FIM 590-002 Group Project:
Prediction of Probability of Charged using Lending Club Dataset

Under the Guidance OF
Dr. Ram Valluru

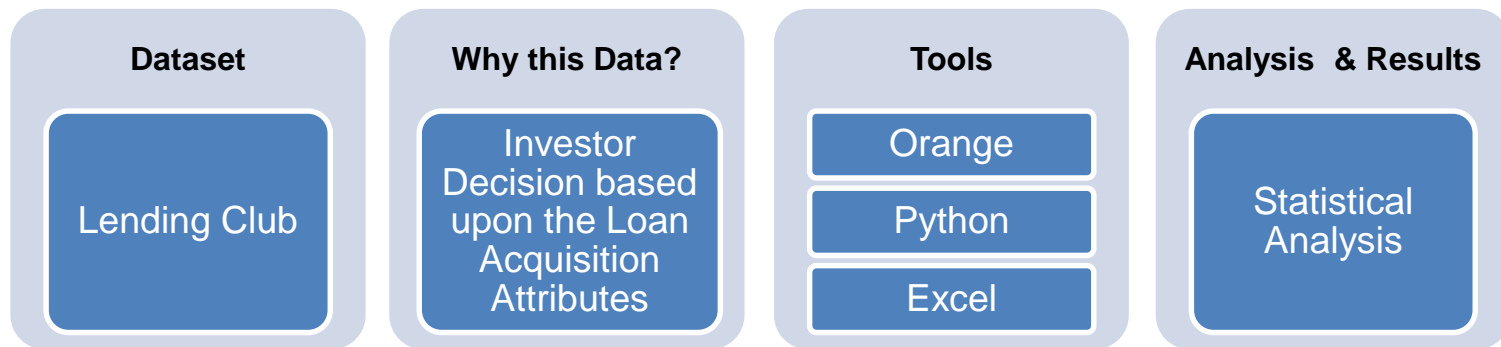**Group Members:**
Jingjing Zhang
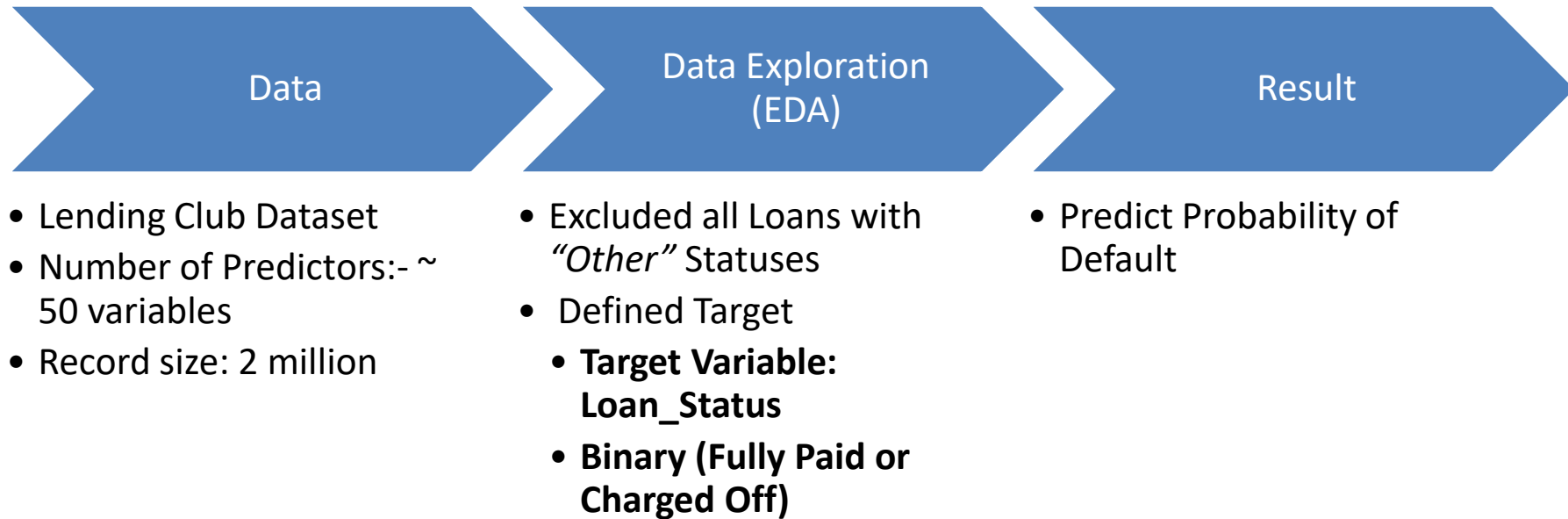Kartheek Manavarthi
Muktanidhi S Dhotrad
Rohit Khurana
Sandeep Girada

1. TOPIC & DATA

- **Project Objective:** If we are given a set of acquisition attributes with respect to a loan on lending club, how can an investor decide to invest or not?

Dataset	Why this Data?	Tools	Analysis & Results
Lending Club	Investor Decision based upon the Loan Acquisition Attributes	Orange Python Excel	Statistical Analysis

2. DATA EXPLORATION & TRANSFORMATIONS

Data	Data Exploration (EDA)	Result

- Lending Club Dataset
- Number of Predictors:- ~ 50 variables
- Record size: 2 million

- Excluded all Loans with *"Other"* Statuses
- Defined Target
  - **Target Variable: Loan_Status**
  - **Binary (Fully Paid or Charged Off)**

- Predict Probability of Default
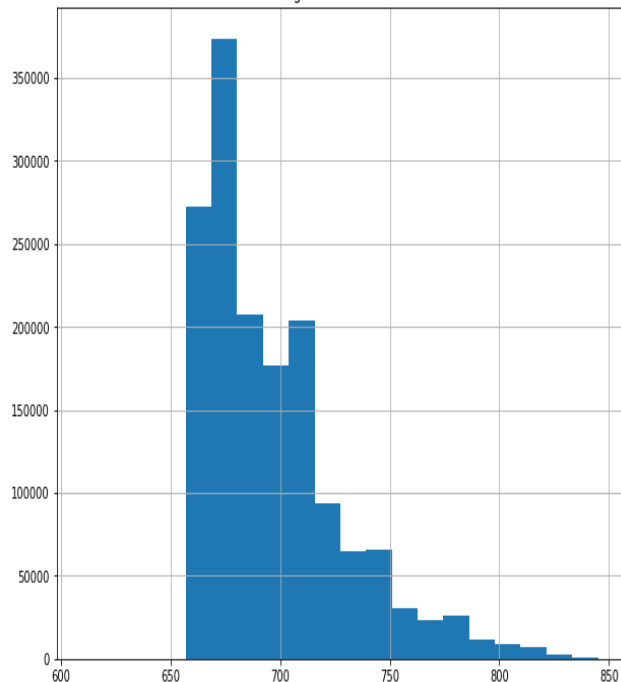
TRANSFORMATIONS AND METHODS
(2. DATA EXPLORATION & TRANSFORMATIONS)

- DTI:
  - Simple mean imputer for DTI
  - Negatives DTI were made zero,

- home_ownership
  - Values like (RENT, MORTGAGE, OWN, OTHER) created dummies for each

- Test and Train divided by *stratified sampling*

- **Data Segregation:** 20% holdout, all performance done on holdout, fitting is done using 5-fold validation
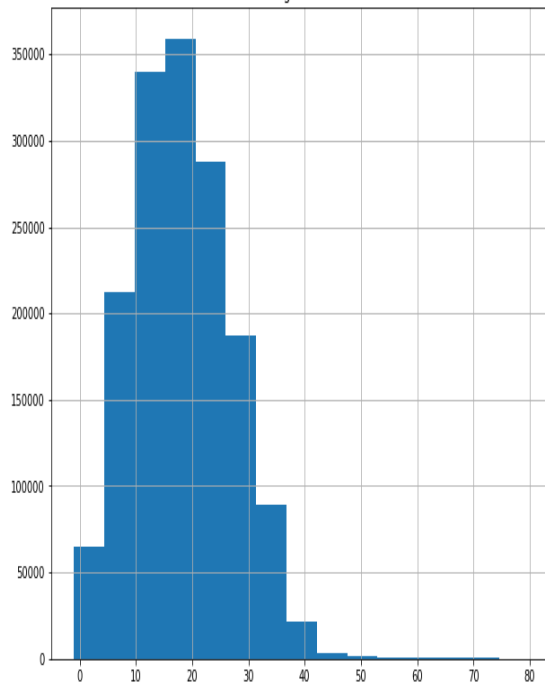
DATA VISUALIZATION – FEATURES
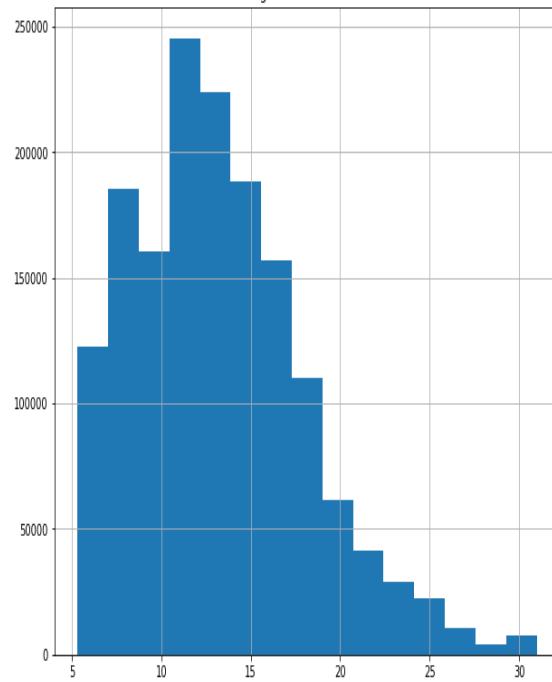## (2. DATA EXPLORATION & TRANSFORMATIONS)

DATA VISUALIZATION – FEATURES AGAINST TARGET VARIABLE
(2. DATA EXPLORATION & TRANSFORMATIONS)
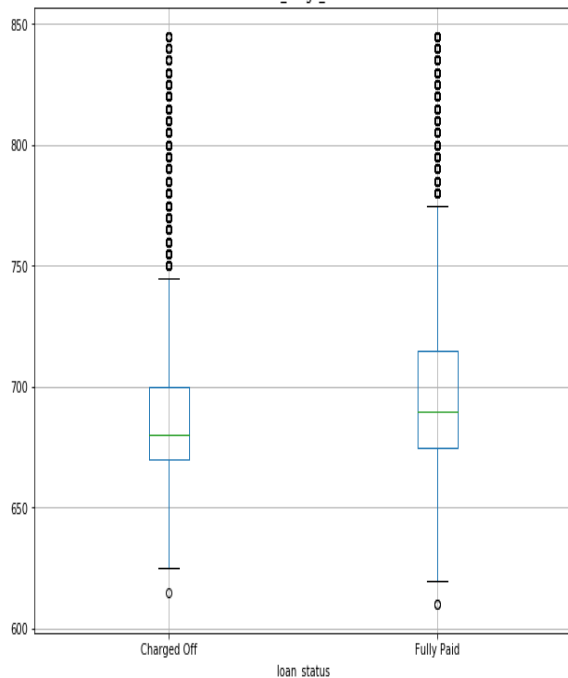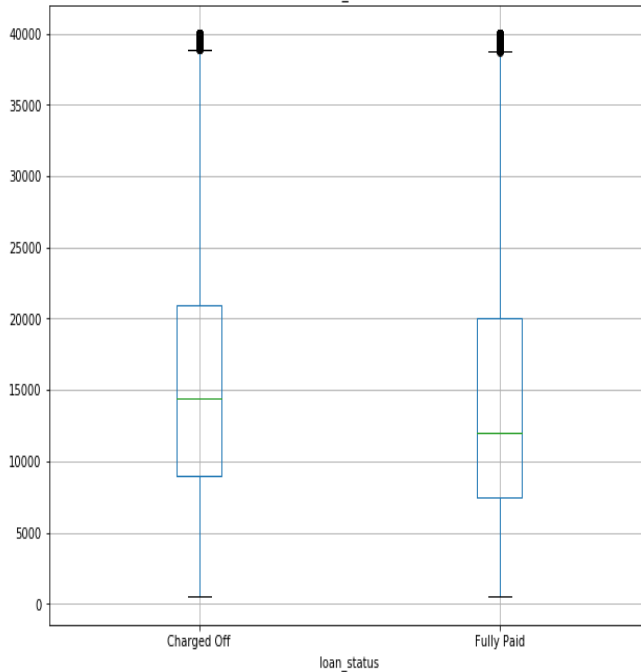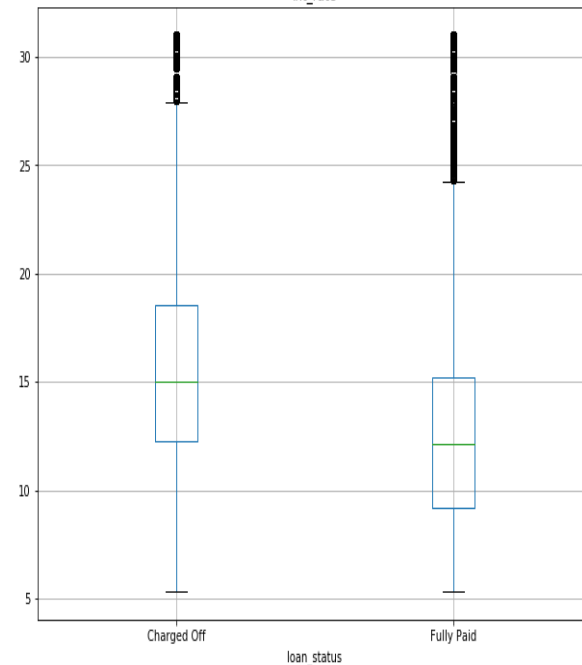
Feature Selection and Target Variable

Independent Variables
- Delinquency in Last 2 Years
- Credit Inquiries in Last 6 Months
- Interest Rates
- Debt to Income (DTI)
- Loan Amount
- Loan Purpose
- FICO_Range_LOW
- Home Ownership

Target Variable
- Loan_Status

3. MODEL BUILDING & ASSESSMENT

Models

Assessment

Gradient Boasting Classifier Algorithm

Model Report

Random Forest

ROC Curve

Logistic Regression

AUC
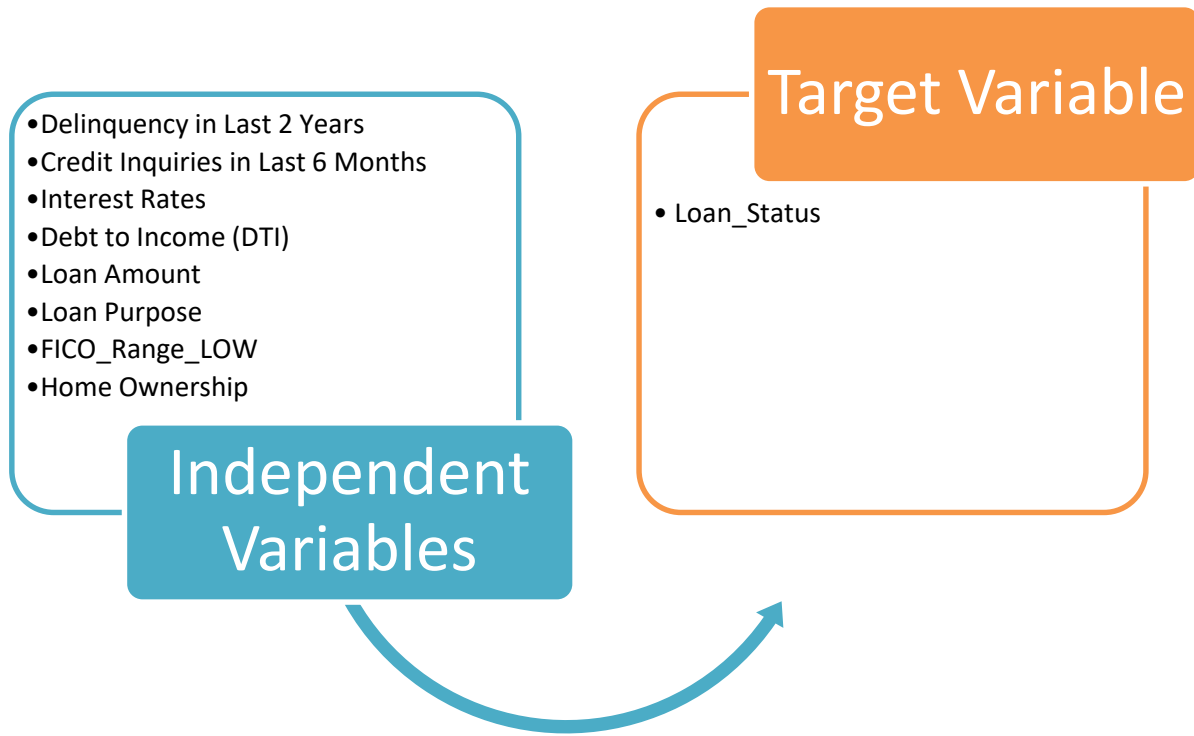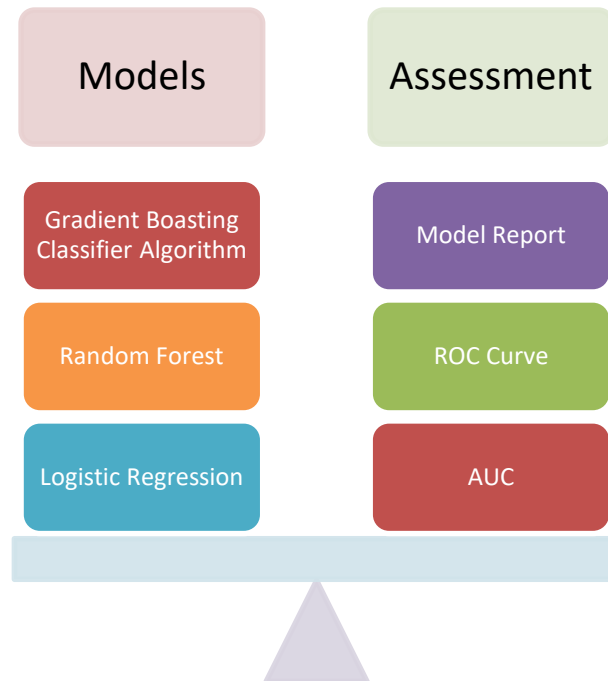
- We have also tried LightGBM and GBC (Gradient Boosting Classifier) Algorithms

MODEL 1: <u>LOGISTICS REGRESSION</u> MODEL ANALYSIS
(3. MODEL BUILDING & ASSESSMENT)

Unbalanced					Balanced				

Unbalanced

	precision	recall	F1 - score	support
0	0.81	0.99	0.89	315389
1	0.50	0.05	0.10	77128

accuracy			0.80	392517
Macro avg	0.65	0.52	0.49	392517
Weighted avg	0.75	0.80	0.73	392517

AUC:	0.69238999

	Predicted 0	Predicted 1
Actual 0	311207	4182
Actual 1	72981	4147

Balanced

	precision	recall	F1 - score	support
0	0.88	0.65	0.75	315389
1	0.31	0.63	0.41	77128

accuracy			0.65	392517
Macro avg	0.59	0.64	0.58	392517
Weighted avg	0.77	0.65	0.68	392517

AUC:	0.69334827

	Predicted 0	Predicted 1
Actual 0	206274	109115
Actual 1	28857	48271

MODEL 2: <u>RANDOM FOREST</u> MODEL ANALYSIS
(3. MODEL BUILDING & ASSESSMENT)

Unbalanced				

	precision	recall	F1 - score	support
0	0.80	0.99	1.00	315389
1	0.60	0.05	0.01	77128

accuracy			0.80	392517
Macro avg	0.70	0.50	0.46	392517
Weighted avg	0.76	0.80	0.72	392517

AUC:	0.699777506

	Predicted 0	Predicted 1
Actual 0	314875	514
Actual 1	76351	777

Balanced				

	precision	recall	F1 - score	support
0	0.88	0.63	0.73	315389
1	0.30	0.66	0.42	77128

accuracy			0.80	392517
Macro avg	0.59	0.64	0.57	392517
Weighted avg	0.77	0.63	0.67	392517

AUC:	0.6997674957

	Predicted 0	Predicted 1
Actual 0	197245	118144
Actual 1	25997	51131

MODEL 2: RANDOM FOREST MODEL ANALYSIS
(3. MODEL BUILDING & ASSESSMENT)
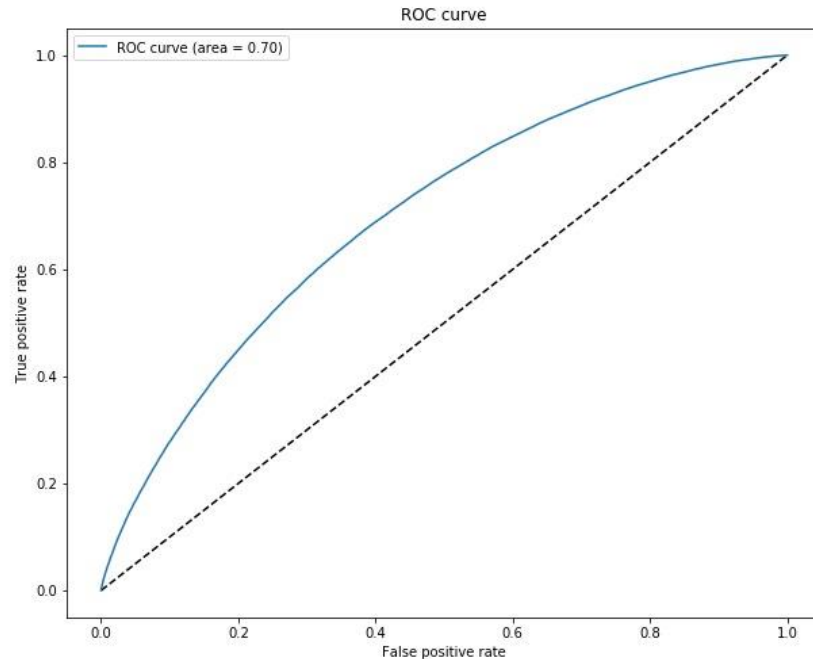
MODEL 3: GRADIENT BOOSTING CLASSIFIER MODEL
(3. MODEL BUILDING & ASSESSMENT)

	Predicted 0	Predicted 1
Actual 0	313466	1923
Actual 1	74658	2470

AUC:	0.699712

	precision	recall	F1 - score	support
0	0.81	0.99	0.89	315389
1	0.56	0.03	0.06	77128

accuracy			0.80	392517
Macro avg	0.68	0.51	0.48	392517
Weighted avg	0.76	0.80	0.73	392517

Cost Function for Comparison

Model	Balancing	False Positive ($0.1)	False Negative ($0.9)	Cost of Prediction
Logistic Regression	Unbalanced	4182	72981	$ 66,106.5
	Balanced	109115	28857	$ 36,882.8
Random Forest	Unbalanced	514	76351	$ 68,767.3
	Balanced	118144	25997	$ 35,208.7
Gradient Boosting	Unbalanced	1923	74658	$ 67,384.5

THANK YOU