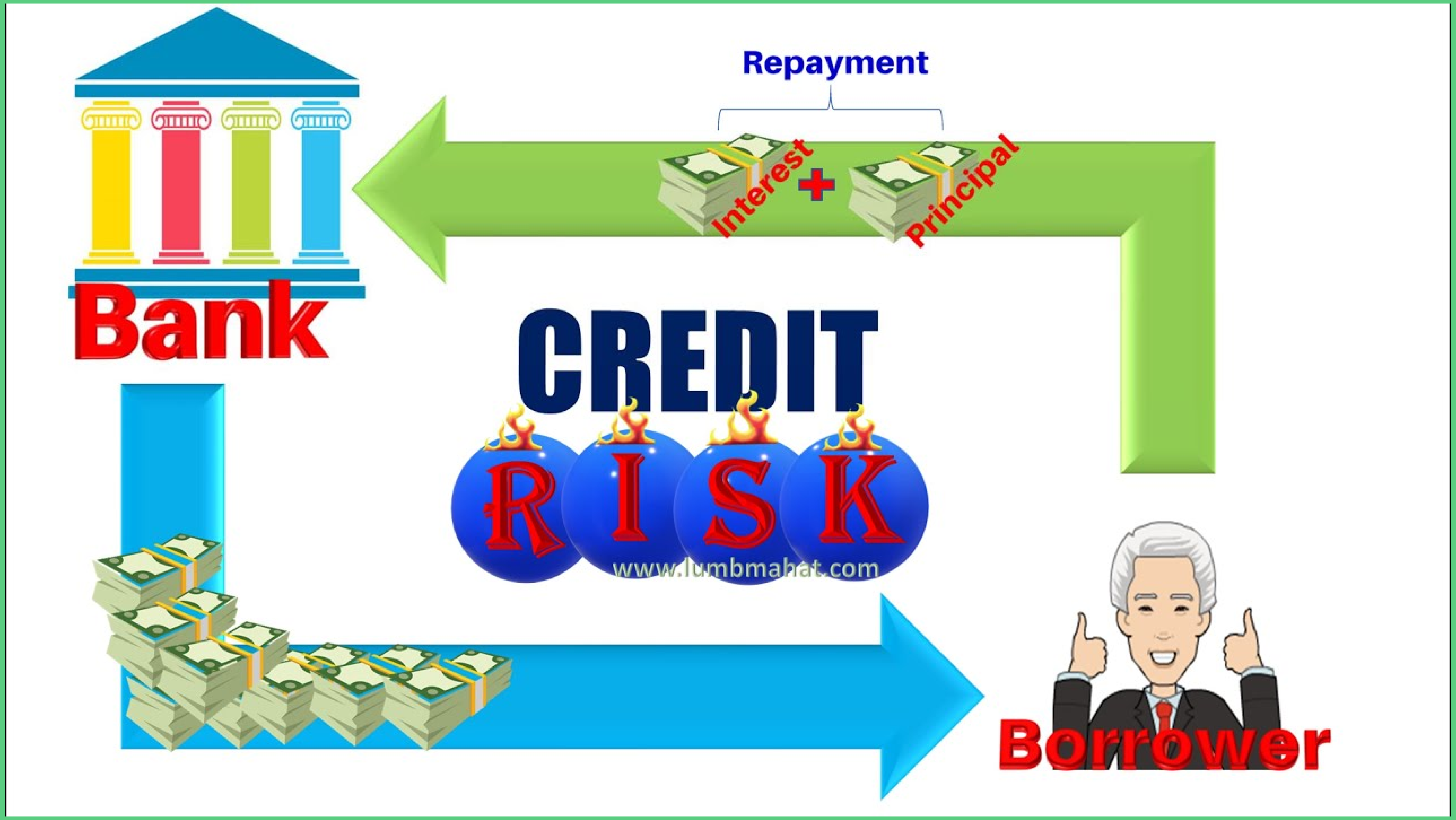
CREDIT RISK MODELING

Ayesha Shariff - 21BDA18

PROBLEM STATEMENT

Implementing statistical methodology to build credit risk models for estimating expected loss.

WHAT IS CREDIT RISK?



WHAT IS EXPECTED LOSS (EL)?

$$\text{PD} \times \text{LGD} \times \text{EAD} = \text{Expected Losses}$$

PD - PROBABILITY OF DEFAULT
LGD - LOAN GIVEN DEFAULT
EAD - EXPOSURE AT DEFAULT

BUILDING CREDIT MODELS

PD MODEL

Methodology to model PD is logistic regression where the dependent variable is whether a customer defaulted or not.

LGD MODEL

LGD is estimated by 2 stage approach - combining logistic regression model and linear regression model

EAD MODEL

Methodology to model EAD is simple linear regression

LendingClub

LENDING CLUB LOAN DATASET 2007 - 2014

- Lending club Loan data from 2007 to 2014 in US.
- It is a lending platform that lends money to people in need at an interest rate based on their credit history and other factors.
- Dataset contains 466285 rows and 74 columns

	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	grade	sub_grade	...
0	1077501	1296599	5000	5000	4975.0	36 months	10.65	162.87	B	B2	...
1	1077430	1314167	2500	2500	2500.0	60 months	15.27	59.83	C	C4	...
2	1077175	1313524	2400	2400	2400.0	36 months	15.96	84.33	C	C5	...
3	1076863	1277178	10000	10000	10000.0	36 months	13.49	339.31	C	C1	...
4	1075358	1311748	3000	3000	3000.0	60 months	12.69	67.79	B	B5	...
...	...	...	...	...	...	...	...	...	...	...	...
466280	8598660	1440975	18400	18400	18400.0	60 months	14.47	432.64	C	C2	...
466281	9684700	11536848	22000	22000	22000.0	60 months	19.97	582.50	D	D5	...

DATA PREPROCESSING

1) Removing strings from emp_length and term variable and converting them to numeric.
2)Calculating months since issue date from issue-date variable and converting them from datetime to numeric.
3)Creating dummy variables for discrete variables.
4)Replacing missing value of

- total_rev_hi_lim by funded_amnt
- annual_income by mean
- other categories by 0

PD MODEL

DATA PREPROCESSING

1) Calculating the weight of evidence (WOE) for each category of independent variable and combining the similar categories into one category.
2)Keeping the category of worst credit risk as reference category (i.e, category with lowest WOE).

The above data preprocessing steps are applied on the following variables:
Discrete variables - grade, home_ownership, addr_state, verification_status, purpose, initial_list_status
Continouos 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

CREDIT SCORECARD

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

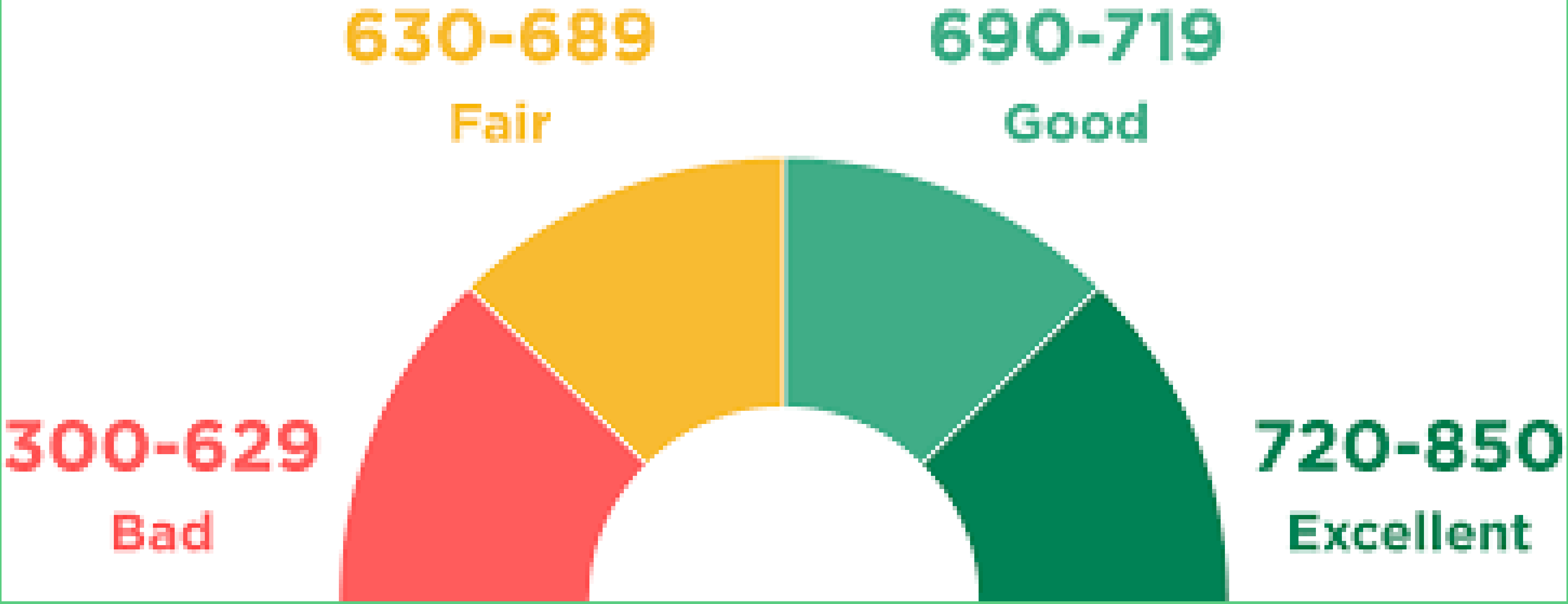
	0
362514	614.0
288564	555.0
213591	580.0
263083	637.0
165001	681.0

FICO Score

- Minimum score = 300
- Maximum score = 850

Rescaling regression coefficients of each dummy variable into scores.

Multiplying each value of the each row of the dummy variable data frame by the corresponding scores for that variable and suming them up.



LGD MODEL

STAGE 1 MODEL

IS RECOVERY RATE 0 OR NOT?

- Independent Variables - Discrete dummy variables and continuous variables
- Dependent Variable - Recovery_ rate_0_1
- Built Logistic Regression to predict if the recovery rate is 0 or not.
- AUC is 65.09%.

STAGE 2 MODEL

IF RECOVERY RATE IS GREATER THAN 0, HOW MUCH EXACTLY IS IT?

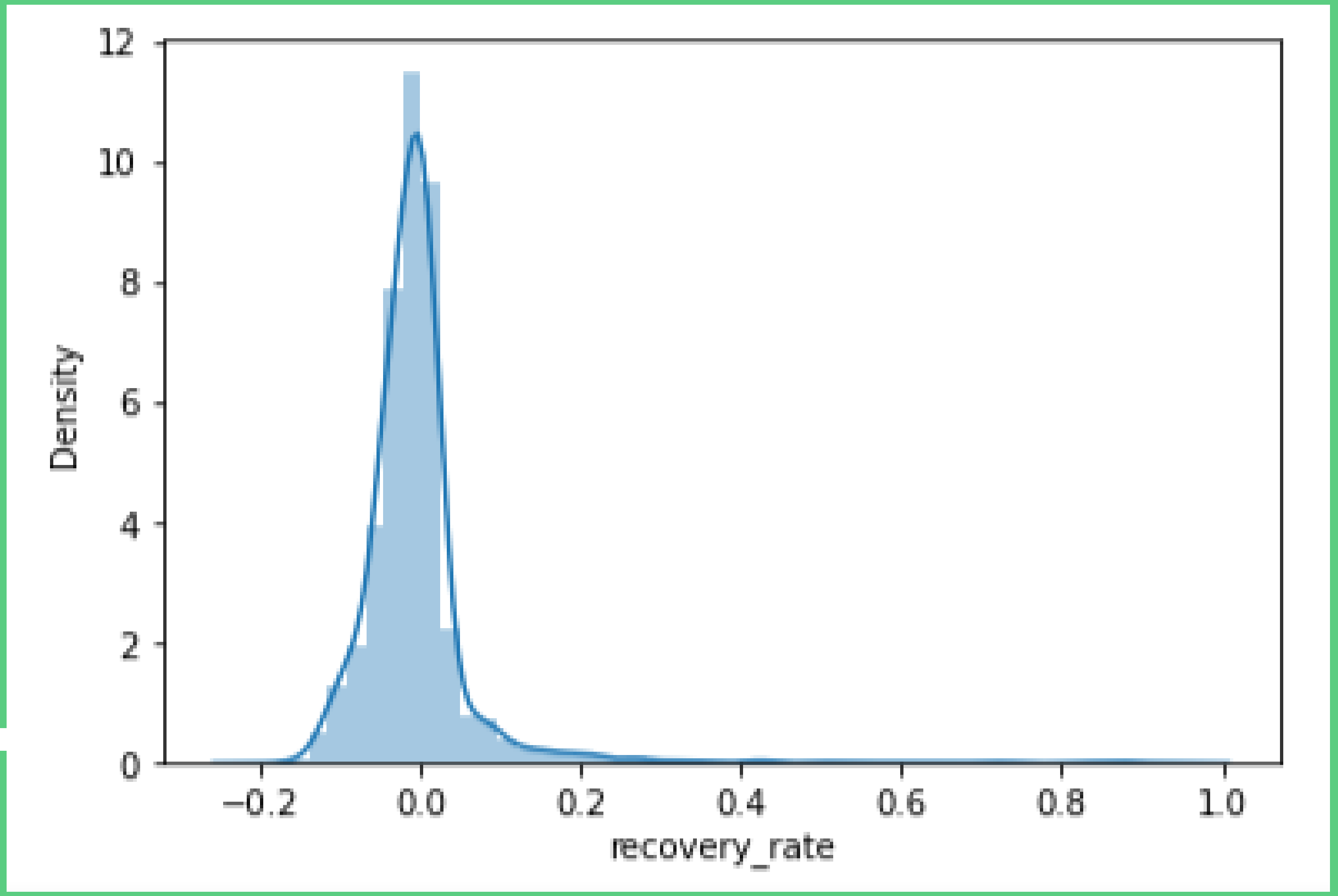
- Independent Variables - Discrete dummy variables and continuous variables
- Dependent Variable - Recovery_ rate
- Built Linear Regression to calculate the exact value.
- The distribution of the residuals resembles normal distribution and most of the residuals are symmetrically distributed.Therefore, it is a good model.

PD MODEL

- Creating a data frame that only contains the preprocessed variables.
- Dependent variable - loan status
- Building Logistic Regression with p values.
- Removing 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.
- Implementing logistic regression again on the final data set and saving the model.
- Testing the model by predicting the probabilities of input test data.
- Model performance is assed by calculating the Area Under the Receiver Operating Characteristic Curve (AUROC).
- AUC is 70.21%

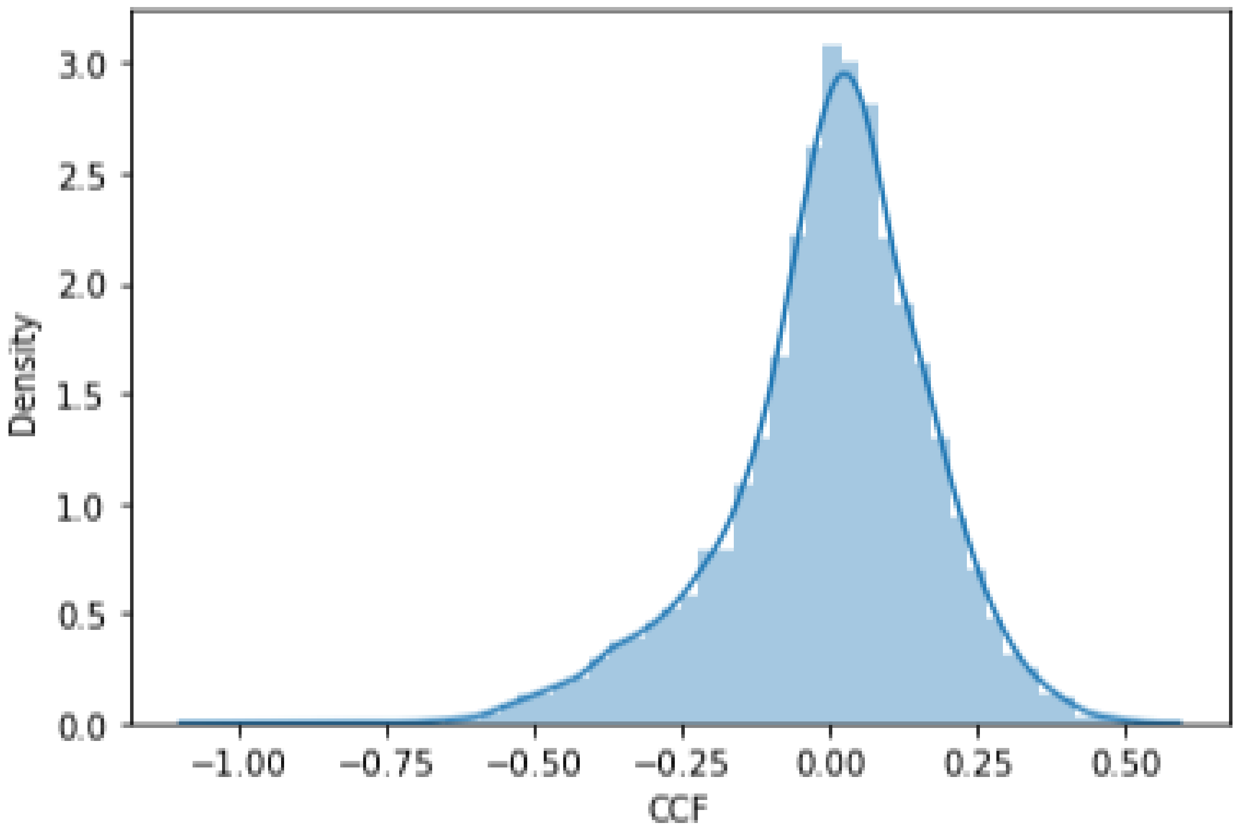
LGD = 1 - RECOVERY RATE

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



EAD MODEL

- Independent Variables - Discrete dummy variables and continuous variables
- Dependent Variable - Credit Conversion Factor (CCF)
- Built Linear Regression to calculate the exact value.
- The correlation between actual and predicted values is more than 0.53 this is moderately strong positive correlation which is good for EAD model.
- The residuals distribution resembles normal distribution and most of the residuals are symmetrically distributed around 0.



- Apply EAD linear regression model on the data frame to predict CCF values.
- EAD = CCF * FUNDED AMOUNT

TOTAL EXPECTED LOSS

EL=LGD*EAD*PD

Multiply the values PD, LGD AND EAD we get the expected loss for each loan.
Total expected loss is sum of expected loss of all the loans.

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

```
502205873.78124875
```

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

```
funded_amnt    6664052450
funded_amnt    6664052450
dtype: int64
```

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

```
funded_amnt    0.07536
funded_amnt    0.07536
dtype: float64
```

TOTAL EL = 502M

CONCLUSION AND USES

- Build a PD Model that helps the lender to know the probability of default of a borrower.
- Calculated the credit scores which the lender can refer to while giving loan to a customer.
- Build models to calculate 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.