greatlearning

# LOAN-STATUS PREDICTION FOR


Under the Guidance of

Mr. Muppidi Srikar

**GROUP 1 – Pune – Nov'19 - DSE**

**Aakanksha Patil**

**Abhishek Deshmukh**

**Sanket Kumar**

**Ravi Mahto**

greatlearning

# LendingClub

- Lending Club is an American peer-to-peer lending company, a marketplace for personal loans that matches borrowers who are seeking a loan with investors looking to lend money and make a return.

- Each borrower fills out a comprehensive application, providing their past financial history, the reason for the loan, and more.

**LendingClub Corporation**

LendingClub

| | |
|---|---|
| Type | Public |
| Traded as | NYSE: LC<br>Russell 2000 Component |
| ISIN | US52603A1097 |
| Industry | Personal finance, Software |
| Founded | 2006; 14 years ago |
| Headquarters | 595 Market Street<br>San Francisco,[1] California, U.S. |
| Key people | Scott Sanborn, CEO & President |
| Products | Peer-to-peer lending |
| Revenue | ▲US$ 500.8 million (2016)[2]<br>▲US$574.5 million (2017) |
| Operating income | ▼US$ -153.4 million (2017) |
| Net income | ▼US$ -153.8 million (2017) |
| Total assets | ▼US$4.641 billion (2017) |
| Total equity | ▼US$922.5 million (2017) |
| Number of employees | 1,530 (2016)[2] 1,837 (2017) |
| Website | lendingclub.com |

# Problem Definition

- Loans are not completely paid off and the borrowers default on the loan. This problem shall be addressed by doing this project.

- The model thus developed shall help to determine if the applicant falls in the category of Non Defaulter or Defaulter by analyzing the details from the comprehensive application form and their past financial history and the details about loan demanded.

- This shall help the business to predict which applicant can default over the time and thus, rejecting the loan application and saving the business from loss.

greatlearning

# Model Building Requirements

- Build a short term and business specific model.
- Realistic model and not a generalized model
- Business oriented model – needs to be calibrated and updated with time and as per business requirement
- Sufficing Banking domain business requirement

greatlearning

# Dataset – Preprocessing & Challenges

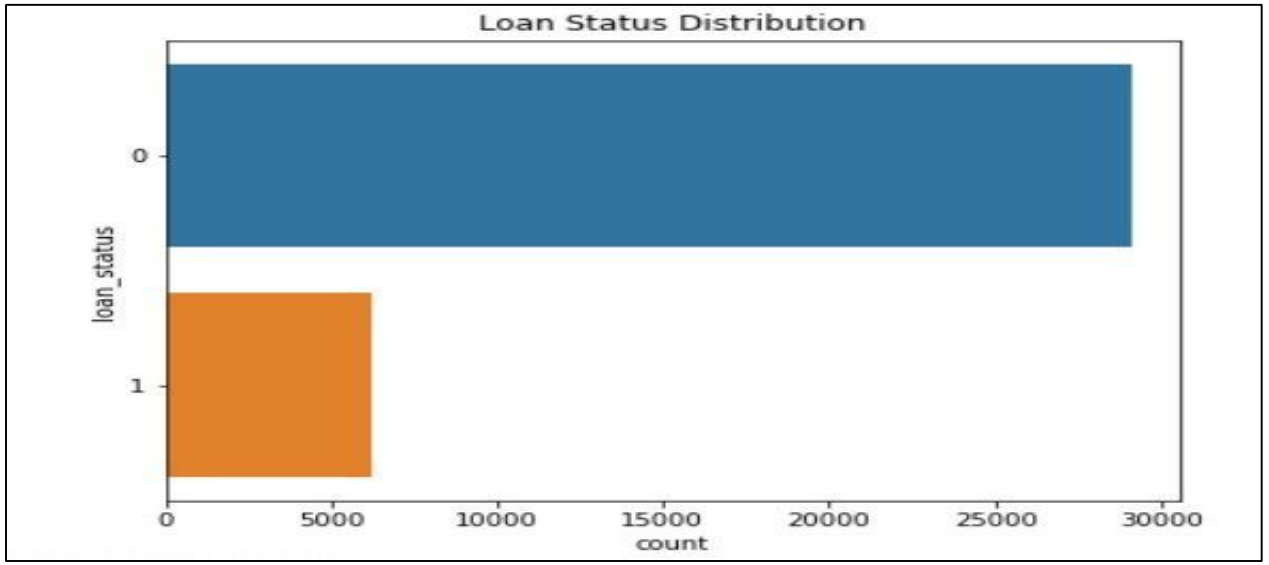
- Rows : 22Lakhs, Columns : 145,  Size : 1.1GB
- Understanding the data
- Selecting the correct sample size as per business requirement and relevance.
- Analyzing & Understanding of 145 domain specific attributes.
- Selecting the most recent data with a more realistic model building approach

# Dataset - Information

- Train : - Rows: 31345 | Columns: 95
- Test : - Rows: 4303 | Columns: 95
- Target :- loan_status
- Numerical: 82 | Categorical: 9
- 39 attributes with missing values , 18 attributes with Outliers
- Feature Categorization

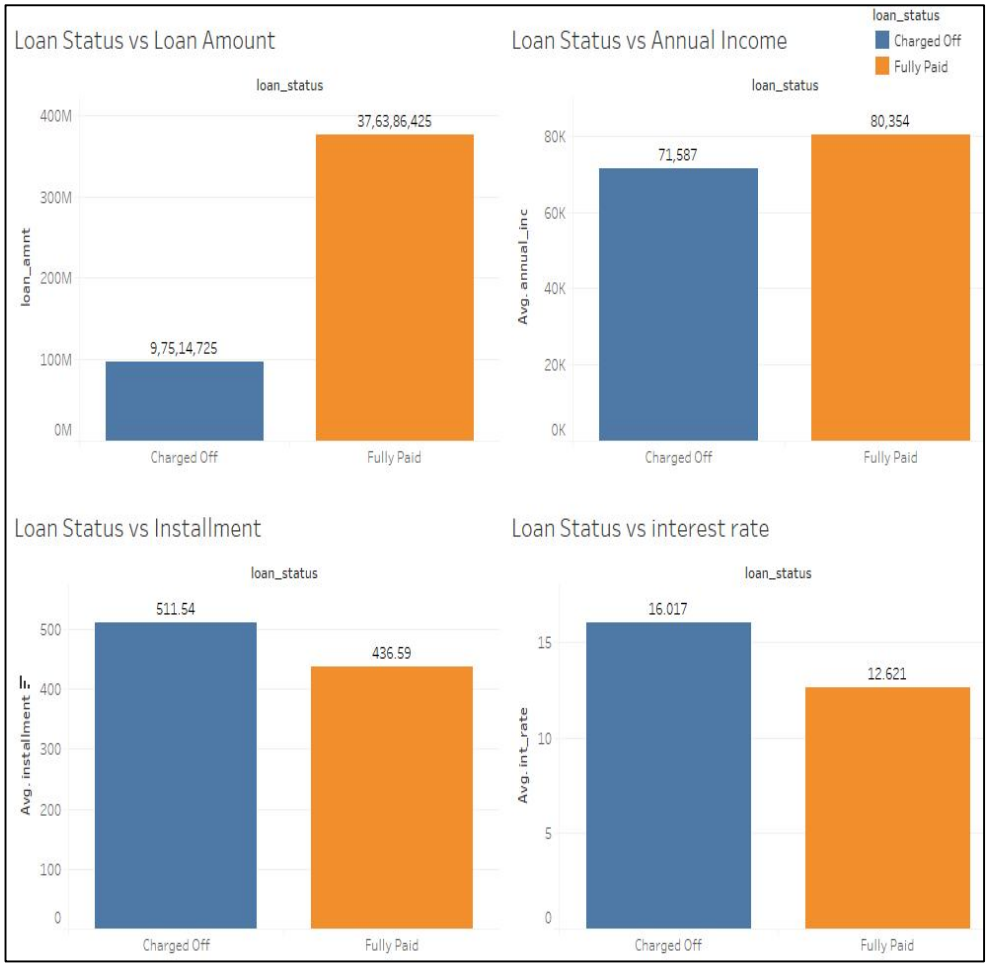
| TYPE | No. of Attributes |
|---|---|
| APPLICATION TYPE DETAIL | 1 |
| GEOGRAPHICAL DETAILS | 1 |
| INDIVIDUAL PROPERTY DETAILS | 1 |
| LISTING STATUS DETAILS | 1 |
| INTEREST DETAILS | 2 |
| INDIVIDUAL EMPLOYMENT DETAILS | 3 |
| INDIVIDUAL INQUIRY DETAILS | 4 |
| LOAN DETAILS | 5 |
| JOINT ACCOUNT DETAILS | 14 |
| INDIVIDUAL - DELINQ/DEROG/RECORD DETAILS | 18 |
| INDIVIDUAL ACCOUNT DETAILS | 45 |

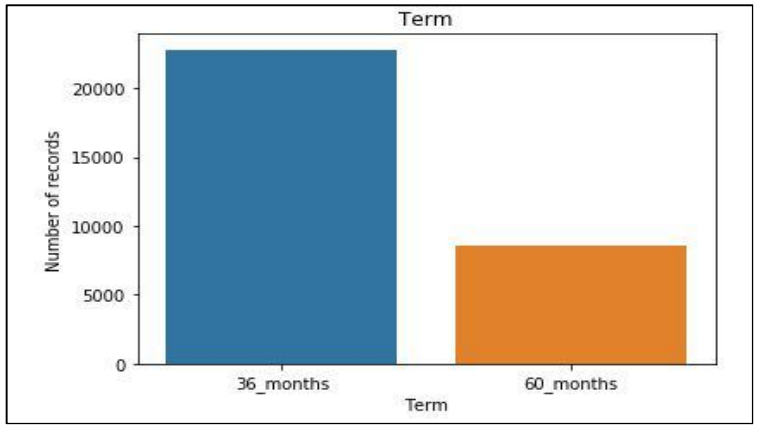
# EDA : Target

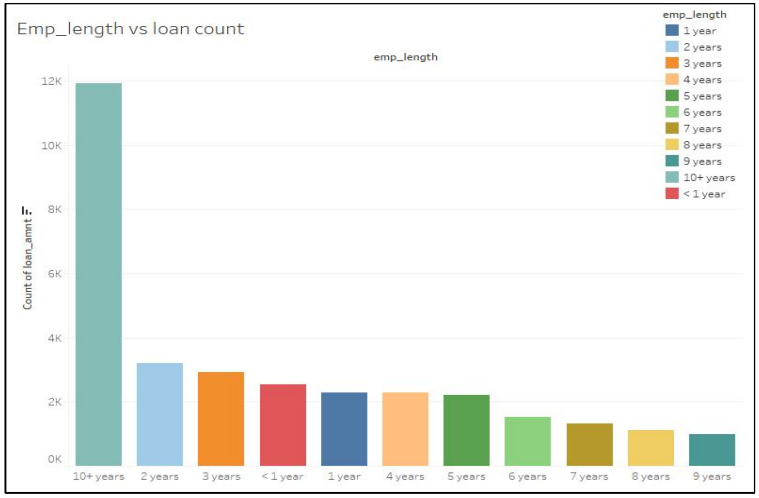



- Loan Status - Target Variable
- 0: Non-Defaulters/Fully paid | 82%
- 1: Defaulters/Charged-off | 18%
- Imbalanced Data
- Calls for Balancing Techniques

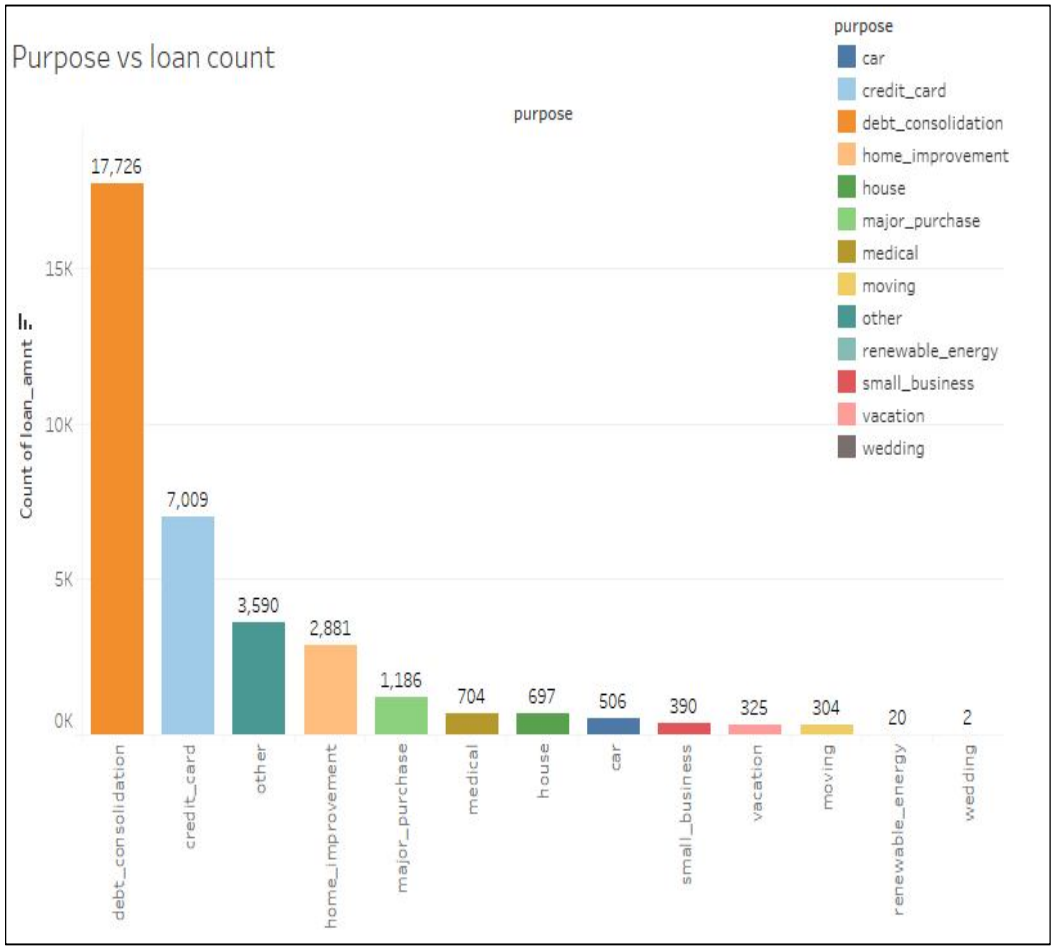
# EDA : Uni-variate Analysis

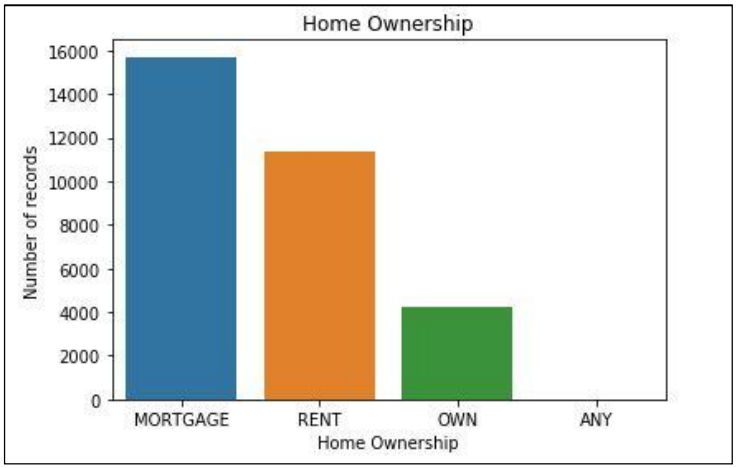


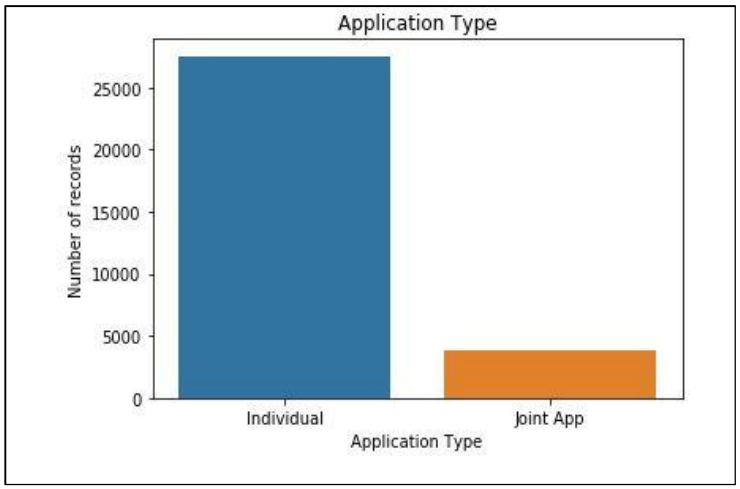



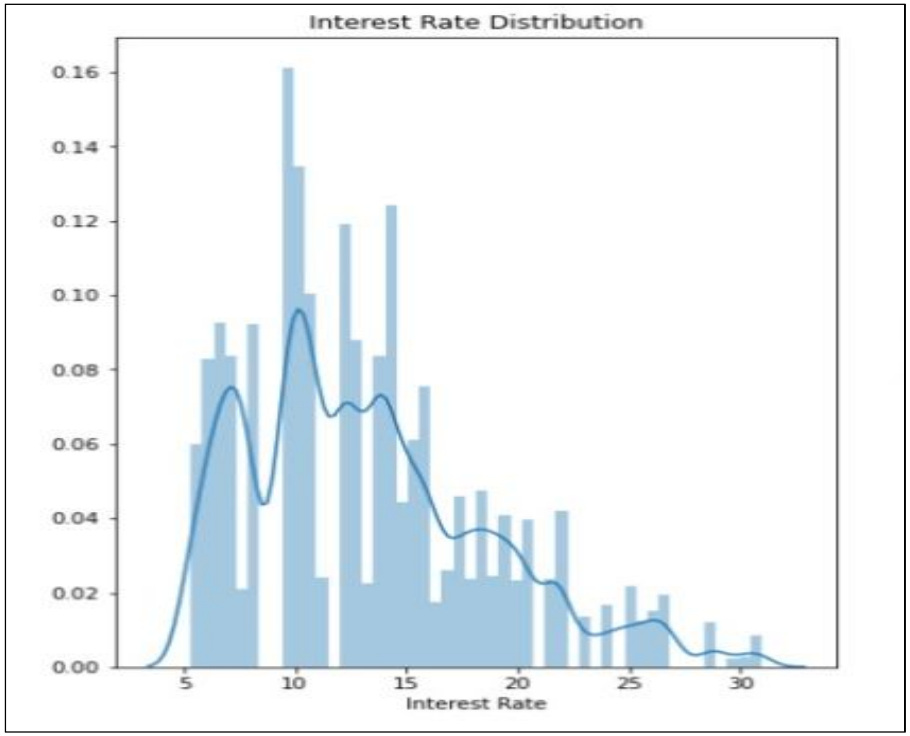
# EDA : Uni-variate Analysis

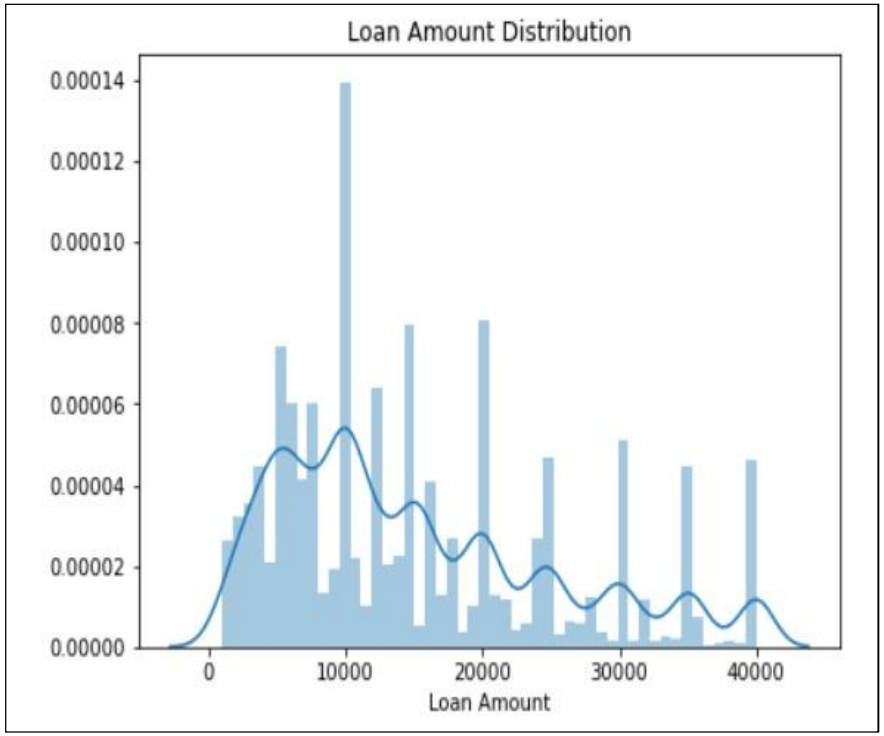


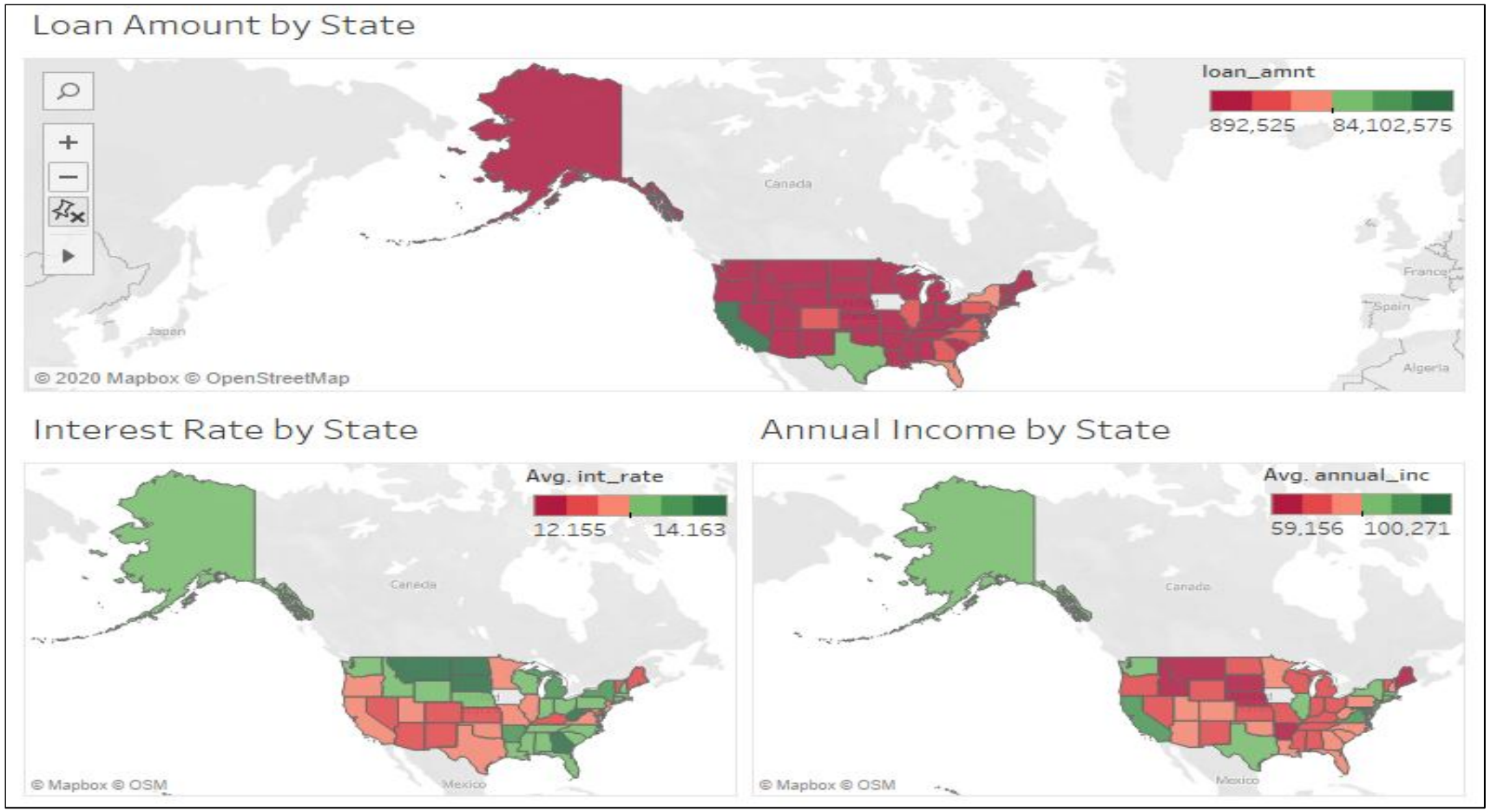



# EDA : Uni-variate Analysis

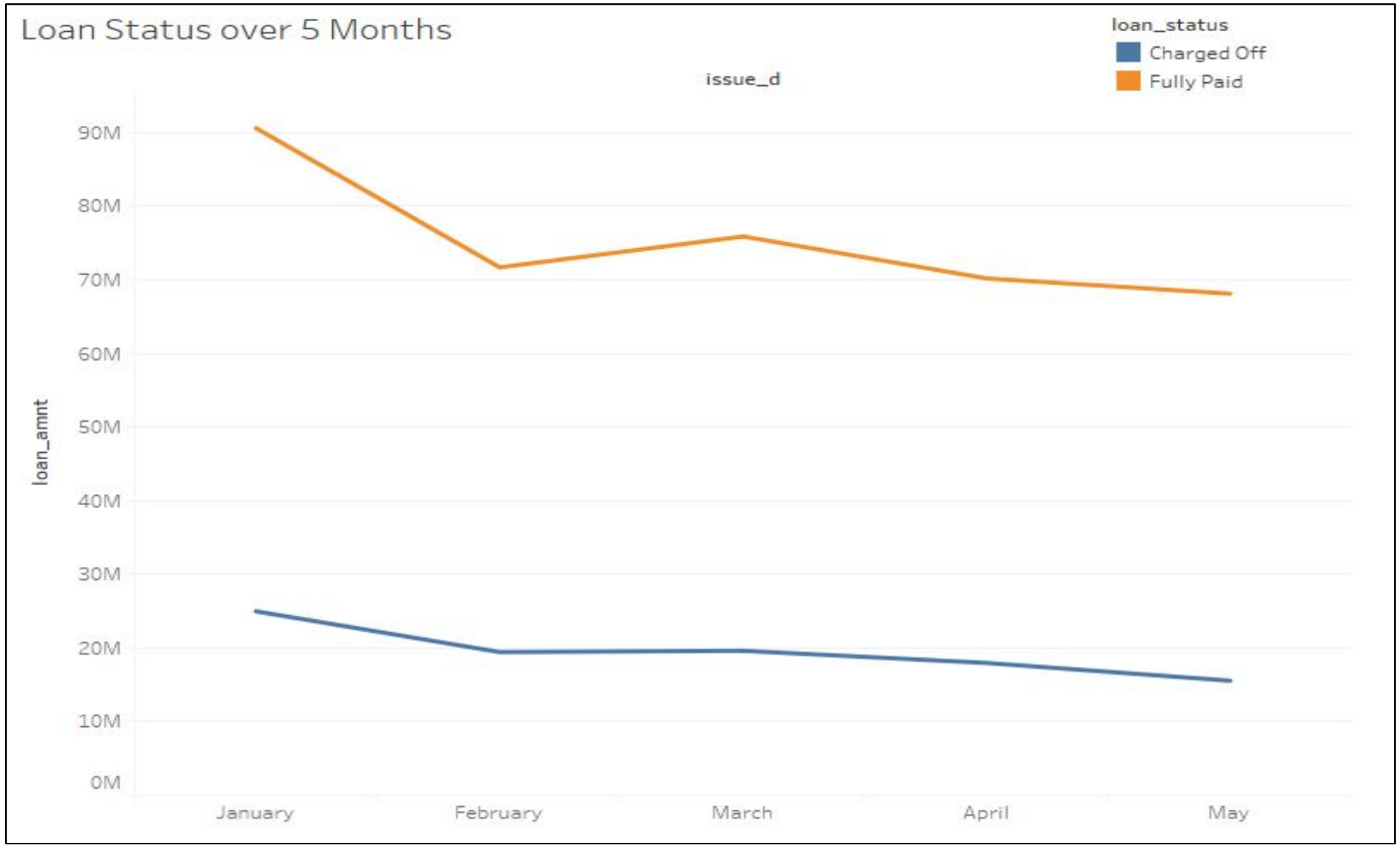


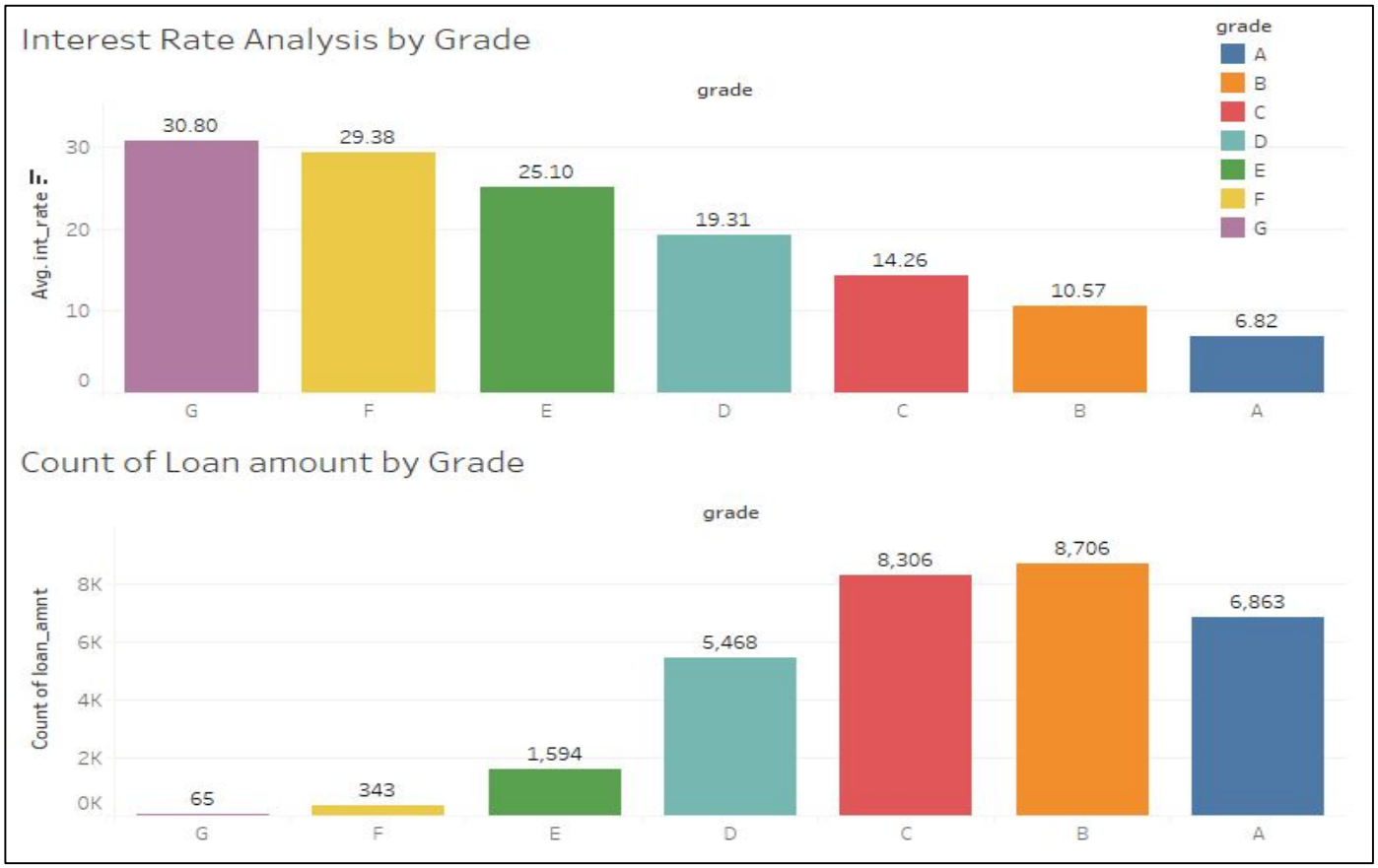
# EDA: Bivariate Analysis



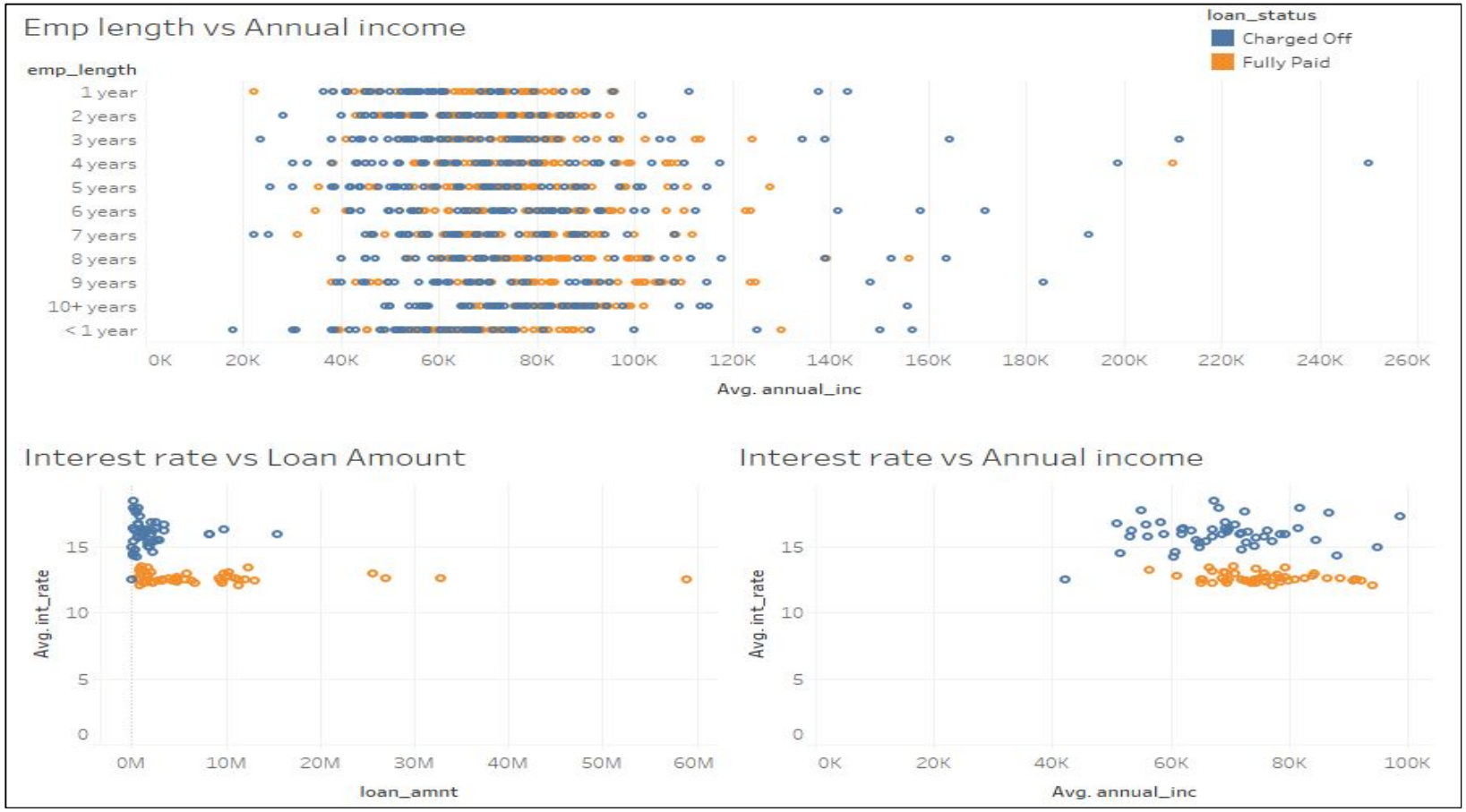
# EDA: Bivariate Analysis



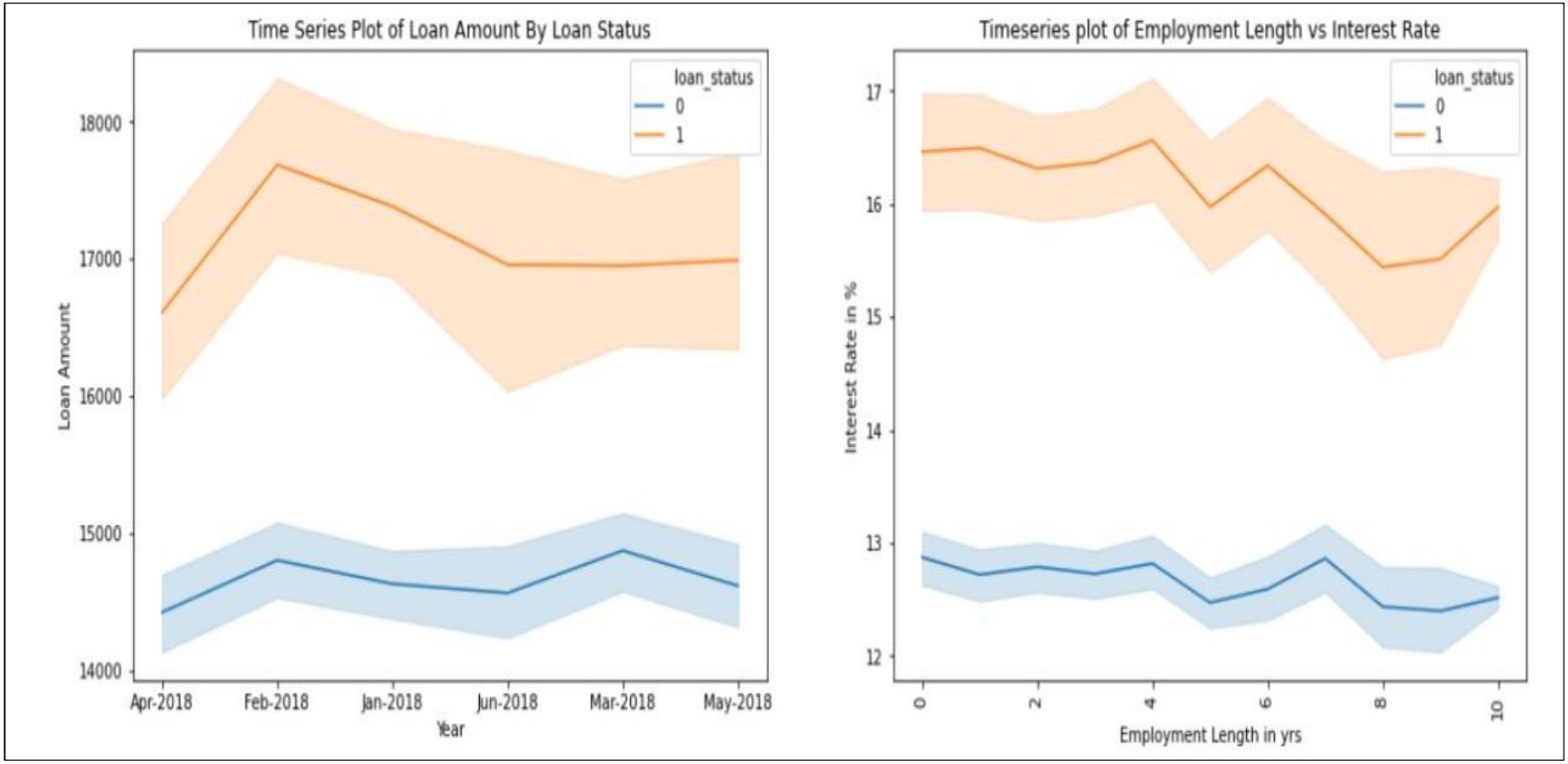
# EDA: Bivariate Analysis



# EDA: Bivariate Analysis

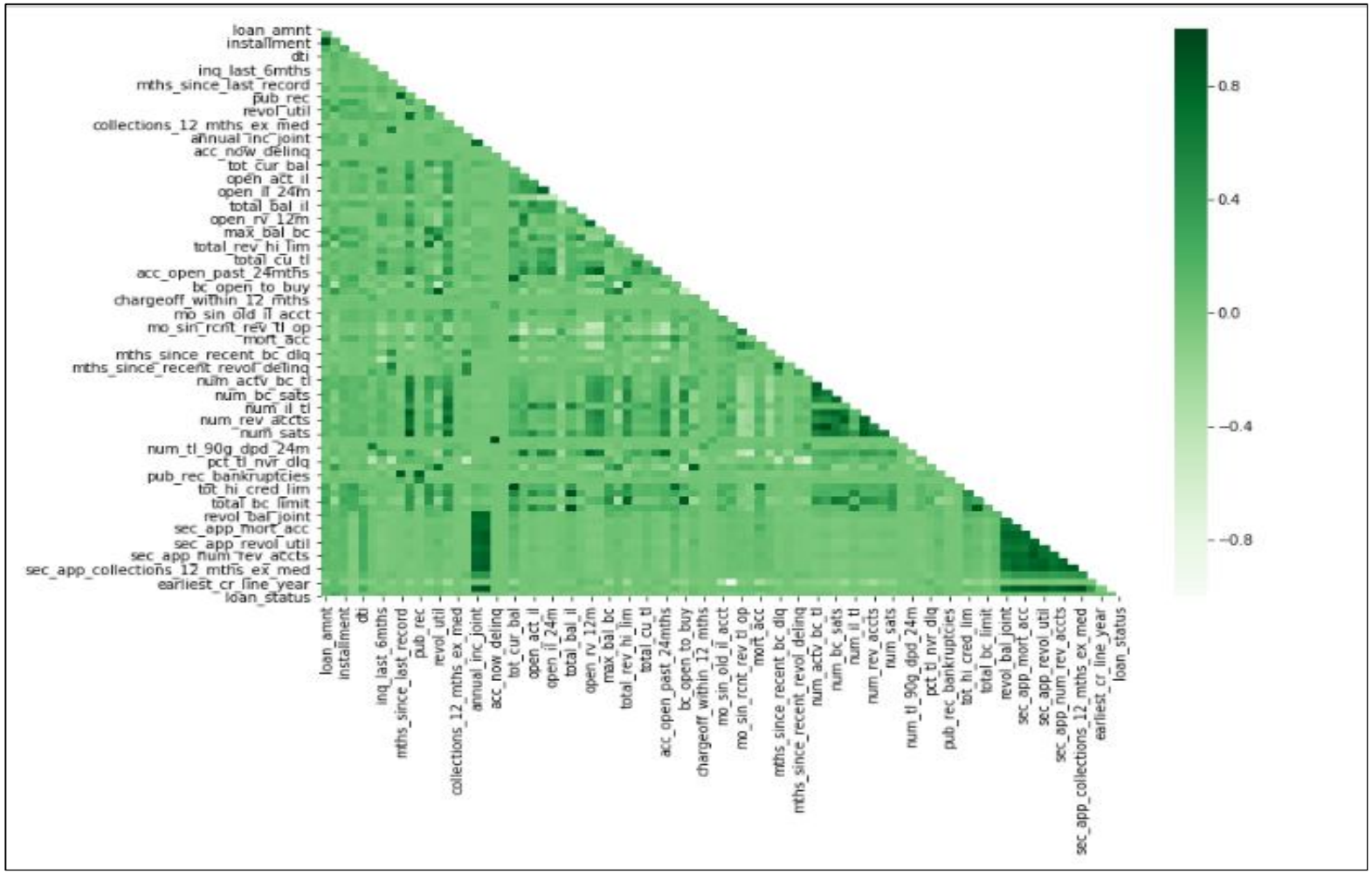


# EDA: Bivariate Analysis



greatlearning

# EDA: Multivariate Analysis



greatlearning

# Data Cleaning

- Outlier treatment – Capping in between 1st to 99th percentile
- Yeo Johnson Transformation – Skewness reduction and standard scaling
- Encoding Categorical Variables – Ordinal Encoding and One Hot Encoding
- Missing Value Imputation
  - 100 % missing values – drop the attribute
  - Non Random Missing Values – Imputation with constant
  - 0.5% missing values – drop the observation
  - Random Missing Values – Imputation using KNN Imputer

greatlearning

# Statistical Testing

- Categorical Attributes:
  - Chisquare – 9 attributes significant

- Numerical Attributes:
  - Independent t Test (normal distributed features) and Mannwhitneyu (non normal features) – 72 attributes significant

# Base Model Fitting

*Linear Model:*

| Feature Selection Technique | Model | F1 weighted | Variance Error |
|---|---|---|---|
| Base Models | LR | 0.76 | 0.0000187 |

*Non-Linear Model:*

| Feature Selection Technique | Model | F1 weighted | Variance Error |
|---|---|---|---|
| Base Models | RFC | 0.75 | 0.0000259 |

greatlearning

# Feature Selection Techniques

- VIF for numerical attributes – 24/82 attributes, VIF greater than 10.
- PCA – Dimension Reduction – 95% variance i.e. 38 PCs considered
- SelectKBest - different feature numbers like 60,70,80,100
- RFE - different feature numbers like 60,70,80,100

greatlearning

# ML Models

- *Linear Models*

| Feature Selection Technique | Model | F1 weighted | Variance Error |
|---|---|---|---|
| Base Models | LR | 0.76 | 0.0000187 |
| KBEST (100 features) | LR | 0.82 | 0.0000135 |
| **KBEST (80 features)** | **LR** | **0.82** | **0.0000091** |
| RFE (57 features) | LR | 0.76 | 0.0000185 |
| Ridge Penalty | LR | 0.76 | 0.0000218 |
| KBEST (57 features) | LR | 0.82 | 0.0000141 |
| Multicollinearity treated (VIF) | LR | 0.76 | 0.0000277 |
| RFE (100 features) | LR | 0.76 | 0.0000174 |
| PCA - 0.95 variance explained | LR | 0.76 | 0.0000236 |
| VIF - Kbest(80 features) | LR | 0.76 | 0.0000187 |
| VIF-RFE(DecisionTree) | LR | 0.76 | 0.0000188 |

greatlearning

# ML Models

- Non Linear Models with SelectKBest , RFE
- Different Models:
  - K Nearest Neighbors Classifier
  - Decision Tree Classifier
  - Random Forest Classifier
  - Bagging Classifier
  - Ada Boost Classifier
  - Gradient Boost Classifier

greatlearning

# Best Model

- Feature Selection Method - KBEST – 80 features
- Model – Logistic Regression
- F1 Weighted Score – 82% (cross validation)
- Variance error - 0.0000091

# Hyper parameter tuning

- Over sampling
- Under sampling
- Focus on Recall – Train 0.69 and Test 0.69

greatlearning

# Business Recommendations

- Linear Model with better interpretability
- Model with good Recall score thus, focusing on eliminating the False Negatives (defaulter getting predicted as non defaulter)
- Top influencing attributes are interest rate, term, employment length, loan amount

greatlearning

# Model Interpretability

- As the loan amount increases, the chances of an applicant might turn out to a defaulter is high
- Higher the interest rate, possibility of applicant to become a default is high
- Lower the annual income, higher the chances of defaulters
- Higher the debt to income ratio, there are the chances that applicants might turn out to a defaulter is high
- As the balance to credit limit ratio increases, possibility of applicant to become a default is high
- More the recent inquiries, more the chance of defaulting.
- More the number of taxes missed, higher the possibility of loan default
- Higher the employment length of the applicant, lower the possibility of being default.
- Higher the term, more the chance of defaulting.

greatlearning

# Future Scope

- Accuracy improvement through fine tuning of model
- End to end Deployment of model using pipeline and web developments frameworks

greatlearning

# Thanks !!