# Predicting Loan Default In Peer-to-Peer Lending

## Udacity Capstone Proposal

François Lemoine
January 31th 2017

## Domain Background

Loan default prediction is a common problem for various financial companies. This is the case for banks (e.g Morgan Stanley, JPMorgan Chase, Goldman Sachs), credit card companies (e.g. Visa, Mastercard), microcredit banks (e.g. Grameen Bank) or FinTech companies (e.g Lending Club). The lending data for Lending Club is available and we will be using it to make predictions about loan default and whether or not we should lend to a customer. Applying machine learning to loan default predictions showcase a useful application of this branch of artificial intelligence to solve real-world and business problems. We will try to build this model with the most transparency possible as to mirror the conditions in which financial institution must disclose this process as to avoid discrimination.

## Problem Statement

If a model is able to identify credit-worthy customers that were not recognized by the traditional FICO credit score while minimizing their risk of default on the loans, this can become a lucrative niche market or micro-market. Thus pushing higher the profit margin of the financial institution. Although, the prospect of more customers seems positive, we need to be careful of not lending to people that will default, this would cause a drop in the earlier stated objective. Thus a somewhat conservative approach and rigorous evaluation metric will be kept in mind. The loan default prediction is a problem of binary classification (do we lend or not) and we will be using logistic regression and the random forest algorithm to build the model.

## Datasets and Inputs

The dataset that we will use is the one provided by Lending Club and is from the years 2007 to 2011. The dataset have 115 columns, some of the variables are not useful and many columns are empty (part of the data cleaning process will include removing these columns). We will try to make sense, remove unhelpful variables, fill in the missing values if possible and create relations between variables.

We will need to be careful with variables that are knowledge about the future as to avoid over fitting on information that is not available at the time of the lending decision (e.g. was the loan repaid on schedule).

## Solution Statement

The solution will come from thorough data cleaning, encoding the data for machine learning processing. From the initial research, the algorithms that will be most useful for this project are random forest and logistic regression. Some work will have to do with the error type (false positive, false negative, true positive and true negative) and this will help guide us to a conservative evaluation of the loan default rate. Some error balancing will have to be done with the model as to make it profitable.

## Benchmark Model

In the solution statement we talked about the need of a conservative evaluation of the default rate. We must also keep in mind that there is a strong imbalance with the target category of loan repayment in the dataset, because about 6 out of 7 loans are repaid. Meaning that we could lend money all the time (always predicting that the borrower would repay) and be right about 85.71% of the time that the loan would be repaid, but that would mean that the model would not be profitable. Say we lend $1000 at 10% interest, we would expect a return of $100 on each loan. But if we run the experiment 7 times, we would earn $600 (6 x $100) and lose $1000 (the defaulter), we are left with a $400 loss. Hardly a profitable enterprise. The benchmark needs to encompass the weight of the defaulter and the optimization between the true positive rate (good borrowers) and the false positive rate (bad borrowers). This implies that we need to ensure a viable machine learning model and predict a higher percentage of potential defaulters in order to avoid lending to them. The benchmark must beat the 85.71% average loan repayment. Although “money is left on the table”, a conservative investor would prefer a steady return on her investment than suffer the 1 in 7 loss.

## Evaluation Metrics

The best metrics to evaluate our algorithm are the recall (true positive rate), the precision and the F1 score (a measure of the relation between the two previous metrics). We will achieve this by training our model on the training dataset and than trying to predict – based on the columns value – the good customers of the testing set. We can than measure if this is practical and realistic. Otherwise, error metric balancing will be handy in recalibrating the model because we know that there is an imbalance

between the repaid loans and defaulted loan. The recall and precision are good metrics for measuring an imbalanced distribution of data points.

## Project Design

We start with data exploration and cleaning. We know that many columns are empty or not useful for our prediction. Afterwards. we'll do features engineering, encoding our variables and applying the machine learning algorithms that we learned about in the nanodegree. Finally, we'll fine tune the model to produce an accurate result on the testing data set. In the data cleaning part we will make sure to deal with missing values and try to create new features. Some visualizations might be used to describe in details correlations and to get a sense of the data. This could be considered the second part of an exploratory analysis. Afterwards, we will try to implement different classifying algorithms to make predictions. These algorithms will include: random forest and logistic regression. Finally, we will test the model on the testing data set to see the results of our work. The goal is to get a better result than the benchmark (greater than 70% of true positive and less than 7% of false positive) and fine tune the model if necessary.

## References

Lending Club (2011) *Lending club statistics*. Available at: https://www.lendingclub.com/info/download-data.action (Accessed: 30 January 2017).