# Stakeholder Report

Emma, Mathias, Michael, and Signe

26th September 2019

## Introduction

This project will look into some interesting data from Lending Club. They are an online peer to peer platform that match lenders with investors for small private and business loans.

This project will analyze data from their platform, to predict which lenders who will not be able to pay their full loan in due time defined here as a "bad loan". So this is esssentially a risk analysis. Lending Club also measure the expected risk for each loan and grade them which they publish in their dataset, but this will not be used for this project, as we want to make our own prediction system.

In this project we´ll use Machine Learning to analyze the dataset. By Machine Learning we´ll try to predict is the loaner will be good or bad.

## The Dataset

The data used in this project is downloaded from kaggle.com provided by Lending Club. The files contain complete loan data for all loans issued through 2007-2018. There is around 150 variables and 2.26 mio. observations. There is variables on credit scores, number of finance inquiries, address including zip codes and state among others. Since this is a very large dataset we reduce the dataset to data only from the year 2013 so there's around 135.000 observations and nine variables - we can know move on to the analysis.

## The Analysis

During the analysis we explore the dataset. From the data we see that contains 9 variables and 134.805 observations.

Down below there's a describtion of the variables there's used for further investegation.

**Purpose**

The variable purpose indicates which purpose the borrower is lending for. This variable will be used to show which kind of loan is riskier than others. The categories are in the variable is: Credit card, debt consolidation, house, car, vacation, home improvement, small business, major purchase, medical, moving, renewable energy and wedding. Here we can see that the purpose for taking a loan from Lending Club is first and foremost debt consolidation (80.000), second credit card (30.000), third home improvement (7400) and then the rest being significantly less.

### Term

The variable term informs about the length of the loan. Term is divided in two numerical values, which is either 36 months (three years) or 60 months (five years). The purpose with this variable, is to check if the length of the loan is affecting on the chance to repay back the loan.

### Loan Status

The variable loan_status are a categorical variable, with the purpose to inform about the status on the loan by the borrower.

### Loan amount

The variable loan_amnt determines the size of the loan by the borrower in USD.

### Interest rate

The variable 'int_rate' indicates which interest rate the borrower is going to pay on the loan. Lending Club has a base rate of all borrowers. In addition to this, there will be a percentage be imposed, which are rated on the borrower's adjustment for risk and volatility.

### Annual income

The variable 'annual_inc' is describing the yearly income of the borrower, calculated in USD.

### Employment length

The variable 'emp_length' indicates the duration of the borrower's employment in years. If the value is given as 'n/a', the individual is currently unemployment. The values between 1 and 9 indicates the years the been the same employment and 10+ indicates if the duration is over ten years.

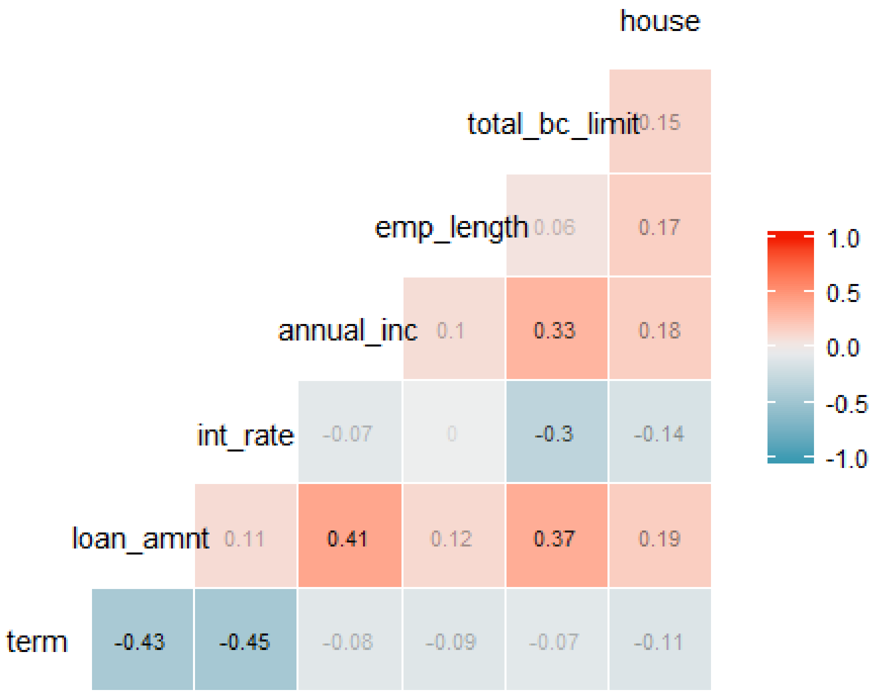
### Home ownership

The 'home_ownership' variable take the following categories: Rent, Own and Mortgage. First, we plot the different home ownership statuses. We have already filtered away any, none and other, so they will not show up.

### Total bankcard limit

The variable 'total_bc_limit' describes the amount of credit the borrower is allowed on the bank account, calculated in USD.

We made a correlation matrix that shows how the different numeric variables are correlated.



## Unsupervised machine learning

Unsupervised machine learning is about finding patterns in the data and visualizing high-dimensional data. We find homogenous subgroups within the larger group. In this case we can see that 3-5 clusters would be good, and in this case we go with 4.

First, we see that the amount of observation in each cluster range from around 18.000 to 46.000, so they are not equal in size, but they are not far from each other.

Cluster 1 and 2 have about the same loan amount of around 20.000 dollars, but cluster 2 have loaners with significantly higher income and bank limit and they also have way lower interest rate which makes great sense as they must have lower risk of being a bad loaner. Cluster 1 is the only cluster where term have a value close to 0. As terms of 36 months take the value 1 and 60 months take the value 0, this means that cluster 1 is the only one that have long loans of 60 months.

Cluster 3 and 4 have about the same loan amount of around 10.000 dollars, and are similar in most variables except that cluster 4 have about 1 year longer employment length and then that the clusters are split 100% on the house variable. This means that the primary difference is that cluster 4 have houses, but cluster 3 does not.

## Supervised Machine Learning

Supervised machine learning is about using machine learning to run different input to output. Here we wanted to predict whether the loaners of Lending Club would categorize as Good or Bad loaners using eight predictors. We ran three different models, a simple logistic model, a decision tree and a random forest. We used 75 percent of our data to train the algorithms and 25 percent to test how they performed. All our three models performed with a Accuracy of 84 percent, which is a high result but be in mind that predicting all loarners as Good gives the same Accurary. Another factor you could focus on is Specificity or True Negatives. Here we are showing the decision tree model.

```
## Confusion Matrix and Statistics
## 
##           Reference
## Prediction   Bad  Good
##       Bad     98   129
##       Good  5160 28313
## 
##                Accuracy : 0.8431
##                  95% CI : (0.8391, 0.8469)
##     No Information Rate : 0.844
##     P-Value [Acc > NIR] : 0.6823
## 
##                   Kappa : 0.0231
## 
##  Mcnemar's Test P-Value : <2e-16
## 
##             Sensitivity : 0.9954
##             Specificity : 0.0186
##          Pos Pred Value : 0.8458
##          Neg Pred Value : 0.4317
##              Prevalence : 0.8439
##          Detection Rate : 0.8401
##    Detection Prevalence : 0.9933
##       Balanced Accuracy : 0.5071
## 
##        'Positive' Class : Good
## 
```

## Conclusion

The models perform well but no better than prediting all loaners as Good, which also gives an Accuracy of 84 percent. So even though we have tested some advanced models, they do not perform much better than prediting all loaners as Good.