Dataset:

	Parameter	Description
0	loan_amnt	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value
1	term	The number of payments on the loan. Values are in months and can be either 36 or 60.
2	int_rate	Interest Rate on the loan
3	installment	The monthly payment owed by the borrower if the loan originates.
4	grade	LC assigned loan grade
5	sub_grade	LC assigned loan subgrade
6	emp_title	The job title supplied by the Borrower when applying for the loan.
7	emp_length	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
8	home_ownership	The home ownership status provided by the borrower during registration or obtained from the credit report. Our values are: RENT, OWN, MORTGAGE, OTHER
9	annual_inc	The self-reported annual income provided by the borrower during registration.
10	verification_status	Indicates if income was verified by LC, not verified, or if the income source was verified
11	issue_d	The month which the loan was funded
12	loan_status	Current status of the loan
13	purpose	A category provided by the borrower for the loan request.
14	title	The loan title provided by the borrower
15	zip_code	The first 3 numbers of the zip code provided by the borrower in the loan application.
16	addr_state	The state provided by the borrower in the loan application
17	dti	A ratio calculated using the borrower’s total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower’s self-reported monthly income.
18	earliest_cr_line	The month the borrower's earliest reported credit line was opened
19	open_acc	The number of open credit lines in the borrower's credit file.
20	pub_rec	Number of derogatory public records
21	revol_bal	Total credit revolving balance

22	revol_util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
23	total_acc	The total number of credit lines currently in the borrower's credit file
24	initial_list_status	The initial listing status of the loan. Possible values are – W, F
25	application_type	Indicates whether the loan is an individual application or a joint application with two co-borrowers
26	mort_acc	Number of mortgage accounts.
27	pub_rec_bankruptcies	Number of public record bankruptcies

Source: Kaggle https://www.kaggle.com/jeandedieunyandwi/lending-club-dataset

It is a labeled data of **4 Million** entries and a total of **28** different parameters, the parameters are a mixture of handwritten data (Purpose of loan / Job position), some numbers (Loan Amount / Int Rate) and categorical variables (House ownership / Verification Status / Loan Status). As shown in the image there are many missing values (81589 Cells) in total, which is almost 0.8% of the total. Some columns have high cardinality values (around 48000 distinct values for title) and many of the columns are highly skewed. Since many of the entries are human entries so we will need a lot of preprocessing to convert entries to numbers. Although this won't require NLP because the entries are in digital formats, we will need to convert it to numeric values since some values are descriptive. Also we are planning to drop some of the features of data, based on the correlation values which we will get by plotting the heat map for our data.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 396030 entries, 0 to 396029
Data columns (total 27 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   loan_amnt             396030 non-null  float64
 1   term                  396030 non-null  object
 2   int_rate              396030 non-null  float64
 3   installment           396030 non-null  float64
 4   grade                 396030 non-null  object
 5   sub_grade             396030 non-null  object
 6   emp_title             373103 non-null  object
 7   emp_length            377729 non-null  object
 8   home_ownership        396030 non-null  object
 9   annual_inc            396030 non-null  float64
 10  verification_status   396030 non-null  object
 11  issue_d               396030 non-null  object
 12  loan_status           396030 non-null  object
 13  purpose               396030 non-null  object
 14  title                 394275 non-null  object
 15  dti                   396030 non-null  float64
 16  earliest_cr_line      396030 non-null  object
 17  open_acc              396030 non-null  float64
 18  pub_rec               396030 non-null  float64
 19  revol_bal             396030 non-null  float64
 20  revol_util            395754 non-null  float64
 21  total_acc             396030 non-null  float64
 22  initial_list_status   396030 non-null  object
 23  application_type      396030 non-null  object
 24  mort_acc              358235 non-null  float64
 25  pub_rec_bankruptcies  395495 non-null  float64
 26  address               396030 non-null  object
dtypes: float64(12), object(15)
memory usage: 81.6+ MB
```

Algorithms

We are planning to implement this model using three main algorithms:

Out of which "Artificial Neural Network" and "XG Boost" will allow a lot of hyper parameter tuning to increase the efficiency which can prove to be useful since we have a lot of parameters out of which some are not that useful and can be ignored and the third algorithm will be Random forest algorithm, we are planning on using this algorithm since we have a large number of samples with a lot of missing entries and Random Forest proves to be the

best choice in this scenario. Although we need to discuss this with the Prof. or TA once before finalizing.

Measuring Success:

Success will be based on the accuracy of the model on the testing data and the precision, recall, F1-Score and Support values on the testing data. We are also planning to check the precision on random samples.

Research Questions:

Should a particular customer be given a loan amount he is requesting for, if yes, what should be the category of the output.(Fully Paid, Charged Off)

Division of Work among team Members:

All the work will be done by all three of us There is a lot of work in preprocessing the data so we are planning that Shilpa and Richa will do the preprocessing for the first half where we will be identifying the the missing values and filling the cells with the most appropriate values generate data profile etc, during which Rahul will be handling the report work for project proposal and milestone 1.

In the later part of preprocessing where we need to convert categorical entries to integer/floating point values, all the work will be done by Richa and Rahul, during which major part of report for milestone 3 will be handled by Shilpa although Rahul and Richa will also contribute towards the documentation.

In the final part we have decided 3 algorithms, out of which Random forest will be implemented by Richa, Neural Networks will be implemented by Shilpa and XG Boost will be implemented by Rahul. Major part of paperwork will be handled by Richa although Rahul and Shilpa will also do the documentation work. And the final PPT will be made by Rahul

What is already Done:

We have already collected the dataset and started working on it, also we have discussed what algorithms we will need to implement based on the biased nature of our dataset, also we have decided what all preprocessing will be needed.

Challenges:

Main challenges might be to convert the skewed columns to normal ones, and also the decision of parameter, since we have a lot of parameters to select from.