Project Proposal

Lending club is the world's largest platform peer-to-peer lending platform where borrowers and investors together, transform the way people access credit. The loan statistics for all the historical loans are provided as 'csv' files on the website. The objective of the project is to apply machine learning techniques to predict the loans that are at a higher risk of defaulting. The data has information about the borrowers for the years 2015 - 2017. Of this 70% of the data is marked as train and 30% test to evaluate the model prediction. Since the target variable- loan status is a binary outcome variable with levels "fully charged" or "charged off ", we plan to use classification models to predict the bad loans. The project is divided into 5 stages (1) Data collection (2) Data cleaning and Imputation (3) Exploratory data analysis (4) Model building and (5) Model evaluation.

Since this is a binary classification problem, we intend to use models like Logistic Regression, Linear Discriminant Analysis model, Naïve Bayes and Random-forest Classifier. As the data set has a lot of features we plan to use dimensionality reduction technique like Principle Component Analysis. Once we fit the model to the training dataset, we predict the outcome on test data and evaluate using AUC, ROC, Confusion matrix (misclassification probability), etc., Since the dataset has various records of borrowers, we could also explore clustering techniques like k-means clustering to group similar records and get more insights on the dataset.

Some more project Details are mentioned below:

Title: Lending club loan default prediction.

Name: Pradeep Krishnan, Abhishek Anand

Details about the database: The data is obtained from lending club website (https://www.lendingclub.com/)

Problem: Predicting the borrowers who will default on loan from the historical data and the default risk probability

Type: Classification and Clustering

Input: Loan amount, Term, Interest rate, Employment length, Home ownership, Annual Income, Total credit line, Installment, etc.,

Output: Loan Status

Reference: https://www.lendingclub.com/info/download-data.action

A sample of records of the dataset and correlation and a few initial visualizations are given in the next page.

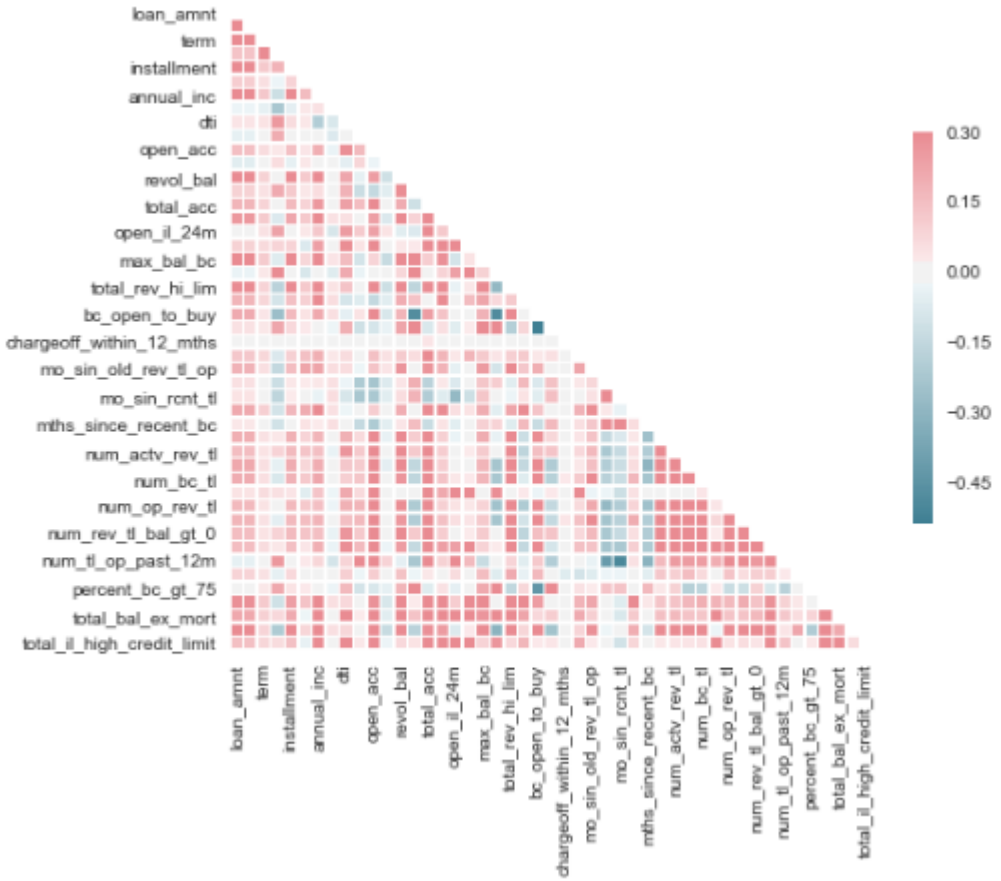
Description of a few of the variables:

	name	description
0	acc_now_delinq	The number of accounts on which the borrower i...
1	acc_open_past_24mths	Number of trades opened in past 24 months.
2	addr_state	The state provided by the borrower in the loan...
3	all_util	Balance to credit limit on all trades
4	annual_inc	The self-reported annual income provided by th...
5	annual_inc_joint	The combined self-reported annual income provi...
6	application_type	Indicates whether the loan is an individual ap...
7	avg_cur_bal	Average current balance of all accounts
8	bc_open_to_buy	Total open to buy on revolving bankcards.
9	bc_util	Ratio of total current balance to high credit/...
10	chargeoff_within_12_mths	Number of charge-offs within 12 months
11	collection_recovery_fee	post charge off collection fee
12	collections_12_mths_ex_med	Number of collections in 12 months excluding m...
13	delinq_2yrs	The number of 30+ days past-due incidences of ...
14	delinq_amnt	The past-due amount owed for the accounts on w...
15	desc	Loan description provided by the borrower
16	dti	A ratio calculated using the borrower's total ...
17	dti_joint	A ratio calculated using the co-borrowers' tot...
18	earliest_cr_line	The month the borrower's earliest reported cre...
19	emp_length	Employment length in years. Possible values ar...
20	emp_title	The job title supplied by the Borrower when ap...
21	fico_range_high	The upper boundary range the borrower's FICO a...
22	fico_range_low	The lower boundary range the borrower's FICO a...
23	funded_amnt	The total amount committed to that loan at tha...
24	funded_amnt_inv	The total amount committed by investors for th...
25	grade	LC assigned loan grade
26	home_ownership	The home ownership status provided by the borr...

Sample of some records and few of the columns are given below:

	term	grade	sub_grade	home_ownership	verification_status	issue_d	loan_status	pymnt_plan	purpose
0	36 months	B	B2	RENT	Verified	Dec-2011	Fully Paid	n	credit_card
1	60 months	C	C4	RENT	Source Verified	Dec-2011	Charged Off	n	car
2	36 months	C	C5	RENT	Not Verified	Dec-2011	Fully Paid	n	small_busines
3	36 months	C	C1	RENT	Source Verified	Dec-2011	Fully Paid	n	other
4	60 months	B	B5	RENT	Source Verified	Dec-2011	Fully Paid	n	other
5	36 months	A	A4	RENT	Source Verified	Dec-2011	Fully Paid	n	wedding
6	60 months	C	C5	RENT	Not Verified	Dec-2011	Fully Paid	n	debt_consolid
7	36 months	E	E1	RENT	Source Verified	Dec-2011	Fully Paid	n	car
8	60 months	F	F2	OWN	Source Verified	Dec-2011	Charged Off	n	small_busines

Correlation plot of some of the features:



Sample visualization of number of loans based on Grade (quality) of loan and purpose: