GA DATA SCIENCE

Lending Club Loan Data

Ann Pattara

Nov, 2018

CONTENT

1 Problem Statement

2 Metrics & Assumptions

3 Data Description

4 Exploratory Analysis

5 Data Preprocessing

6 Model Analysis

7 Conclusion & Next Steps

PROBLEM STATEMENT

Background

- A major problem that P2P lending platforms face today is the issue of delinquency
- Although P2P is a profitable business, the odds of the occasional delinquency makes this business highly risky and potentially unattractive

Objective

- To analyze factors that predict customers delinquency and develop a predictive model to identify customers with a potential tendency of delinquency using historical data

METRICS & ASSUMPTIONS

Assumptions

1. The dataset is accurate
2. All the datapoints are available and if not, appropriate measures can be taken to include the missing values
3. The dataset is comprehensive and avoid any bias and reflect the population as much as possible

Metrics

1. Accuracy greater than 85%
2. Precision
3. Recall

DATA DESCRIPTION

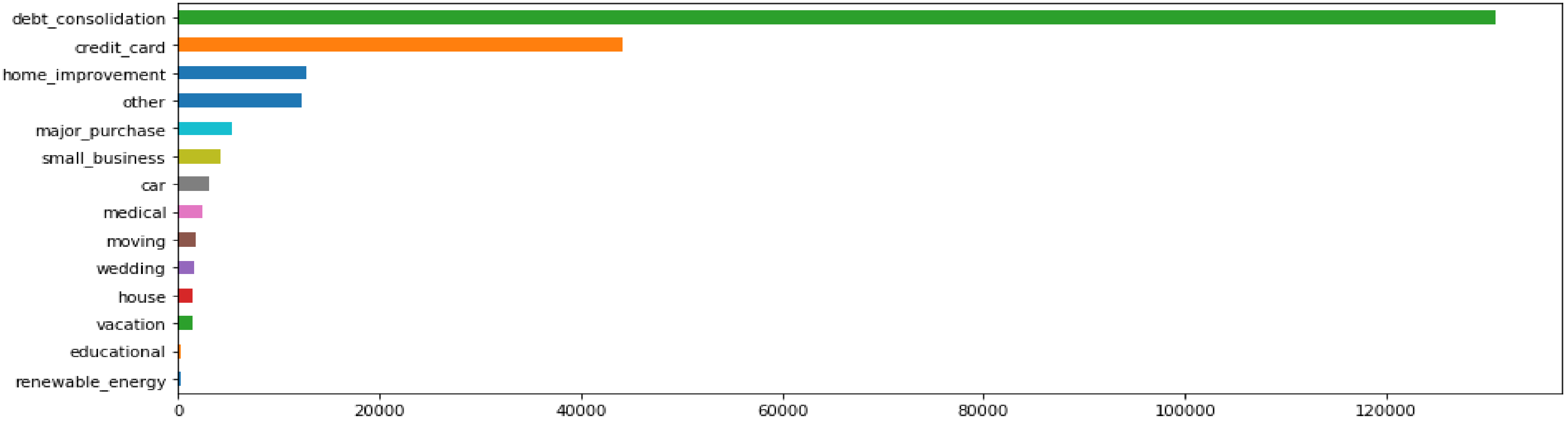
Lending Club loan dataset contains 890 thousand observations and 75 variables from year 2007-2015

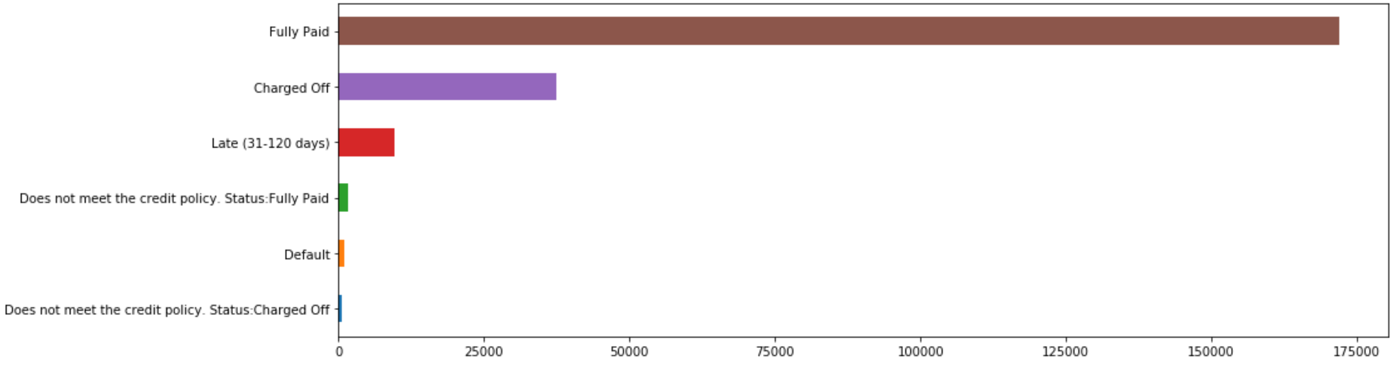
# id	# loan_amnt amount of money requested by the borrower	A term	# int_rate The interest rate on the loan	# installment	A grade
1077501	5000.0	36 months	10.65	162.87	B
1077430	2500.0	60 months	15.27	59.83	C

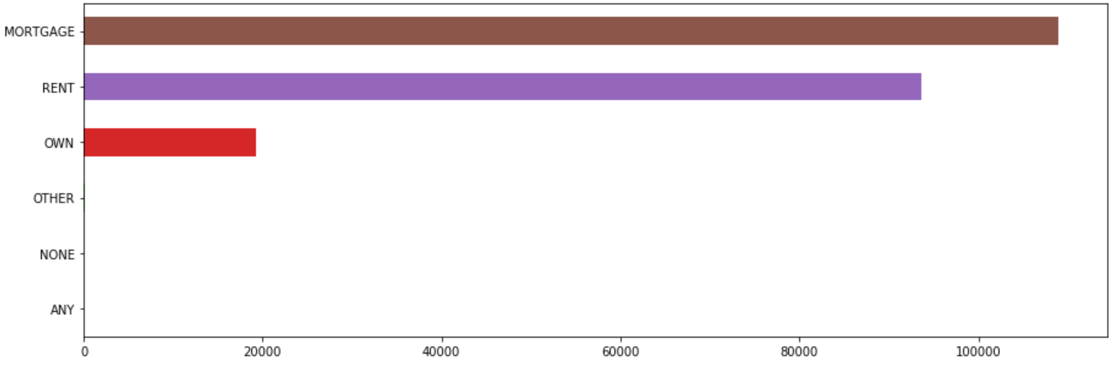
A emp_length	A home_ownership	# annual_inc	A verification_status	issue_d	A loan_status
10+ years	RENT	24000.0	Verified	Dec-2011	Fully Paid
< 1 year	RENT	30000.0	Source Verified	Dec-2011	Charged Off

A loan_status	A pymnt_plan	A purpose	A addr_state	A application_type
Fully Paid	n	credit_card	AZ	INDIVIDUAL
Charged Off	n	car	GA	INDIVIDUAL

EXPLORATORY ANALYSIS







DATA PREPROCESSING

The main objectives of the data preprocessing phase is Dimensionality reduction and Data cleansing

1. Created a subset for only those loans issued after Dec-2009 (222076 observations)
2. Converted date fields into numeric years
3. Excluded those columns that had more than 90% NA observations
4. For others, replaced NA with Zero or other appropriate values
5. Encoded categorical values into numeric values
6. Identified Target Variable as Loan Status. Changed to Loan Class – Good(173404) and Bad(48672) loan) - 80:20 ratio

Replace Date Observations with Numbers

Exclude Columns with >60% NA Observations

Replace NA Values Appropriately

Exclude Low Variance & High Correlation Columns

Use Business Sense for Any Other Exclusions

MODEL ANALYSIS – Decision Tree

- Each model is designed for a distinct loan purpose (eg: Credit Card, Home Improvement etc)
- Created training and test datasets with an initial split of 80-20
- Fit Decision Tree model to cleansed data with 19 variables on training and test sets
- Analyze variable importance, accuracy, precision, recall

Measure	Value
Accuracy	70.3%
Precision	54.2%
Recall	54.5%

Variable Name	Importance
loan_amt	0.07563348
int_rate	0.06811373
grade	0.05345723
emp_length	0.04583799
home_ownership	0.01292297
annual_inc	0.08381747
term	0.01103021
dti	0.01103021
iniq_last_6mths	0.10459646
mnths_since_last_delinq	0.02309473
mths_since_last_record	0.03727288
open_acc	0.01721132
revol_bal	0.05505993
revol_util	0.07645808
total_acc	0.09223143
mths_since_last_major_derog	0.06763209
tot_cur_bal	0.01810915
total_rev_hi_lim	0.08048443
ver_stat	0.06830906

MODEL ANALYSIS – Random Forest

- Each model is designed for a distinct loan purpose (eg: Credit Card, Home Improvement etc)
- Created training and test datasets with an initial split of 80-20
- Fit Random Forest model with number of trees = 500. Number of threads = 2
- Analyzed variable importance, accuracy, precision, recall

Measure	Value
Accuracy	80.5%
Precision	68.3%
Recall	52.6%

Variable Name	Importance
term	0.07331508
dti	0.08486268
home_ownership	0.03303834
open_acc	0.04577928
tot_cur_bal	0.01679431
mnths_since_last_delinq	0.08442444
mths_since_last_record	0.01523126
emp_length	0.09364918
grade	0.02661408
revol_bal	0.03913852
mths_since_last_major_derog	0.01757569
int_rate	0.05563736
ver_stat	0.08284123
loan_amt	0.08479649
revol_util	0.06721789
total_rev_hi_lim	0.02060364
annual_inc	0.07765079
total_acc	0.07082148
iniq_last_6mths	0.01000824

CONCLUSION & NEXT STEPS

1. The Decision Tree Model is quick and easy to implement, but delivers an accuracy of only 70% with poor precision and recall values
2. The Random Forest Model performs better than Decision Tree but has an accuracy of only 80% and fairly poor precision and recall values
3. The main cause of poor model performance is the highly skewed dataset (80-20)

As next steps,

1. Use SMOTE, an external library that helps under sample/ over sample the underlying data can help balance the data skewedness and improve performance
2. Try other machine learning models like logistic regression, support vector machines, XG Boost etc to improve model performance