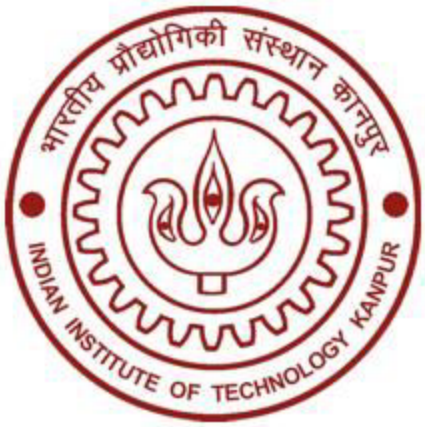


IME672A
Data Mining & Knowledge Discovery

Project Report
LOSS GIVEN DEFAULT

Group Number: 3
GROUP MEMBERS:
Manan Maheshwari - 180400
Harshit Jain - 180285
Priyam Sareen - 180555
Kadavath premsingh - 180336
Ikjot Singh - 180303

Acknowledgments

We wish to express our sincere gratitude to our instructor **Dr. Faiz Hamid** for his valuable advice and guidance in completing this project. The way he presented each and every topic in the class made the topics very interesting and understandable, which helped a lot in making our project possible.

Introduction

Loss-given default (LGD) is the amount of money a bank or other financial institution loses when a borrower defaults on a loan. It is a very important parameter in risk models used by banks to calculate their economic capital and expected losses.

In this project, we aim to predict the loss of an asset(lgd_time) to a bank. In the given dataset we have multiple parameters which will be used in various models to project this loss.

The data set has been kindly provided by a European bank and has been slightly modified and anonymized. It includes 2,545 observations on loans and LGDs. Key variables are:

- LTV: Loan-to-value ratio, in %
- Recovery_rate: Recovery rate, in %
- lgd_time: Loss rate given default (LGD), in %
- y_logistic: Logistic transformation of the LGD
- lnrr: Natural logarithm of the recovery rate
- Y_probit: Probit transformation of the LGD
- purpose1: Indicator variable for the purpose of the loan; 1 = renting purpose, 0 = other
- event: Indicator variable for a default or cure event; 1 = event, 0 = no event

Our Approach in Brief

After visualizing and preprocessing the data, we split the data into training and testing data using stratified sampling, splitting it into a 7:3 ratio. We used the training dataset for training the following models:-

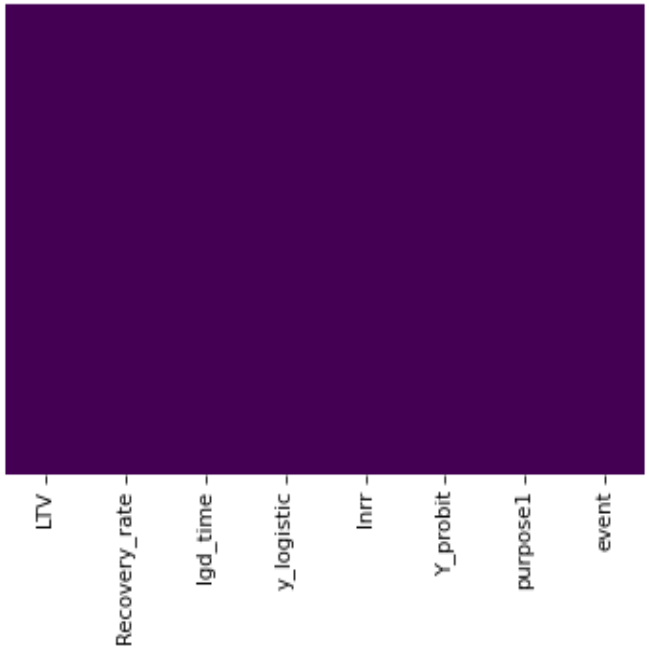
1. **Linear Regression**
2. **Transformed Linear Regression**
3. **Probit Transformed Regression**
4. **Decision Tree Regression**
5. **Support Vector Regression**
6. **Random Forest Regression**
7. **Tobit Regression**

We tested these models on the testing dataset and compared their performance on 2 factors:
b. Mean-Square Error Value
c. R^2 Value

Salient Features Of Dataset

- The given dataset has 2545 rows(data points) and 8 columns (attributes).
- LTV is a continuous-valued numeric attribute.
- lgd_time is a numeric attribute with values between 0 and 1.
- Recovery_rate is 1 - lgd_time.
- y_logistic, Y_probit and lnrr are transformations of the lgd_time variable.
- purpose1 is a binary variable that takes value 1 if the loan is for renting purposes and for other purposes 0.
- event is also a binary variable which is 1 when the borrower defaults the loan and else 0

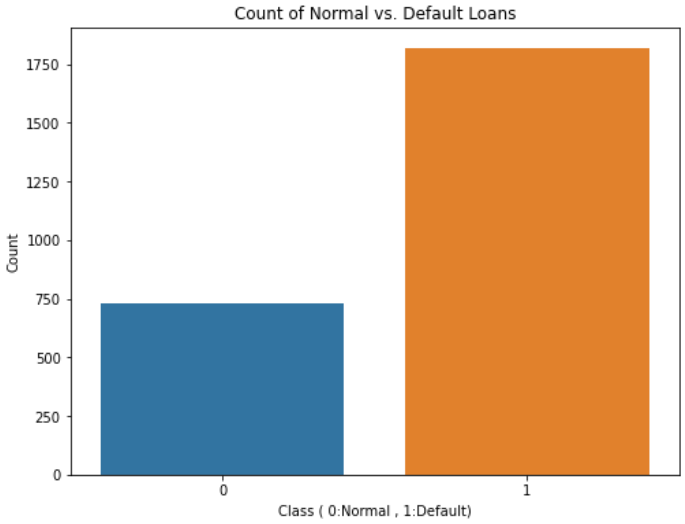
Finding the NULL values



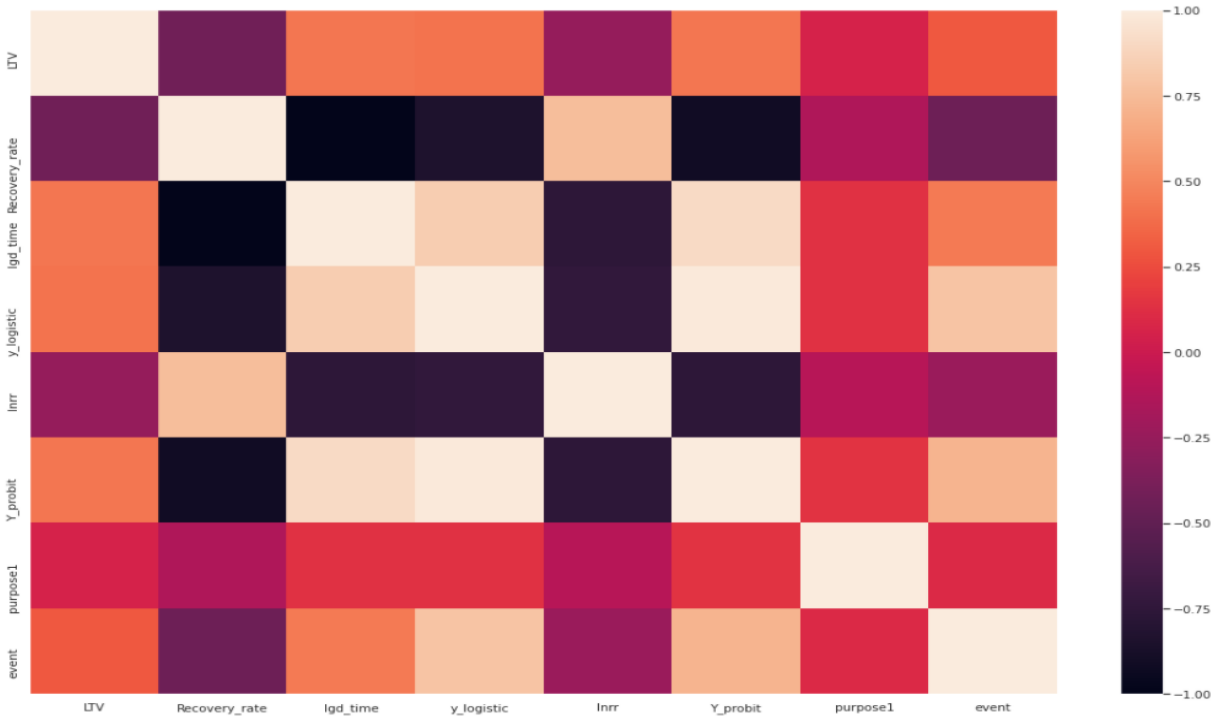
- There are no NULL values and missing data present in our dataset. So we didn't have to do any data preprocessing.

DATA VISUALISATION

- Default loans Distribution
- In our dataset 1817 loans are defaulted while the other ones are non - default loans or the loans which were cleared on time
- **It is clear that for non-default loans that the recovery_rate will be 1 hence lgd_time will be 0.**
- We then visualized the distribution of various variables through box plots, violin plots.



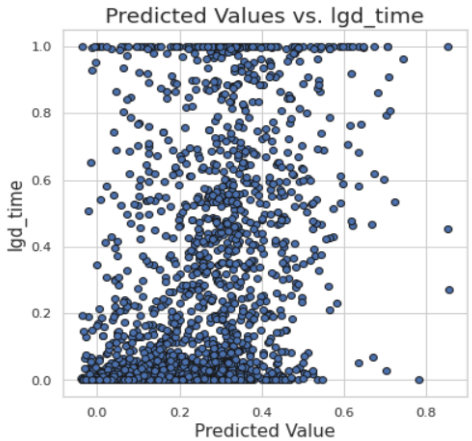
CORRELATION MATRIX



Models

- **Linear Regression**

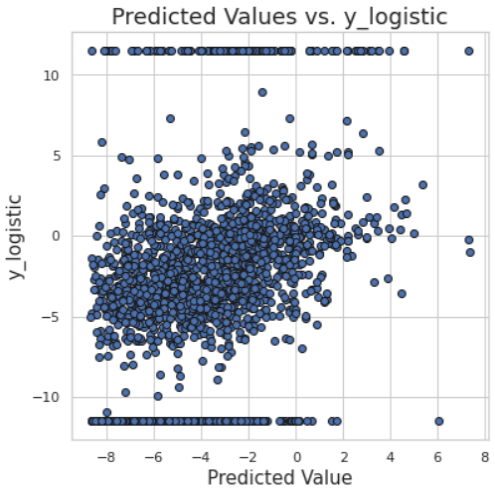
$$LGD_i = \beta' x_i + \epsilon_i$$



Mean Square Error : 0.108 R^2 Score : 0.193

- **Logistic Transformed Linear Regression**

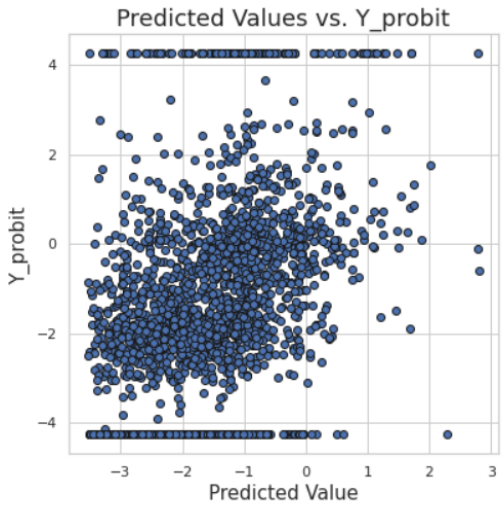
$$\ln \frac{LGD_i}{1 - LGD_i} = \boldsymbol{\beta}' \boldsymbol{x}_i + \epsilon_i$$



Mean Square Error : 36.88 R^2 Score : 0.182

- **Probit Transformed Linear Regression**

$$\Phi^{-1}(LGD_i) = \boldsymbol{\beta}' \boldsymbol{x}_i + \epsilon_i$$



Mean Square Error : 5.30 R^2 Score : 0.197

- **Random Forest Regression -** MSE : 46.47

- **Decision Tree Regression -** MSE : 52.22

- **Support Vector Regression -** MSE: 30.39
- **Tobit Regression -** MSE: 52.22

Results and Interpretation

- The Simple Linear Regression using statsmodels OLS gives a mse of 0.10 and a low R^2 value of 0.19. The correlation between the dependent variable and the independent features is weak.
- Logistic and Probit Transformed Regressions give a pretty solid relation with a similar R^2 value but they seem to fit the residuals well.
- Other regressions like Decision Tree, Support Vector, and Random Forest give a similar mean squared error and also do not seem to capture the relation between the features.
- The best model which fits the data is the Probit Transformed Regression as it captures the residuals properly and also provides a better relation between the variables.