Course 5 Task 3 – CreditOne Report

Customer Default Identification Report that addresses:

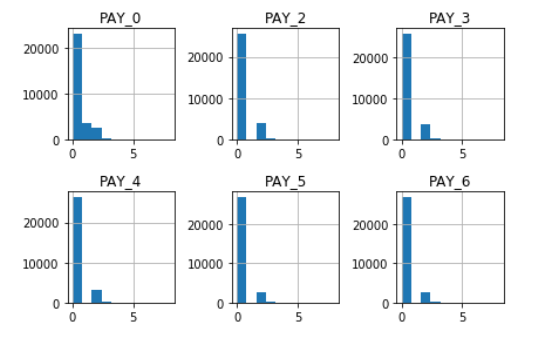
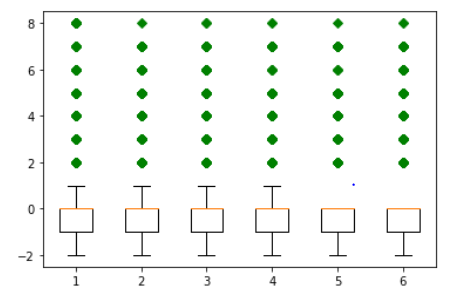
**Problem:**

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.

**Analysis:**

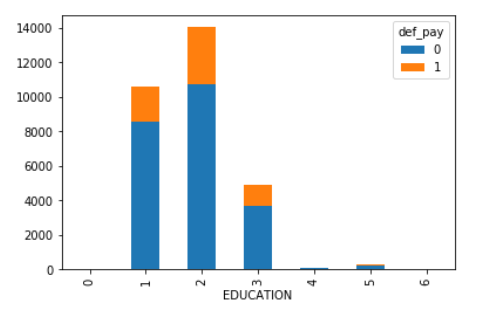
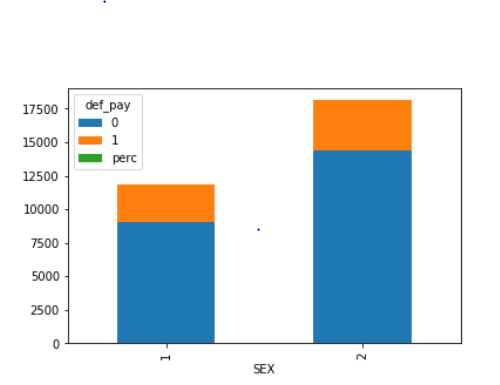
1. How do you ensure that customers can/will pay their loans?

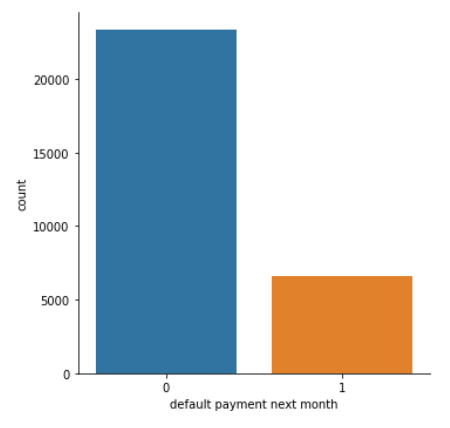
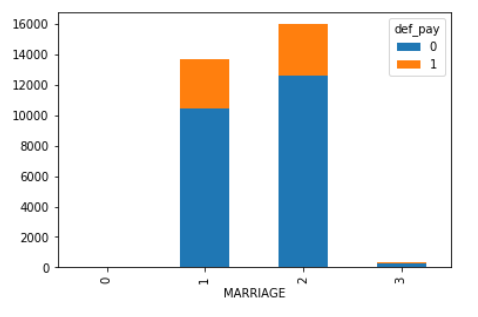
Based on the features / Attribute analysis we need to make sure the customers are not late in payments more than 2 months. Their payment amount history vs bill amount is also a strong considering factor.



1. Can we approve customers with high certainty?

Analysis of certain features such as age, sex, married/single, limit\_balance and their payment habits will help CreditOne to ensure certainty of customer not defaulting





As you progress through the task, begin thinking about how to solve the company's problem.

Here are some lessons the company learned from addressing a similar problem last year:

1. We cannot control customer spending habits

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

We must focus on the problems we can solve:

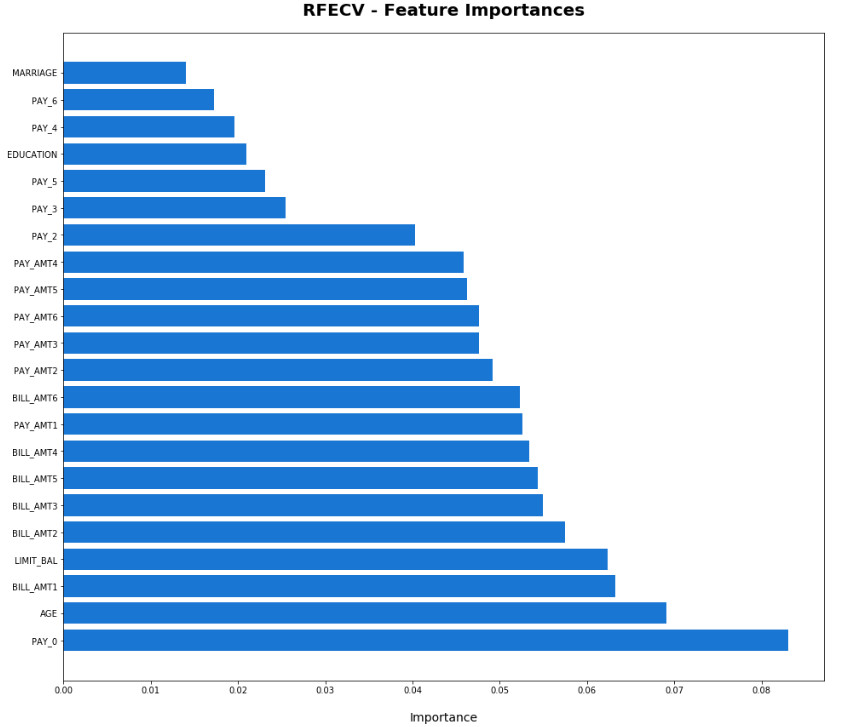
* 1. Which attributes in the data can we deem to be statistically significant to the problem at hand?

History of the past payments, marriage, Sex, Age and pay amount deemed to be statistically significant

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

The strongest predictors of default are the PAY\_X (ie the repayment status in previous months), the LIMIT\_BAL & the PAY\_AMTX (amount paid in previous months). This will help in defaulter prediction.

Recursive feature analysis to identify concrete information by feature importance.



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

Exploratory Data Analysis with visualization to derive initial data understanding then used five different analysis algorithms/methods to uncover more information.

A bank needs to evaluate if the customer can reliably repay the loan or default. The bank determines the probability of default for customers based on multidimensional attributes about the customers such as credit utilization rate, income level, age, education, etc., and macroeconomic information such as the unemployment rate, real/nominal GDP, bankruptcy filings, etc.

Based on this multi-dimensional data on customers and the economy, predictions need to be made to answer the question - which customer has higher probability of default?

* What are the main lessons you've learned from this experience?

Observations:

* Defaults have a higher proportion of Lower LIMIT\_BAL values
* NonDefaults have a higher proportion of Females (Sex=2)
* NonDefaults have a higher proportion of MoreEducated (Education=1 or 2)
* NonDefaults have a higher proportion of Singles (Marriage=2)
* NonDefaults have a higher proportion of people 30-40years
* NonDefaults have a MUCH higher proportion of zero or negative PAY\_X variables (this means that being current or ahead of payments is associated with not defaulting in the following month). This is a strong relationship as the distribution are more separated - so we expect the PAY\_X to be important!

Lessons Learned:

* Understand the default dataset thoroughly
* Identify key features and it’s correlation with the dependent variables
* Remove nonrelated data

At a high level, EDA is used to understand and summarize the contents of a dataset, usually to investigate a specific question or to prepare for more advanced modeling. EDA typically relies heavily on visualizing the data to assess patterns and identify data characteristics that the analyst would not otherwise know to look for. It also takes advantage of a number of quantitative methods to describe the data.

I would recommend performing comprehensive credit risk analysis predicting for default rate based on correlated data. Data analysis reveals, men married with lower education have higher chance of defaulting on the credit so company should focus scrutinizing those customers more.

Following has been done using Jupyter Notebook which has been added to the GitHub account:

1. Cleaning and [Pre-processing](http://scikit-learn.org/stable/modules/preprocessing.html#preprocessing) – Jupyter Notebook added in GitHub
2. [Covariance Estimation](http://scikit-learn.org/stable/auto_examples/index.html#covariance-estimation) - Jupyter Notebook added in GitHub
3. EDA - Jupyter Notebook added in GitHub
4. [Feature Engineering](http://scikit-learn.org/stable/auto_examples/index.html#feature-selection) (either PCA or RFE) and [Dimensionality Reduction](http://scikit-learn.org/stable/modules/decomposition.html#decompositions)
   1. RFE - Jupyter Notebook added in GitHub
5. [One-Hot Encoding](http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html)(if needed) – Not needed for this task as all the features are integers
6. [Classification](http://scikit-learn.org/stable/supervised_learning.html#supervised-learning) (Build three model and choose the best) - Jupyter Notebook added in GitHub
7. [Model Tuning](http://scikit-learn.org/stable/modules/grid_search.html) (Tune at least two parameters for each model you build) - Jupyter Notebook added in GitHub
8. Model Evaluation - Jupyter Notebook added in GitHub

* Even though the tool is very helpful and serves its purpose,
* one major improvement that can be added to this tool would
* be ﬁxing a time schedule to solve every problem. Because,
* the user might be having multiple activities running at the
* same time so, giving a time frame to solve every problem
* will help the user track and solve all the problems within
* stipulated time. Adding this function to the tool will make
* it m