Capstone Project 1: Milestone Report

Questions:

Define the problem:

The problem is predicting the chances whether a given proposal will be a good loan(paid off) or a bad loan(charged off/write off)

Identify your client:

The client is any lending club investor.

Describe your data set, and how you cleaned/wrangled it.

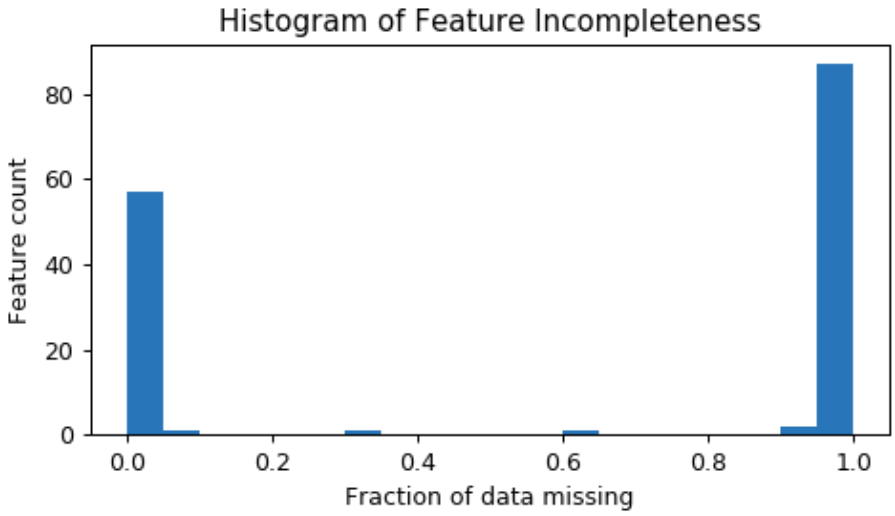
- The dataset was directly obtained from the url www.lendingclub.com . Initially the shape of the data showed only 1 column which was due to a single heading present in the cvs file. After the deletion of that row the data set represented approx 43000 rows and 150 columns
- Several columns were renamed to make them readable
- The '%' sign from the interest rate column was removed and the column was converted into float data type for the ease of computaion/visualisation.
- The employee length had two rows of object data type due to boolean / arithmatic operators(<,+) present in it. It was modified for the sake of calculations.
- In the complete dataset, 15% of the loans were classified as bad loans which gives us an imbalanced dataset.This problem will be dealt on later stage.
- "Charged Off", "Default", "Does not meet the credit policy" Charged Off", "In Grace Period","Late (16-30 days)", "Late (31-120 days)" were classified as **Bad Loans** and all others as **Good Loans** which was done under the column loan_condition . This column was our target variable.
- Logarithmic transformation of annual income was done due to large values.

• Their were 30,000 unique employee titles/designation which would be of minimal usage hence, this column was dropped.

• column “id” was made the index

• Columns having more than 30% of the values missing/Nan were dropped which resulted in 58 columns left for model training.

• The following histogram shows the incompleteness:



From the above histogram, we see there's a large gap between features missing "some" data (<20%) and those missing "lots" of data (>40%). Because it's generally very difficult to accurately impute data with more than 30% missing values, we drop such columns.

List other potential data sets you could use.

There were many other latest quarterly(Q1,Q2,Q3,Q4) dataset present on the lending club website but as the number of columns were very less hence the large dataset was used.

Also , several other cleaned datasets were also present on Kaggle but using the dataset directly from the website was much more of a realistic approach.

Explain your initial findings.

- The initial findings seem to suggest a positive correlation between several features of the applicant with the loan status(target variable) which can be utilised for the model building.
- Also, the verification status did not have any significant impact on the loan target variable
- This shows that 41.3% of the bad loans were not verified.
- The fico_score mean of the good loans is more than 700 and most of the values of fico_score lies near 700 score hence the fico_score for good loans should be closer to 700.
- The majority of loans is either graded as B or C — together these correspond to more than 50% of the loan population. While there is a considerable amount of A graded or “prime” loans (~17%), there is a small amount of E graded, or “uncollectible” loans (~0,06%). Which is a good sign for Lending Club.
- Number of bad loans were less in the category A jobs and it had more number of good loans as compared to the other category.