Tayfun Ayazma, Ph.D.

tayfunayazma@gmail.com

940-595-8413

LENDINGCLUB LOAN ANALYSIS

February 26, 2020

Table of Contents

About the LendingClub: ***2***

Problem Statement: ***3***

About the Dataset: ***4***

Data Preprocessing ***5***

Exploratory Data Analysis (EDA) ***7***

Feature Selection ***16***

Model Building / Machine Learning ***19***

Imbalanced Data: **19**

Balanced Data: **20**

Business Recommendations ***21***

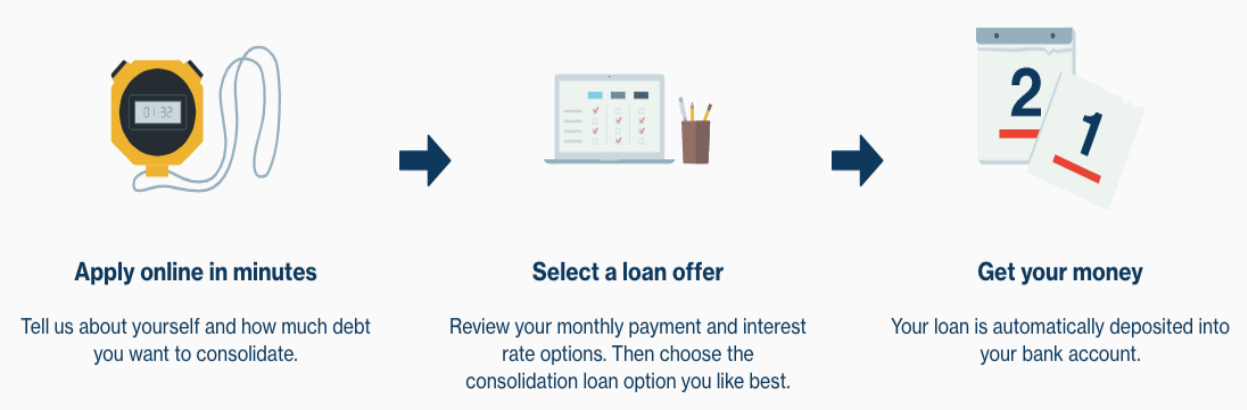
Conclusion ***21***

About the LendingClub:

LendingClub is an American peer-to-peer lending company, headquartered in San Francisco, California. The company offers loan trading on a secondary market by bringing borrowers and investors together. The company claims that $15.8 billion in loans has issued for millions of people between the years of 2007 and 2015. LendingClub offers a variety of loan option for consumers and businesses including personal loans, business loans, patient solutions, and K-12 education loans.

How it Works:

Individual consumers can apply online for a personal loan through LendingClub. They can select a loan offer after reviewing their monthly payment and interest rate options. Once approved, the loan is deposited into the consumers' bank account. Personal loans can range from $1,000 to $40,000 depending upon the information provided and credit report. Two active personal loans are allowed at the same time.



Investors, on the other hand, put funds in SEC-registered securities called Member Payment Dependent Notes, which form fractions of loans instead of making loans directly to the borrowers. Investors receive monthly payments out of principal and interest payments. LendingClub makes money by charging borrowers an origination fee and investors a service fee.

Problem Statement:

Loan default is the failure to pay back a debt according to an initial arrangement. It causes huge losses to the banks and lenders, so they pay much attention on loan defaults and apply various methods to detect and predict default behaviors of their customers. Machine learning algorithms which have a good performance on this purpose, are widely used in banking and lending.

In this project, I will develop various machine learning models in order to correctly detect and predict good (low risk) and bad (high risk) loans using the LendingClub Loan Data. The machine learning models would bring added value by minimizing the associated risks and the LendingClub can decide whether a person is fit for a loan or not in the future. An exploratory data analysis (EDA) will be initially performed to develop insight about the data and features. After getting the data ready for the analysis, various machine learning models will be developed to correctly identify and predict good and bad loans. The analyses will also provide insightful recommendations and future steps that the LendingClub managers and investors might want to take into consideration.

About the Dataset:

The data for the analyses were acquired from Kaggle.com - Lending Club Loan Data. The data contain about 890 thousand observations and 75 variables including loan amount, funded loan amount, interest rate, credit scores, employment length, homeownership, annual income, address including zip codes, and state, and collections. A data dictionary is provided in a separate file. The dataset used for this project can be acquired using the following link: https://www.kaggle.com/wendykan/lending-club-loan-data

Data Preprocessing

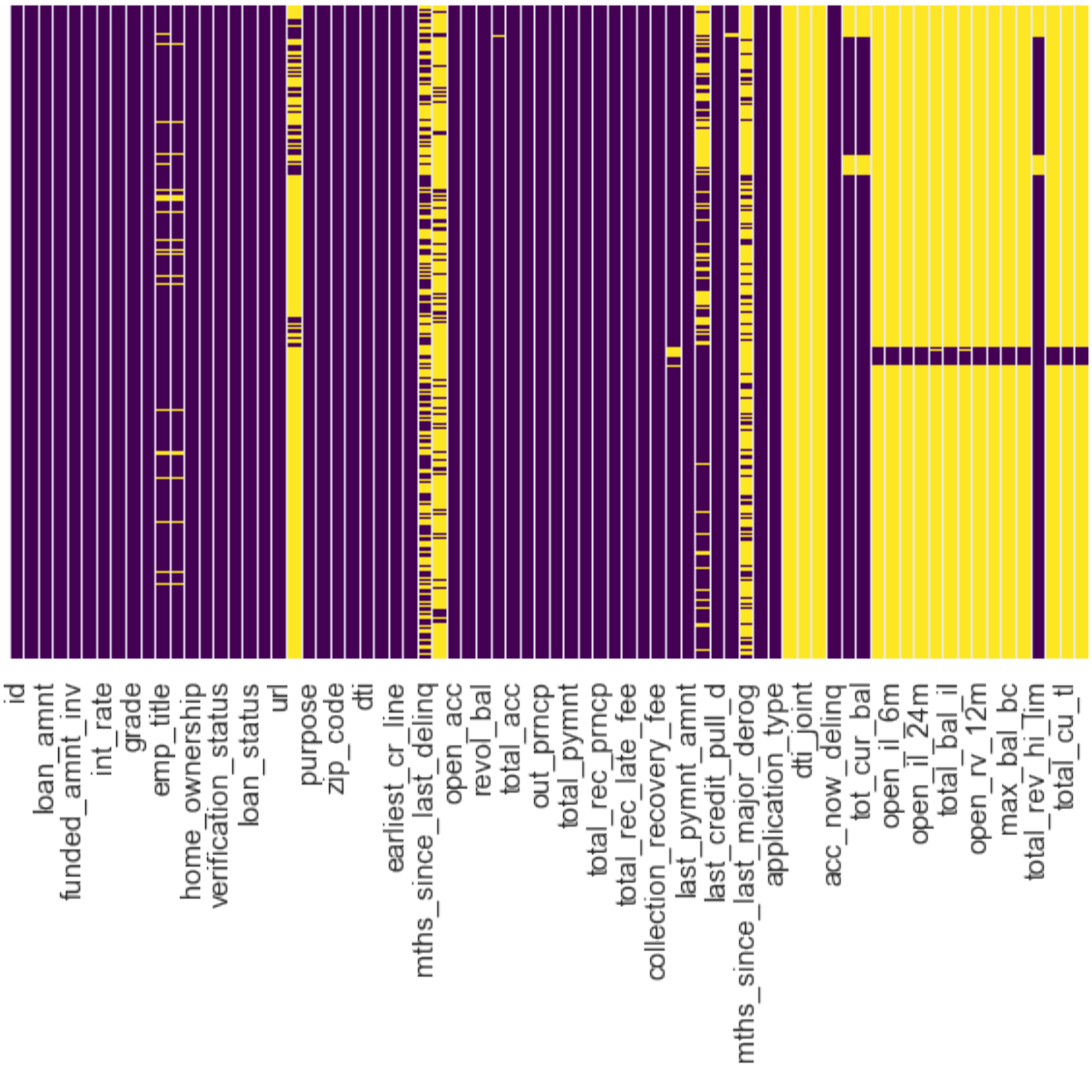
Data preprocessing is an extremely important step in machine learning as the quality of data affects the ability of the models to learn and predict. Hence, before leaping into performing analysis and running algorithms, various data preprocessing steps were taken to inspect the data, to look for the missing values, and to clean the data to get ready for further analysis. The following steps were taken in this section to get the data ready before feeding it into the machine learning models:

Steps taken to get the data ready for further analysis:

1. All the necessary packages that will be used throughout the project were imported.
2. LendingClub loan dataset was read into a dataframe.
3. A general insight about the dataset was developed through checking the head, shape, info, and description of the dataframe. The dataframe consists of a total of 887,379 rows (entries) and 74 columns (features). A preliminary analysis of the head and info of the dataframe showed that there are some columns with huge numbers of missing values (only a few pieces of information were entered). Checking the description of the dataframe revealed that the numerical attributes looks normal (within accepted ranges) and no outliers were detected. Unique values of id attribute were counted to see how many unique individuals/borrowers we are dealing with. In this dataset, there were 887,379 unique individuals that indicate that each entry is unique to an individual.

	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	grade	sub_grade	...
0	1077501	1296599	5000.0	5000.0	4975.0	36 months	10.65	162.87	B	B2	...
1	1077430	1314167	2500.0	2500.0	2500.0	60 months	15.27	59.83	C	C4	...
2	1077175	1313524	2400.0	2400.0	2400.0	36 months	15.96	84.33	C	C5	...
3	1076863	1277178	10000.0	10000.0	10000.0	36 months	13.49	339.31	C	C1	...
4	1075358	1311748	3000.0	3000.0	3000.0	60 months	12.69	67.79	B	B5	...

4. To gain insight about the missing values, a heatmap that visually shows the missing values was produced. The visual representation of the missing values through the heatmap was supported by producing a new dataframe that shows the number and percentages of the missing values in a particular column.

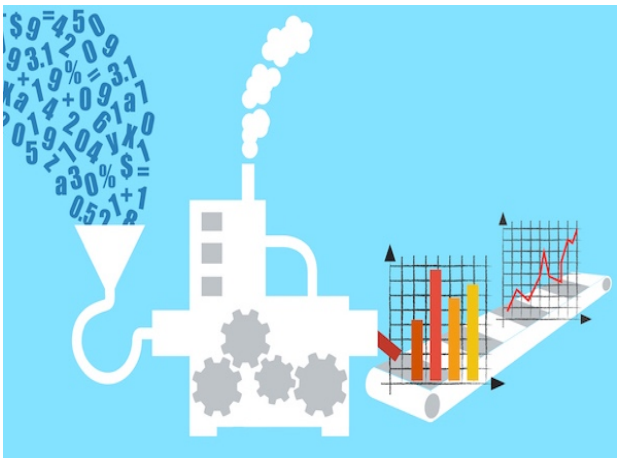


5. Columns that have 80% or more missing values were dropped. (80% is a conventional way of setting a threshold to drop features with missing values).
6. After dropping those features with 80% or more missing values, 52 features were left. Further analysis of the data showed that some features have either low variance or a unique value for each entry which can lead to overfitting. Hence, those features were removed from the dataset as well.

7. The final dataframe consists of 887,379 rows and 46 columns.

The following section provides the details of various exploratory data analyses which were performed to understand the emerging themes and relationships.

Exploratory Data Analysis (EDA)

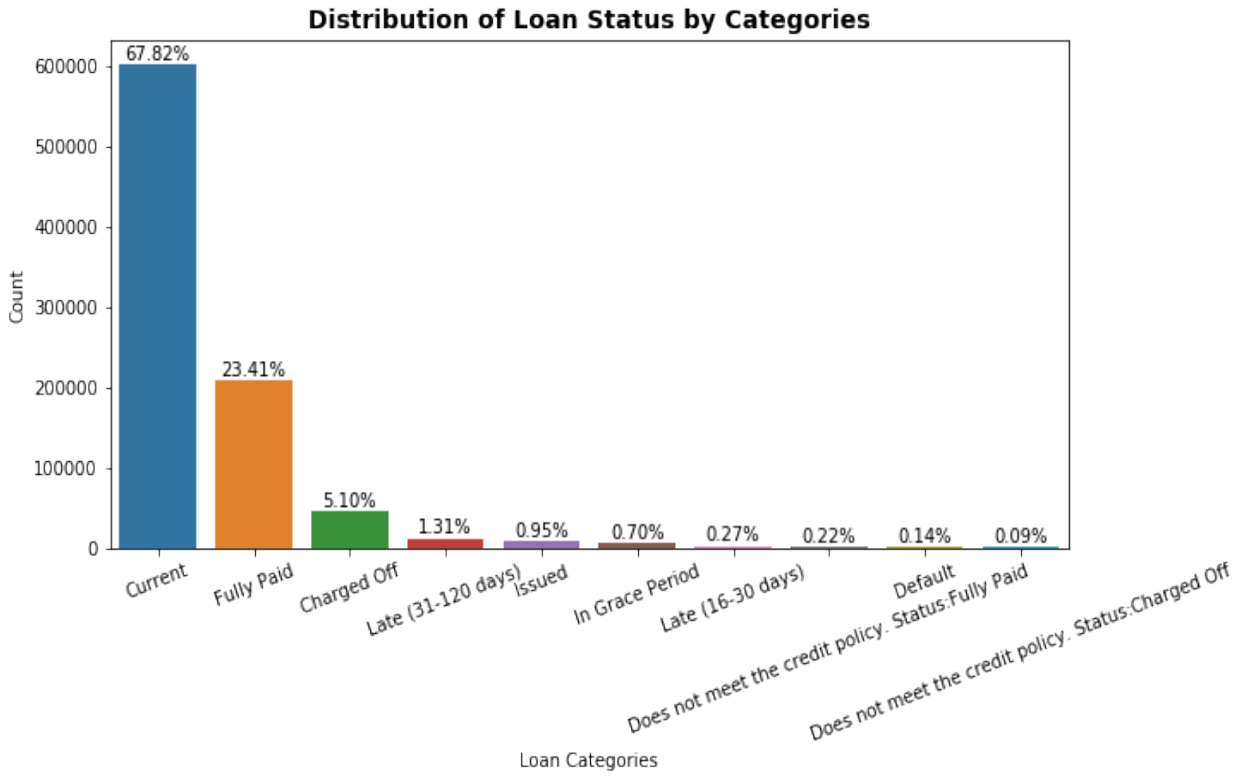


This section focuses on exploratory data analysis which allows us to understand emerging themes and relationships between the features before moving onto deeper and complex data analysis. We got a good glimpse of data in the previous section and will explore the data with graphs for the following themes and relationships. Loan status is the target variable in this project, so that the following graphs were produced to explore the relationship between the loan status feature and some other important features.

1. Distribution of Loan Status by Categories
2. Distribution of Total Loans Issued by Year
3. Distribution of Loan Amount by Borrower Purposes
4. Distribution of Loan Status by Borrower Purposes
5. Distribution of Interest Rate by Grade Category
6. Loan Status by Grade Category
7. Good vs. Bad Loan by Interest Rate
8. Distribution of Homeownership
9. Loan Status by Homeownership
10. Loan Status by Employment Length
11. Correlation between Features

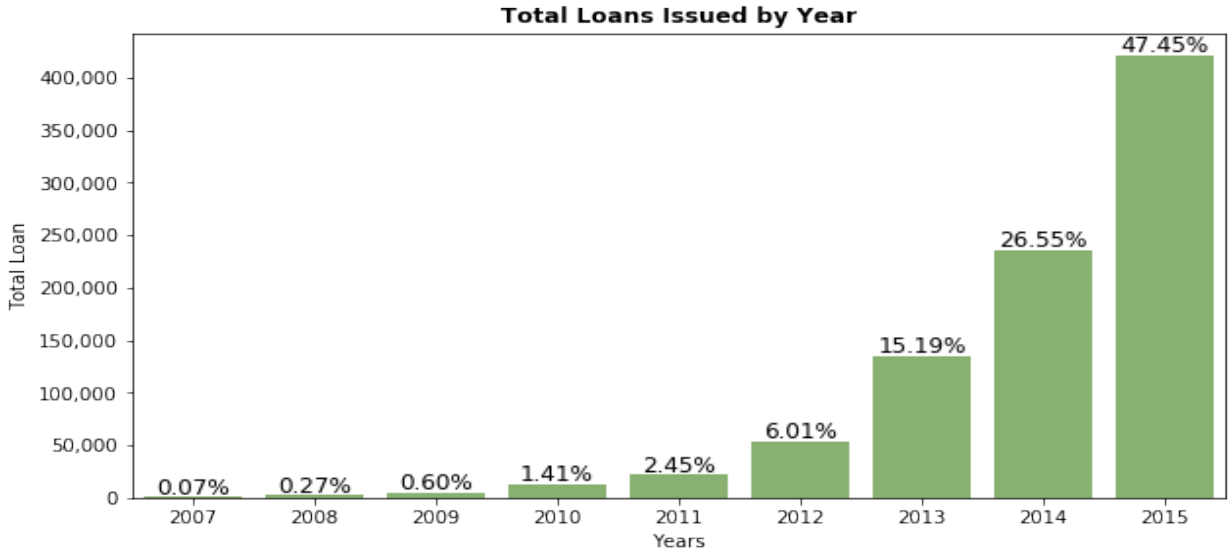
Distribution of Loan Status by Categories

The following bar plot displays the distribution of loan status by categories. This bar plot gives us insight into how the data are distributed across loan status. A big percentage of loans are currently being paid while a fair amount of loans was fully paid. The other categories, on the other hand, have relatively smaller percentages. The categories of "Charged Off" and "Default" corresponds to respectively %5.10 and 0.14% of the entire loans issued by the LendingClub. These numbers indicate that the data are imbalanced which needs to be taken into account when building our predictive models.



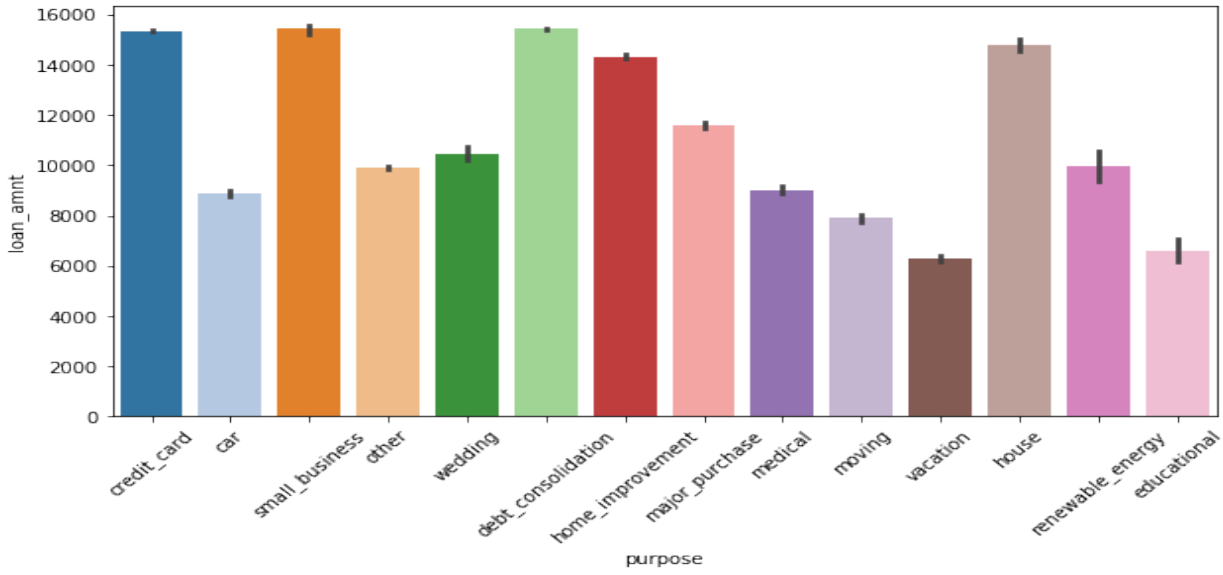
Distribution of Total Loans Issued by Year

The count of total loans issued by the LendingClub over the years displayed below. As can be seen from the chart, the number of loans issued is increasing over the years with the highest in 2015.



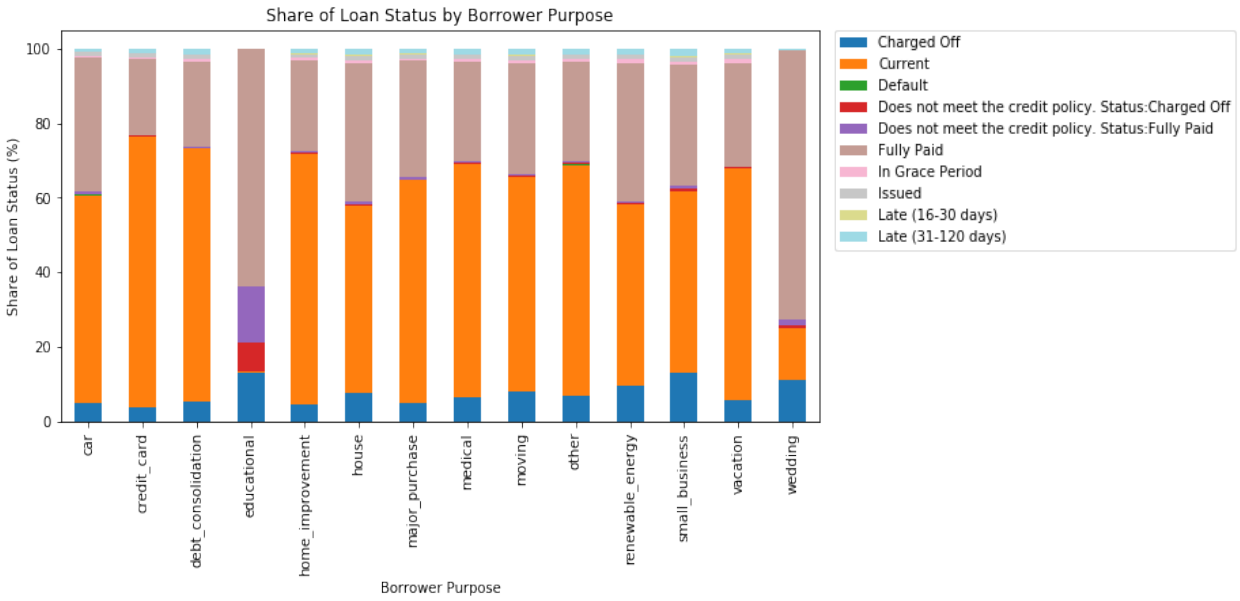
Distribution of Loan Amount by Borrower Purposes

The amount of loans issued by the LendingClub varies across the borrowers' purposes listed below. The largest amount of loan was issued respectively for the borrowers with credit card, debt consolidation, and small business purposes.



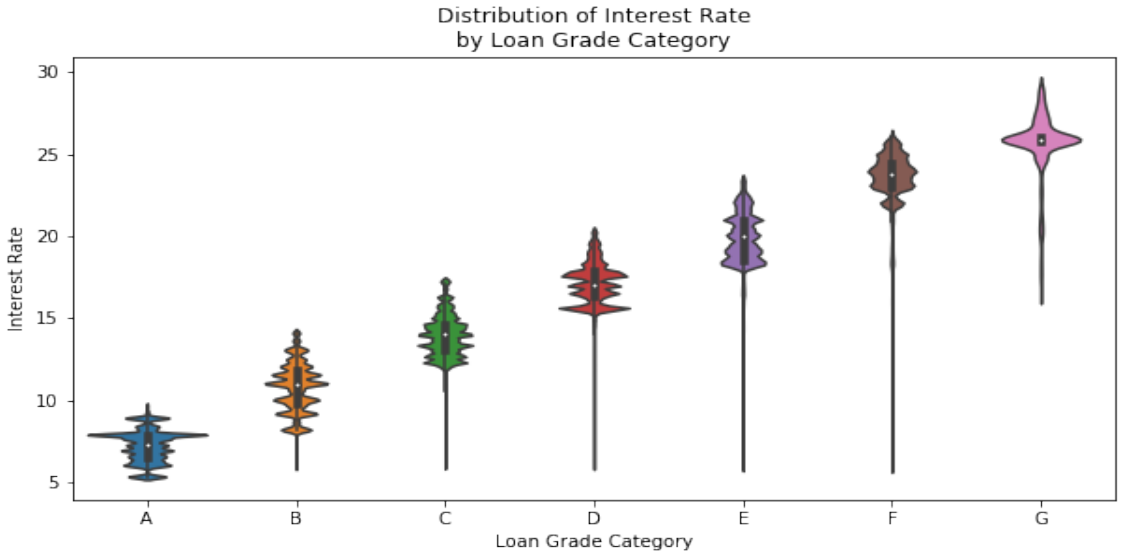
Distribution of Loan Status by Borrower Purpose

An additional chart was produced below to provide insight into the share of loan status by borrower purpose. Borrowers with educational, small business and wedding purposes have a greater percentage to be charged off than the borrowers with other purposes.



Distribution of Interest Rate by Grade Category

The violinplot below clearly illustrates that interest rate increases as loan grade category moves from A to G. It means that borrowers with a bad loan grade category are likely to have higher interest rate for the loans than the borrowers with a good loan grade category. The violinplot also shows that the average interest rate for the borrowers with "A" loan grade is 7.24, while it is 25.63 for the borrowers with "G" loan grade category. It also must be noted that the variation in interest rate is mostly increasing as the loan grade category moves from "A" to "F".



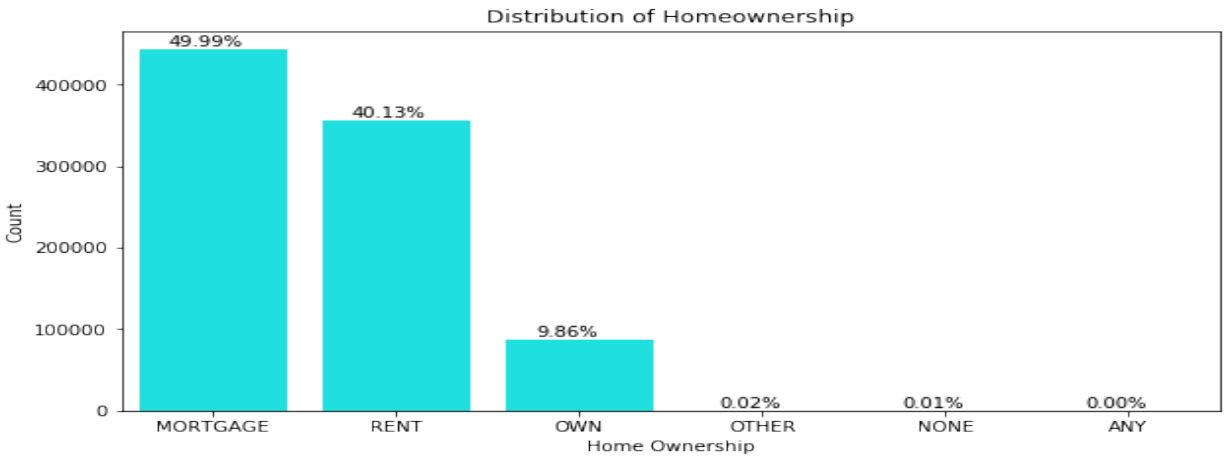
Loan Status by Grade Category

The cross-tabulation of loan status and loan grade categories was displayed below. The table shows that almost 70% of the borrowers with "A" loan grade category are currently paying their loans, while almost 27% of them fully paid their loans. The borrowers with a "G" loan grade category have a high percentage of being currently paying their loans. When we look at the loan defaults and charged offs, individuals with "G" loan grade category have relatively higher percentages to be defaulted or charged off than the other loan grade categories. However, the percentage for loan defaults is relatively small only at 0.4%.

grade	A	B	C	D	E	F	G
loan_status							
Charged Off	1.77	3.74	5.14	7.51	8.85	12.73	14.43
Current	69.72	67.47	69.62	65.92	66.56	58.96	53.07
Default	0.03	0.08	0.15	0.22	0.28	0.34	0.4
Does not meet the credit policy. Status:Charged Off	0.01	0.03	0.06	0.14	0.22	0.4	1.31
Does not meet the credit policy. Status:Fully Paid	0.06	0.11	0.2	0.35	0.53	0.67	2.22
Fully Paid	26.77	26.14	21.43	21.51	18.28	20.51	20.88
In Grace Period	0.25	0.49	0.77	1.01	1.28	1.54	1.71
Issued	0.98	0.99	1.01	0.85	0.84	0.84	0.71
Late (16-30 days)	0.09	0.16	0.28	0.41	0.52	0.67	0.78
Late (31-120 days)	0.33	0.79	1.36	2.07	2.62	3.33	4.48

Distribution of Homeownership

The distribution of homeownership shows that almost 50% of the borrowers have Mortgage, followed by the borrowers who are paying rent with a percentage of 40.13%. The borrowers who own homes correspond to 9.86% of the all the individuals applied for the loan from the LendingClub.



Loan Status by Homeownership

The cross-tabulation of loan status and homeownership features was provided below. The table shows that almost 70% of the borrowers with Mortgage is currently paying for their loans to the LendingClub, while almost 24% of them fully paid their loans. A big percentage of the borrowers who pay rent for their homes are similarly currently paying their loans. However, they have greater percentage to be charged off for their loans than the borrowers with Mortgage.

home_ownership loan_status	ANY	MORTGAGE	NONE	OTHER	OWN	RENT
Charged Off	0	4.48	14	14.84	4.6	5.98
Current	66.67	68.48	4	1.65	70.93	66.26
Default	0	0.11	0	0	0.13	0.17
Does not meet the credit policy. Status:Charged Off	0	0.08	2	6.04	0.06	0.1
Does not meet the credit policy. Status:Fully Paid	0	0.2	8	14.84	0.16	0.26
Fully Paid	33.33	23.66	72	62.64	20.53	23.77
In Grace Period	0	0.64	0	0	0.73	0.78
Issued	0	0.95	0	0	1.19	0.9
Late (16-30 days)	0	0.25	0	0	0.3	0.28
Late (31-120 days)	0	1.13	0	0	1.39	1.51

Loan Status by Employment Length

The length of employment is another factor considered before issuing loans by the loan companies. However, as the cross-tabulation table indicates, there is no clear pattern between the loan status categories and the length of employment.

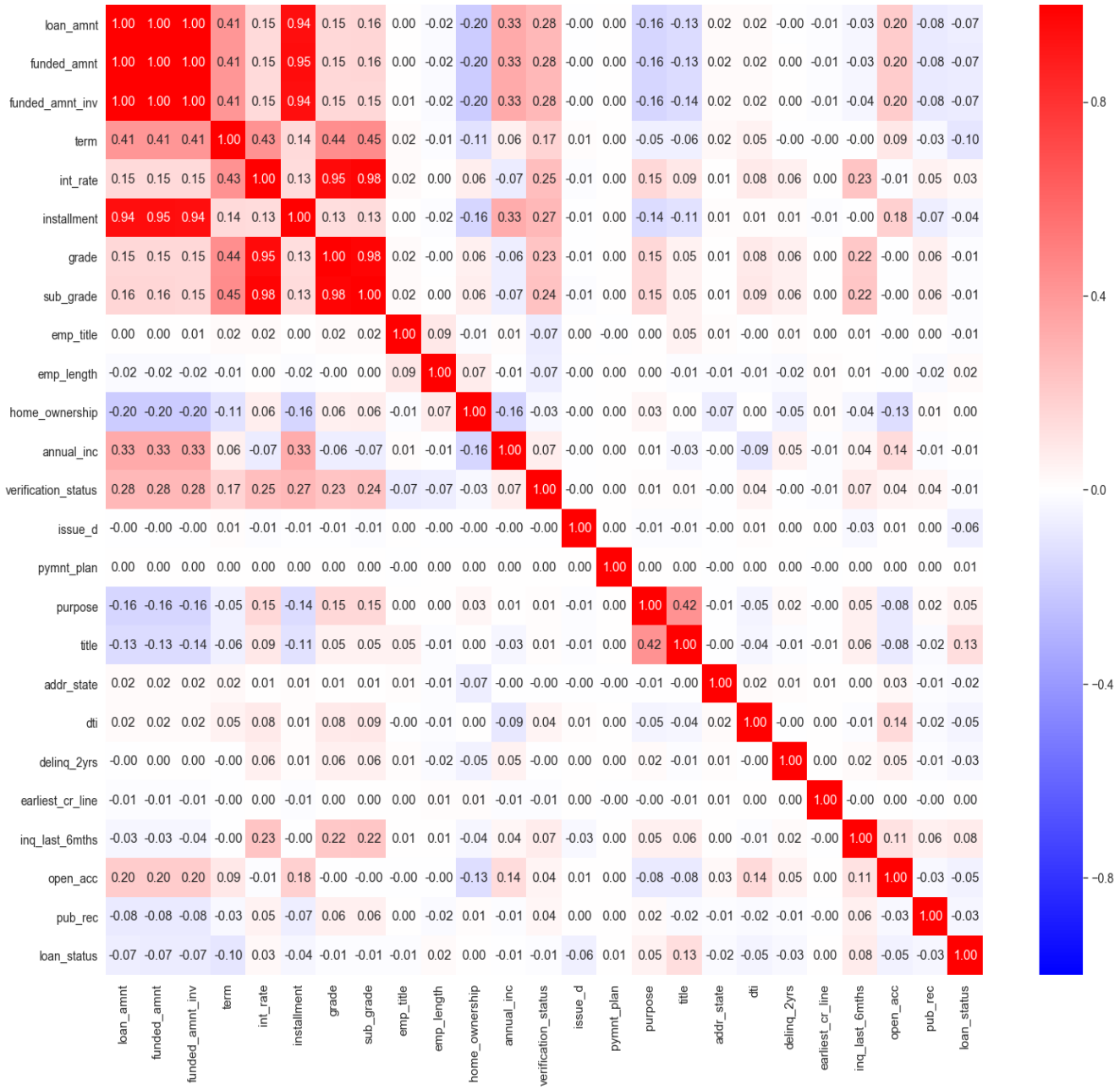
emp_length loan_status	1 year	10+ years	2 years	3 years	4 years	5 years	6 years	7 years	8 years	9 years	< 1 year
Charged Off	5.19	4.5	5.11	5.05	5.28	5.75	6.27	5.83	4.9	5.13	5.46
Current	66.39	70.25	66.36	66.99	65.45	64.05	62	65.14	69.39	68.81	66.03
Default	0.17	0.13	0.11	0.15	0.13	0.14	0.15	0.16	0.17	0.14	0.13
Does not meet the credit policy. Status:Charged Off	0.16	0.05	0.11	0.1	0.11	0.09	0.1	0.07	0.07	0.06	0.16
Does not meet the credit policy. Status:Fully Paid	0.45	0.11	0.34	0.28	0.28	0.22	0.24	0.15	0.17	0.18	0.51
Fully Paid	24.33	21.86	24.76	24.06	25.55	26.67	28.07	25.75	22.06	22.48	24.12
In Grace Period	0.73	0.68	0.72	0.8	0.69	0.66	0.75	0.71	0.76	0.73	0.74
Issued	0.96	0.97	0.89	0.97	0.93	0.92	0.75	0.59	0.97	0.85	1.09
Late (16-30 days)	0.26	0.24	0.27	0.27	0.3	0.28	0.24	0.26	0.25	0.32	0.32
Late (31-120 days)	1.36	1.21	1.34	1.33	1.28	1.23	1.42	1.33	1.26	1.31	1.44

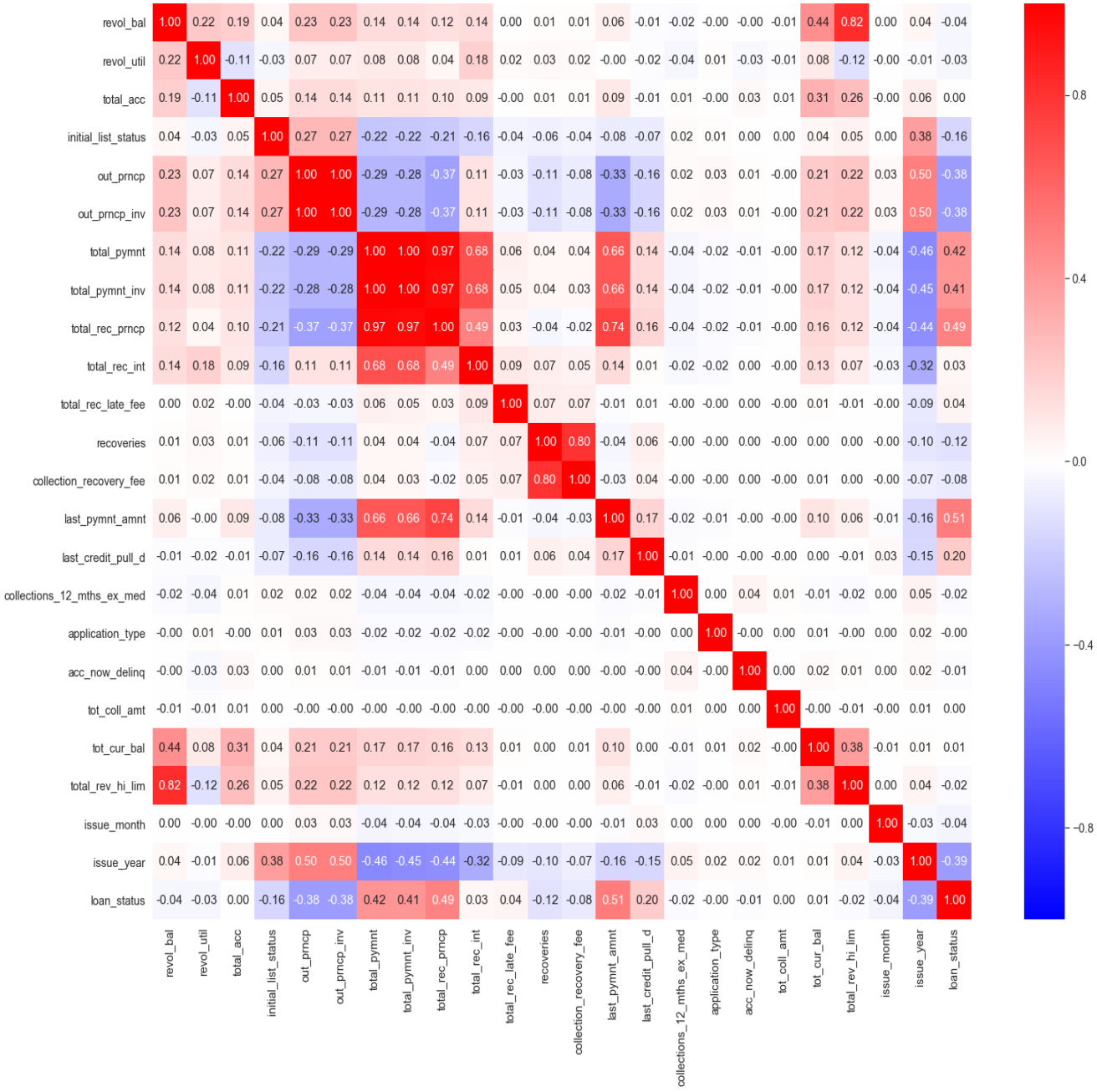
Correlation Between Features

Due to the existence of so many features in our dataset, the entire dataset were divided into two parts in order to create a correlation matrix using the heatmap below. Features with object data type were converted into categorical data type and categorical codes were assigned to include those features into the correlation matrix below.

The correlation heatmap clearly illustrates the relationships between the variables of interest. While the red color indicates a positive correlation between the two variables, blue color indicates negative correlation. The strength of correlations has also been added on the heatmap above. As an example, there is a positive strong correlation between loan amount and installment as expected. It also must be noted that correlation does not imply causation.

It is necessary to remove highly correlated features from the dataset to improve our models. Highly correlated features were dropped from the dataset in the feature selection section (>=70).




revol_bal 1.00 0.22 0.19 0.04 0.23 0.23 0.14 0.14 0.12 0.14 0.00 0.01 0.01 0.06 -0.01 -0.02 -0.00 -0.00 -0.01 0.44 0.82 0.00 0.04 -0.04
revol_util 0.22 1.00 -0.11 -0.03 0.07 0.07 0.08 0.08 0.04 0.18 0.02 0.03 0.02 -0.00 -0.02 -0.04 0.01 -0.03 -0.01 0.08 -0.12 -0.00 -0.01 -0.03
total_acc 0.19 -0.11 1.00 0.05 0.14 0.14 0.11 0.11 0.10 0.09 -0.00 0.01 0.01 0.09 -0.01 0.01 -0.00 0.03 0.01 0.31 0.26 -0.00 0.06 0.00
initial_list_status 0.04 -0.03 0.05 1.00 0.27 0.27 -0.22 -0.22 -0.21 -0.16 -0.04 -0.06 -0.04 -0.08 -0.07 0.02 0.01 0.00 0.00 0.04 0.05 0.00 0.38 -0.16
out_prncp 0.23 0.07 0.14 0.27 1.00 1.00 -0.29 -0.28 -0.37 0.11 -0.03 -0.11 -0.08 -0.33 -0.16 0.02 0.03 0.01 -0.00 0.21 0.22 0.03 0.50 -0.38
out_prncp_inv 0.23 0.07 0.14 0.27 1.00 1.00 -0.29 -0.28 -0.37 0.11 -0.03 -0.11 -0.08 -0.33 -0.16 0.02 0.03 0.01 -0.00 0.21 0.22 0.03 0.50 -0.38
total_pymnt 0.14 0.08 0.11 -0.22 -0.29 -0.29 1.00 1.00 0.97 0.68 0.06 0.04 0.04 0.66 0.14 -0.04 -0.02 -0.01 -0.00 0.17 0.12 -0.04 -0.46 0.42
total_pymnt_inv 0.14 0.08 0.11 -0.22 -0.28 -0.28 1.00 1.00 0.97 0.68 0.05 0.04 0.03 0.66 0.14 -0.04 -0.02 -0.01 -0.00 0.17 0.12 -0.04 -0.45 0.41
total_rec_prncp 0.12 0.04 0.10 -0.21 -0.37 -0.37 0.97 0.97 1.00 0.49 0.03 -0.04 -0.02 0.74 0.16 -0.04 -0.02 -0.01 -0.00 0.16 0.12 -0.04 -0.44 0.49
total_rec_int 0.14 0.18 0.09 -0.16 0.11 0.11 0.68 0.68 0.49 1.00 0.09 0.07 0.05 0.14 0.01 -0.02 -0.02 0.00 -0.00 0.13 0.07 -0.03 -0.32 0.03
total_rec_late_fee 0.00 0.02 -0.00 -0.04 -0.03 -0.03 0.06 0.05 0.03 0.09 1.00 0.07 0.07 -0.01 0.01 -0.00 -0.00 0.00 -0.00 0.01 -0.01 -0.00 -0.09 0.04
recoveries 0.01 0.03 0.01 -0.06 -0.11 -0.11 0.04 0.04 -0.04 0.07 0.07 1.00 0.80 -0.04 0.06 -0.00 -0.00 0.00 -0.00 0.00 0.00 -0.00 -0.10 -0.12
collection_recovery_fee 0.01 0.02 0.01 -0.04 -0.08 -0.08 0.04 0.03 -0.02 0.05 0.07 0.80 1.00 -0.03 0.04 -0.00 -0.00 0.00 -0.00 0.01 0.00 -0.00 -0.07 -0.08
last_pymnt_amnt 0.06 -0.00 0.09 -0.08 -0.33 -0.33 0.66 0.66 0.74 0.14 -0.01 -0.04 -0.03 1.00 0.17 -0.02 -0.01 -0.00 -0.00 0.10 0.06 -0.01 -0.16 0.51
last_credit_pull_d -0.01 -0.02 -0.01 -0.07 -0.16 -0.16 0.14 0.14 0.16 0.01 0.01 0.06 0.04 0.17 1.00 -0.01 -0.00 -0.00 -0.00 0.00 -0.01 0.03 -0.15 0.20
collections_12_mths_ex_med -0.02 -0.04 0.01 0.02 0.02 0.02 -0.04 -0.04 -0.04 -0.02 -0.00 -0.00 -0.00 -0.02 -0.01 1.00 0.00 0.04 0.01 -0.01 -0.02 0.00 0.05 -0.02
application_type -0.00 0.01 -0.00 0.01 0.03 0.03 -0.02 -0.02 -0.02 -0.02 -0.00 -0.00 -0.00 -0.01 -0.00 0.00 1.00 -0.00 0.00 0.01 -0.00 0.00 0.02 -0.00
acc_now_delinq -0.00 -0.03 0.03 0.00 0.01 0.01 -0.01 -0.01 -0.01 0.00 0.00 0.00 0.00 -0.00 -0.00 0.04 -0.00 1.00 0.00 0.02 0.01 0.00 0.02 -0.01
tot_coll_amt -0.01 -0.01 0.01 0.00 -0.00 -0.00 -0.00 -0.00 -0.00 -0.00 -0.00 -0.00 -0.00 -0.00 -0.00 0.01 0.00 0.00 1.00 -0.00 -0.01 -0.00 0.01 0.00
tot_cur_bal 0.44 0.08 0.31 0.04 0.21 0.21 0.17 0.17 0.16 0.13 0.01 0.00 0.01 0.10 0.00 -0.01 0.01 0.02 -0.00 1.00 0.38 -0.01 0.01 0.01
total_rev_hi_lim 0.82 -0.12 0.26 0.05 0.22 0.22 0.12 0.12 0.12 0.07 -0.01 0.00 0.00 0.06 -0.01 -0.02 -0.00 0.01 -0.01 0.38 1.00 0.00 0.04 -0.02
issue_month 0.00 -0.00 -0.00 0.00 0.03 0.03 -0.04 -0.04 -0.04 -0.03 -0.00 -0.00 -0.00 -0.01 0.03 0.00 0.00 0.00 -0.00 -0.01 0.00 1.00 -0.03 -0.04
issue_year 0.04 -0.01 0.06 0.38 0.50 0.50 -0.46 -0.45 -0.44 -0.32 -0.09 -0.10 -0.07 -0.16 -0.15 0.05 0.02 0.02 0.01 0.01 0.04 -0.03 1.00 -0.39
loan_status -0.04 -0.03 0.00 -0.16 -0.38 -0.38 0.42 0.41 0.49 0.03 0.04 -0.12 -0.08 0.51 0.20 -0.02 -0.00 -0.01 0.00 0.01 -0.02 -0.04 -0.39 1.00
revol_bal
revol_util
total_acc
initial_list_status
out_prncp
out_prncp_inv
total_pymnt
total_pymnt_inv
total_rec_prncp
total_rec_int
total_rec_late_fee
recoveries
collection_recovery_fee
last_pymnt_amnt
last_credit_pull_d
collections_12_mths_ex_med
application_type
acc_now_delinq
tot_coll_amt
tot_cur_bal
total_rev_hi_lim
issue_month
issue_year
loan_status
0.8
0.4
0.0
−0.4
−0.8

Feature Selection

Dealing with the Rows Which Contain Missing Values

The missing values in our dataset need to be dealt before building the machine learning models. The rows that contain missing values were dropped since we have sufficient amount of data. 769,128 rows were left after dropping the rows with the missing values.

Converting Loan Status into Good or Bad Loan

The target feature of this project is "Loan Status" which has 8 different categories after dropping the rows with missing values. For the purpose of this project, we needed to convert the loan status feature into "Good Loan" or "Bad Loan" based on the description of loan status categories. The categories of "Current", "Fully Paid", and "Issued" were considered as a good loan. The category of "In Grace Period" can be considered to be good or bad depending upon the strictness of investor. It was treated as a good loan in this project. The categories of "Late (16-30 days)", "Late (31-120 days)", "Charged Off" and "Default" were converted to a bad loan.

The dataset is highly imbalanced in the way that the number of bad loans is only 46,539 out of 769,128 which corresponds to 6.05% of the entire dataset.

Feature Selection for Object Data Type

According to the descriptive statistics of the features that have object data type, several of the features needs to be dropped due to the fact that they have a large number of unique values which leads to memory error. The features with high correlation were also dropped because only one of them will be sufficient to feed the machine learning models. Lastly, some features were dropped from the dataset since they do not provide valuable info for further analysis.

One Hot Encoding

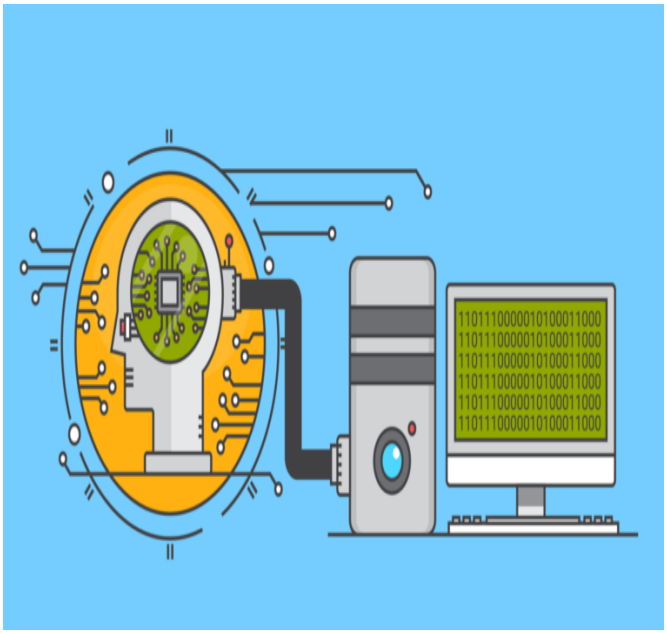
After dropping the features with low variance and high correlation, we ended up with 7 features that have "object" data type. These features need to be converted into numerical data type because machine learning algorithms will not work with categorical variables. To convert these features, we could either use "Label Encoder" available in Python or "One Hot Encoder". Since most of the features with "object" data type have multiple categories, it would be inappropriate to use the "Label Encoder". Instead, we went with "One Hot Encoder" to convert these features into numerical data type.

Feature Name	Description
loan_amnt	The listed amount of the loan applied for the borrower
int_rate	Interest rate on the loan
installment	The monthly payment owed by the borrower if the loan originates
annual_inc	The self-reported annual income provided by the borrower during the registration
loan_status	Current status of loan
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
delinq_2year	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years
inq_last_6mths	The number of inquiries in past 6 months (excluding auto and mortgage inquiries)
open_acc	The number of open credit lines in the borrower's credit file
pub_rec	Number of derogatory public records
revol_bal	Total credit revolving balance

revol_util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit
total_acc	The total number of credit lines currently in the borrower's credit file
out_prncp	Remaining outstanding principal for total amount funded
out_prncp_inv	Remaining outstanding principal for portion of total amount funded by investors
total_pymnt	Payments received to date for total amount funded
total_pymnt_inv	Payments received to date for portion of total amount funded by investors
total_rec_int	Interest received to date
total_rec_late_fee	Late fees received to date
recoveries	post charge off gross recovery
collection_recovery_fee	Post charge off collection fee
collections_12_mths_ex_med	Number of collections in 12 months excluding medical collections
acc_now_delinq	The number of accounts on which the borrower is now delinquent
tot_coll_amt	Total collection amounts ever owed
tot_cur_bal	Total current balance of all acounts
grade	LC assigned loan grade
emp_length	Employment length in years
homeownership	The home ownership status provided by the borrower during registration or obtained from the credit report
payment_plan	Indicates if a payment plan has been put in place for the loan
purpose	A category provided by the borrower for the loan request
initial_list_status	The initial listing status of the loan
application_type	Indicates whether the loan is an individual application or a joint application with two co-borrowers

Now, we selected our features and got the data ready to build our machine learning models.

Model Building / Machine Learning



The ultimate purpose of this project was to correctly detect and predict good (low risk) and bad (high risk) loans using various machine learning models which would bring added value by minimizing the associated risks. The dataset is highly imbalanced in the way that the number of bad loans was only 46,539 out of 769,128 which corresponds to 6.05% of the entire dataset. Hence, we wanted to see if balancing the data makes difference in predicting good and bad loans. Two separate section were developed for machine learning models: model building using imbalanced data and model building using balanced data.

Imbalanced Data:

Our first approach to build models was based upon imbalanced dataset. The following machine learning algorithms were used in this section to predict good and bad loans:

1. Logistic regression classifier
2. Decision tree classifier
3. Random forest classifier

After splitting the data into train (75%) and test (25%) sets and a data normalization process was performed on numeric continuous features, we did 3-fold cross-validation to see average accuracy score. Among the models given above, the logistic regression classifier achieved the highest accuracy score of more than 98% whereas the decision tree classifier performed the worst accuracy score of 96%. However, given that this dataset was imbalanced, we also checked some other model evaluation reports such as confusion matrix and classification report

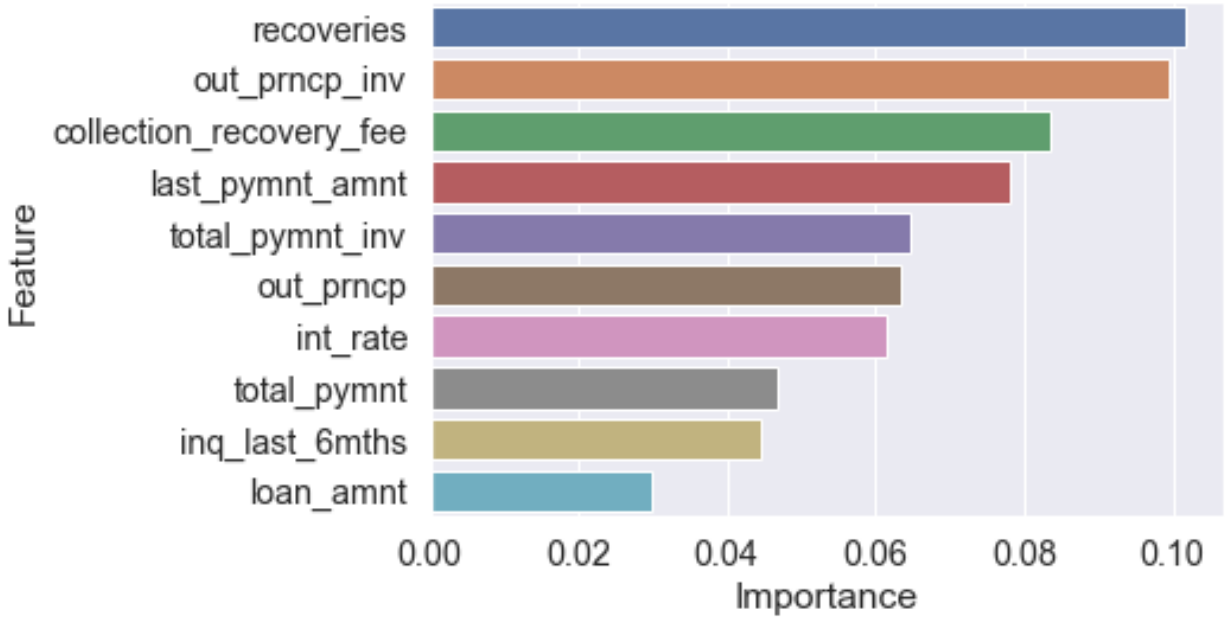
which was used to evaluate and compare our models. Again, the logistic regression classifier performed better in terms of higher F1 score and better prediction results.

Balanced Data:

Now, we used a technique called Synthetic Minority Oversampling Technique (SMOTE) to oversample the minority class of bad loans to 300,000 from 46,539. The same machine learning algorithms again were used with 3-fold cross validation. Across models, even though the decision tree classifier provided the highest accuracy score, it was not doing as well as the logistic regression and random forest classifiers in predicting good and bad loans based on the results of the confusion matrix and classification report.

Feature Importance

Since the random forest classifier is built upon decision tree, it enables us to plot feature importance which tells us how useful a feature is in predicting an outcome. The top 10 features with highest importance in predicting good and bad loans were produced below. “recoveries” was the most important feature followed by “out_prncp”, and “collection_recovery_fee” features.



Business Recommendations

This section provides several recommendations for the LendingClub based on our analyses of the loan dataset:

1. Across different borrower purposes, those with educational purposes, small business purposes, and wedding purposes have a greater probability to be charged off than the borrowers with any other purposes. The LendingClub might want to take it consideration before issuing loans in order to minimize the risks of loan defaults and charge offs.
2. The machine learning models that we have developed here successfully detect and predict good and bad loans. These algorithms can help the LendingClub identy the risky loans. More importantly, while we were working on this project, the LendingClub revealed another dataset for the next 3 years. This new dataset could be used by the LendingClub to validate our predictions.
3. The LendingClub might want to use the most important features in predicting good and bad loans given above as a metric to identify individuals with higher risk to be defaulted or charged off. For example, the number of inquiries in the last 6 months is highly important in predicting the bad loans. Hence, the LendingClub could prepare a metric using the highly important features in predicting good and bad loans.

Conclusion

The ultimate purpose of this project was to predict good (low risk) and bad loans (high risk) using the loan dataset revealed by the LendingClub. The processes of data wrangling, exploratory data analysis, feature selection, and model building were performed on the loan dataset. The insights gained as part of these processes were highlighted and several business recommendations were provided for the LendingClub.

The code associated with this report is available at:

https://github.com/tayfunayazma/lendingclub_loan_analysis