THE UNIVERSITY OF SYDNEY

School of Information Technologies
Faculty of Engineering & IT

Classifying “Bad” Loans Using Lending Club Loan Data

Unit of Study: INFO3406, Introduction to Data Analytics

Assignment name: Project Stage 1

Tutorial time: Monday, 11am

Tutor name: Dr. Ali Anaissi

DECLARATION

I declare that I have read and understood the *University of Sydney Academic Dishonesty and Plagiarism in Coursework Policy*, and except where specifically acknowledged, the work contained in this assignment/project is my own work, and has not been copied from other sources or been previously submitted for award or assessment.

I understand that failure to comply with the the *Academic Dishonesty and Plagiarism in Coursework Policy*, can lead to severe penalties as outlined under Chapter 8 of the *University of Sydney By-Law 1999* (as amended). These penalties may be imposed in cases where any significant portion of my submitted work has been copied without proper acknowledgement from other sources, including published works, the internet, existing programs, the work of other students, or work previously submitted for other awards or assessments.

I realise that I may be asked to identify those portions of the work contributed by me and required to demonstrate my knowledge of the relevant material by answering oral questions or by undertaking supplementary work, either written or in the laboratory, in order to arrive at the final assessment mark.

Student ID: 480490915

Student name: Ryan Dunham

Signed: Ryan Dunham **Date:** 05/09/18

SIT Building, J12
The University of Sydney
NSW 2006 Australia

T +61 2 9351 3423
F +61 2 9351 3838
E sit.info@sydney.edu.au **sydney.edu.au/it**

ABN 15 211 513 464 CRICOS 00026A

THE UNIVERSITY OF SYDNEY

Problem Definition

I will be exploring the Lending Club loan database for my data analytics project, which I acquired from Kaggle. This dataset includes 887,379 entries, with 74 different attributes, and details each individual loan created through Lending Club.

Lending Club is a financial technology company that specializes in peer-to-peer lending; they connect investors and borrowers to create unsecured personal loans of up to $40,000. These unsecured loans are a type of alternative investment for investors, as they make interest on the loans while borrowers are supplied with their desired capital. Lending Club eventually buys the issued loans, and makes payments to the investors minus service fees, while it collects payments from the borrower.

A major problem that Lending Club faces is the risk of default, or the risk that a borrower will not pay back a portion or the entire loan amount. While Lending Club heavily screens its potential borrowers, declining approximately 90% of loan applications, the risk of default still exists. Not only does loan default cause problems for Lending club, but other delinquencies do as well. Loans that are late or charged off are problematic, as they may default in the future and require the company to attempt to collect, which costs Lending Club time, resources, and money.

Problem Approach

In the dataset there are 8 different loan statuses:

Loan Status	Description
Issued	New loan that has passed review, received funding and has been issued
Current	Loan is up to date on all outstanding payments
Fully paid	A loan that has been repaid in full including all principal and interest payments.
In grace period	The loan is past due, but within the 15-day grace period
Late (16-30 days)	Loan has not been current, or up to date on outstanding payments for 16 to 30 days
Late (31-120 days)	Loan has not been current, or up to date on outstanding payments for 30 to 120 days
Default	Loan has not been current, or up to date on outstanding payments for 121 days or more
Charged off	Loan has been deemed uncollectible, no longer reasonable expectation of further payments

In order to better understand Lending Club customers and their risk of default, I plan to perform exploratory data analysis in order to create a classifier that will predict whether a loan will be “bad”, based on the various attributes in the dataset, and ultimately test the accuracy of this model. A “bad” loan, is one in which the loan status is default, charged off, late (16-30) days, late (31-120) days or in grace period. A “good loan” is one in which the loan status is fully paid, issued, or current. My target variable in this project is a “bad” loan.

While there are several attributes in this dataset, I will focus my analysis on just a handful of features. These features are summarized by the chart below:

THE UNIVERSITY OF SYDNEY

Features	Description
loan_amnt	The loan amount requested by the borrower
term	The length of the loan term, can be 36 months or 60 months
addr_state	The state of residence of the borrower
int_rate	Interest rate
grade	Grade assigned to loan by Lending Club (A, B, C, D, E, F, G)
purpose	The purpose of the loan provided by borrower
emp_length	Length of employment of borrower in years
home_ownership	Home ownership status. (Rent, own, mortgage, other)
annual_inc	Annual income of borrower
dti	Debt to Income ratio
installment	Monthly payment of the borrower
delinq_2yrs	The number of 30+ day past due delinquencies of the borrower in the past 2 years
inq_last_6mths	Number of inquiries in the past 6 months
open_acc	The number of open credit lines in borrowers file
pub_rec	Number of derogatory records
revol_bal	Total credit revolving balance, portion of credit card balance that goes unpaid at the end of a billing cycle
revol_util	Revolving line utilization rate, amount of credit borrower is using relative to total revolving credit
total_acc	Total number of credit lines currently in borrowers file

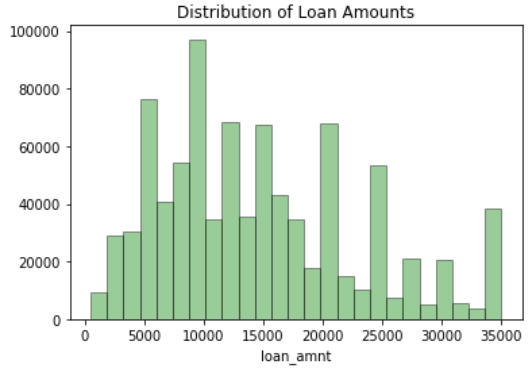
Although I will explore each of these variables in order to create a classifier, there are a few attributes that I will analyze more thoroughly. I will focus my exploratory analysis and visualization primarily on the following features in order to better understand Lending Club users: int_rate, dti, addr_state, purpose, loan_amnt and grade. Analyzing these attributes will provide more in depth information into the demographics of customers, their financial strength, and what they use the loan for.

Data Exploration

This dataset was initially very large and quite messy. At first there were 74 different attributes, many of which were mostly incomplete or redundant. I began my data transformation process by reading the data (csv file) into a pandas dataframe, using the loan id as an index. I utilized the python libraries pandas, numpy, seaborn and matplotlib in this section of my project. Upon looking at the data, I noticed that many of the features (columns) in the dataframe were mostly null. I then ran tests to determine how much of the column was null, and removed features that had a majority null values (>50%). Next, I analyzed how many unique values each column had, and

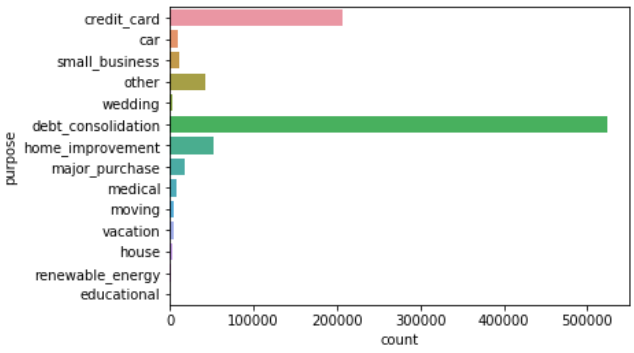
focused this analysis on categorical features. I decided to remove redundant categorical features that had too many unique values. For instance, I removed the column url, because it had 887,379 unique values which were of no use to me. After removing these extraneous features, my dataframe now consisted of 43 features. My cleaning process then shifted to null values, in which I filled with either 0 for numerical values, or N/A for categorical values. I decided to clean the emp_length column so that the number of years was displayed without the word "years" and designated 10+ years experience as 10. The final step in my data transformation process was creating a target variable column, in which a loan was labeled either "good" or "bad" based on the criteria described before.

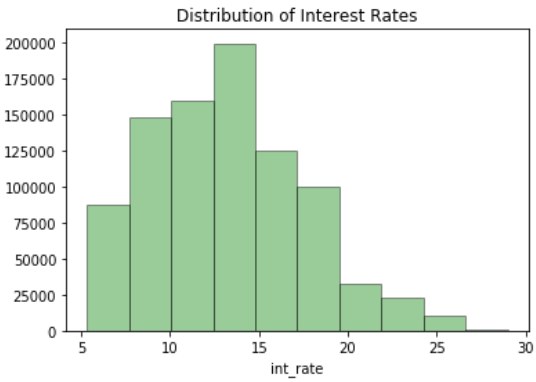
Through my initial exploratory analysis of the various attributes, I noticed some interesting information about the various distributions. I found it extremely interesting that the mode for the purpose attribute was for debt consolidation. Additionally, the next most common purpose was for credit cards. A majority of loans created through Lending club are for paying back other debts, an interesting fact. I am curious to find out more about this, and whether the rate of "bad loans is higher for debt consolidation purposes. The dti or debt to income ratio measures an individual's ability to manage monthly payments and pay debts. When attempting to plot this data, I noticed something strange. The data was barely visible due to extremely high outliers. While the median dti is 17.65 and the standard deviation is 17.19 , there were several data points over 100 and two over 1000. These could prove problematic, as they are several standard deviations above the mean. Further exploration will be needed to determine whether these data points are anomalies or errors. Most of the other features were distributed rather normally.



Descriptive Statistics (Loan Amounts in $):

count 887379.000000
mean 14755.264605
std 8435.455601
min 500.000000
25% 8000.000000
50% 13000.000000
75% 20000.000000
max 35000.000000
Name: loan_amnt, dtype: float64





References

https://www.kaggle.com/wendykan/lending-club-loan-data
https://www.lendingclub.com/
https://help.bitbond.com/article/20-the-10-loan-status-variants-explained