Overview:
This project analyzes data from the peer-to-peer lending platform Lending Club. The goal is to predict whether a borrower will pay back their loan.

Audience and Problem:
The hypothetical client of this analysis is either Lending Club itself, or any other lending institution that does not have enough of its own data yet with which to do a similar analysis. Predicting loan defaults is of the major problems banks are facing, and achieving high accuracy has an immediate effect on their bottom line. The resulting model is used to decide whether to accept a customer's application for a new loan – and if so, what loan amount to offer and at which interest rate. In addition, the results could also be used to target marketing material at specific audiences that are most profitable.

Note that, while gaining the last bit of additional accuracy gets increasingly hard in any machine learning application, it also yields disproportionate results in a case such as this: If we are able to predict the default risk more accurately than our competitors, we will be able to identify customers which are particularly profitable: Not necessarily those whose default risk is low, because most of those will likely take their business elsewhere unless we offer them a commensurable low interest rate. Rather, the most profitable customers will be those whose default risk is low relative to how most of our competitors view them. In order to attract those customers, we will want to offer them a slightly lower interest rate than they would get elsewhere, but this discount will be small enough to still yield a greater expected return to the lender. Conversely, we will also identify a segment of customers whose risk of default is larger than implied by the models of our competitors. While we do not want to turn those customers away, we will want to offer them a commensurately higher interest rate, which will drive most of them away to competitors who priced their risk differently.

Data Source:
A subset of the data is publicly available on Lending Club's website, but in order to access all the data I had to create an account. This data set is very comprehensive, it spans about 10 years (from 2007 – shortly after the firm's founding – to the present) and contains roughly 1 million usable observations for 150 variables.

Project Outline:
The first step is to identify variables whose values were not known at the time the loan was issued (e.g., whether the parties agree to a settlement plan, etc.) and delete those columns. Next, I will identify columns that have too many missing values and thus need to be dropped. These missing values will later be imputed using techniques such as SVM and MICE. Likewise, it is not clear that keeping the earliest data – which date back to virtually the beginning of the platform, and thus might be unrepresentative. I will look for evidence whether to discard the earliest data by examining whether the default rate was initially higher due to Lending Club's credit risk models not being fine-tuned yet.

After performing a standard EDA, I will use Principal Component Analysis to reduce the number of variables without using too much information.

The main part of the analysis will then use machine learning techniques to build a predictive model. I will start with individual models such as regularized logistic regression, support vector machines, and random forests. I will both try using the principal components as the

predictors, as well as using the original predictors in combination with stronger regularization. Cross-validation will allow us to tune the hyperparameters for each model, and later to compare the performance of these different models. Ultimately, however, the goal is to then combine these models into an ensemble, which should give even higher accuracy.

Deliverables:
The whole analysis will be documented in a Jupyter Notebook. In addition, I will produce a slide deck as well as a final report.