1. Abstract

Marketplace loans are loans made in P2P consumer platforms. P2P lending is an emerging Internet-based application where individuals can directly borrow money from each other. This market is an important fintech market. The past decade has witnessed the rapid development and prevalence of online P2P lending platforms, examples of which include Prosper, Lending Club, and Kiva. Meanwhile, extensive research has been done that mainly focuses on the studies of platform mechanisms and transaction data. In this project, I first discuss about the evolution and the state of the market (i.e., volumes, borrower characteristics, loan characteristics). Then, I analyze the real-world data from this market from LendingClub.com and provide insights into default behavior (i.e., what borrower characteristics and environmental conditions lead to default) using logistic regression analysis. The model predicts/estimates whether the loan would be a default or no (y-variable) based on various factors(x-variable). The findings illustrate how, for different factors such as interest rate, term, etc plays roles in predicting loan default for a given data. The model has predictability accuracy of 73% which can help lenders to estimate the probability of default in the given dataset. Then, I analyze loan portfolio (i.e., how portfolio given different returns depending on borrower characteristics). Factors explaining default are loan purpose, annual income, current housing situation, interest rate, credit history, indebtedness and interest rate. The grade assigned by the site is the most predictive factor of default, but the accuracy of the model is improved by adding other information, especially the borrower's debt level. Finally, I propose my opinions on the prospects of P2P lending and suggest some future research directions in this field.

2. Introduction

Peer-to-peer (P2P) lending is the practice of lending money to individuals or businesses through online services that match lenders directly with borrowers. Since the P2P lending companies offering these services operate entirely online, they can run with lower overhead and provide the service more cheaply than traditional financial institutions. Therefore, lenders can acquire higher returns as compared to investments and saving products given by banks and other institutions. This money lending process is also known as "crowd-lending". Some of the forms of peer to peer lending are; student loans, commercial and real estate loans, payday loans, secured business loans, leasing and factoring. The interest rates can be set by lenders who compete for the lowest rate on the reverse auction model or fixed by the intermediary company based on an analysis of the borrower's credit. The lender's investment in the loan is not normally protected by any government guarantee.

Global Peer to Peer Lending (P2P) Market is valued at USD 34.16 Billion in 2018 and expected to reach USD 589.05 Billion by 2025 with a CAGR of 50.2% over the forecast period. The regions covered in this market report are North America, Europe, Asia-Pacific, Latin America and Africa. Based on country level, the market of peer to peer lending is sub divided into U.S., Mexico, Canada, U.K., France, Germany, Italy, China, Japan, India, South East Asia, Middle East Asia (UAE, Saudi Arabia, Egypt) GCC, Africa, etc. Key Players for Peer to Peer Lending Market Report- Some major key players for P2P lending market are LendingClub Corporation, Funding Circle Limited, Daric, RateSetter, Prosper Marketplace, Inc., Zopa Limited, Welendus,

MarketInvoice and others. I am analyzing the data from LendingClub in this project to provide insights into default behavior (i.e., what borrower characteristics and environmental conditions lead to default) using regression analysis. I have also discussed about the evolution and the state of the market (i.e., volumes, borrower characteristics, loan characteristics) in the section 3.

3. Background Study

3.1 Evolution

The first P2P lending platform (i.e., Zopa3) was established in 2005, a company in United Kingdom. Since then more and more different types of P2P lending platforms have emerged (e.g., Prosper4, LendingClub, Kiva5 and Renrendai6). These platforms work under different mechanisms, including trading rules and risk managements. Funding Circle, launched in August 2010, became the first significant peer-to-business lender and offering small businesses loans from investors via the platform. In the US it started in February 2006 with the launch of Prosper Marketplace, followed by Lending Club. Early P2P platforms had few restrictions on borrower eligibility, which resulted in adverse selection problems and high borrower default rates. In addition, some investors viewed the lack of liquidity for these loans, most of which have a minimum three-year term, as undesirable. In 2008, the U.S. Securities and Exchange Commission (SEC) required that P2P companies register their offerings as securities, pursuant to the Securities Act of 1933. The registration process was an arduous one; Prosper and Lending Club had to temporarily suspend offering new loans while others, such as the U.K.-based Zopa Ltd., exited the U.S. market entirely. Both Lending Club and Prosper gained approval from the SEC to offer investors notes backed by payments received on the loans. LendingClub.

The peer-to-peer companies are also required to detail their offerings in a regularly updated prospectus. The SEC makes the reports available to the public via EDGAR (Electronic Data-Gathering, Analysis, and Retrieval). More people turned to peer-to-peer companies for borrowing following the financial crisis of 2007–2008 because banks refused to increase their loan portfolios. This market also faced increased investor scrutiny because borrowers' defaults became more frequent and investors were unwilling to take on unnecessary risk. Rapid increase of P2P market worldwide during the past years, now show that Global Peer to Peer Lending (P2P) Market is valued at USD 34.16 Billion in 2018 and expected to reach USD 589.05 Billion by 2025 with a CAGR of 50.2% over the forecast period. A research report on financial technologies from the International Organizations of Securities Commissions (IOSCO) pinpoints four supply and demand factors that have supported the growth of P2P lending: 1) Reduced technology costs; 2) Underserved market segments; 3) Low interest rates; 4) Risk diversification.

3.2 Features

One of the main advantages of P2P lending for borrowers can sometimes be better rates than traditional bank rates can offer. The advantages for lenders can be higher returns than obtainable from a savings account or other investments, but subject to risk of loss, unlike a savings account. The interest rates may also have a lower volatility than other investment types. Peer-to-peer lending also attracts borrowers who, because of their credit status or the lack thereof, are unqualified for traditional bank loans. Because past behavior is frequently indicative of future

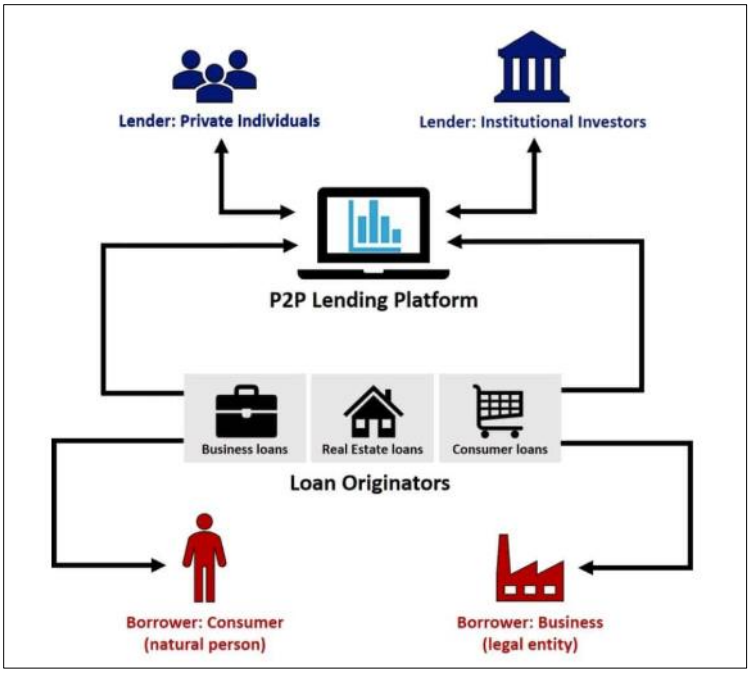
performance and low credit scores correlate with high likelihood of default, peer-to-peer intermediaries have started to decline a large number of applicants and charge higher interest rates to riskier borrowers that are approved. It seemed initially that one of the appealing characteristics of peer-to-peer lending for investors was low default rates, e.g. Prosper's default rate was quoted to be only at about 2.7 percent in 2007. Since inception, Lending Club's default rate ranges from 1.4% for top-rated three-year (term=36 months) loans to 9.8% for the riskiest loans. Because, unlike depositors in banks, peer-to-peer lenders can choose themselves whether to lend their money to safer borrowers with lower interest rates or to riskier borrowers with higher returns, in the US peer-to-peer lending is treated legally as investment and the repayment in case of borrower defaulting is not guaranteed by the federal government (U.S. Federal Deposit Insurance Corporation) the way bank deposits are. For investors interested in socially conscious investing, peer-to-peer lending offers the possibility of supporting the attempts of individuals to break free from high-rate debt, assist persons engaged in occupations or activities that are deemed moral and positive to the community, and avoid investment in persons employed in industries deemed immoral or detrimental to community.

3.3 Current state

The global P2P lending market has shown extraordinary growth rates since 2013, where systematic data collection in all regions of the world began. In 2017, the global P2P lending market funded loans for than $400 billion, but an important observation is that this loan volume is very much concentrated on a few large countries/regions. According to the data from the Cambridge Center for Alternative Financing, P2P lending is absolutely dominated by China and the Americas region (mainly the United States). China accounted for 87.0% of the overall market followed by the Americas region with 10.2%. However, the Chinese P2P lending industry has experienced massive turmoil in recent years with even more frauds and scandals surfacing since the 2017.

3.4 How an actual loan take place in P2P market

Loan originators take care of the demand-side (in blue) by providing loans to the platform, which enables platforms to focus their marketing only on attracting lenders/investors (the supply-side) (in red) (Fig. 1). Since the loan originator's loans are facilitated on the platform's marketplace, it is possible for the platform to remove the loan originator if it provides bad returns and instead try to find someone more reliant. This could happen if, for example, the borrowers the loan originator provides to the platform repeatedly do not pay back their loans. This will lead to investors losing money, which will force the platform to react because it must make sure investors see good returns to keep them on the platform.



Source: https://p2pmarketdata.com/p2p-lending-explained/

Fig. 1 P2P lending platform's role as an intermediator

Many marketplace lenders will let your check your rate and apply online. Typically, applying will only take a few minutes. Each lender will have different requirements. For personal loans, this includes your credit score, debt-to-income ratio, salary, employment status and credit history. For business loans, this includes your time in business, personal and business credit score, your debt service coverage ratio, revenue and profits. However, most lenders will only make loans to borrowers who are at least 18 years old and reside in a state they serve. You will also need a verifiable bank account and a Social Security Number. Once you submit the application, a lender may present you with a variety of loan offers. If you select one of these offers, you will generally need to submit to a hard credit check, which can affect your credit score. Most peer-to-peer lenders are quick to give you a loan decision, either same day or within a few days. Funding is also quick, with most borrowers receiving funds within two to 14 days.

3.5 Lending Club

LendingClub is now the world's largest online P2P lending marketplace, headquartered in San Francisco, California. The platform not only provides personal loans but also facilitates business loans and financing for elective medical procedures. LendingClub adopts a similar working and trading mechanism with Prosper's. The interest rates range from 5.6%-35.8%, depending on the loan term and borrower rating. The default rates vary from about 1.5% to 10% for the more risky borrowers. LendingClub enables borrowers to create unsecured personal loans between $1,000 and $40,000. The standard loan period is three years (36 months). Investors can search and browse the loan listings on LendingClub website and select loans that they want to invest in based on the information supplied about the borrower, amount of loan, loan grade, and

loan purpose. Investors make money from interest. LendingClub makes money by charging borrowers an origination fee and investors a service fee. Though viewed as a pioneer in the fintech industry and one of the largest such firms, LendingClub experienced problems in early 2016, with difficulties in attracting investors, a scandal over some of the firm's loans and concerns by the board over CEO Renaud Laplanche's disclosures leading to a large drop in its share price and Laplanche's resignation.

4. Literature Review

Given the rapid development of P2P lending and the availability of its transaction data, many research works have been done in the past.[1] Explaining loan default is an important issue because in P2P lending individual investors bear the credit risk, instead of financial institutions, which are experts in dealing with this risk.[2]. Factors explaining default are loan purpose, annual income, current housing situation, credit history and indebtedness.

In the last years several empirical studies have been made using data from P2P lending platforms. Ruiqiong and Junwen [3] perform a recent revision on empirical research. Factors explaining successful funding of loans is a widely researched topic. Lin, Prabhala and Viswanathan [4] study if borrowers' online friendships increase the probability of successful funding and its role in lowering ex post default rates. But they do not analyze the predictive capability or the accuracy of the model. Emekter, Tu, Jirasakuldech and Lu [5] evaluate the credit risk of P2P online loans, using Lending Club data, but they do not provide the model's accuracy. Gonzalez and Loureiro [6] study the impact of borrower profiles, focusing on borrowers' photographs and their results support the 'beauty premium' effect. Weiss, Pelger and Horsch [7] study credit bid's funding success, with similar results. They also study the factors explaining loan final interest rate. They study P2P loan bidding and find that the most important factor lenders use to allocate funds is the rating assigned by the P2P lending site. Traditional banks rely on risk analysts who approve hundreds of operations. By contrast, P2P borrowers and lenders are involved in a social network [8]. Lenders themselves analyze and select borrowers. Lee and Lee [9] and Zhang and Liu [10] analyze lenders behavior in P2P lending, finding strong evidence of herding behavior among lenders. Jiang, Wang and Wang [11] analyzed default prediction method for P2P lending combined with soft information related to textual description and their method can improve loan default prediction performance compared with existing methods based only on hard information.

The table (Fig. 2) shows studies specifically on default risk on lending club data by various authors and their brief.

Authors	**Data Used**	**Significant variables**	**Metrics used**
SerranoC, GutiérrezB, LópezL,2015 [12]	Lending Club data from 2008-2014	loan purpose, annual income, current housing situation, credit history, grade and indebtedness	Univariate means tests and survival analysis. Logistic regression model is developed to predict defaults
Jie Wana,b, Heng Zhanga,c , Xiaoqian Zhua	Lending Club data from 2016-2018	interest rate, loan amount, verification status, the ratio of monthly debt	Cox proportional hazard model to investigate the

, Xiaolei Suna , Gang Li, 2019 [13]		payments divided by monthly income	influential factors of P2P network loan prepayment risk
Michal Polena, Tobias Regner (2016) [14]	Lending Club data	The debt-to-income ratio, inquiries in the past 6 months and a loan intended for a small business are positively correlated with the default rate. Annual income and credit card as loan purpose are negatively correlated	Binary logistic regression, backward stepwise elimination

Fig. 2 Literature Review brief on Loan Default

This project also relates to finding default risk by predicting the default using logistic regression analysis and determining important factors for the same. Moreover, the project includes analyzing risk-return profile of loan portfolio which was not done in the above studies.

5. Data Analysis

5.1 Data description

In this project, the primary dataset has been taken from LendingCLub using Kaggle.com which provides data for loans accepted and loans rejected. I have specifically used the 'Accepted Loans' data. The dataset ranges from 2007 to 2018 and contains quite detailed information regarding the loans that LendingClub has issued during these years. P2P lenders suffer a severe problem of information asymmetry, because they are at a disadvantage facing the borrower. For this reason, LendingClub provides potential lenders with information about borrowers and their loan purpose. Apart from the unique ID of each loan it has variables (n=150) that states the loan amount, term, grade, home ownership, verification status, etc. It also tells the loan status which is vital for this project. It gives information such as if loan has been 'Fully Paid', 'Charged Off', 'Default', 'Late', etc. This whole dataset is a snapshot of a period and hence contains some loans as completed and some as currently going on.

5.1.1 Data Cleaning

Since The Lending Club was launched in 2012, there may have been significant changes in its loan portfolios since then. It is seen that the loan issuance started escalating after 2012 (Fig. 3) when the Lending Club was launched. Moreover, since it is required to analyze data for at least 5 years, I choose to train my model for 2012-2017 and test it for 2018.

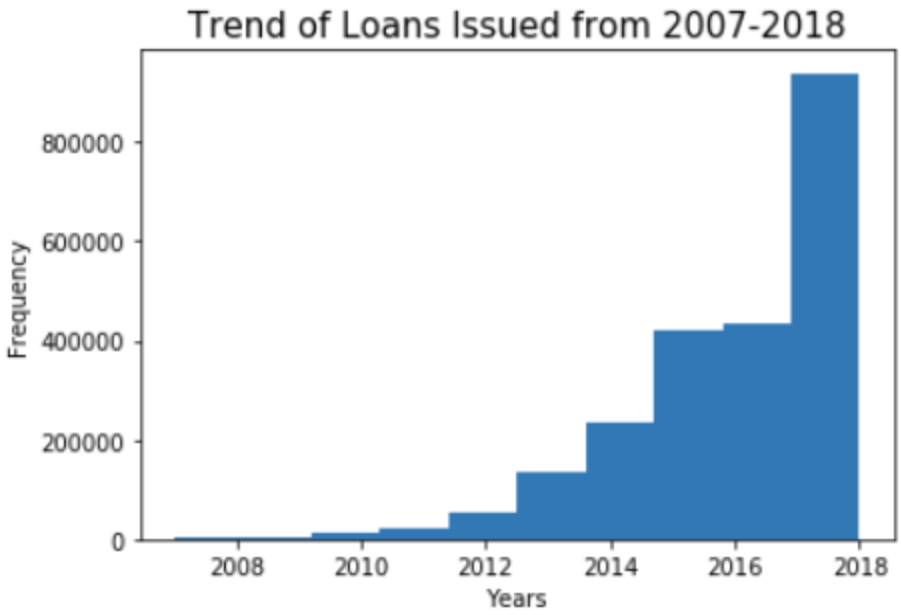


Fig. 3 Trend of Loans Issued from 2007-2018

Using the above figure, I clean data based on null values in year and filter it based on period 2012-2018. Coming to status of loans, the dataset contains both completed and intermediate loans. If one misses a payment, he gets 15 days period without fee i.e. grace period. After that he will see late. If he misses several payments, they classify him as default. If after they feel he won't pay anything it is charged off. For default prediction I take only the Fully Paid, Charged Off and Default as loan status since they are the final state. A loan must either be Fully Paid or be Charged Off/Default (if it takes too long to pay back). A loan payoff can be late but if at the moment of record, it is fully paid then it is fully paid. Therefore, data is filtered based on above 3 loan status only. Hence all the analysis so on is based on complete loans only.

Since the dataset had 151 variables in total, containing many irrelevant variables for this analysis, I first remove columns manually based on understanding them. I decide to drop several columns because they are a linear combination of others, or they contain too little information. For eg. for Zip Code, we only have the first 3 numbers, so this feature is not helpful. After that, I remove columns based on the percentage of null values in them. I keep threshold as 20% and on carefully analyzing them, I see none of those variables being useful. I hence remove those too. Rest of the variables (n=77) will be filtered based on either EDA or recursive feature elimination.

After filtering the data, 77 variables can be categorized into 2 types of variables, one, categorical variables such as grade, term, employment title, etc and second, numeric variables such as loan amount, interest amount, etc. The data description of some of the important categorical variables (Fig. 4) below shows us some useful information such as the unique values in the data. There are 11 categories for employment length, 7 for grades and 2 types of term. Moreover, data description of some of the important numeric variables (Fig. 5) below shows us some useful information such as the mean, standard deviation, quartile values in the data. It can be seen that

the average interest charges was 13.27%, whereas max interest rate charged was 30.99%. Such information gives a quick snapshot of a big dataset.

	count	unique	top	freq
grade	1228126	7	B	358326
id	1228126	1228126	14680062	1
term	1228126	2	36	926277
emp_title	1221711	356748	Teacher	21263
emp_length	1228126	11	10+ years	433310
home_ownership	1228126	6	MORTGAGE	612051
verification_status	1228126	3	Source Verified	496310
loan_status	1228126	3	Fully Paid	986102
purpose	1228126	14	debt_consolidation	719314
addr_state	1228126	51	CA	179406

Fig.4 Data description of categorical variables

	count	mean	std	min	25%	50%	75%	max
loan_amnt	1228126.0	14705.168912	8763.735751	1000.00	8000.00	12500.00	20000.00	40000.00
funded_amnt	1228126.0	14704.793279	8763.447110	1000.00	8000.00	12500.00	20000.00	40000.00
int_rate	1228126.0	13.270808	4.794273	5.31	9.75	12.74	16.02	30.99
annual_inc	1228126.0	78160.485692	71272.694759	0.00	48000.00	65000.00	93000.00	10999200.00
fico_range_low	1228126.0	695.540714	31.349915	660.00	670.00	690.00	710.00	845.00
fico_range_high	1228126.0	699.540844	31.350538	664.00	674.00	694.00	714.00	850.00
installment	1228126.0	446.591825	263.125163	4.93	255.14	382.55	592.73	1719.83

Fig. 5 Data description of numeric variables

5.2 Exploratory data analysis

Exploratory data analysis (EDA) can be useful to understand the trend of the variables and help to select the variables for the regression model. Therefore, EDA will be divided into two categories.

5.2.1 Understanding the data

Interest rate being an important factor in analyzing issuance and default of loan, I look at the change of rates through time. Each loan receives a grade that ranges from G (being highest interest rate) through A (being lowest interest rate). Starting November 2017, grade F and G are no longer used. The code will plot the average interest rate according to grate through time.

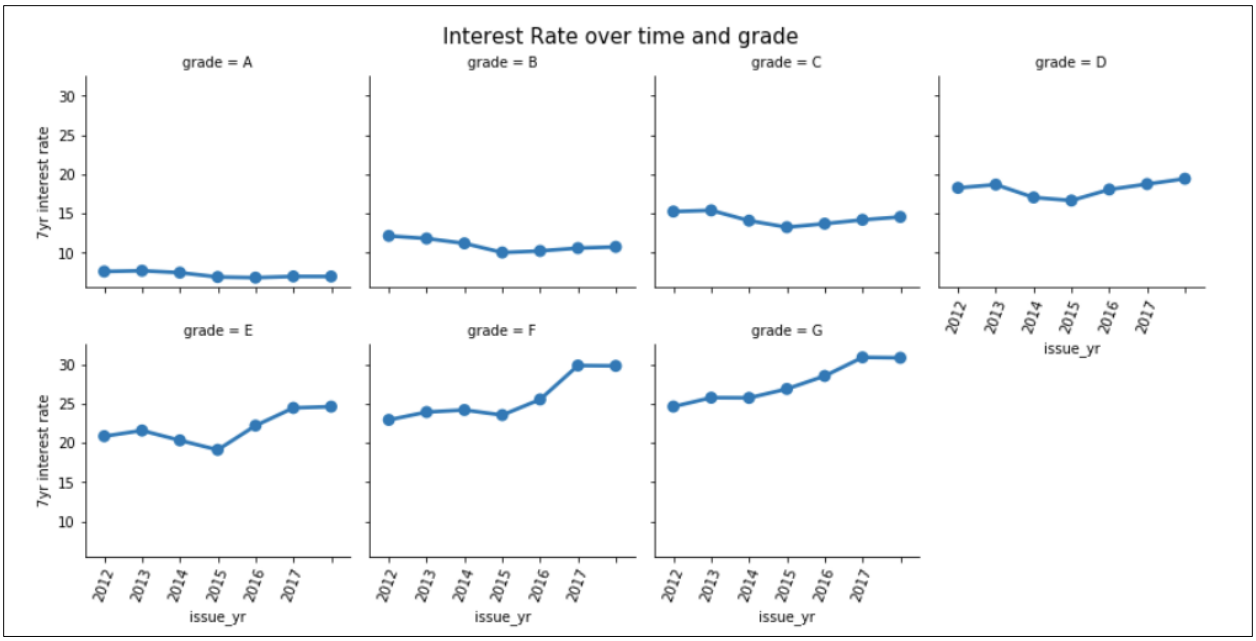


Fig. 6 Interest Rate over time and grade

From the chart above (Fig. 6), it can be analyzed that from year 2012 to 2018 (7 years) that interest rate of Grade G loan has been highest in 2018 and has been showing increasing trend. It is followed by Grade D, E and F too. We can also see that interest rate for these grades increase quickly from 2014. Whereas, Grade A loans have always been the lowest interest rate loans. This is an important factor which can depend upon borrower's credit history, nature of loan, etc. I have very few data points from 2012 to 2014. I take a closer look to see if this increase in average rate is not due to the small number of observations. The below table (Fig. 7) shows that grade G loans were issued in very less number in years 2012, 2013 and 2018. That can be one of the reasons for higher average rate for Grade G.

	int_rate						
grade	A	B	C	D	E	F	G
issue_yr							
2012	10753	16805	9902	5088	795	103	24
2013	17057	40313	24693	14505	3231	608	15
2014	35333	53460	44042	20510	7066	1980	179
2015	70118	91725	77423	32711	9443	1361	245
2016	47289	80413	66594	27671	7977	1951	466
2017	26326	40991	37788	16625	5295	1012	514
2018	12282	11901	9928	5724	1360	57	20

Fig. 7 Number of loans issues over time and grades

Coming to another variable, purpose of loan, I look at the number of loans issued categorized by its purpose (Fig. 8). Only 1 loan was given for educational purpose, whereas most of the loans issued lied under debt_consolidation. Debt consolidation refers to the act of taking out a new loan to pay off other liabilities and consumer debts, generally unsecured ones. Multiple debts are combined into a single, larger piece of debt, usually with more favorable payoff terms. It is also evident that sometimes people get stuck in the circle of taking one loan to pay another and never come out of it. Purpose of loan is another important factor to determine the default.

```
Number of loans issued categorised by its purpose
debt_consolidation    761666
credit_card           290148
home_improvement       84522
other                  73876
major_purchase         27239
medical                14861
small_business         13585
car                    13037
moving                  8897
vacation                8684
house                   6872
wedding                 1346
renewable_energy         830
educational                1
Name: purpose, dtype: int64
```

Fig. 8 Number of loans issued categorized by its purpose

I then analyze the issuance of loan based on employment length (Fig. 9). It can be determined that most of the loans have been issued to those people who had employment length of 10 or more years. Therefore, employment length is one of the factors used by LendingClub to issue loan. Further moving to employment title, I suppose it could have been one of the factors, but on analyzing, I find 30,6976 types of employment title and it cannot be efficiently used for modelling. Hence remove it.

```
Loan Issuance based on Employment Lenght in years
10    510748
2     117357
3     103504
0     103475
1      85248
5      80868
4      77114
6      60504
8      59219
7      57849
9      49678
```

Fig. 9 Loan Issuance based on Employment Length in years

5.2.2 Feature Selection based on logistic default variable

To run the logistic model, I first choose the 'Y' variable which the dependent variable and taken as loan status which has 3 elements Fully Paid, Charged Off and Default. I convert this variable into logistic variable as 0 & 1 where, '1' represents loan as 'No Default' i.e. Fully Paid and '0' represents loan as 'Default' i.e. Charged Off and Default. To choose the 'X' variables which are the independent variables for my model I further perform EDA. I transform a few variables into the date type for easier analysis such as issue date and earliest credit line to extract years form them.

Another important factor, that weather a borrower would be able to repay the loan or not is related to its annual income. Interest rate on loan is also dependent on income history of the borrower. Therefore, I look at the median of incomes in each grade (Fig. 10). Highest median income borrower came under Grade A loan which totally makes sense. Although the other grades do not show much variance, but this factor can be one of the important factors. To check it, I look at the average annual income for default & no default (Fig. 11).

grade	annual_inc
A	75000.0
B	65000.0
C	62000.0
D	60000.0
E	61000.0
F	64000.0
G	65000.0

Fig. 10 Median of incomes in each grade

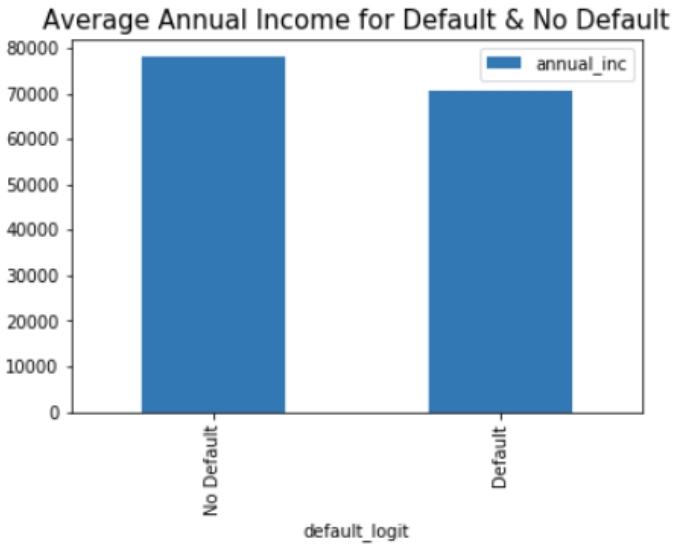


Fig. 11 Average Annual Income for Default & No Default

From the above figure (Fig. 11), it is clear that annual income ($77,956.48) of borrowers with no default was higher than the annual income of borrowers with default ($70,570.82). Therefore, it is an important factor for our regression model. I further analyze the trend of default state wise. I find the percentage of default as per total loans issued in each state. The below table (Fig. 12) show top and bottom 5 states with default rate. The state Iowa (IA) shows only 2 loans issues, therefore gives false information due to a smaller number of data. Top state is Mississippi with 26.1% and the bottom state is District of Columbia with 13.6% default rate. But on running the model, states give no results in predicting default and therefore, I ignore this variable for prediction. It can be used to just look at the trend or for some other analysis.

State	Default_cases	Total_issued	Perct_default
IA	1	2	50.000000
MS	1716	6569	26.122698
NE	900	3581	25.132645
AR	2392	9801	24.405673
AL	3872	16161	23.958913
OK	2842	11981	23.720891
State	Default_cases	Total_issued	Perct_default
NH	915	6277	14.577027
OR	2293	15959	14.368068
VT	364	2598	14.010778
ME	281	2027	13.862852
DC	444	3261	13.615455

Fig. 12 Percentage of default as per total loans issued in each state

Since loans are also based on the borrower's FICO scores which is an indicator of his credit history, I analyze the average high and low FICO scores they had based on default and no default

(Fig 13 & 14). It is clearly seen that the borrowers who defaulted had bother FICO high and FICO low scores, lower than as compared to borrower who did not default.

	default_logit	fico_range_low	fico_range_high
0	0	697.654548	701.654713
1	1	687.512104	691.512157

Fig. 13 Table of Average FICO score for Default & No Default

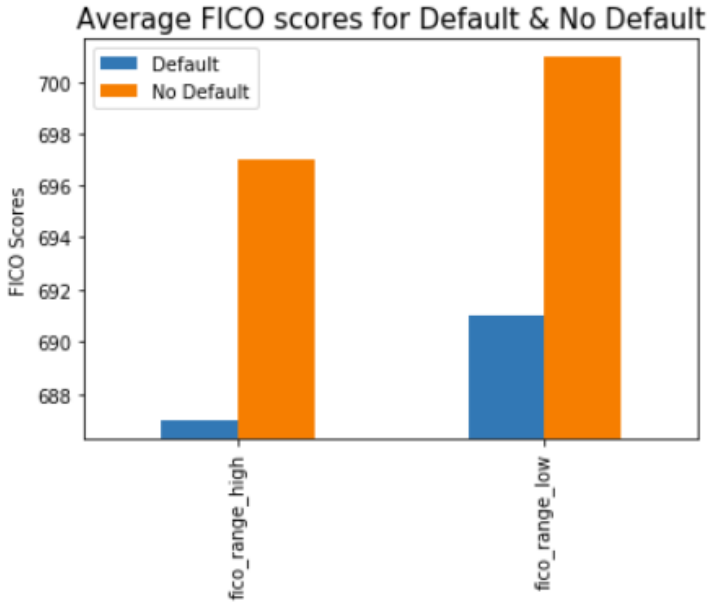


Fig. 14 Graph of Average FICO score for Default & No Default

Since there is a very clear difference between the scores, it can be one of the important variables for the model. Another factor analyzed is 'Term' of the loan. As LendingClub's standard loans are 3 months loan i.e. 36 months loan, we see 60 months loan too in the dataset. I therefore look at the percentage of Default & No Default cases by Term out of the total loans issued (Fig. 15).

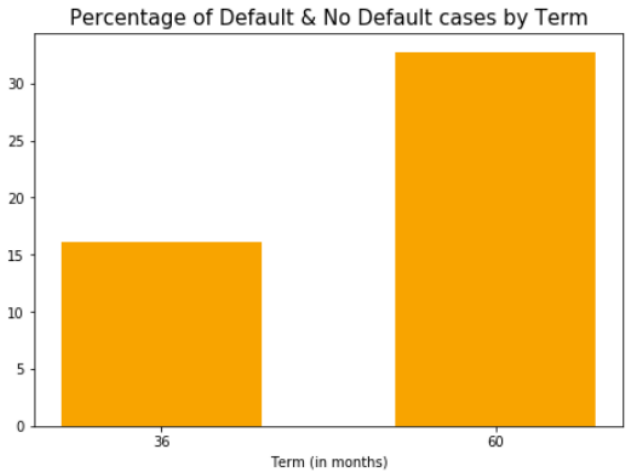


Fig. 15 Percentage of Default & No Default cases by Term

From the above graph (Fig. 15) it can be seen that 32% of the loans issued for 60 months of term were defaulted, whereas 16% of the loans issued for 36 months of term were defaulted. Therefore, longer term resulted in more defaults which can be another important variable for the model. Looking at the defaults, it is also important to see default cases based on grades since grades is an important factor. The table below (Fig. 16) show the number of default cases based on grades. For example out of total loans under Grade A (n=212,714), 200,475 (94.2%) loans were fully paid loans and if we look at loans under grade F & G, around 50% of the loans defaulted. Therefore grade is very essential to calculate default probability.

	default_logit	grade	count
0	0	A	200475
1	0	B	311707
2	0	C	273490
3	0	D	127722
4	0	E	52489
5	0	F	16052
6	0	G	4167
7	1	A	12239
8	1	B	46619
9	1	C	77543
10	1	D	55129
11	1	E	32984
12	1	F	13307
13	1	G	4203

Fig. 16 Default cases based on grades

5.3 Regression & Results

Through EDA, I could figure out some most useful variables for my regression model, but for the rest I used my own understanding of data and data documentation to filter out the variables. I used Logistic regression analysis to model default (Y variable) as a function of 14 variables (X variables): ('loan_amnt', 'term', 'int_rate', 'default_logit', 'grade', 'emp_length', 'annual_inc', 'home_ownership', 'verification_status', 'purpose', 'issue_yr', 'mort_acc', 'fico_range_low', 'fico_range_high', 'pub_rec'). Since some of the variables such as term, grade, purpose, verification status and home ownership are categorical variables, I have converted them to dummy variables to perform regression. I then finally drop the null values. The descriptive statistics (Fig. 17) of the final variables of the data sample is as below. It shows some of the useful information such as information about employment length of borrower, FICO scores, loan amount, interest rate, etc It shows the average interest rate was 13%, although highest interest rate charges was 31%. There is also a huge gap between average annual income ($74649) and highest annual income ($10999200). These factors do relate to default, high interest rate, low income, etc.

	count	mean	std	min	25%	50%	75%	max
loan_amnt	1305564.0	14517.140236	8734.198924	1000.00	8000.00	12000.00	20000.00	40000.00
int_rate	1305564.0	13.276636	4.792159	5.31	9.75	12.74	16.02	30.99
default_logit	1305564.0	0.201391	0.401040	0.00	0.00	0.00	0.00	1.00
emp_length	1305564.0	6.234525	3.702452	0.00	3.00	7.00	10.00	10.00
annual_inc	1305564.0	76469.076710	70092.326382	0.00	46000.00	65000.00	90044.25	10999200.00
issue_yr	1305564.0	2015.113276	1.438911	2012.00	2014.00	2015.00	2016.00	2018.00
mort_acc	1298069.0	1.670769	2.000441	0.00	0.00	1.00	3.00	51.00
fico_range_low	1305564.0	695.611950	31.547759	660.00	670.00	690.00	710.00	845.00
fico_range_high	1305564.0	699.612093	31.548436	664.00	674.00	694.00	714.00	850.00
pub_rec	1305564.0	0.220158	0.608895	0.00	0.00	0.00	0.00	86.00

Fig. 17 Descriptive statistics of variables/features.

Two things are to be done in regression analysis, first is to estimate/train the model for the time period of 2012-2017 and then testing this model for the data of 2018. Therefore I divide the data into two parts based on issue year.Before training the model for 2012-2017, I checked for the distribution of values of Logistic Default variable which is 0 (No Default) and 1 (Default). It can be seen (Fig. 18) that distribution is not balanced/scaled. Out of total, 988912 (79.6%) entries are 0 i.e. No Default and rest 252839 (21.4%) entries are 1 i.e. Default. It is therefore necessary to rebalance the distribution for better analysis; else the logistic model will interpret everything as 0 i.e. No Default.

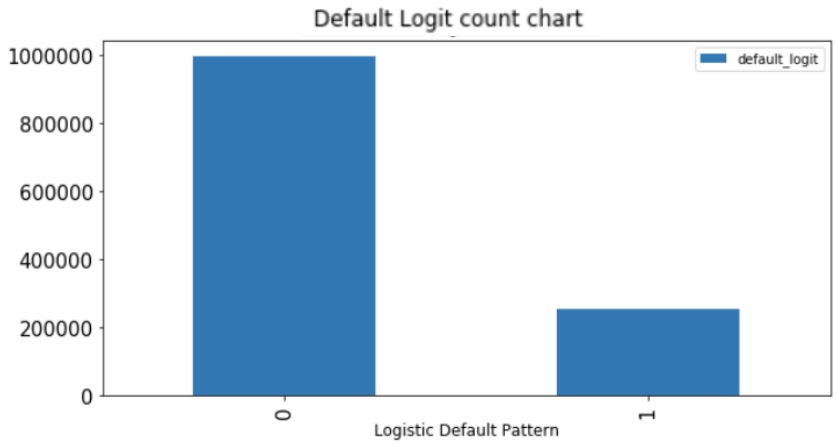


Fig. 18 Default Logit count chart

To rebalance the above discussed variable, I used imblearn library and use Synthetic Minority Oversampling Technique (SMOTE) function from it. SMOTE involves duplicating examples in the minority class i.e. 1 in our logistic variable, although these examples don't add any new information to the model. Instead, new examples can be synthesized from the existing examples. This is a type of data augmentation for the minority class. After using SMOTE, I have the balanced sample and below (Fig. 19) is the description of the same.

```
Length of oversampled data is  1977824
Number of no default in oversampled data 988912
Number of default 988912
Proportion of no default data in oversampled data is  0.5
Proportion of default data in oversampled data is  0.5
```

Fig. 19 Description of Logistic Default (Y- variable) after using SMOTE

Now I have a perfect balanced data. I over-sampled only on the training data, because by oversampling only on the training data, none of the information in the test data is being used to create synthetic observations, therefore, no information will bleed from test data into the model training. After rebalancing I use Recursive Feature Elimination (RFE) to eliminate the variables that are not of much importance for my model. RFE is based on the idea to repeatedly construct a model and choose either the best or worst performing feature, setting the feature aside and then repeating the process with the rest of the features. This process is applied until all features in the dataset are exhausted. The goal of RFE is to select features by recursively considering smaller and smaller sets of features. RFE gives results in True or False and also ranks the variables from low (most important) to high (least important). Since after making dummy variables, I had 41 variables, I only keep 25 for my model using RFE results. Although RFE did not give the variables I found relevant through EDA, I manually add 2 other variables (loan_amt & int_rate) to the final results. Therefore, the final 27 variables are:

'fico_range_low' 'fico_range_high'

’grade_A'	'purpose_car'
'grade_B'	'purpose_credit_card'
'grade_C'	'purpose_debt_consolidation'
'grade_D'	'purpose_home_improvement'
'grade_F'	'purpose_major_purchase'
'term_36'	'purpose_medical'
'term_60'	'purpose_moving'
'home_ownership_MORTGAGE'	'purpose_other'
'home_ownership_OWN'	'purpose_small_business'
'home_ownership_RENT'	'purpose_vacation'
'verification_status_Not_Verified'	'loan_amnt'
'verification_status_Source_Verified'	'int_rate'
'verification_status_Verified'	

After obtaining the balanced and the filtered sample, to train my model, I use sklearn library of python which fits the model and estimates the best representative function for the data points (could be a line, polynomial or discrete borders around). Since, sklearn does not give a summary of regression for r^2 or p-values, I just find intercept as [0.03152347] and coefficients as [-6.01082357e-02, 6.59002698e-02, -7.24831940e-01, -1.16094303e+00, -9.14533363e-01, -8.61929279e-01, -2.24946611e-01, -1.72301273e+00, -3.64125936e-01, -2.21558353e+00, -6.27963302e-01, -8.70880052e-01, -1.46761319e+00, -1.33125759e+00, -1.42563262e+00, -6.79549597e-02, -1.12756153e+00, -1.43573440e+00, -4.19701484e-01, -1.30709632e-01, -7.46807379e-02, -5.03212106e-02, -3.83772792e-01, -7.28954199e-02, -4.58774109e-02, 9.34810968e-07, 6.75865433e-02] for final variables in order above respectively. Coefficient tells us how much the dependent variable (y variable- Default) is expected to happen when that independent variable (x variables) increases by one/changes, holding all the other independent variables constant. With that representation and model, I can calculate new data points or say predict the y variable for given data points. I have used __ variables as x-variables and Logistic Default (0 & 1) as y-variables to train the model. I then use the model to predict the y-variable of the data for 2018 using the same x-variables of 2018 dataset and compare it with actual y-variable points for the dataset. I get the accuracy of logistic regression on test set of 2018 as 0.73 or 73%. It can be looked at the summary (Fig. 20) below.

	precision	recall	f1-score	support
0	0.88	0.78	0.83	43694
1	0.25	0.41	0.31	7690
accuracy			0.73	51384
macro avg	0.57	0.60	0.57	51384
weighted avg	0.79	0.73	0.75	51384

Fig. 20 Summary of Logistic Regression

Since the accuracy is not 100% i.e. the model does not predict fully we look at the confusion matrix (Fig. 21) for the different cases that would have happened and it can be clearly seen below that 34,295 (67%) of the total data points were true positive and 3,124 (6%) of the data points were true negatives i.e. were estimated correctly whereas rest of the others were false.

34,295 (True Positives)	9,399 (False Negatives)
4,566 (False Positives)	3,124 (True Negatives)

Fig. 21 Confusion Matrix

6. Analysis of a Loan portfolio

6.1 Return on randomly selected portfolio, How LendingClub's portfolio works

The return of lending is the interest of the capital lent. Interest rate depends on various characteristics/factors of the borrower such as annual income of borrower, FICO scores, type of loan, employment length, etc. Creditworthiness is also assessed with a statistical credit scoring model which classifies borrowers into different grades (from A to G) classes as discussed in above. The grades also reflects the probability of default like grade G has highest rate of interest as compared to grade A. Investing in loan portfolios is like any other investments which consists of risk-return. Apart from getting profit from interest on loans, monthly interests and repayments can be re-invested too earning more on that. On the other hand, investor must consider that all investment activities include a risk of partial or total loss of capital. Investing in consumer and business loans do not make an exception. Investment risk is a variation of the received returns compared to the expected return. The key risks of investing in loan portfolio is the payment default of a borrower (credit risk), received return differing from the expected return (prepayment risk), risk related to the time which is needed to sell a loan on the secondary market (liquidity risk), the foreign exchange risk when a borrower is living in a different currency zone and that Fellow Finance stops providing the service (market risk). One of the best ways to reduce credit risk is to diversify capital in numerous loans. A good rule of thumb is to acquire a loan portfolio containing at least 100 loans. This means that 1% of the total investment capital should be invested in one loan application. Diversification does not decrease the expected return but reduces volatility.

In this project, I analyze the risk-return of a loan portfolio from this market from an investor's perspective. From the dataset, loan can be categories in 4 categories based on type of

payment that was made for 2 major categories of 'Fully Paid' and 'Default'. Default has been categorized as default only, whereas 'Fully Paid' can be categorized as 'Prepayment' when loan was fully paid before the due date, resulting in interest loss; 'No prepayment' when loan was paid on time as installments were due, resulting in gaining full interest; 'Late payment' when some payments were made after due date with late fees but no principal amount was lost. The below graph (Fig. 22) shows the data based on these 4 categories. 64.6% of the loans lies under the category of prepayment, meaning there is loss is loss of interest in these cases. 19.7% of the loans were defaulted loans, 12.5% of the loans were fully paid with no prepayment and only 3.1% of the loans were under late payment.

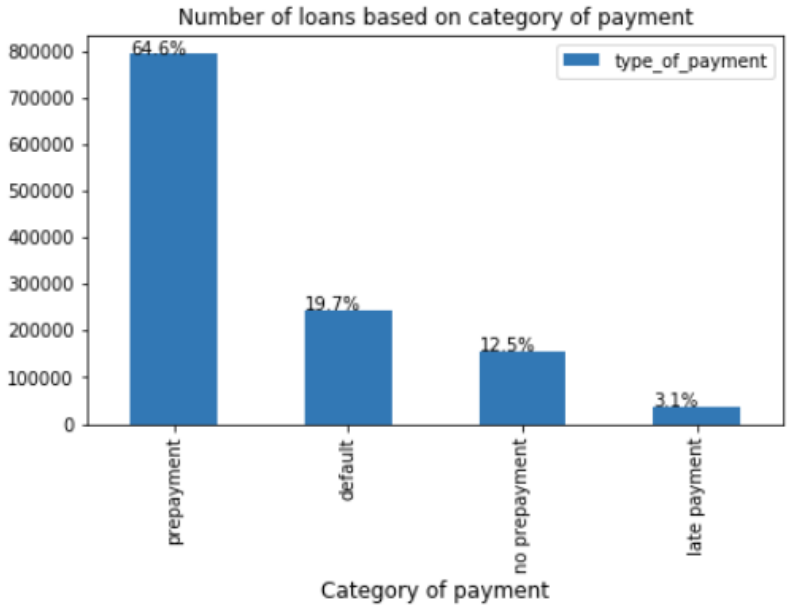


Fig. 22 Number of loans based on category of payment

I plan to construct a portfolio of randomly selected 100 loans. The characteristics of the portfolio consists of different grades, interest rates, etc. The risk/loss of/from the portfolio can be assessed based on 2 factors. First, the loans that have been defaulted will result in losing interest as well as principle amount. Second, the loans that have been prepaid will result in losing interest. The profit from portfolio will be the interest and the amount that would be received as installment can be reinvested again in the other loans. For analyzing returns of the portfolio, I calculate return on investment for each loan. It is calculated as taking out NPV of the cashflows where discounting rate is taken as the prime rate and dividing by funded amount. Negative returns will automatically take into account the loss due to default or loss of interest. The weights of the loans are based on the funded amount of the loan. Below graph (Fig. 23) show the returns of randomly selected 100 loans and their returns classified by grades. It is clear that in each grade returns are negative. Reason being most the loans were prepaid and includes defaults too. One important thing to note is that as Grade went from A to G, the returns varied largely which is understandable too.

	grade	ret_on_inv
0	A	-28.775458
1	B	-38.614792
2	C	-52.266827
3	D	-57.171535
4	E	-75.348867
5	F	-91.287076
6	G	-91.098028

Fig. 23 Returns of randomly selected 100 loans portfolio

It is also known that investor may lose more from default loan than what he can earn from good loans. Investor should see that he does not miss opportunity to give loan to good people (make profit from the interest) and on the other side does not give loan to people who will default (you lose interest + principle and how much you lose in bad loan i.e. after few payments or last payment or so on). But you could earn/charge more interest in such doubtful loans. Let's say, in bad loan you lose 100$ and in good loan u make 25$. Therefore, you are much more worried about bad loans and you can give 3 good loans if u have 1 bad loan. Investor needs to see this balance/proportion in his portfolio. Therefore, it is advisable to choose loans based on certain conditions such as specific grades, employment length, etc. In such portfolio, investor can do diversification and invest more percentage on possibly good loans to have safe/secured returns and also invest some percentage on loans such as grade D,E,F or G loans to get higher returns. An example for same can be seen in the chart below (Fig. 24). This portfolio consists of 100 loans based 3 filters, 'Grade' as 'A', 'term' as '36' and 'verification_status' as 'Verified'. The filters are based on EDA done previously. The returns differ very clearly and can be compared to the randomly selected loans based on no condition. Although this return to may not seem very high as loans in this portfolio too consists of some prepayments and defaults. Portfolio's risk and return can always differ from investor to investor.

	grade	ret_on_inv
0	A	5.944358

Fig. 24 Returns of conditionally selected 100 loans portfolio

6.2 How LendingClub's portfolio works

LendingClub allows investor to build its own portfolio. Instead of investing in an entire loan, they can invest in pieces of loans in $25 increments. When investor invests in a piece of a loan, LendingClub will issue them a Note in the amount of their investment with a stated interest rate. Investors may choose to spread their investment across hundreds or even thousands of loans in order to create a diversified portfolio. Using loan grades or other selection criteria, investor can balance the risks and returns of their investments and build a portfolio that suits

their goals. In addition to loan grade, investor can choose the term length (3 or 5 years), loan purpose, borrower location, and any other available criteria.

There are several methods for building one's own portfolio. Investor may use the Browse Loans feature to manually review and select loans currently available on the platform and build the portfolio Note by Note. One may also qualify to use LendingClub's Automated Investing service, an automated investing tool driven by investment criteria that he selects. Investor can change his investment criteria and pause or cancel Automated Investing at any time. Currently, there is no minimum balance on standard investing accounts. Investors may invest as little as $25 in a loan facilitated by LendingClub. Diversification, spreading investments evenly across many Notes, will result in more stable returns. LendingClub makes monthly principal and interest payments to investors as borrowers make payments on their loans. It is easy to reinvest the cash investor receives from monthly payments using Automated Investing or by manually placing orders for new Notes. Investor may withdraw his available cash at any time by scheduling a transfer to your bank account. Investing in loan portfolio is like investing in any other security portfolio. It is all about risk and return you get from the portfolio.

7. Conclusion

Through the literature study and data analysis, it can be concluded that lender can use various factors/variables that can help them to analyze the profit/loss on giving a loan. The regression model giving accuracy of 73% can help them analyze if the loan given will end as a default or no. Results show that, the higher the interest rate, the higher the probability of default is. The grade assigned by the LendingClub company is another important default predictor. Loan characteristics such as loan purpose; borrower characteristics like annual income, current housing situation, credit history and borrower indebtedness are related to default. However, other common drivers in default studies, such as loan amount or length of employment, have not a significant relationship with default within the data analyzed. Moreover, investing in a loan portfolio can be done as investing in any other security portfolio. Investor can choose his risk/return profile based on several condition while choosing his loans in the portfolio. Loans chosen wisely can help earn even more than securities such as investing in bonds. Collectively with all the research, it can be said that, loan market is as emerging market and investing in such markets is profitable too. Further research/modelling may include simulating an investment series by calculating cash flows for each month, set up triggers i.e. if you get certain amount of cash from the loans, reinvest on another loan. This way there can be continuous investment and earnings.

8. Resources:

1. Hongke Zhao, Yong Ge, Qi Liu, Guifeng Wang, Enhong Chen, and Hefu Zhang. 2017. P2P Lending Survey: Platforms, recent advances and prospects. ACM Trans. Intell. Syst. Technol. 8, 6, Article 72 (July 2017), 28 pages.
2. Serrano-Cinca C, Gutiérrez-Nieto B, López-Palacios L (2015) Determinants of Default in P2P Lending. PLoS ONE 10(10): e0139427. https://doi-org.ezproxy.rit.edu/10.1371/journal.pone.0139427

3. GAO Ruiqiong, FENG Junwen. An Overview Study on P2P Lending. Int Bus Manage. 2014;8(2): 14–18.
4. Lin M, Prabhala NR, Viswanathan S. Judging borrowers by the company they keep: friendship networks and information asymmetry in online peer-to-peer lending. Manage Sci. 2013;59(1): 17–35
5. Emekter R, Tu Y, Jirasakuldech B, Lu M. Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending. Appl Econ. 2015;47(1): 54–70.
6. Gonzalez L, Loureiro YK. When can a photo increase credit? The impact of lender and borrower profiles on online peer-to-peer loans. J Behav Exp Finan. 2014;2: 44–58.
7. 38. Weiss GN, Pelger K, Horsch A. Mitigating adverse selection in P2P lending: empirical evidence from Prosper.com; 2010. Available: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1650774.
8. Yum H, Lee B, Chae M. From the wisdom of crowds to my own judgment in microfinance through online peer-to-peer lending platforms Electron Commer R A. 2012;11(5): 469–483
9. Lee E, Lee B. Herding behavior in online P2P lending: An empirical investigation. Electron Commer R A. 2012;11(5): 495–503.
10. Zhang J, Liu P. Rational herding in microloan markets. Manage Sci. 2012;58(5): 892–912.
11. Jiang, C., Wang, Z., Wang, R. et al. Loan default prediction by combining soft information extracted from descriptive text in online peer-to-peer lending. Ann Oper Res 266, 511–529 (2018)
12. Serrano-Cinca C, Gutiérrez-Nieto B, López-Palacios L (2015) Determinants of Default in P2P Lending. PLoS ONE 10(10): e0139427.
13. Jie Wana,b, Heng Zhanga,c, Xiaoqian Zhua, Xiaolei Suna, Gang Li (2019) Research on Influencing Factors of P2P Network Loan Prepayment Risk Based on Cox Proportional Hazard
14. Michal Polena, Tobias Regner (2016) Determinants of borrowers' default in P2P lending under consideration of the loan risk class
15. https://www.kaggle.com/wordsforthewise/lending-club
16. https://www.marketwatch.com/press-release/at-502-cagr-p2p-peer-to-peer-lending-market-size-raising-to-usd-58905-billion-by-2025-2020-04-28#:~:text=Global%20Peer%20to%20Peer%20Lending,50.2%25%20over%20the%20forecast%20period.
17. https://en.wikipedia.org/wiki/Peer-to-peer_lending
18. https://en.wikipedia.org/wiki/LendingClub
19. https://www.fellowfinance.com/for-investor/risk-and-return
20. https://www.kaggle.com/shubhampali1993/default-loan-eda
21. https://www.lendingclub.com/investing/investor-education/what-is-the-process-to-begin-investing#:~:text=Build%20your%20portfolio.,with%20a%20stated%20interest%20rate.
22. https://p2pmarketdata.com/p2p-lending-explained/