Lending Club Loan Data Analysis – Milestone Report

Lending Club is no new name in the lending industry. It has shifted the lending paradigm with the use of technology devising efficient, convenient and the smartest way about investing and borrowing.

With Lending Club, an investor can invest in a portfolio of loans. But how will those loan perform, and which loans to pick from? In order to say anything meaningful about what loans to choose, we must first estimate how loans will do over time. What percentage of loans will default? What percentage of loans will get paid off in full?

Through this project we are interested in making a guess at the probability of default, which is intended to support company's decision in approving/not approving the loans given the features that are collected from the lenders through loan application.

What important fields and information does the data set have?

Lending Club dataset came with useful features related to loan as well as applicant. Loan itself has some intrinsic characteristics like loan_status (paid, defaulted, late, in-grace categories) loan term (36/60 months), grade (A...F). For each applicant, company collects information regarding his/her employment length, annual_inc, home_ownership_status etc. On the basis of the information received for each of the mentioned variables, loan approval criteria is determined. These fields will be very useful in our analysis.

What are its limitations i.e. what are some questions that you cannot answer with this data set?

This dataset is limited to year 2014-2015. Although mentioned in the data dictionary, yet there are a few fields like FICO scores is missing from the data file which could have served as one of the most important features in predicting loan defaults.
Also, this dataset does not have demographic and geographic details that could have been a nice addition to our analysis.

What kind of cleaning and wrangling did you need to do?

Original data came with 656732 rows and 122 columns. Eliminating features was the first task in hand. There were fields which were not making intuitive sense for learning algorithms like "ID", "Member ID", "URL", "month the last payment was received" etc. Thus we removed such fields from the dataset. We also dropped rows which were populated with "nan" values. And lastly we removed columns having missing values for greater than 50% of the rows.

To label the dataset, we classified any loan that “defaulted” or were “charged off” as negative examples (0)/ “bad loans”, while we classified any loan that was “fully paid” as positive examples (1)/”good loans”. We eliminated loans which were “late” and “in-grace” period or current from the dataset as due to their reversible state.

Wrangling was required in these areas.

Null Removal
Dropping irrelevant fields.
Assigning numeric values to categorical features

We cleaned each of the input variables for nulls, N/A, any non-printable characters. Summary of each one can be found below.

Cleaning “loan_status”

Action	loan_status_category
1	‘Fully Paid'
0	'Charged Off' and Default
Remove	'Current’,‘In-GracePeriod’, ‘Late(16-30days)','Late (31-120 days)' or blank

Cleaning “verification status”

Action	Verification_status
1	'Verified', 'Source Verified'
0	'Not Verified'
Remove	All other keywords or blank

Cleaning “emp_length”

Action	Verification_status
< , +, Years	Replace with ‘ ‘
n/a	'Not Verified'
Remove	Nothing

Cleaning “home_ownership”

Action	loan_status_category
1	‘OWN
2	'MORTGAGE’
3	‘RENT’
REMOVE	‘OTHER’

Cleaning “term”

We will assign numeric values to the term values:

Action	term_clean
1	“36 months”
0	“60 months”

Are there other datasets you can find, use and combine with, to answer the questions that matter?

We are planning to take 2 years’ worth of data for our analysis. We are going ahead with the loan data for the year 2014 and 2015 and combining them together in one data frame and call it loandata.

Any preliminary exploration you’ve performed and your initial findings.

- There are 256,714 “Fully paid” rows. & “75,603” bad loans. Hence, 29 % of the loans turns out to be bad loans in our dataset.
- 21% (16202/75603) of the bad loans were not verified in our dataset.
- By analyzing emp_length, we found out, though intuitive enough, yet it was okay to assume that loan payments would continue as long as individual continues to work however, there are significant number of defaults found in the range (1-3 years) of emp_length. This feature seems to have some predictive strength which could be investigated in the modelling section.
- Exploring “grade”, we found out that many low grade loans have been allotted against 36 months. Also, we see there are good number riskier loans allotted against 60 months term which got defaulted.

  18 % of the 36 months term loans have turned bad vs 36 % of loan with 60 months term. We will check its predictive power in the modeling section.