**Project report on**

**“P2P REVENUE OPTIMIZATIONFOR LENDERS”**

**Submitted towards partial fulfilment of the criteria**

**for award of PGPBABI by Great Lakes Institute of Management**

**Submitted By**

**Group No. 10 [Batch: 2019-20 Location: Bangalore]**

**Group Members**

|  |
| --- |
| 1. **Saurav Banerjee - UMD99XSVGR** |
| 1. **Raunak Rudra - N7OLAWANQZ** |
| 1. **Shraddha Chaudhari - BM7D62U49L** |
| 1. **Thokchom Joychandra Singh - E8DLVOV0YE** |

**Research Supervisor:**

**Nimesh Marfatia, Mumbai**

**Great Lakes Institute of Management**



**ABSTRACT**

This study focuses on understanding and optimizing the factors affecting loan default in P2P Lending. There is a need for focus on this issue as lenders in P2P Lending bears the credit risk. P2P Lenders suffers a severe issue of information asymmetry and they are at a disadvantage while lending to borrowers. To solve this issue P2P Lending sites provides investors with relevant information related to borrowers and their loan purpose. A grade is also assigned to each borrower on the basis of their credit worthiness. The study is based on loans’ data collected from Lending Club from 2007 to 2017. Factor affecting loans default are Loan Amount, Loan Term, instalment, home ownership, annual income, credit history and indebtedness. Classification models are used to predict the default from the data and optimization technique is used to improve the lender’s performance.

Keyword: P2P Lending, personal loans, debt accumulation loans, credit risk

* Techniques: Classification Techniques, Probability of Default Credit Risk Model
* Tools: Python, Excel
* Domain: Lendingclub.com

**Acknowledgement**

We wish to place on record our deepest appreciation for the guidance and help provided to us by our Mentor Mr. Nimesh Prashant Marfatia, Mumbai. Mr. Nimesh Marfatia helped us narrow down on the choice of the Project as well as the scope and focus area of the Project. He gave us valuable feedback at every stage to enhance the process and the output.

We would also like to place on record our appreciation for the guidance provided by P. V. Subramaniam for giving us valuable feedback and being a source of inspiration in helping us to work on this project.

We certify that the work done by us for conceptualizing and completing this project is original and authentic.

Date: April 01, 2021 Saurav Banerjee

Place: Bangalore Raunak Rudra

Shraddha Chaudhari

Thokchom Joychandra Singh

**Certificate of Completion**

I hereby certify that the project titled “**P2P Revenue Optimization for Lenders**”for case resolution was undertaken and completed under my supervision by Saurav Banerjee, Raunak Rudra, Shraddha Chaudhari and Thokchom Joychandra Singh of Post Graduate Program in Business Analytics and Business Intelligence (PGPBABI).

Nimesh Prashant Marfatia

**Date: 31-03-2021**

**Place: Mumbai**

**Table of Contents**

[1. Introduction 7](#_Toc68187771)

[2. Scope & Objectives 9](#_Toc68187772)

[3. Data Sources and Description 10](#_Toc68187773)

[4. Data Pre-processing 11](#_Toc68187774)

[5. Exploratory Data Analysis 16](#_Toc68187775)

[6. Modelling Approach 24](#_Toc68187776)

[7. Creating Probability of Default (PD) Risk Based Model 33](#_Toc68187777)

[8. Investment Portfoliousing Optimization Strategies 35](#_Toc68187778)

[9. Actionable Insights and recommendations 37](#_Toc68187779)

[10. References and Bibliography 38](#_Toc68187780)

**List of tables and graphs**

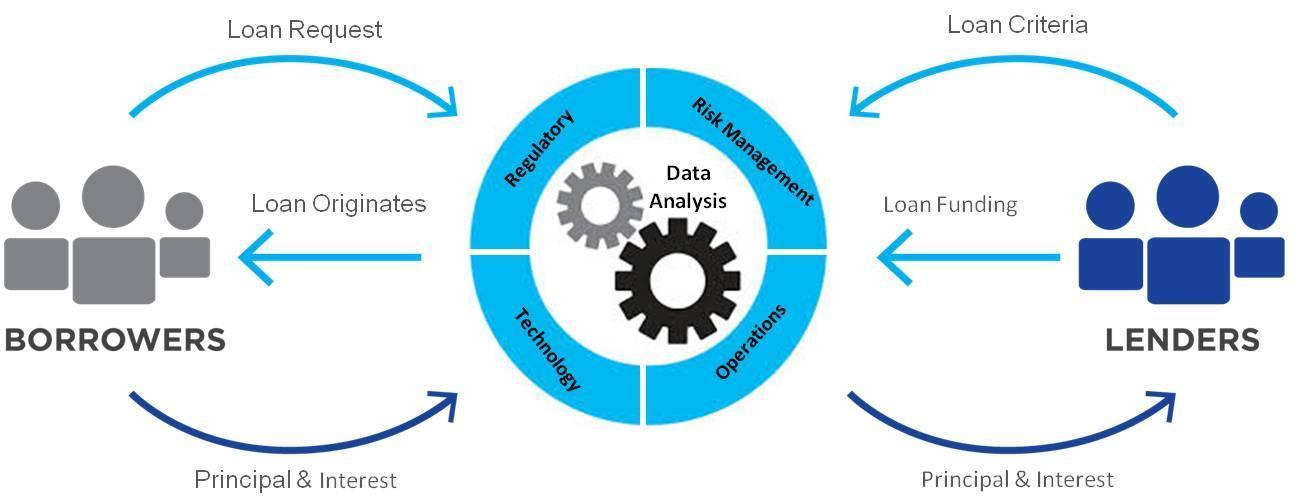
|  |  |  |
| --- | --- | --- |
| Fig No. | Graph Name | Page No |
| 1.1 | P2P Lending Cycle | 8 |
| 4.1 | Multi-Correlation Plot | 13 |
| 5.1 | Loan trends over years | 17 |
| 5.2 | Loan Amount Distribution | 17 |
| 5.3 | Loan Amount by Loan Status | 18 |
| 5.4 | Loan Term vs. Loan Status Plot | 18 |
| 5.5 | Interest Rate Histogram | 19 |
| 5.6 | Interest Rate vs. Loan Status Plot | 19 |
| 5.7 | Employee Length Vs Loan Status Plot | 20 |
| 5.8 | Home Ownership vs. Loan Status Plot | 20 |
| 5.9 | Annual Income vs. Loan Status | 21 |
| 5.10 | Verification Status vs Loan Status Plot | 21 |
| 5.11 | Purpose Plot | 21 |
| 5.12 | Grade vs. interest rate, dti | 22 |
| 5.13 | State-wise loan distribution | 23 |
| 5.14 | State wise Avg Interest rate | 23 |
| 6.1 | AUC plot for Decision Tree | 27 |
| 6.2 | AUC plot for LR | 28 |
| 6.3 | AUC plot for Random Forest | 29 |
| 6.4 | AUC plot for Naïve Bayes | 30 |
| 6.5 | LDA Plot | 30 |
| 6.6 | AUC plot for LDA | 31 |

|  |  |  |
| --- | --- | --- |
| Table No. | Table Name | Page No |
| 4.1 | Data Description | 13 |
| 4.2 | Data Imbalance (Before Balancing) | 15 |
| 4.3 | Data Imbalance (After Balancing) | 16 |
| 6.1 | Pearson Correlation Table | 25 |
| 6.2 | VIF Result | 26 |
| 6.3 | Confusion Matrix for Decision Tree | 27 |
| 6.4 | Confusion Matrix for Logistic Regression | 28 |
| 6.5 | Confusion Matrix for Random Forest | 29 |
| 6.6 | Confusion Matrix for Naïve Bayes | 30 |
| 6.7 | Confusion Matrix for LDA | 32 |
| 6.8 | Model Performance Comparison | 33 |
| 7.1 | Calculated Probability | 34 |
| 7.2 | Calculated score | 35 |
| 8.1 | Optimization Parameters | 36 |

# Introduction

**Peer-to-peer** lending, also abbreviated as **P2P** lending, is the practice of [lending](https://en.wikipedia.org/wiki/Loan) money to individuals or businesses through online services that match lenders with borrowers. Peer-to-peer lending companies often offer their services online, and attempt to operate with **lower**[**overhead**](https://en.wikipedia.org/wiki/Overhead_(business)) and provide their services more **cheaply** than traditional financial institutions. As a result, lenders can earn higher returns compared to savings and investment products offered by banks, while borrowers can borrow money at lower interest rates, even after the P2P lending company has taken a fee for providing the match-making platform and credit checking the borrower. There is additional risk of the borrower defaulting on the loans taken out from peer-lending websites.

**P2P Lending Cycle**2



**Fig 1.1: P2P Lending Cycle**

In India, the leading credit scoring agency is Credit Information Bureau India Limited - CIBIL3.Other credit score providers in India are Equifax4, Experian5 and CRIF High Mark[[1]](#footnote-2). **Most of the lenders in India consider CIBIL score as the standard for providing credit to a customer.** With traditional banks and financial institutions denying more and more personal and small business loans, people are becoming more interested in non-traditional avenues of finance such as online peer to peer lending. For example, if your bank denies you a loan due a procedural and theoretical approach you could go to an online P2P Lending platform7 to get assessed and get a quick and hassle free loan.P2P Lending platforms offer help to a segment of the population that otherwise might not be able to get a loan the traditional way, without any hidden costs.

This is an industry that’s gaining popularity rapidly in India as there are already **21 platforms operating** in India which has been granted licenses by the Reserve Bank of India8. The reason for the massive popularity of these platforms is that P2P lending is a great investment opportunity for people looking to invest their savings and earn great returns, also a great way for creditworthy borrowers to secure quick and easy personal loans.

TransUnion CIBIL Limited is a [credit information company](https://en.wikipedia.org/wiki/Credit_bureau) operating in [India](https://en.wikipedia.org/wiki/India) which maintains credit files on **600 million** individuals only and P2P lending can help worthy borrowers even if they do not have credit files. We are taking US lending data for this study and approximately **14%** of the population (331 million) in USA10 has no credit score whatsoever, and is labelled as credit invisible. As a result, these under banked individuals will have difficulty obtaining new lines of credit but can be greatly benefited from peer-to-peer lending platforms.

Peer to peer lending offers a **5-12% return**9 for the average retail investor, depending on their level of activity, skill, and risk tolerance but there is no consensus as it is affected by external factors such as economic condition and platform’s maturity and structural changes. The Overall interest rate charged for a loan can be seen as the loan's overall interest rate when the base interest rate and origination fees are added together. Base interest rate is determined based on various socio-economic factors of the borrower and drives monthly loan payment whereas origination fee is the fee that P2P lending platform charges when a loan is availed through them, and it is the main way they make money as a company.

A break-up of the various factors that constitute the P2P lending rate below:

1. ***Past Performance*:** Past repayment behaviour of borrower, his diligence in making utility payments and past loans may determine his intention and ability to repay.
2. ***Financial ability*:** Markers like income, sources of income, dependants, earning members and expenditure patterns show the financial ability.
3. ***Macro-economic variables*:** Economic condition, unemployment rates, gross domestic product (GDP) and consumer price Index.
4. ***Other Factors*:** Other factors such as credit utilization, Maximum current balance owed on all revolving accounts, demography etc.

From the above listed parameters, it can be observed that there are opportunities where based on borrower’s profile a model can be derived to optimize the loan parameters such as loan duration, interest rate and amount to **maximize the ROI for Lender**.

**Our study is motivated by need to cluster and profile the borrowers based on various socio-economic factors to reduce delinquency and loan default rates and to maximize the returns for the lenders.**

-----------------------------------------------------------------------------------------------------------------------------------------------

8https://www.rbi.org.in/Scripts/BS\_NBFCList.aspx

9https://www.lendingmemo.com/

10https://www.valuepenguin.com/

# Scope & Objectives

The discussion about the interest rate settings mechanism in crowdfunding (P2P) platforms is interesting from a risk-return perspective. There are two models which are widely popular - posted prices pricing and reverse auction models.

In **posted prices pricing models**, crowdfunding platforms also set the interest rate for each borrower based on the traditional risk-return perspective and by trying to tailor the price based on the risk of the borrowers. The other method is **reverse auction** where borrowers create loan listings, specifying the amount of money they want to borrow, and the highest rate they are willing to pay. Potential lenders vet the various borrowers for credit worthiness and establish (but don’t divulge) the lowest rate at which they would fund the loan. The auction starts at the lender’s reserve rate and continues as lenders bid lower and lower — in effect, a reverse auction. In a reverse auction process[[2]](#footnote-3)however, risk and return are evaluated by the investors and it is not clear if these, usually not professional investors, correctly evaluate the risk-return trade-off in their financial decisions.

**Scope for the project**:

* The data used for this project is based on the Lending Club[[3]](#footnote-4) dataset for the year 2007-2017
* Loan Default rates and their subsequent impact on the model is limited to the information available in the dataset and does not take into consideration extraneous economic factors.

**Outside Scope**

* This model is based on US data and fields, it would not apply to India and other markets. This can be extended to Indian market by considering local P2P platform’s data.

The **Objectives of project** are:

* To understand the **relationship between individual data** such as credit score, length of employment, job title, gross income and location and their propensity to default.
* Create a classification model for loan defaulters vs. non defaulters.
* Developing a model which will **maximize the Return of Investments for lenders** and lending platforms.
* Grouping borrowers into risk buckets to maximize the return for lenders based on factors such as FICO range, loan term (36- or 60-month loans), loan purpose and loan amount, etc.

# Data Sources and Description

All our unique borrower-specific data stem from lending club, the 2nd biggest player in the US P2P lending market and has over $59 billion dollars in loans issued as of March 2020. 67.42% of Lending Club borrowers report using their loans to refinance existing loans or pay off their credit cards.

We collected the data for Lending Club through their public APIs. We carry out the EDA analysis for our full sample for year 2007-2017. The sample for this dataset consists of information on approximately 1.7 million loans to private individuals granted between January 2007 and December 2017 and we are using this data for EDA.

To train the classification models we will use 2007-15 data as training data set and 2016-17 as a testing data set.

# Data Pre-processing

The dataset comes with 151 features but we will consider total of 20 features out of which there are 20 predictor variables and one dependent variable which is loan\_status. These features include loan information, application type and borrower’s financial and demographic information.

Below is the data dictionary for all the 151 variables

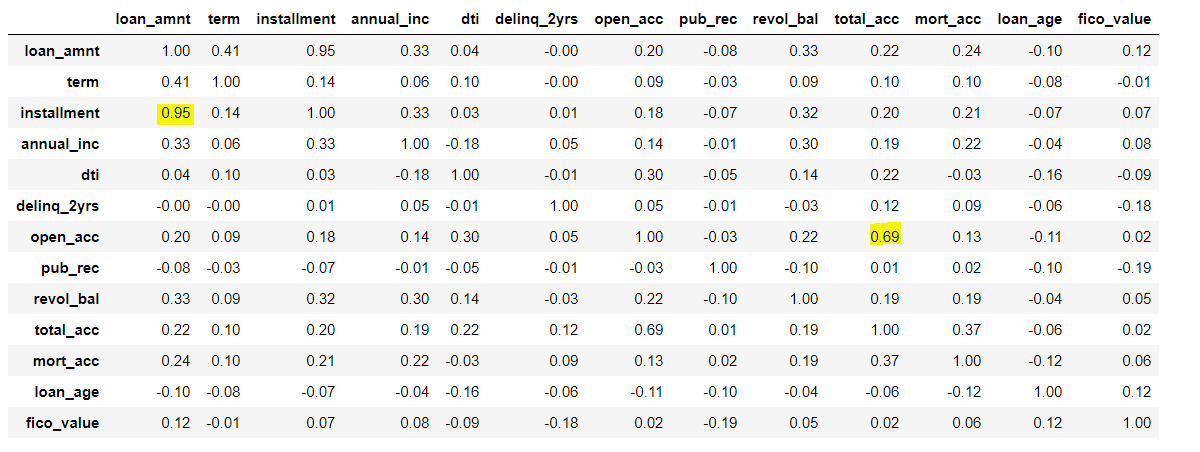


|  |  |
| --- | --- |
| **Data Field** | **Description** |
| addr\_state | The state provided by the borrower in the loan application. |
| annual\_inc | The self-reported annual income provided by the borrower during registration. |
| delinq\_2yrs | The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years. |
| fico\_range\_high | The upper boundary range the borrower’s FICO at loan origination belongs to. |
| fico\_range\_low | The lower boundary range the borrower’s FICO at loan origination belongs to. |
| home\_ownership | The home ownership status provided by the borrower during registration. Our values are: RENT, OWN, MORTGAGE, OTHER. |
| Installment | The monthly payment owed by the borrower if the loan originates. |
| int\_rate | Interest Rate on the loan |
| loan\_amnt | The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value. |
| mort\_acc | Number of mortgage accounts. |
| Purpose | A category provided by the borrower for the loan request. |
| Term | The number of payments on the loan. Values are in months and can be either 36 or 60. |
| total\_acc | The total number of credit lines currently in the borrower's credit file |
| verification\_status | Indicates if income was verified by LC, not verified, or if the income source was verified |
| issue\_d | The month which the loan was funded |
| loan\_status | Current status of the 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. |
| 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 |
| initial\_list\_status | The initial listing status of the loan. Possible values are – W, F |
| application\_type | Indicates whether the loan is an individual application or a joint application with two co-borrowers |

**Table 4.1: Data Description**

1. Data Leakage

Data Leakage is a scenario where some information which will not be available in the real-life prediction is used to train the model, allowing a model or machine learning algorithm to be unrealistically accurate with predictions. This would mean that a variable in the data set is highly correlated with the target variable.



**Fig 4.1: Multi-Correlation Plot**

When a variable is highly correlated to the target variable, a model using this attribute is expected to lead to a highly predictive model. However, at the time of real model usage, this variable might not be available rendering it useless. In the lending club data, attributes related to payment that has been made till date are correlated to the loan status. So, a model using these attributes will likely result in a strong model performance.

In the current dataset, the total payments made on the loan will be highly correlated to other attributes such as loan default and loan early repayment. For example, if the loan is defaulted, the total payments will be less and hence this could be a strong predictor variable. However, while predicting in real time, this attribute will not be available as the loan has not been issued yet for any payments to be made. We will be excluding all payments related columns in the dataset.

1. Removed id, member id, URL column as they would not have any significant bearing on the model.
2. We have removed the Grade and Sub grade fields before running the model since we did not want the classification from Lending club of applicants to bias our independent classification model. The thought process is to take this as a future scope wherein we can perform clustering based on applicant’s demographic and financial data to create our own grades which then can be included in model.
3. Delete columns having 50% of missing data. This brought number of columns to 90.
4. Dropped rows having loan status as Current, In Grace Period, Late (31-120 days), Late (16-30 days), Default(only 3 records) as we wanted to focus on only 2 status Fully paid & Charged off.

Also replaced ‘Does not meet the credit policy. Status: Fully Paid status’ to Fully Paid

and ‘Does not meet the credit policy. Status: Charged Off’ to Charged Off in loan\_status column.

1. Created new column for loan age and dropped issue\_d column. Age has been calculated as difference between issue date of loan and January 1st 2021.
2. Convert fico score by taking average of fico\_range\_high and fico\_range low from original dataset.
3. Data type conversion

Each column in the source data has an associated data type such as numeric, categorical, or date-time feature. Python automatically detects the data type of each attribute. However, sometimes the data types inferred by Python are not appropriate. Ensuring data types are correct is important as several downstream steps depend on the data type of the features, for example: [missing value imputation](https://www.pycaret.org/missing-values) for numeric and categorical features should be performed separately. We ensured that assigned data types are correct.For example, ‘pub\_rec’, ‘open\_acc’, ‘total\_acc’, ‘mort\_acc’ were converted to integer data type from the original float data type

1. Data treatment

Columns such as out\_prncp, out\_prncp\_inv, policy\_code have only one unique value so dropping those columns. Also dropped insignificant columns based on EDA and other factors. These data treatments brought down the number of features to 20.

* loan\_amnt
* term
* installment
* home\_ownership
* annual\_inc
* verification\_status
* loan\_status (predictor variable)
* purpose
* addr\_state
* dti
* delinq\_2yrs
* open\_acc
* pub\_rec
* revol\_bal
* total\_acc
* initial\_list\_status
* application\_type
* mort\_acc
* loan\_age
* fico\_value

1. Handling data imbalance

An imbalanced classification issue is an example of a prediction problem where the distribution of examples across the known classes is biased or tilted towards one class. The distribution can vary from a small bias to a high level of disproportion. As the lending data deals with defaults, we know that there is bias in the data. Only 17.8% of the dataset relates to ‘Charged off’ and rest have been fully paid off. For the purpose of this analysis the Class of Interest is – ‘Charged Off’

Before balancing:

|  |  |
| --- | --- |
| Fully Paid | 82% |
| Charged Off | 18% |

**Table 4.2: Data Imbalance (Before Balancing)**

The next step is to remove the bias. We will be using the SMOTE from ‘Imbalance Learn’ package (imblearn) to remove bias. Using this, the minority class (loan default) is replicated to balance the data. This balances the data by randomly oversampling the minority class.

After Balancing

|  |  |
| --- | --- |
| Fully Paid | 50% |
| Charged Off | 50% |

**Table 4.3: Data Imbalance (After Balancing)**

# Exploratory Data Analysis

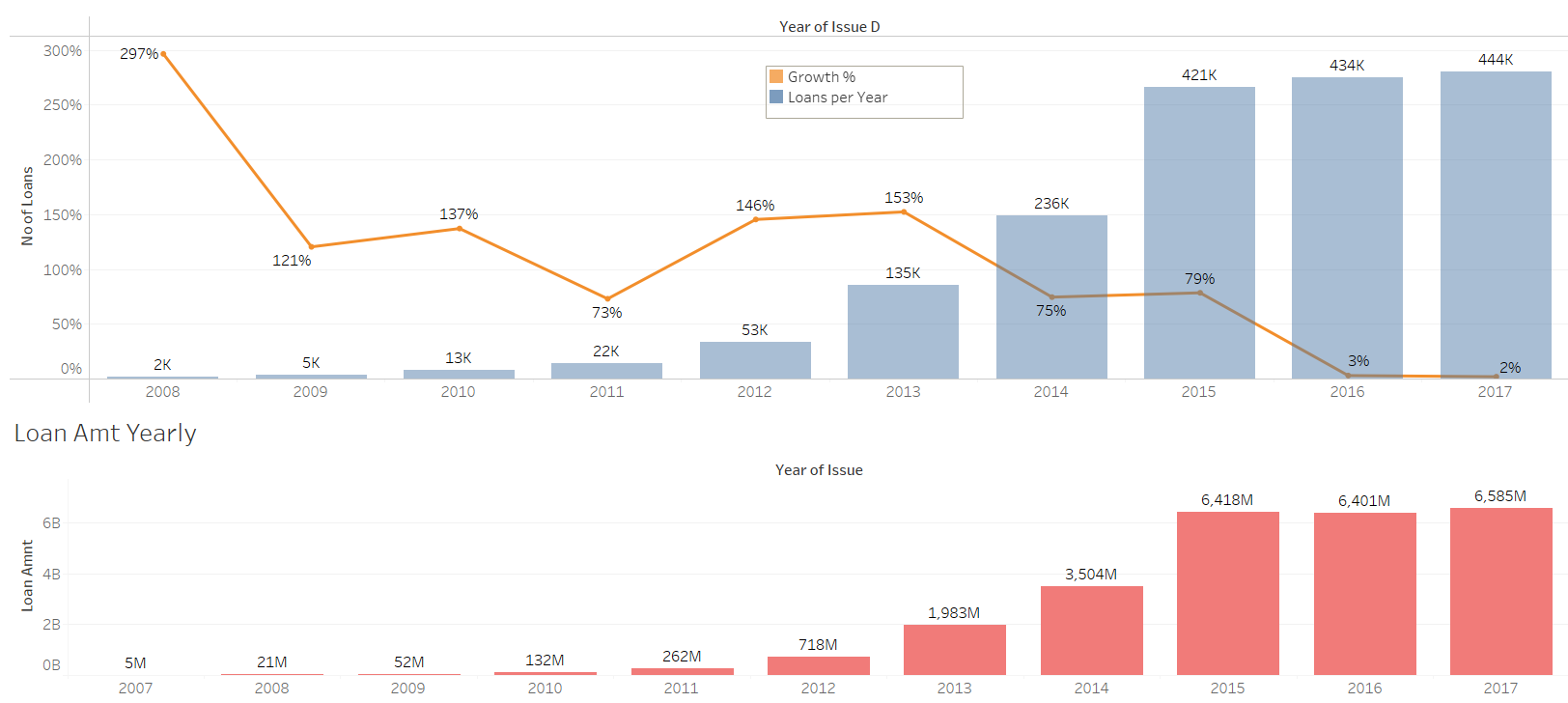
**Exploratory Data Analysis:**

A basic EDA is conducted to understand the structure of the data and identify useful insights.

EDA can be found at Tableau Public URL

<https://public.tableau.com/profile/raunak.rudra#!/vizhome/LendingClubEDA/EDA>

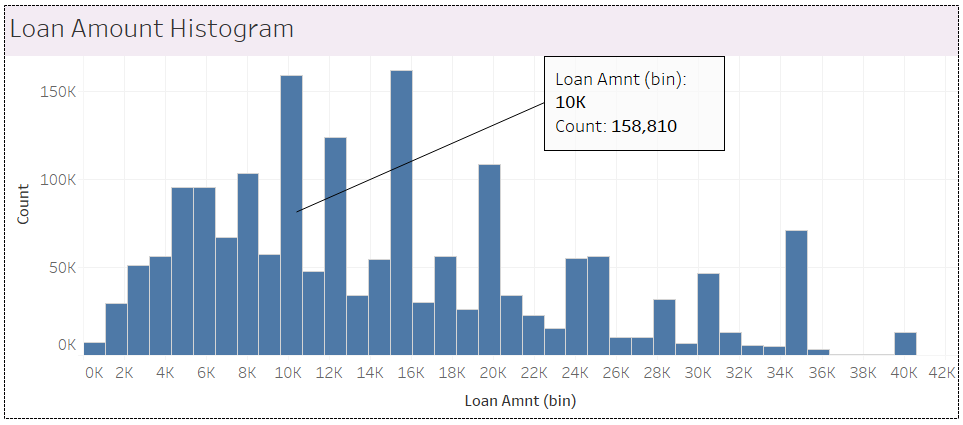
**Increase in loans per year but growth percentage has flattened**

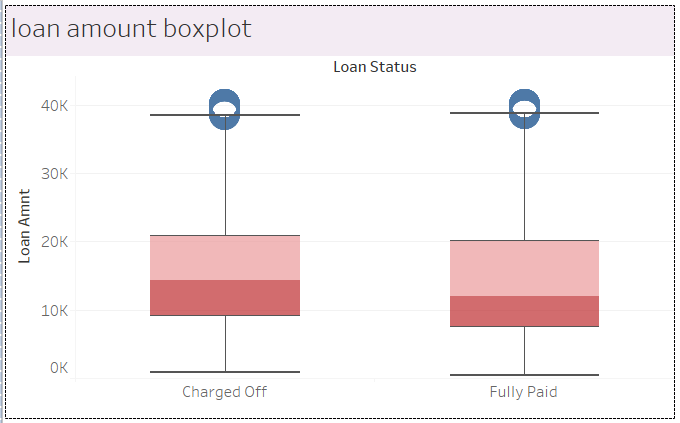
****

**Fig 5.1: Loan trends over years**

**Loan Amount:**

Loan amounts range from 1000$ to $40000 with a median of $12000 for fully paid loans and $14975 for charged off loans

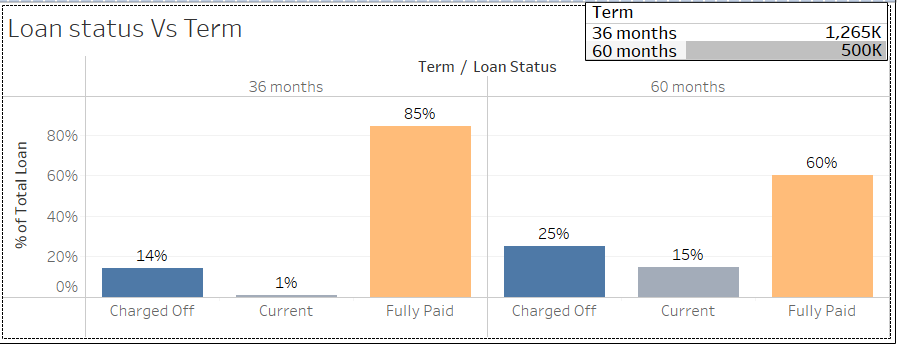
**Fig 5.2: Loan Amount Distribution**

****

**Fig 5.3: Loan Amount by Loan Status**

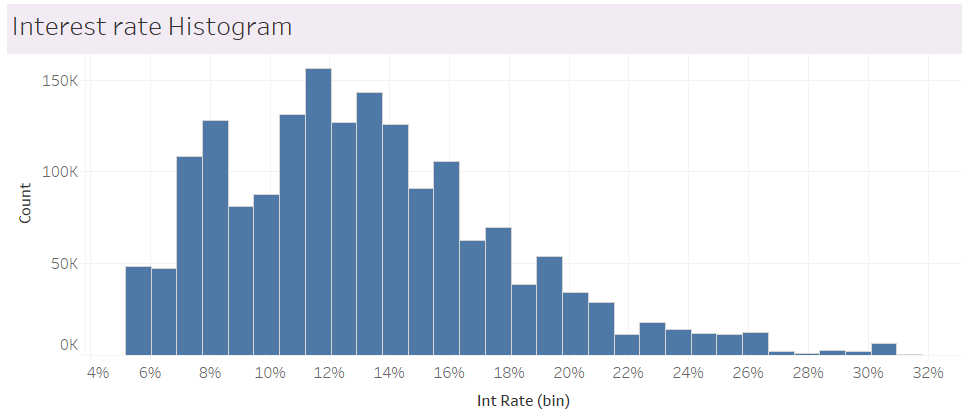
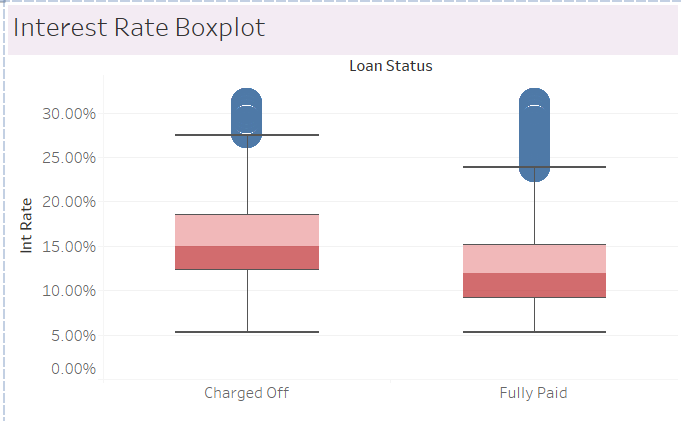
**Loan Term:**

Out of the total 1765K loans being granted, 1265K are of 36 months term and remaining 500K have taken loans for a term period of 60 months. There is slightly higher charge-off for 60 months term.

**Fig 5.4: Loan Term vs. Loan Status Plot**

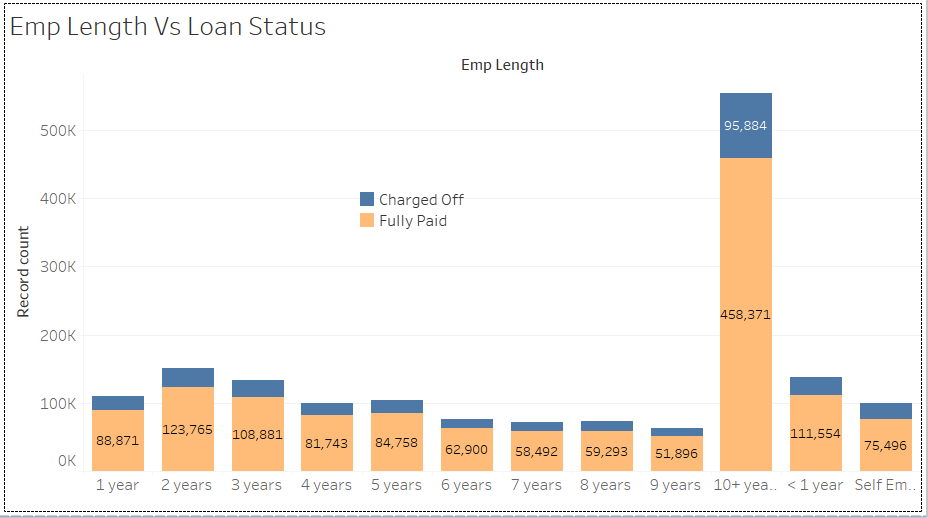
**Interest Rate:**

The interest rate of the loans ranges from 5% to 30 % with a median of 12% for fully paid loans and 15% for charged off loans

**Fig 5.5: Interest Rate Histogram****Fig 5.6: Interest Rate vs. Loan Status Plot**

**Employment Length:**

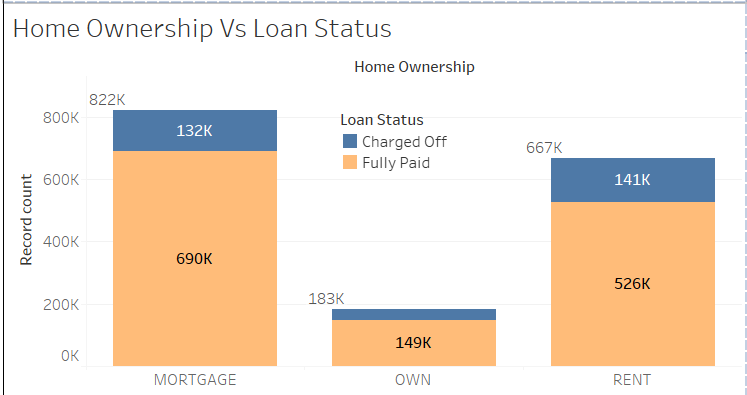
Highest number of loan takers has more than 10+ years of work experience. Employment time span doesn’t seem to have a significant correlation between employment lengths and charged off.

****

**Fig 5.7: Employment Length Vs Loan Status Plot**

**Home Ownership:**

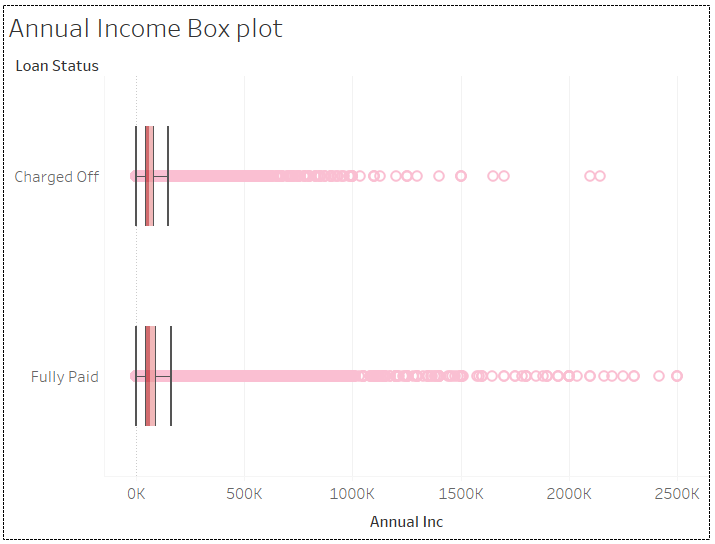
Home ownership here represents the home ownership status provided by the borrower during registration or obtained from the credit report. The values are: Mortgage, Own and Rent



**Fig 5.8: Home Ownership vs. Loan Status Plot**

Renters and mortgage payers constitute a major chunk of the borrower while home owner only constitute a small portion of the population.

**Annual Income:**

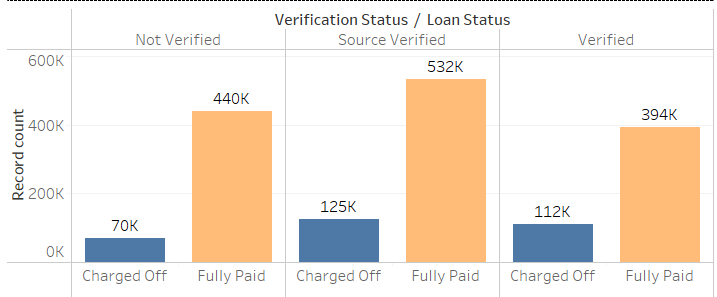
****

**Fig 5.9: Annual Income vs. Loan Status**

Annual income doesn’t have lot of impact on loan paying capabilities.

**Verification Status:**

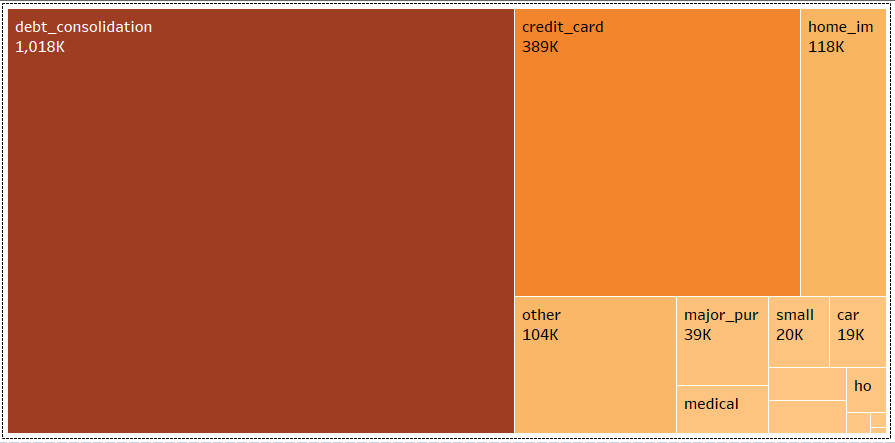
Verification status indicates if income was verified by [Lending Club], not verified, or if the income source was verified.



**Fig 5.10: Verification Status vs Loan Status Plot**

Charge off and verification does not seem to have any relation i.e., borrowers whose income source as well as income were verified still default.

**Purpose:**

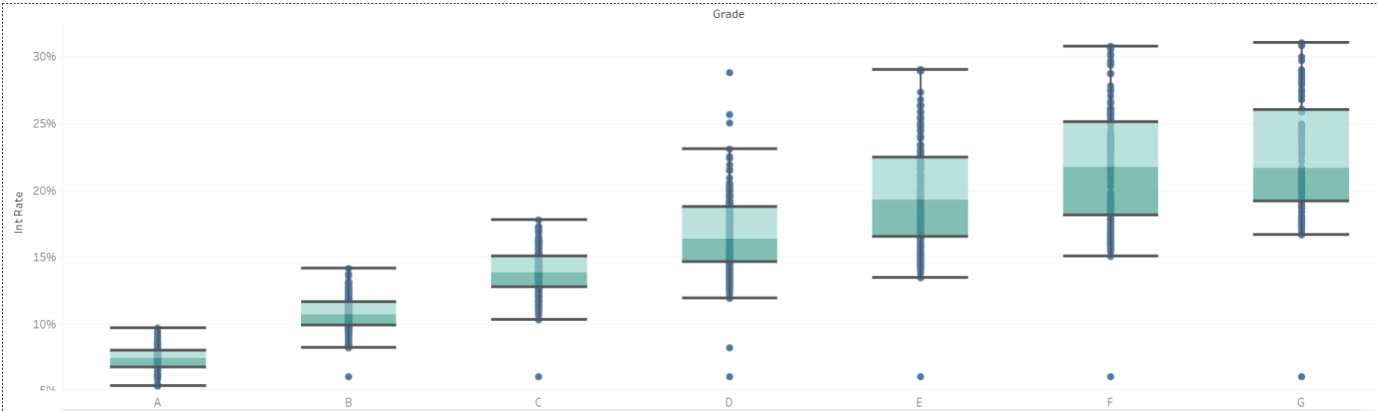
****

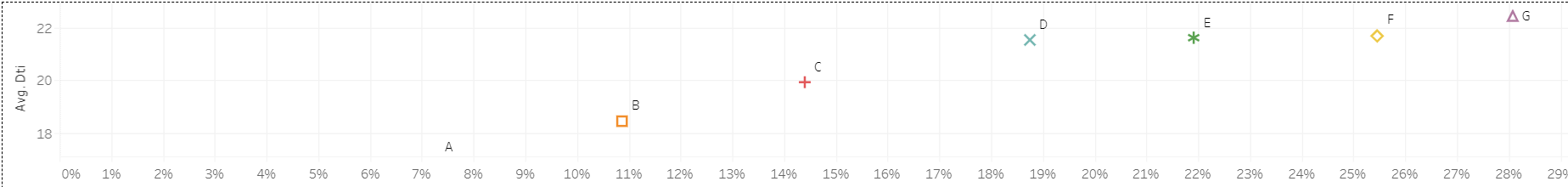
**Fig 5.11: Purpose Plot**

Most of the borrowers in the data have borrowed money for debt consolidation purposes

**Grade &dti box-plot**

Lower the grade, higher is the interest rate & debt to income ratio



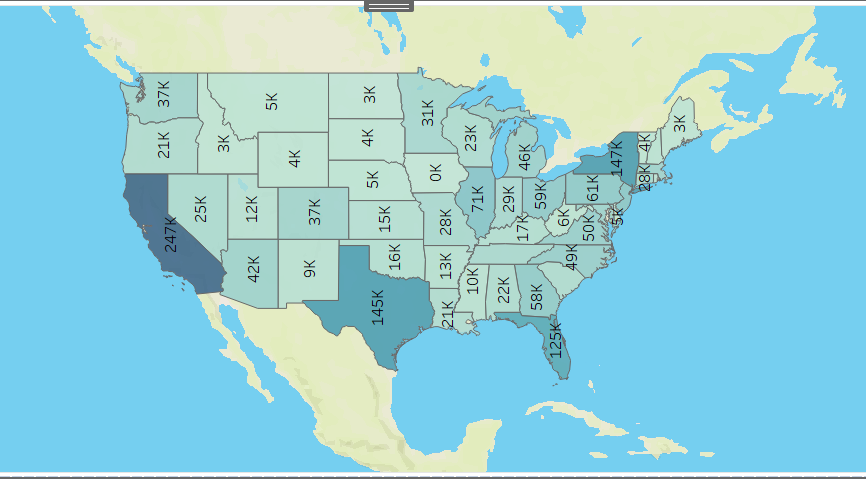


**Fig 5.12: Grade vs. interest rate, dti**

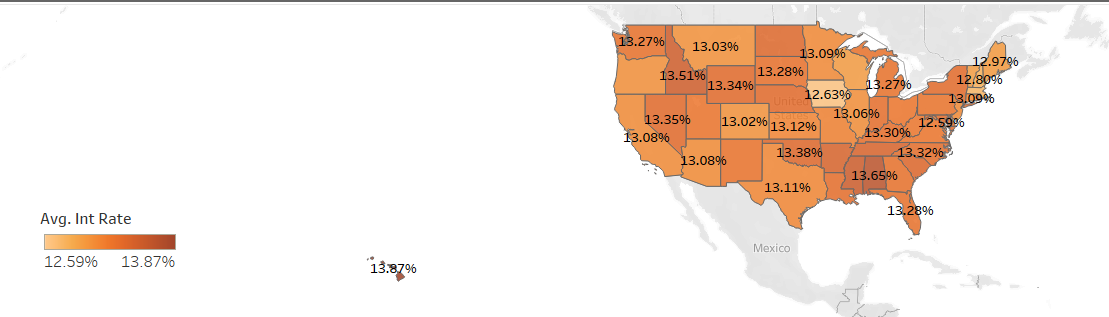
Grade B constitutes the highest number of borrowers with 62623 observations belonging to the grade. Whereas the number of defaulters in grade G is the highest with 37.33 % of the borrowers (420 out of 1125) defaulting which is the highest among all the grades

**State:**

California has the highest number of borrowers with a total of 385K borrowers. Hawaii has highest average interest rate of 13.77%



**Fig: 5.13: State-wise loan distribution**

****

**Fig 5.14: State wise Average Interest rate**

# 

# Modelling Approach

Using the data provided, we would implement models to predict the classification for loan defaults. We want to try the following models: decision tree, random forest, logistic regression, Naive Bayes and Linear Discriminant Analysis. As the dataset size was very large, fitting a Neural Network model was a resource intensive exercise and thus was not attempted.In this section we will explain how we selected different models and each model/modelling procedure was evaluated.

In discussing the evaluation of classification models, we will see how well do the model's estimates of class probability actually order the loans by their likelihood of default? This is measured by the area under the ROC curve (AUC), the AUC measures the following. Given two randomly selected loans, one that defaults and one that does not, what is the probability that the model will assign a higher default probability to the defaulting loan? A model that can perfectly discriminate default from non-default would have an AUC of 1.0.

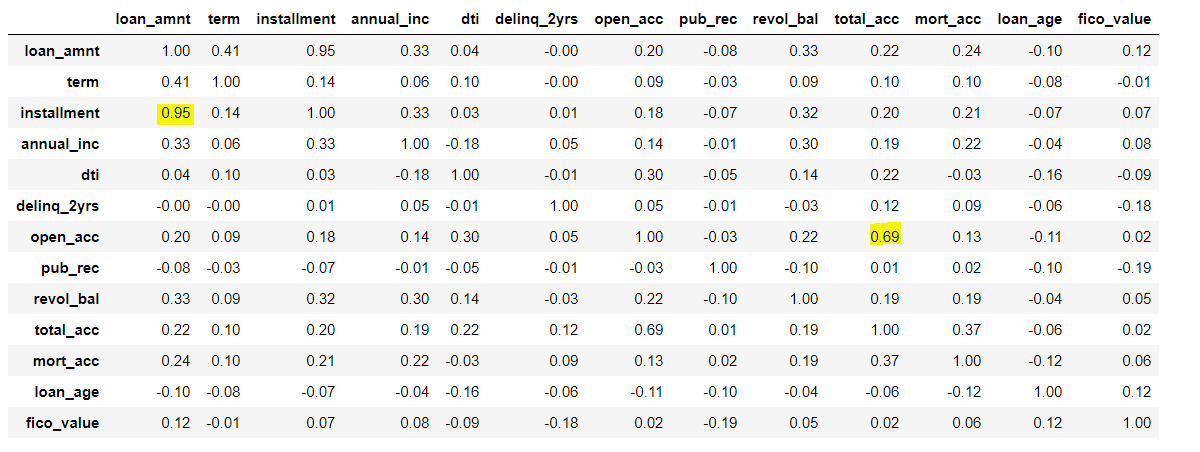
In doing so, we will have determined the optimal subset for prediction of default, which we can then use to determine default probability, a measure essential to our analysis of lending strategies

**Balancing the Dataset** - The robust dataset overall provided an excellent base with which to run the tests outlined above. However, an important feature of the dataset is its skewed nature. As shown in EDA 82% of the loans are fully paid whereas only 18% are charged off using the training dataset if we run the classifiers on this skewed dataset, our classifiers simply always classify loans as fully paid because it maximizes accuracy, thereby nullifying any valuable information about feature importance. We used over-sampling to translate the number of Fully Paid loans to be equal to the number of Charged Off Loans so the dataset used for classification consists of 50% Fully Paid Loans and 50% Charged Off Loans.

**Validating Assumptions before running the classification Model**

Following tests has been performed on the data.

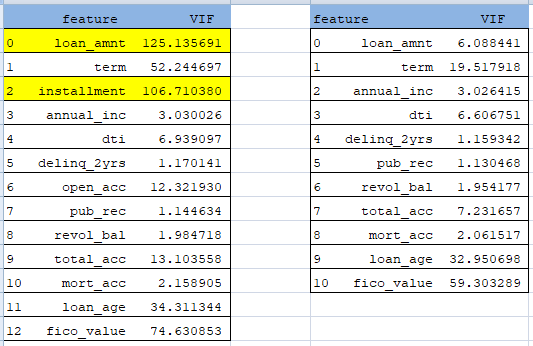
1. **Pearson Correlation**: Performed this test to check correlation between data and it can be seen that open\_acc and total\_acc are positively correlated with correlation 69% and loan\_amnt and instalments are negatively correlated with 95%



**Table 6.1: Pearson Correlation Table**

1. Variance Inflation Factor (VIF) was also performed to test for multicollinearity.

Result of VIF



**Table 6.2: VIF Result**

It seems higher values for loan\_amnt and installment So VIF after removing installment. Loan\_amnt and installment has been removed for Naïve Bayes, Logistic regression and LDA because they are impacted by multicollinearity.

1. Normality Test (K-S test): Performed K-S tests to check whether the distributions of each feature of the two classes (“Fully paid” or “charged-off”) are same or different. If p-value is high, then the distribution of two samples is same and if the p-value < 0.05 then distribution of two samples is different. For dataset K-S test p value for all features are less than 0.05,so all of these features have different distributions for Fully Paid/Charged Off.

The Null Hypothesis and Alternative Hypothesis are given below:

* Null Hypothesis (H0): Distributions of the two samples are equal.
* Alternative Hypothesis (Ha): Distributions of the two samples are unequal.

1. Homogeneityof Variance: Performed Levene’s test to test the homogeneity of variance, if p-value has greater than 0.05, it shows all samples have equal variances.

For dataset Levene’s test p-value for all features are significant > 0.05 so all these features have equal variances.

The Null Hypothesis and Alternative Hypothesis are given below:

* Null Hypothesis (H0): Variances of the two samples are equal.
* Alternative Hypothesis (Ha): Variances of the two samples are unequal.

Assumptions for statistical models:

1. Assumption of appropriate outcome structure:

The main assumptions for the classification models are the appropriate structure of the outcome variable. In given dataset dependent variables “Fully Paid” and “Changed Off” are converted into binary “0” and “1” respectively.

1. Assumption of independence of observation:

Assumed the observations to be independent of each other. In other words, the observations should not come from repeated measurements or matched data. In given dataset each observation is independent.

1. Assumption of absence of multicollinearity:

Certain classification models like Logistic regression and Naïve Bayes require there to be little or no multicollinearity among the independent variables. This means that the independent variables should not be too highly correlated with each other. In given dataset correlation between loan\_amount and installment is high (95%) and also between total\_acc and open\_acc is 69%.So,we have removed installment and open\_acct features. Tree based models are not impacted by multicollinearity.

1. Assumption of large data size:

Machine learning models typically requires a large sample size. Given dataset has large sample size almost 887442 records and is therefore large (n> 30)

**Classification Models**

**Decision Tree:**

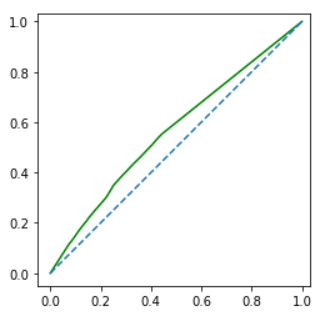
Decision trees tend to over fit, especially when the tree is deep and leaf nodes comprise too few attributes. Limiting the maximum depth or the minimum leaf node attributes not only reduces over fitting, but also speeds up training significantly, as decision tree model builds numerous decision trees before taking the average of their predictions. The model decision tree was created using the **DecisionTreeClassifier** and ‘**entropy**’ criterion with a max depth of 18 was used to build an appropriate decision tree for selecting the best splitter.

**Performance of Decision Tree:**

Confusion Matrix

|  |  |  |  |
| --- | --- | --- | --- |
|  | Model Prediction | | |
| Actual loan status |  | No Default | Default |
| No Default | 434595 | 207275 |
| Default | 85956 | 64044 |

**Table 6.3: Confusion Matrix for Decision Tree**

****

**Fig 6.1: AUC plot for Decision Tree**

**AUC = 56.397**

Accuracy =0.6296980564991729

Sensitivity =0.42696

Specificity =0.6770763550251608

**Logistic Regression:**

Logistic regression belongs to the class of generalized linear model and it measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities. We will be using close to 20 predictor variables to estimate the probability of loan default and using that for classification.

**LogisticRegression** from sklearn library was used creating the logistic regression model and SMOTE from imblearn was used to remove the data imbalance. One hot encoding was done to classify the states into 4 regions West, Mid-West, North-East and South.

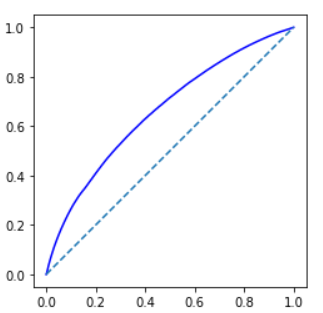
Probability of default was kept as 0.48 for cut-off.

**Performance of Logistic Regression:**

Confusion Matrix

|  |  |  |  |
| --- | --- | --- | --- |
|  | Model Prediction | | |
| Actual loan status |  | No Default | Default |
| No Default | 394386 | 247484 |
| Default | 57625 | 92375 |

**Table 6.4: Confusion Matrix for Logistic Regression**

****

**Fig 6.2: AUC plot for LR**

AUC = 66.0756

Accuracy=0.6146981196408502

sensitivity=0.6158333333333333

specificity=0.614432829077539

**Random Forest:**

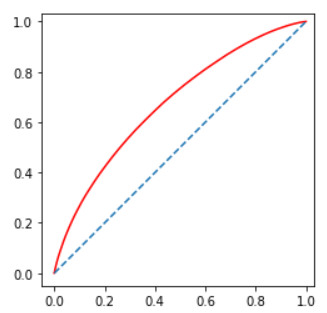
Random forest is an ensemble-style modelling technique, bootstrapping, random feature selection. We decided to tackle the loan classification problem by approaching the simple problem of whether a loan will be fully paid or charged off at completion, simplifying our initial approach to a binary classification problem. We used an equal number of fully paid and charged off loans, and randomly assigning into training and test sets. Increases in the number of trees did not significantly improve our results. Ensemble of Decision Trees (**RandomForestClassifier** in sklearn): It extends the concept of decision trees to build an ensemble of trees, with each tree built using a sample of predictors to mitigate over-fitting. Random Forest model was created with ‘gini’ criterion and min\_samples\_leaf=1, min\_samples\_split=2

Type of random forest: classification

**Performance of Random Forest:**

|  |  |  |  |
| --- | --- | --- | --- |
| Confusion Matrix | Model Prediction | | |
| Actual loan status |  | No Default | Default |
| No Default | 387771 | 254099 |
| Default | 53682 | 96318 |

**Table 6.5: Confusion Matrix for Random Forest**

****

**Fig 6.3: AUC plot for Random Forest**

AUC = 67.408

Accuracy=0.6113238284061778

sensitivity=0.64212

specificity=0.6041270039104492

**Naive Bayes:**

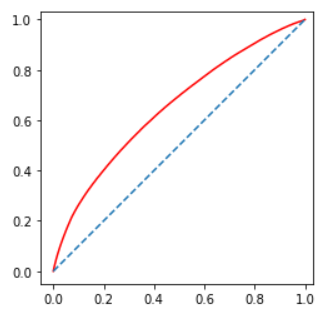
Naive Bayes is a technique based on Bayes theorem of conditional probability for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. There is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. The default loan probability is explained by the 20 variables that were considered as independent features. GaussianNB from sklearn package was used for building the model

**Performance of Naive Bayes:**

Confusion Matrix

|  |  |  |  |
| --- | --- | --- | --- |
|  | Model Prediction | | |
| Actual loan status |  | No Default | Default |
| No Default | 371542 | 270328 |
| Default | 55350 | 94650 |

**Table 6.6: Confusion Matrix for Naïve Bayes**

****

**Fig 6.4: AUC plot for Naive Bayes**

AUC = 65.0199

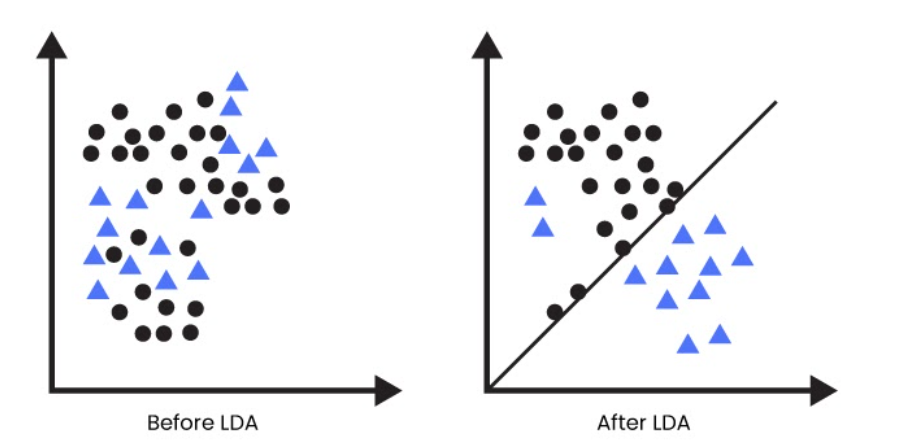
Accuracy=0.5887228964350208

sensitivity=0.631

specificity=0.5788430679109478

**Linear Discriminant Analysis:**

Linear Discriminant Analysis (LDA) is the most commonly used dimension reduction technique in supervised learning. It is basically a pre-processing step for pattern classification and machine learning applications. It projects the dataset into moderate dimensional space with a genuine case of separable features that minimize over fitting and computational costs. It is widely used for modelling data into categories i.e., distributing varieties into two or more classes



**Fig 6.5: LDA Plot**

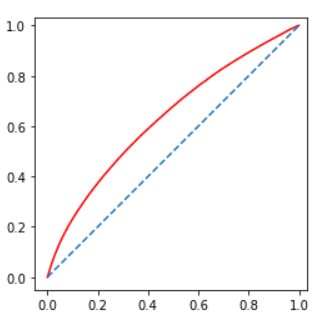
Assumption:  Independent variables are normal for each level of the grouping variable. The data is Gaussian that each variable is shaped like a bell curve when plotted.

**Performance of LDA:**

Confusion Matrix

|  |  |  |  |
| --- | --- | --- | --- |
|  | Model Prediction | | |
| Actual loan status |  | No Default | Default |
| No Default | 410217 | 231653 |
| Default | 67181 | 82819 |

**Table 6.7: Confusion Matrix for LDA**

****

**Fig 6.6: AUC plot for LDA**

AUC = 63.194

Accuracy=0.6226224001414373

Sensitivity=0.5521266666666667

Specificity=0.639096701824357

**Model Performance Comparison:**

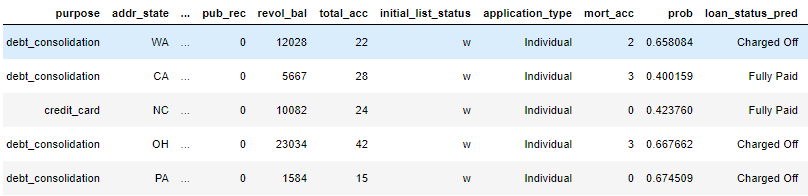
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Description** | **Decision Tree** | **Naïve Bayes** | **Logistic regression** | **LDA** | **Random Forest** |
| sklearn Package | DecisionTreeClassifier | GaussianNB | Logistic Regression | LinearDiscriminantAnalysis | RandomForestClassifier |
| No of Predictors | 20 | 20 | 20 | 20 | 20 |
| Train Data – Year | 2007-15 | 2007-15 | 2007-15 | 2007-15 | 2007-15 |
| Train Data – No of Observations | 883,760 | 883,760 | 883,760 | 883,760 | 883,760 |
| Test Data – Year | 2016&2017 | 2016&2017 | 2016&2017 | 2016&2017 | 2016&2017 |
| Test Data – No of Observations | 791,870 | 791,870 | 791,870 | 791,870 | 791,870 |
| Accuracy | 62.96% | 58.87% | 61.46% | 62.26% | 61.13% |
| Sensitivity | 42.69% | 63.7% | 61.58% | 55.21% | 64.21% |
| Specificity | 67.70% | 57.88% | 61.44% | 63.90% | 60.41% |
| AUC | 56.397 | 65.01 | 66.07 | 63.19% | 67.40 |

**Fig 6.8: Model Performance Comparison**

# Creating Probability of Default (PD) Risk Based Model

As Logistic regression is giving the best performance out of 5 classification models, we used the model output to create a Risk model based on Probability of Default (PD). Probability of Default can also be estimated using historical data and statistical techniques. Generally, the higher the default probability, the higher the interest rate the lender will charge the borrower. We have used the probability of default to categorize the loan applicants into three categories High Risk, Medium Risk and Low Risk.

We used pickle package to serialize the Logistic Regression model and save the serialized format to a file. In our next algorithm for predicting the Probability of Default and categorizing the applicant we import and load the model as an object called ‘Pickled\_LR\_Model’. This object was applied to test data to predict both the loan status Fully Paid (class 0) and Charged Off (class 1).



**Table 7.1: Calculated Probability**

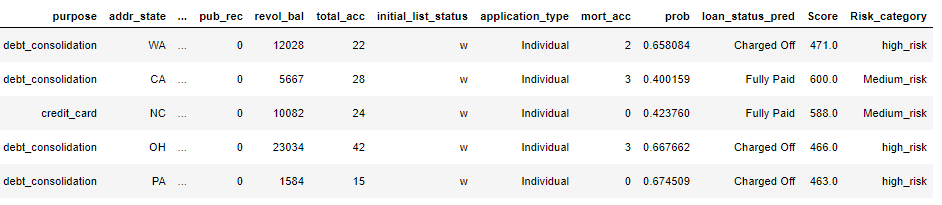
The computed probability of default was then used to calculate the risk score using the min- max scaling. Minimum risk score is assigned as 300 and maximum risk score is assigned as 800.

Risk Scoring Logic:

* We have used probability predictions from Logistic Regression as input for the Probability of Default model
* The Class of Interest is the “Charged Off’ class and therefore the higher the probability output from Logistic Regression, higher the chances of loan default.
* Therefore, for the risk score calculation, we have used 1-Probability of default (1-p) as we wanted lower risk for the investor. So, lower the risk score, higher the probability of default.
* Since the probability always lies between 0 and 1, we have used the min-max scaler to scale the computed probability for the applicants between 300-800 score range which is used as the risk score for the loan applicant.

The above computed scores are used to classify loan applicant into three categories. Scoring classes are given below:

* Low Risk: Score range of 300-500
* Medium Risk: Score range of 500-650
* High Risk: Score range of 650-800



**Table 7.2: Calculated Score**

# Investment Portfolio using Optimization Strategies

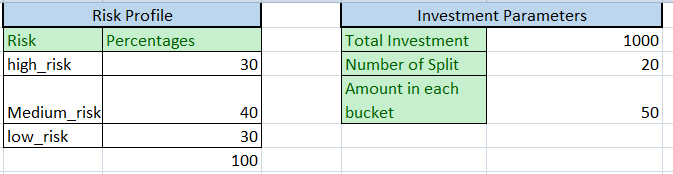
We have as part of the scope of this project, attempted to create a Microsoft Excel based utility considering random samples of data from the test data set. An investor will seek a portfolio which has the highest possible return, subject to investment constraints which can be – risk profile of investor, investment amount, and diversification requirements (e.g., high risk loans will be only 10% of the total portfolio).

**Investment Strategy**:

The overall objective is to balance the risk vs returns for each investor based on their risk appetite. To achieve this objective, we have suggested that each investor invests into multiple loans across the Low, Medium and High-risk score applicant profiles and the allocation of the overall investment amount is done based on the risk appetite selected by the investor.

The aforementioned model can be used to categorize the applicants of loan into High, Medium and Low risk and then the entire pool of applicants can be available for investors to choose based on their risk appetite. A web portal can be integrated with Lending club platform where investors can assign percentages to high, medium and low risk loans as per their preference and an optimization model will show the list of applicants in which they can invest.

In our scope we have tried to create a prototype using Excel Solver using a random sample of 60 loan applicants where investors choosepercentage of High, Medium and Low Risk based on their risk appetite. The number of investment buckets and total investment can be chosen by Investor to suit their profile.



**Table 8.1: Optimization Parameters**

Based on user’s selection the solver maximizes the total revenue keeping the constraint of percentage selection of risk profile loans selected by user and at the same time provides best loan applicants which investors can choose. The total profit for the investor is considered as the total profit generated from interest rate, and does not consider charges to be paid in-lieu to Lending Club.



As we can see above that optimization model has chosen 5 High risk loans, 3 Medium risk loans and 2 Low risk loans and at the same time maximizing the profit to $174.52annually in investment of $1000 of $100 into each loan.

# Actionable Insights and recommendations

**Recommendations** that we want to provide from this project are

* Probability of default that an applicant will default
* Categorize the applicants into high risk, medium risk and low risk
* Provide a recommendation engine to investors on how to diversify their portfolio of P2P loans

The recommendation will be made in a way that the risk is minimized (and returns are maximized) for the lender. Currently the default rates are very high in P2P lending. For example, the lending club data shows that the average default rate across all categories is between 9% and 11% on average and there are some categories where the default rate is over 30%.

The key challenge of P2P lending model revolves around the ability to assess risk for borrowers. This presents the following changes:

* Higher interest and rejection rates - inability to perform the right level of risk assessment end up with higher rejection rates or interest rates. This keeps many potential borrowers away from the system
* Perceived high risk of investments - while the P2P model in general have higher risk relative to some of the traditional lending models, relative lack of regulations and knowledge of the model makes the situation worse for investors so far few of them want to invest in P2P lending

**Applications of the project are:**

* increasing the number of lenders and borrowers on the P2P lending platform
* increase profits for the lenders by reducing the risk of default and maximizing returns
* decrease rejection rates and optimize interest rates for borrowers

In developing and emerging markets such as India, a significant portion of the market does not have access to formal borrowing methods such as banks. Informal lending is very expensive and keeps away many potential borrowers and hence the related economic activity. P2P lending platforms aim to solve this problem and any improvements to the methods they employ currently would result in a direct benefit not only to lenders and borrowers but to the entire ecosystem increasing the economic activity and the benefits around the same.

**Future Scope**

* One of the future scopes of this project would be to collect India’s P2P platforms data and learn from Lending club’s modelling done as part of this exercise and use that to profile loan applicants in terms of probability of default and use that to optimize the investment for lenders.
* Interest rate optimization to reduce loan default and customer churn.
* Use of social parameters for predicting loan default

# References and Bibliography

* + - 1. JESSICA FOO, LEK-HENG LIM, and KEN SZE-WAI WONG, “Macroeconomics and Fintech: Uncovering Latent Macroeconomic Effects On Peer-To-Peer Lending”, Department of Statistics, University of Chicago, November 2017.
      2. Andreas Dietricha, RetoWernlia, “Credit Determinants of Interest Rates in the P2P Consumer Lending Market: How Rational are Investors?”, April 17, 2019.
      3. Kevin Tsai, SivagamiRamiah, and Sudhanshu Singh.Peer Lending Risk Predictor. CS 229 Autumn 2014.Stanford University.
      4. Kingma, D., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv:1412.6980
      5. Shunpo Chang, Simon Dae-Oong Kim, GenkiKondo.Predicting Default Risk of Lending Club. CS 229
      6. Autumn 2015. Stanford University.
      7. Riza Emekter, Yanbin Tu, Min Lu Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending
      8. Carlos Serrano-Cinca, BegoaGutierrez-Nieto, Luz Lopez-Palacios Determinants of Default in P2P Lending
      9. Milad Malekipirbazari, VuralAksakalli Risk assessment in social lending via random forests
      10. Balyuk, T. (2018). “Financial Innovation and Borrowers: Evidence from Peer-to-Peer Lending.” Rotman School of Management Working Paper.
      11. Berger, S.C., and F. Gleisner. (2009). “Emergence of Financial Intermediaries in Electronic Markets: The Case of Online P2P Lending.” BuR Business Research Journal 2(1).
      12. 7 Problems that Keep P2P Lending a Niche Investment. <http://www.lendingmemo.com/peer-lending-problems/>.
      13. Default Rates at Lending Club and Prosper: When Loans Go Bad – S. Cunningham, <http://www.lendingmemo.com/lending-club-prosper-default-rates/>
      14. H. Ghasemkhani, Y. Tan, and A. Tripathi. The Invisible Value of Information Systems: Reputation Building in an Online P2P Lending System. Working Paper, 2013.

**Abbreviations**

CIBIL - CreditInformation Bureau (India) Limited

US – United States of America

USA – United States of America

GDP – Gross DOmESTIC PRODUCT

ROI – Return on Investment

FICO - Fair Isaac Corporation

EDA – Exploratory Data Analysis

API – Application Programming Interface

LC – Lending Club

URL – Uniform Resource Locator

SMOTE – Synthetic Minority Over-sampling Technique

DTI – Debt to Income

ROC – Receiver Operating Characteristic

AUC – Area Under Curve

VIF – Variance Inflation Factor

LDA – Linear Discriminant Analysis

LR – Logistic Regression

PD – Probability of Default

1. 1https://en.wikipedia.org/

   2https://www.datadriveninvestor.com/

   3https://www.cibil.com

   4https://www.equifax.co.in

   5https://www.experian.in

   6https://www.crifhighmark.com

   7https://medium.com/ [↑](#footnote-ref-2)
2. https://en.wikipedia.org/wiki/Reverse\_auction [↑](#footnote-ref-3)
3. https://www.lendingclub.com/ [↑](#footnote-ref-4)