**CREDIT ONE DATA ANALYSIS**

Credit One has come to us seeking insight and recommendations to solving the age-old question: Can trustworthiness be measured or predicted, specifically pertaining to matters of money and credit?

Recently, Credit One has reported an increase in credit defaults among their customers. We have been tasked with examining the data in an attempt to pinpoint demographic features that may predict default in advance, and thus allow Credit One to make more informed decisions of credit approval and credit limits.

Analysis of the available data is preformed keeping these main questions in mind:

1. Can we approve customers with high certainty?

> What factors available in the data can be used to decide if credit should be approved and  
 what that credit limit should be?

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

> Can customer default be predicted based on data available?

**EXPLORATORY ANALYSIS** \*While a full Exploratory Data Analysis was preformed and its report completely separately, included  
 here is a brief re-visitation of the process and results.

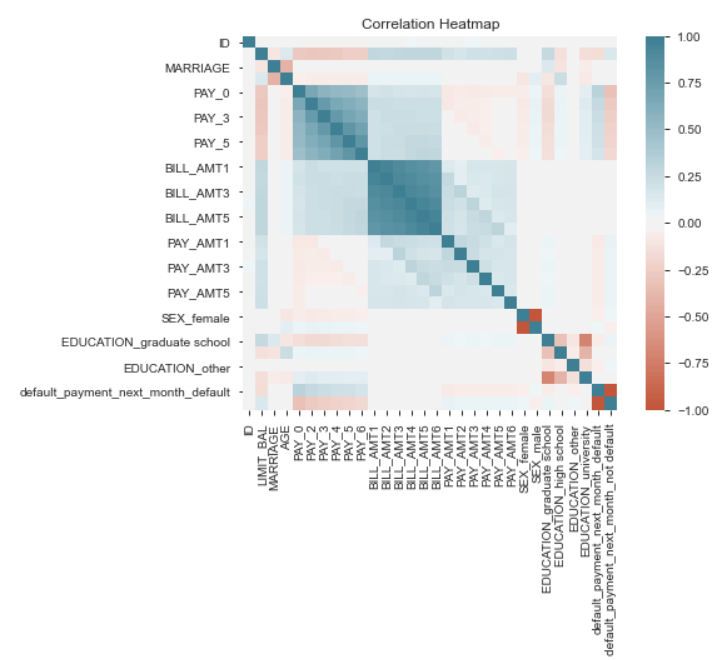
We began by collecting and mining the data captured by Credit One’s third party server for analysis. The data collected is comprised of 30,000 unique customer records categorized into multiple variables including:

* Age
* Gender
* Marital Status
* Education Level
* Credit Limit
* 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)

In order to inspect and interpret the data offered, we first prepare the information so that it is combined, cleaned, and preprocessed to produce a manageable data set. Next, we group and compare variables of the data to reach conclusions to the questions presented – most notably examining for relationships and correlations with Credit Limit and Default Status.

Of the 30,000 records available, we find that 6636 have defaulted on payment – about 22% total. Unfortunately beyond that, little useable insight was found in the exploratory stage. As described in more detail in the full Exploratory Data Analysis Report, there were simple observations made concerning demographics such as gender, age, and education, yet no high correlations were found among the data variables to determine which factors are most prevalent among customers who default. Nor were any correlations found to ensure customers can/will repay their credit loans.

The relationships observed concerning demographics were weak and forced at best, as proven by their low correlations (shown by the heatmap below), and their superficiality leads to questions of validity.

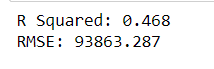
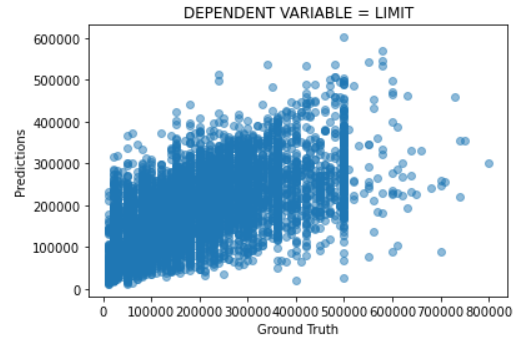


Regardless of this lack of any obvious correlation, we continue exploring the data for possible relationships and patterns in hopes of explaining and solving the issue of reported increase of credit default among Credit One customers. As the exploratory analysis provided little actionable insight, we attempt to delve deeper into the data available.

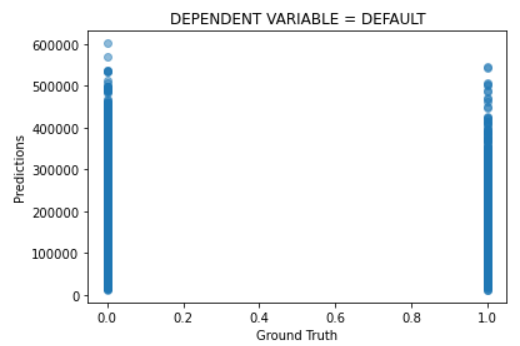
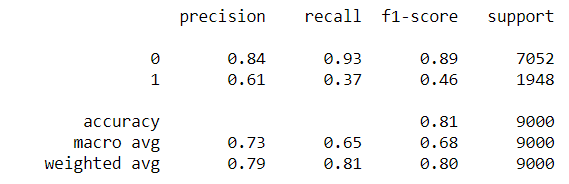
**IN DEPTH ANALYSIS**

Still using the same data, now cleaned and preprocessed as done so in our EDA, we move on to perform an in-depth analysis using machine learning for modeling and predictions. We use both regression and classification for supervised machine learning, along with train/test/split and cross validation for evaluation. The data is divided into training/testing sets with dependent variable and independent variables selected to statistically compare influence and determine patterns to create predictive models.

First, we use a regression algorithm with ‘Credit Limit’ as our dependent variable. Unfortunately, our modeling and predictions perform poorly, unable to find any reliable patterns in the data to accurately make predictions. Our best model delivers an accuracy of a mere 46% (R Squared), as well an outrageous expectation of error with the RMSE score. Simply flipping a coin would provide us with better results.



Next, we focus on ‘Default’ as our dependent variable with a classification algorithm. Again, our models leave much to be desired. While the prediction report shows quite an improvement in accuracy/quality scoring, the model provides no expressed pattern or correlation of useable predictions to provide actionable insight. There’s nothing to see here but two straight vertical lines, hardly anything of use.



**CONCLUSION**

After much exploration and analysis of the data, we revisit the questions originally posed:

1. Can we approve customers with high certainty?

In short, we can’t. At least not with the data available to us. Trustworthiness and responsibility cannot be measured or predicted with simple demographics. It is simply impossible come to a conclusion of approval with any certainty using factors as age, marital status, or education level while not capturing and considering valuable information such as income, employment status, equity, or longer-term credit history.

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

Again, we can’t. All in all, customer spending habits are unpredictable and difficult to determine accurately with any confidence ever. The most likely indicators available to us in this data that may lead to default still come well after credit has already been issued. Thus, leaving Credit One “holding the (empty) bag”.