Credit Risk Modeling

A Thesis
Submitted to the Faculty of

St. Joseph's University, Bangalore

by

Ayesha Shariff

in partial fulfilment of the
requirements for the degree of

Master of Science
in
Big Data Analytics

June 2023

I WOULD LIKE TO EXTEND MY SINCERE GRATITUDE TO THE ALMIGHTY GOD, MY PARENTS, MY TEACHERS AND MY PROJECT MENTOR.

Date: 22-06-2023

TO WHOMSOEVER IT MAY CONCERN

This is to certify that Ayesha Shariff has successfully completed a three-month project under my supervision. The project focused on building a credit risk model to estimate the expected loss on a loan.

During her tenure, Ayesha displayed commendable dedication and enthusiasm towards her assigned responsibilities. She consistently exhibited a strong work ethic and diligently tackled the tasks at hand. Her commitment to achieving objectives and meeting deadlines were quite impressive.

Praveen Kumar J
Systems Engineer
Tata Consultancy Services

Declaration of the Candidate

I hereby declare that this work entitled, "Credit Risk Modeling" has been originally carried out by me, Ayesha Shariff under the guidance and supervision of Mr. Praveen Kumar J. This work has not been submitted elsewhere for the award of any other degree or diploma certificate.

Bengaluru
June 2023

Ayesha Shariff
Register No.: 21BDA18
M.Sc. in Big Data Analytics

ACKNOWLEDGEMENT

This internship research work carried out as part of my fourth semester M.Sc. Big Data Analytics coursework couldn't have been completed without the support of the following people.

I sincerely thank Mr. Thawar Chand Gehlot, Hon'ble Governor of Karnataka and Chancellor of Universities as well as Professor Lingaraja Gandhi, Hon'ble Vice Chancellor, Bengaluru City University, for providing excellent infrastructure required to carry out research in the field of Big Data Analytics.

I thank Rev. Dr. Fr. Victor Lobo, SJ, Principal of St. Joseph's College (Autonomous), who constantly motivates and believes that every Josephite has the potential to reach world standard. I would also like thank him for always extending his support to the Department of Big Data Analytics. This has motivated me to perform this research work.

I thank Fr Denzil Lobo SJ, Dean of St Joseph's Institute of Information Technology for his support to the Department of Big Data Analytics.

I thank Prof. Jayati Bhadra, Head of the Department of Advanced Computing at St. Joseph’s University, Bengaluru for her ongoing support and advice.

It gives me immense pleasure to thank Dr. Srinivas Bhogle, Visiting Professor, Department of Big Data Analytics. I feel proud and blessed to be his student. He has been instrumental in helping me excel in my Analytics skills from a mathematical point of view. I will forever be grateful to him for making me an independent learner. Once again my sincere thanks to him.

I would like to express my heartfelt gratitude to my supervisor, Mr. Praveen Kumar, for generously dedicating his valuable time to assist me in completing this project. Without his extensive knowledge and guidance, this work would not have reached the level of achievement it has attained

I would like to express my gratitude to all the faculty and adjunct faculty members of MSc. Big Data Analytics Department for their guidance throughout my postgraduate studies.

I thank each of my M.Sc. classmates for their continuous direct and indirect support.

I would like to thank my family and friends who have always been the constant source of motivation.

I thank the Almighty for blessing me with the opportunity to carry out this research work.

Ayesha Shariff.

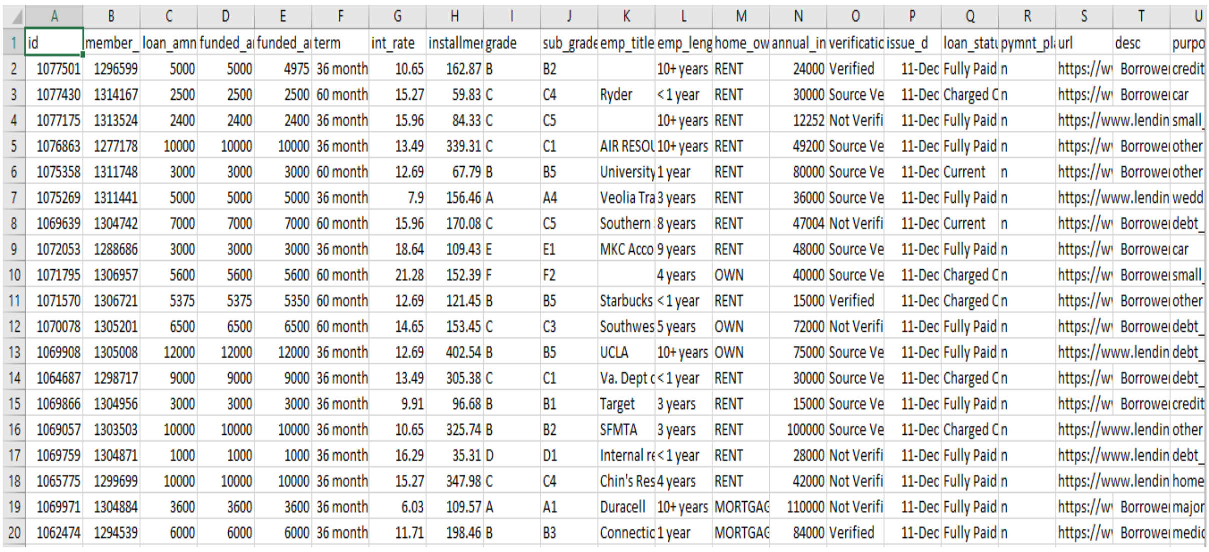
CONTENTS

1. INTRODUCTION

Credit risk is the risk of a borrow defaulting on a debt and that the lender may lose the principal of the loan or associated interest. When a bank receives a loan application, it has to make a decision as to whether to approve the loan or not based on the applicant’s profile. If the bank deems the applicant to have bad credit risk, it means the applicant is not likely to repay the loan and approving the loan could result in financial loss to the bank. The event of a borrower not being able to repay the debt is called default. Therefore, to protect themselves against borrower defaults lenders must assess credit risk associated with each borrower. Lenders incur credit losses from every portfolio or exposure over a given period of time i.e, lenders know there is a certain amount of credit risk with every borrower. To estimate the credit risk of each borrower we need to calculate the expected loss of each borrower. Lenders failure to estimate the borrower’s probability of default can have great consequences for lenders and society in general. Lending to borrower with high probability of default is the main reason for serious financial crisis. Expected loss can be calculated by multiplying probability of default (PD), loan given default (LGD) and exposure of default (EAD). I have created logistic model to calculate the probability of default which predicts the probability that a borrower might fail to repay the loan. For non tech people like credit agents and front office workers I have simplified the statistical model for estimating credit risk as scorecards. Scorecards represent coefficient of PD model in simplified way as scores. These scores are used to decide whether an applicant is good or bad. I have calculated LGD by combining logistic and linear regression models. EAD is calculated by simple linear regression model. Finally, multiplying the values of PD, LGD and EAD for each borrower would give us the expected loss.

2. ABOUT DATASET

The dataset considered here is Lending Clubs Loan dataset from the year 2007 to 2014 in United States. Lending club is a lending platform that lends money to people in need at an interest rate based on their credit history and other factors. They match people looking to invest money with people looking to borrow money. Lending Club is an online financial service that connects borrowers with loans that fit their needs. It works best for borrowers with good to excellent credit who are looking for personal loans or debt consolidation. Lending Club began offering loans in 2007 as a peer-to-peer online lender and has lent more than $85 billion in the years since. Dataset contains 466285 rows (i.e, number of borrowers/ loans given) and 74 columns (i.e, information about each loan given). These loans are consumer loans. Most of the columns have null values, so I have not considered those columns for building the models.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	id	member_	loan_amn	funded_a	funded_a	term	int_rate	installme	grade	sub_grade	emp_title	emp_leng	home_ow	annual_in	verificatic	issue_d	loan_statu	pymnt_pl	url	desc	purpo
2	1077501	1296599	5000	5000	4975	36 month	10.65	162.87	B	B2		10+ years	RENT	24000	Verified	11-Dec	Fully Paid	n	https://w	Borrowe	credit
3	1077430	1314167	2500	2500	2500	60 month	15.27	59.83	C	C4	Ryder	< 1 year	RENT	30000	Source Ve	11-Dec	Charged C	n	https://w	Borrowe	car
4	1077175	1313524	2400	2400	2400	36 month	15.96	84.33	C	C5		10+ years	RENT	12252	Not Verifi	11-Dec	Fully Paid	n	https://www.lendin		small_
5	1076863	1277178	10000	10000	10000	36 month	13.49	339.31	C	C1	AIR RESOU	10+ years	RENT	49200	Source Ve	11-Dec	Fully Paid	n	https://w	Borrowe	other
6	1075358	1311748	3000	3000	3000	60 month	12.69	67.79	B	B5	University	1 year	RENT	80000	Source Ve	11-Dec	Current	n	https://w	Borrowe	other
7	1075269	1311441	5000	5000	5000	36 month	7.9	156.46	A	A4	Veolia Tra	3 years	RENT	36000	Source Ve	11-Dec	Fully Paid	n	https://www.lendin		wedd
8	1069639	1304742	7000	7000	7000	60 month	15.96	170.08	C	C5	Southern	8 years	RENT	47004	Not Verifi	11-Dec	Current	n	https://w	Borrowe	debt_
9	1072053	1288686	3000	3000	3000	36 month	18.64	109.43	E	E1	MKC Acco	9 years	RENT	48000	Source Ve	11-Dec	Fully Paid	n	https://w	Borrowe	car
10	1071795	1306957	5600	5600	5600	60 month	21.28	152.39	F	F2		4 years	OWN	40000	Source Ve	11-Dec	Charged C	n	https://w	Borrowe	small_
11	1071570	1306721	5375	5375	5350	60 month	12.69	121.45	B	B5	Starbucks	< 1 year	RENT	15000	Verified	11-Dec	Charged C	n	https://w	Borrowe	other
12	1070078	1305201	6500	6500	6500	60 month	14.65	153.45	C	C3	Southwes	5 years	OWN	72000	Not Verifi	11-Dec	Fully Paid	n	https://w	Borrowe	debt_
13	1069908	1305008	12000	12000	12000	36 month	12.69	402.54	B	B5	UCLA	10+ years	OWN	75000	Source Ve	11-Dec	Fully Paid	n	https://www.lendin		debt_
14	1064687	1298717	9000	9000	9000	36 month	13.49	305.38	C	C1	Va. Dept c	< 1 year	RENT	30000	Source Ve	11-Dec	Charged C	n	https://w	Borrowe	debt_
15	1069866	1304956	3000	3000	3000	36 month	9.91	96.68	B	B1	Target	3 years	RENT	15000	Source Ve	11-Dec	Fully Paid	n	https://w	Borrowe	credit
16	1069057	1303503	10000	10000	10000	36 month	10.65	325.74	B	B2	SFMTA	3 years	RENT	100000	Source Ve	11-Dec	Charged C	n	https://www.lendin		other
17	1069759	1304871	1000	1000	1000	36 month	16.29	35.31	D	D1	Internal re	< 1 year	RENT	28000	Not Verifi	11-Dec	Fully Paid	n	https://www.lendin		debt_
18	1065775	1299699	10000	10000	10000	36 month	15.27	347.98	C	C4	Chin's Res	4 years	RENT	42000	Not Verifi	11-Dec	Fully Paid	n	https://www.lendin		home
19	1069971	1304884	3600	3600	3600	36 month	6.03	109.57	A	A1	Duracell	10+ years	MORTGAG	110000	Not Verifi	11-Dec	Fully Paid	n	https://w	Borrowe	major
20	1062474	1294539	6000	6000	6000	36 month	11.71	198.46	B	B3	Connectic	1 year	MORTGAG	84000	Verified	11-Dec	Fully Paid	n	https://w	Borrowe	medic

2.1 DATA PREPROCESSING

- Removed strings from emp_length and term variable and converting them to numeric.
- Calculated months since issue date from issue-date variable and converting them from datetime to numeric.
- Created dummy variables for discrete variables.
- Replaced missing value of total_rev_hi_lim by funded_amnt, annual_income by mean and other categories by 0 for simplifying the model.

The final columns that are considered for building models are:

Discrete variables - grade, home_ownership, addr_state, verification status, purpose, initial_list_status

Continuous variables - term_int, emp_length_int, mths_since_issue_d, int_rate, funded_amnt, delinq_2yrs, inq_last_6mths, open_acc, pub_rec, total_acc, acc_now_delinq, total_rev_hi_lim, installment, annual_inc, mths_since_last_delinq, dti, mths_since_last_record

3. PROBLEM STATEMENT

Implementing statistical methodology to build credit risk models for estimating expected loss of the loans given to borrowers that would help banks or any finance companies to make better decisions while giving out loans in future. To estimate the expected loss, probability of default model, loan given default model and exposure at default model is build. Using, regression coefficient of the probability of default model, calculate credit score for each loan. Calculate total expected loss and its proportion to the amount funded by the bank which shows how well the bank is functioning.

4. MODELING

Banks consider two types of models while dealing with loans. First, Application models are used to estimate firms credit ratings at the moment of application i.e, when a new borrower asks for loan. The estimated credit ratings in turn are the basis on which banks decide whether to grant a loan or not. They also use it to price the loans like what interest rates should be charged for the respective loan. Riskier the loan is, higher will be the interest rate charged to the customer. To know how riskier the loan can be credit scores are seen for the respective loan.

Second, Behavior models are used to calculate the probability of default and respectively expected loss after a loan is granted. Banks also use this model to decide whether to give an additional loan to an existing customer.

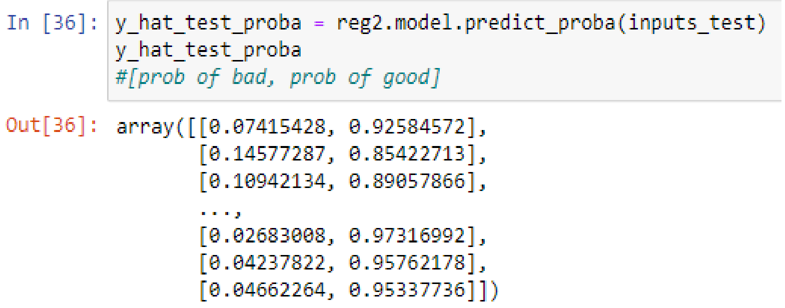
For calculating the expected loss, I have built three models:

1. Probability of Default (PD) model
2. Loan Given Default (LGD) model
3. Exposure at Default (EAD) model

4.1 PD MODEL

Probability of default models interpretability is extremely important as it is required by the regulators. The model must be very easy to understand and apply. For people who have never heard of statistical analysis should be able to work with it. Therefore, all the independent variables in the PD model are converted to dummy variables. The statistical methodology to model PD is logistic regression where the dependent variable is precisely whether a customer defaulted or not. Default here is defined based on the delinquency of the borrower measured in days past the payment due date i.e, more than 90 days overdue. The model was built as follows

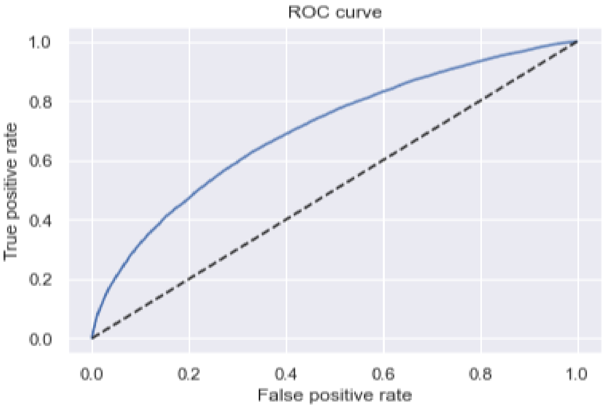
1. Preprocessed the data by calculating the weight of evidence (WOE) i.e, to what an extent an independent variable would predict a dependent variable, for each category of independent variable and combined the similar categories into one category. This is done because too many dummy variables are not needed for modeling so I have grouped them. Keeping the category of worst credit risk as reference category (i.e, category with lowest WOE).
2. Created a data frame of only the preprocessed data that will be used to build he model.
3. Building Logistic Regression with p values.
4. Removed the variables that have p value greater than 0.05 because they are not statistically significant - delinq_2yrs, open_acc, pub_rec, total_rev_hi_lim, total_acc.
5. Implemented logistic regression again on the final data set and saved the model using pickle.
6. Tested the model by predicting the probabilities of input test data.



```
In [36]: y_hat_test_proba = reg2.model.predict_proba(inputs_test)
         y_hat_test_proba
         #[prob of bad, prob of good]

Out[36]: array([[0.07415428, 0.92584572],
                [0.14577287, 0.85422713],
                [0.10942134, 0.89057866],
                ...,
                [0.02683008, 0.97316992],
                [0.04237822, 0.95762178],
                [0.04662264, 0.95337736]])
```

7. Model performance is assessed by calculating the Area Under the Receiver Operating Characteristic Curve (AUROC).
8. AUC is 70.21%

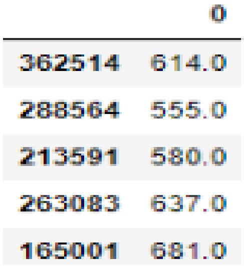


4.1.1 CREDIT SCORE CARD

Scorecards represent coefficient of PD model in simplified way as scores. These scores are used to decide whether an applicant is good or bad. For deciding what will be the range of the scores I have considered the standard Fair Isaac Corporation (FICO) score that has minimum score as 300 and maximum score as 850. The regression coefficient of PD model can be simplified as scores by implying the formula.

$$\text{Variable score} = \frac{\text{C x (Max Score – Min Score)}}{\text{Max sum of coeff – Min sum of coeff}}$$

Here, C is the regression coefficient.



	0
362514	614.0
288564	555.0
213591	580.0
263083	637.0
165001	681.0

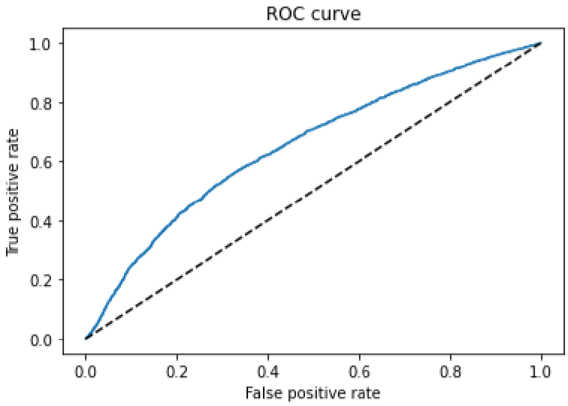
4.2 LGD MODEL

Loan given default is the proportion of the amount that was lost which cannot be recovered after the borrower defaulted. This proportion of amount is nothing but the amount left after the recoveries done. LGD can be calculated by using recovery rate i.e, LGD = 1 – Recovery Rate. Therefore, the dependent variable is recovery rate and the independent variables are the preprocessed variables but here I have only taken discrete variables as dummy variables and keep continuous variables as it is.

More than half of the observations have the recovery rate equal to 0. So for estimating LGD I have performed 2 stage approach and then finally combined both the models output to get the LGD value.

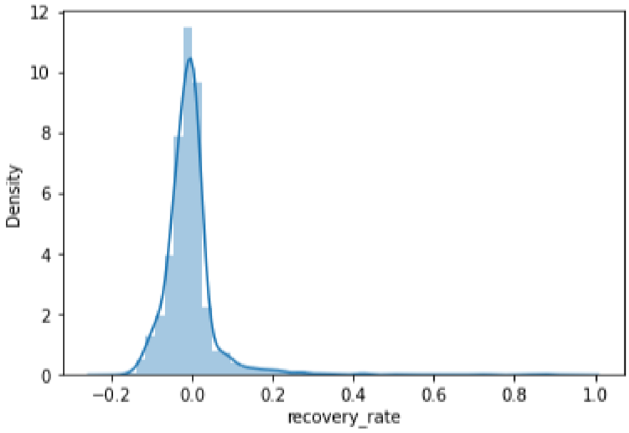
4.2.1 STAGE 1: Is recovery rate 0 or not?

1) Independent Variables - Discrete dummy variables and continuous variables
2) Dependent Variable - Recovery_ rate_0_1
3) Built Logistic Regression to predict if the recovery rate is 0 or not.
4) Model performance: AUC is 65.09%. This accuracy is enough for an LGD model.



4.2.2 STAGE 2: If recovery rate is greater than 0, how much exactly is it?

1) Independent Variables - Discrete dummy variables and continuous variables
2) Dependent Variable - Recovery_ rate
3) Built Linear Regression to calculate the exact value.
4) Model performance: Plot the distribution of residuals (i.e, difference between actual and predicted values)



The distribution of the residuals resembles normal distribution and most of the residuals are symmetrically distributed. Therefore, it is a good model

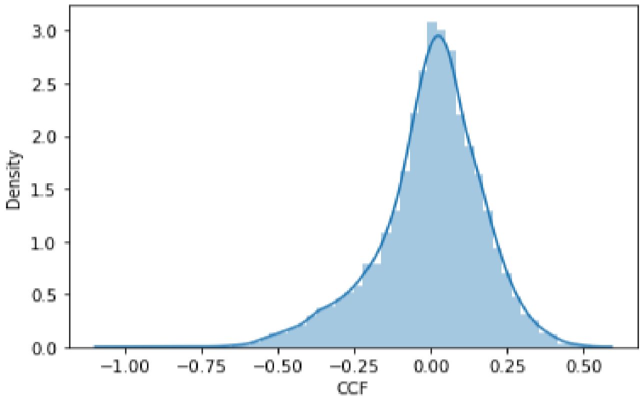
4.2.3 COMBINING STAGE 1 AND STAGE 2

1) Apply stage 1 logistic model on data frame.
2) Apply stage 2 linear model on data frame.
3) Recovery rate = stage 1 output * stage 2 output
4) LGD = 1- Recovery rate

4.3 EAD MODEL

Exposure at default is the total loss of amount a bank is exposed to at the moment the borrower defaults i.e, maximum amount a bank may lose on a loan. This data can be seen only for accounts that are defaulted. So I have included all the defaulted accounts for building the model. Here, the borrower has defaulted in proportion of original funded amount. This proportion is dependent variable for EAD model. This proportion is called Credit Conversion Factor (CCF).

1) Independent Variables - Discrete dummy variables and continuous variables
2) Variable - Credit Conversion Factor (CCF)
3) Built Linear Regression to calculate the exact value.
4) Model Performance: The correlation between actual and predicted values is more than 0.53 this is moderately strong positive correlation which is good for EAD model.



Model Performance: The residuals distribution resembles normal distribution and most of the residuals are symmetrically distributed around 0.

5) Apply EAD linear regression model on the data frame to predict CCF values.
6) EAD = CCF * Funded Amount

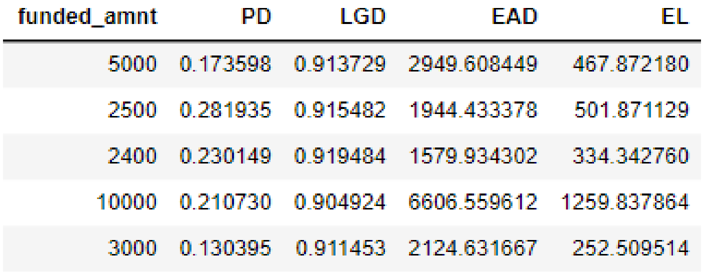
4.4 EXPECTED LOSS

Expected loss can be due to borrower specific loss, economic environment or both. It is the expected credit loss, the amount a lender might lose by lending loan to a borrower. It is calculated by multiplying the probability of default, loan given default and exposure at default of a loan.

EL = PD*LGD*EAD

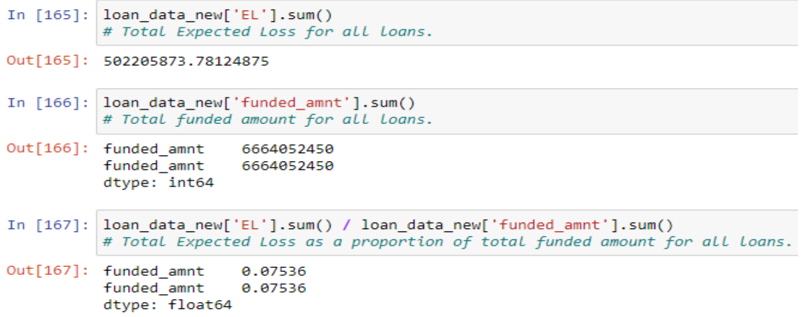
Banks don’t care about the loss they’ll experience from a single borrower, it is negligible compared to their overall exposure. So, I have found the total expected loss across all the borrowers. The total expected loss is the sum of the expected losses of all the loans in a portfolio.

1) Calculating the LGD and EAD by applying the models on preprocessed dataset.
2) PD model is applied on separate dataset because it requires dummy variables for continuous variables.
3) Concatenate the data frames of PD and LGD_EAD
4) EL = PD * LGD * EAD for each borrower or loan.



funded_amnt	PD	LGD	EAD	EL
5000	0.173598	0.913729	2949.608449	467.872180
2500	0.281935	0.915482	1944.433378	501.871129
2400	0.230149	0.919484	1579.934302	334.342760
10000	0.210730	0.904924	6606.559612	1259.837864
3000	0.130395	0.911453	2124.631667	252.509514

5) Calculating the total expected loss and the total Expected Loss as a proportion of total funded amount for all loans.



```
In [165]:  loan_data_new['EL'].sum()
            # Total Expected Loss for all loans.

Out[165]:  502205873.78124875

In [166]:  loan_data_new['funded_amnt'].sum()
            # Total funded amount for all loans.

Out[166]:  funded_amnt    6664052450
            funded_amnt    6664052450
            dtype: int64

In [167]:  loan_data_new['EL'].sum() / loan_data_new['funded_amnt'].sum()
            # Total Expected Loss as a proportion of total funded amount for all loans.

Out[167]:  funded_amnt    0.07536
            funded_amnt    0.07536
            dtype: float64
```

Here the proportion of total expected loss to total funded amount for loan portfolio of lending club from 20017 to 2014 is 7.5%.

Banks usually hold 10% of its assets as capital. Therefore, expected loss on its loan portfolio should be less than its capital. Observed expected loss are anywhere between 2% to 10%. Depending on this exposure the bank management can plan about how to give out the loans in future.

5. CONCLUSION

- Build a probability of default model that helps the lender to know the probability that a borrower might not repay the loan.
- Calculated the credit scores which the lender can refer to while giving loan to a new customer or an additional loan to an existing customer.
- Build models to get the loan given default and exposure at default.
- Calculated the expected loss of each loan that a bank or a finance company might face for each borrower.
- Calculated the proportion of total expected loss to funded amount, which will help the banks in planning about how to give loans in future.

6. FUTURE SCOPE

This project could be enhanced in the following ways:

- Developing a user interface that showcases the expected loss, probability of default and credit score when the details of the customer are entered.
- Use other machine learning algorithms for prediction to check whether they give better accuracy.
- Apply beta regression for LGD and EAD models. Beta regression could not be used in the project since python doesn’t support it.