidence. With further exploration we may be able to identify the riskier demographics, but currently there is little information to be gleaned from the data as-is without deeper examination and manipulation. It is also important to note that the strongest correlations are not going to determine who is likely to default, or even who to issue credit to, until **after credit is already issued**.

**Lessons Learned**

***What Worked?***

* Data cleaning and validation were the biggest hurdles in this instance. Reviewing the data time and again proved to be advantageous. Multiple headers and mixed data variables created several obstacles to overcome.
* Focusing on the standard demographics of age, gender, marital status, and education have allowed us to focus on the largest groups of those customers who default as well as those who do not. Understanding those demographics may well prove fruitful as this analysis moves forward.
* I am unfortunately not completely confident in the accuracy of the records as collected and recorded, but at least I am aware that those issues persist and can continue to keep that in mind moving ahead.

***What Didn’t?***

* Comparing the various columns/variables to the default status proved to be more difficult than expected. In theory comparing a binary column to a binary column should be simple enough, yet I struggled with this is practice. It has been quite the challenge to compare the variables in a useful way, particularly multiple variables of the same category separated into multiple columns.
* From the comparisons I was able to make, I noticed that the majority of the highest numbers of instances of default come from groups with larger numbers of records overall, so ratio or percentage-wise, are we seeing realistic assumptions? I think those may still be plausible areas to examine for further possible identifying factors considering these are the demographics most often seeking credit/lending as well.
* The strongest correlations are found among variables highly dependent on one another, such as credit limit, bill amount, and payment amount, while we find no strong correlations related to default specifically.
* I expect the best indicator of default comes only after credit is issued. Obviously as a customer racks up bill amounts, the greater the debt becomes and the likelihood of default also increases. So unfortunately, that expected predictor leaves us days late and thousands of dollars short.
* Lastly, I am not completely confident that these records were collected/recorded correctly. For example, default status may not be correctly displayed – there are several instances of “no consumption of credit” or “paid in full” records ending with record of default. How can that be? We have no choice other than to use the data we are provided, but if the data collected by the source is incorrect, any assumptions, predictions, and/or recommendations that follow will also be incorrect.