Lending Club Loan Analysis

Group 1

Benjamin Allen <beallen@syr.edu>, Harper He <xhe128@syr.edu>, Abhiram Gopal <agopal@syr.edu>

Agenda

- Project description
- Methods Summary
- Results Summary
- Problems encountered
- Team member contribution

Project Description

Lending Club

P2P Lending

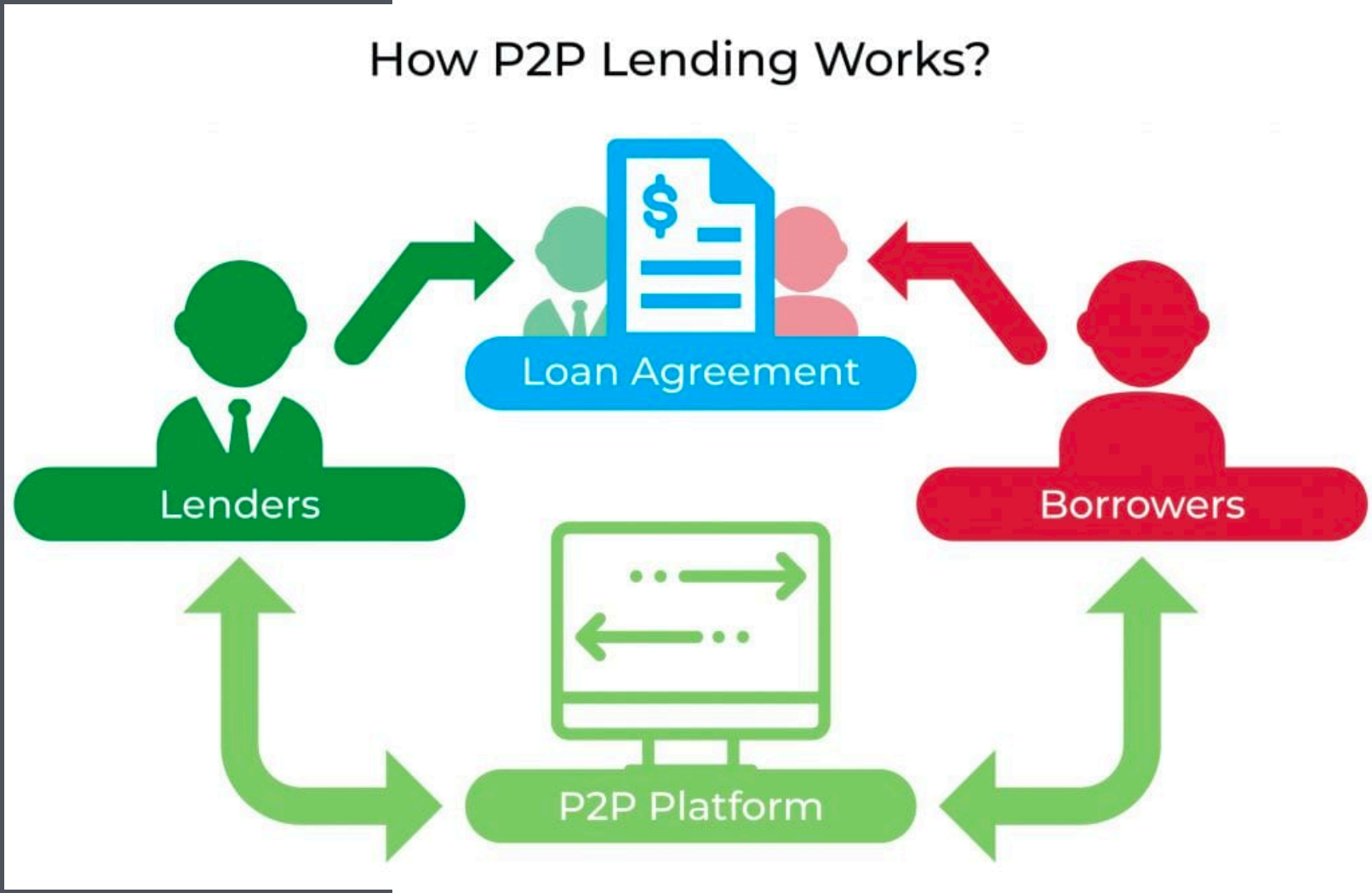
Peer-to-peer (P2P) lending is a Practice of lending/investing or borrowing money from one private individual to another private individual.

LendingClub

Lending Club

P2P Lending

Peer-to-peer (P2P) lending is a Practice of lending/investing or borrowing money from one private individual to another private individual.



LendingClub

Problem Statement

Risks to Consider

Lending Club are devoted to examining and finding some of the relevant factors to if the current loans would be charged off so as to avoid loss.

Investors

Investors purchase Notes, which correspond to fractions of loans, to potentially earn competitive returns.

LendingClub

LendingClub screens borrowers, facilitates the transaction, and services the loans.

Borrowers

Borrowers use loans to consolidate debt, improve their homes, finance major purchases, and more.

Loans are issued via WebBank, member FDIC

Problem Statement

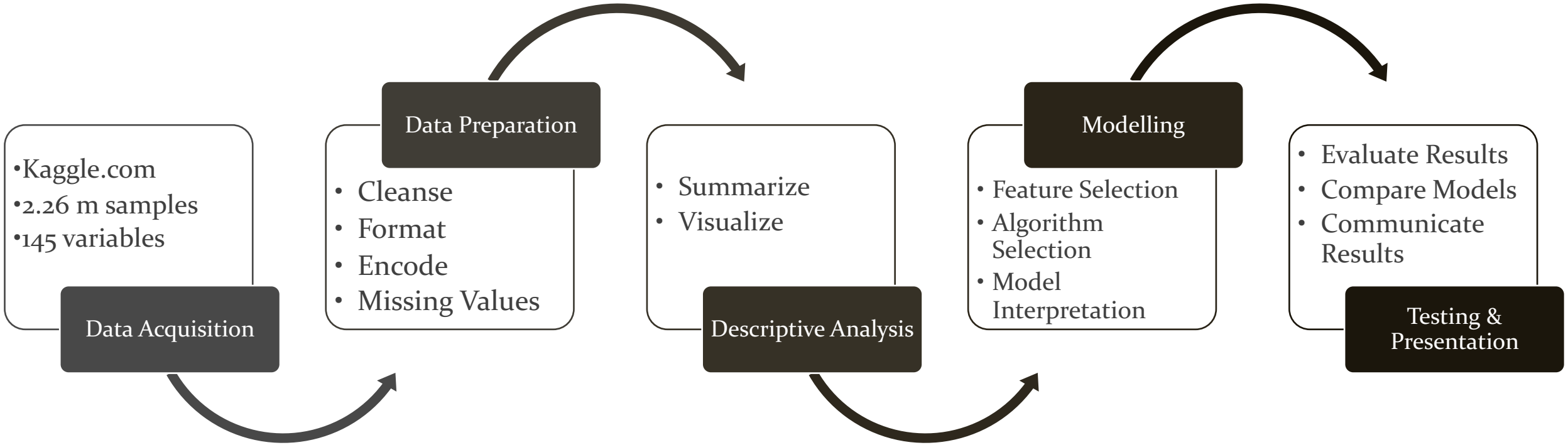
Predict if a current loan will be charged off

- Monitor the loans in progress
- Define people with bad debt habits
- Alert investors
- Avoid loss
- Save Lending Club's reputation

Methodology

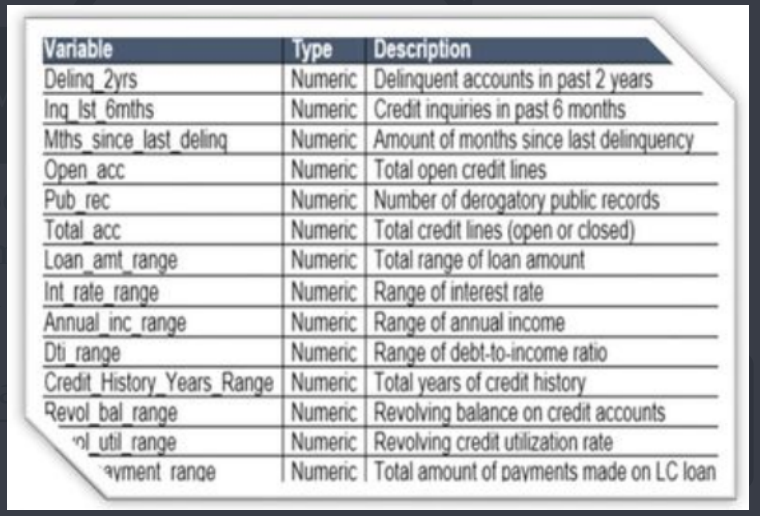
2

Approach


•Kaggle.com
•2.26 m samples
•145 variables
Data Acquisition
Data Preparation
• Cleanse
• Format
• Encode
• Missing Values
• Summarize
• Visualize
Descriptive Analysis
Modelling
• Feature Selection
• Algorithm Selection
• Model Interpretation
• Evaluate Results
• Compare Models
• Communicate Results
Testing & Presentation

Approach

- Time Span:
  11 Years (2007-2018)
- Size:
  2.26 million observations * 145 variables
- Information
  Borrower's demographic information
  Borrower's past credit history
  Latest payment information
  Current loan Information

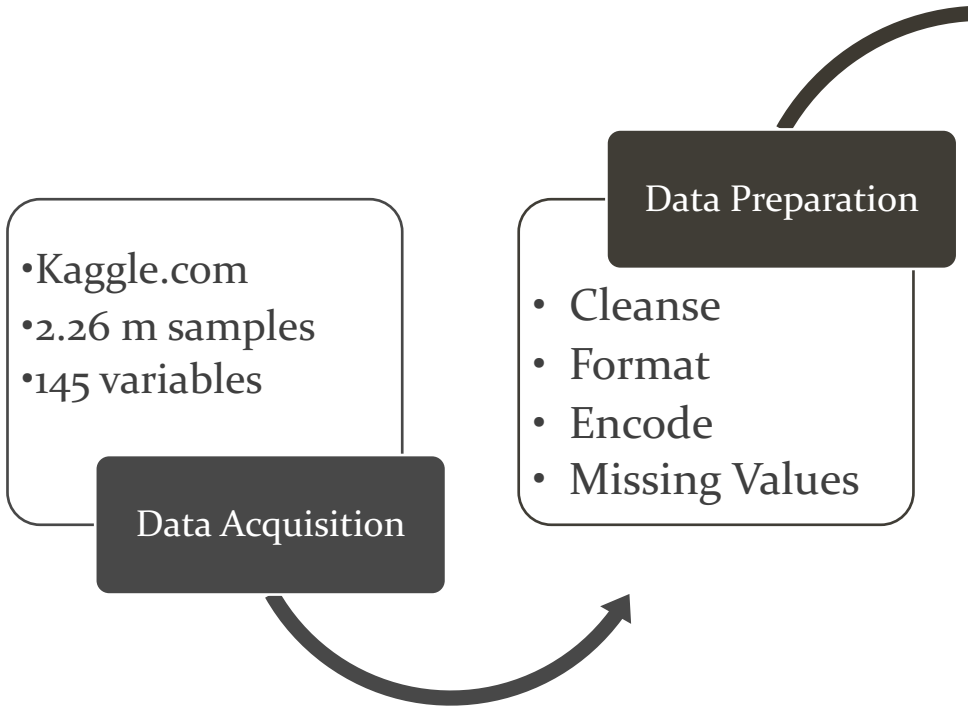


Variable	Type	Description
Delinq_2yrs	Numeric	Delinquent accounts in past 2 years
Inq_lst_6mths	Numeric	Credit inquiries in past 6 months
Mths_since_last_delinq	Numeric	Amount of months since last delinquency
Open_acc	Numeric	Total open credit lines
Pub_rec	Numeric	Number of derogatory public records
Total_acc	Numeric	Total credit lines (open or closed)
Loan_amt_range	Numeric	Total range of loan amount
Int_rate_range	Numeric	Range of interest rate
Annual_inc_range	Numeric	Range of annual income
Dti_range	Numeric	Range of debt-to-income ratio
Credit_History_Years_Range	Numeric	Total years of credit history
Revol_bal_range	Numeric	Revolving balance on credit accounts
ol_util_range	Numeric	Revolving credit utilization rate
ayment_range	Numeric	Total amount of payments made on LC loan

- Kaggle.com
- 2.26 m samples
- 145 variables

Data Acquisition

Approach

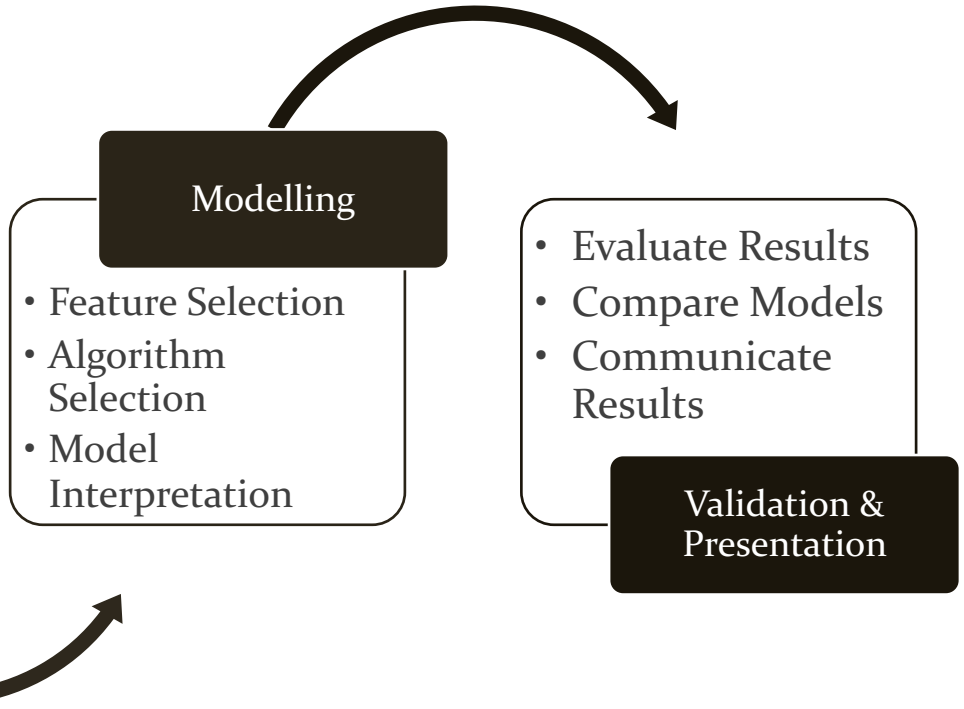


- Missing Value:
  - Remove variables with more than 50% NAs
  - Replace with mean (numeric)
  - Replace with mode (categorical)
- Numeric variables
  - Winsorize (limit outliers)
  - Standard Scale
- Categorical variables:
  - String Indexer
  - One Hot Encode
- Size:
  - 1.20 million observations * 72 variables

Approach

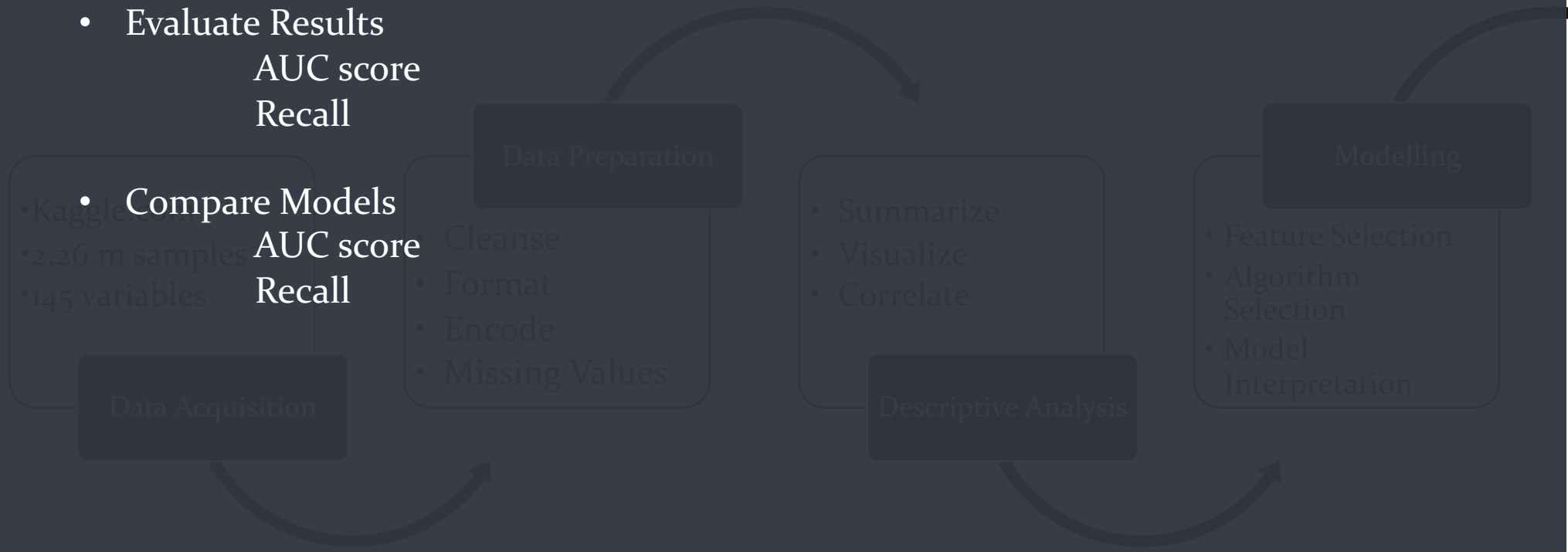
- Feature Selection:
  - Correlation Matrix
  - Random Forest Top 20 Features

- Algorithm Selection:
  - Random Forest
    - MaxDepth
    - MaxBins
  - Logistic Regression
    - NumBins
    - ElasticNetParam
    - RegParam
  - GBT Model
    - MaxDepth
    - MaxBins



Approach

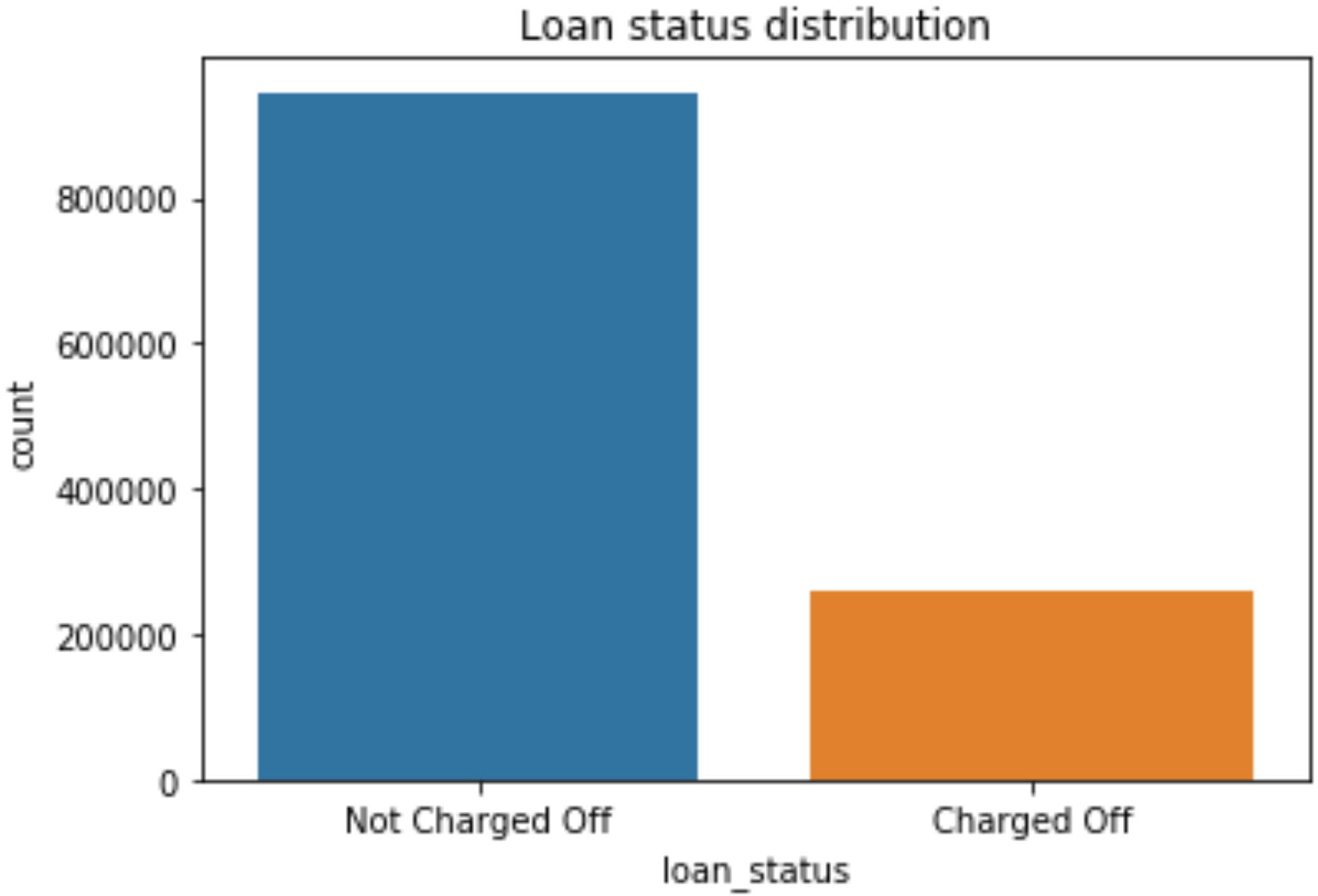
- Evaluate Results
  - AUC score
  - Recall
- Compare Models
  - AUC score
  - Recall



Descriptive Analysis

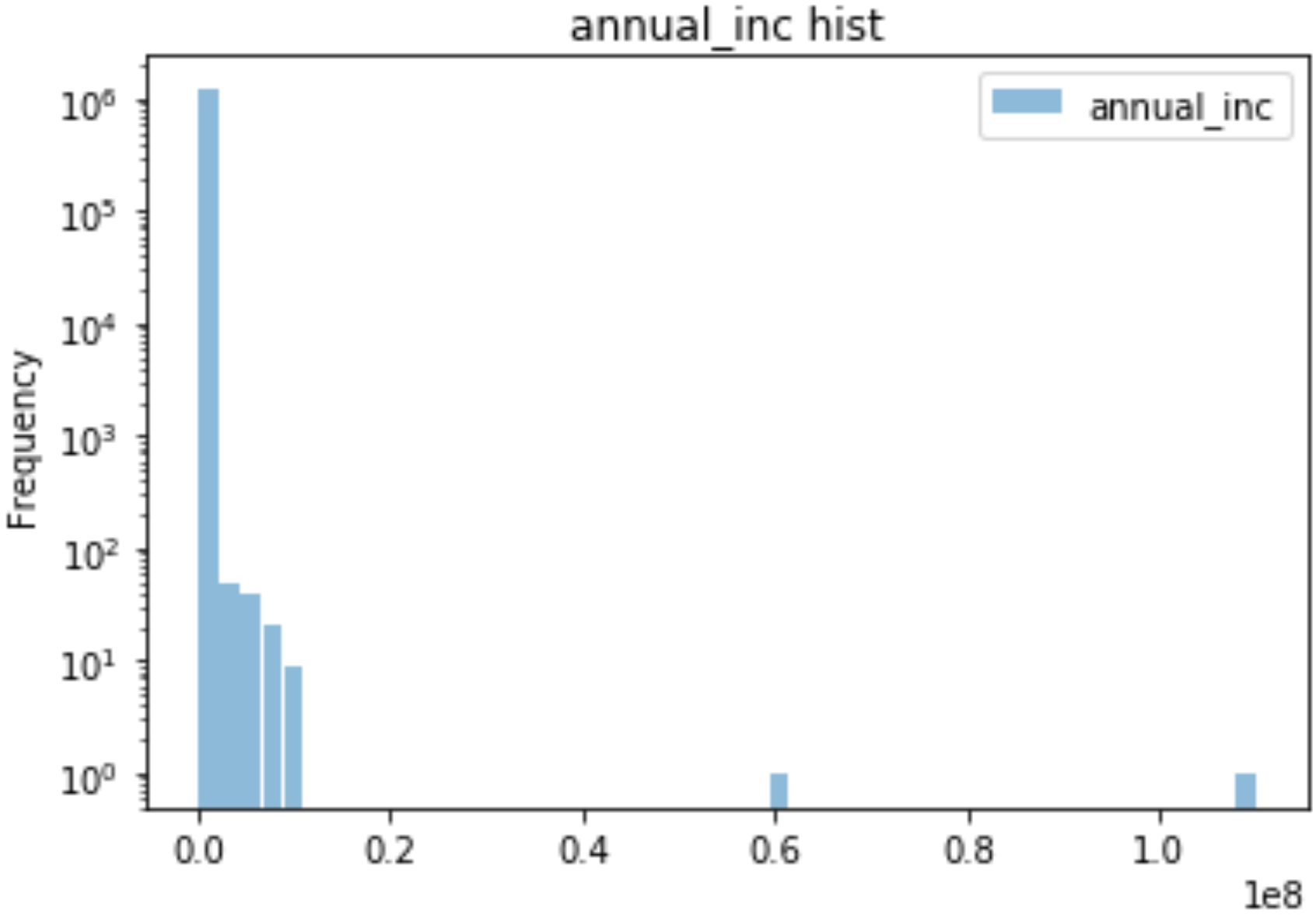
3

Loan Status Distribution



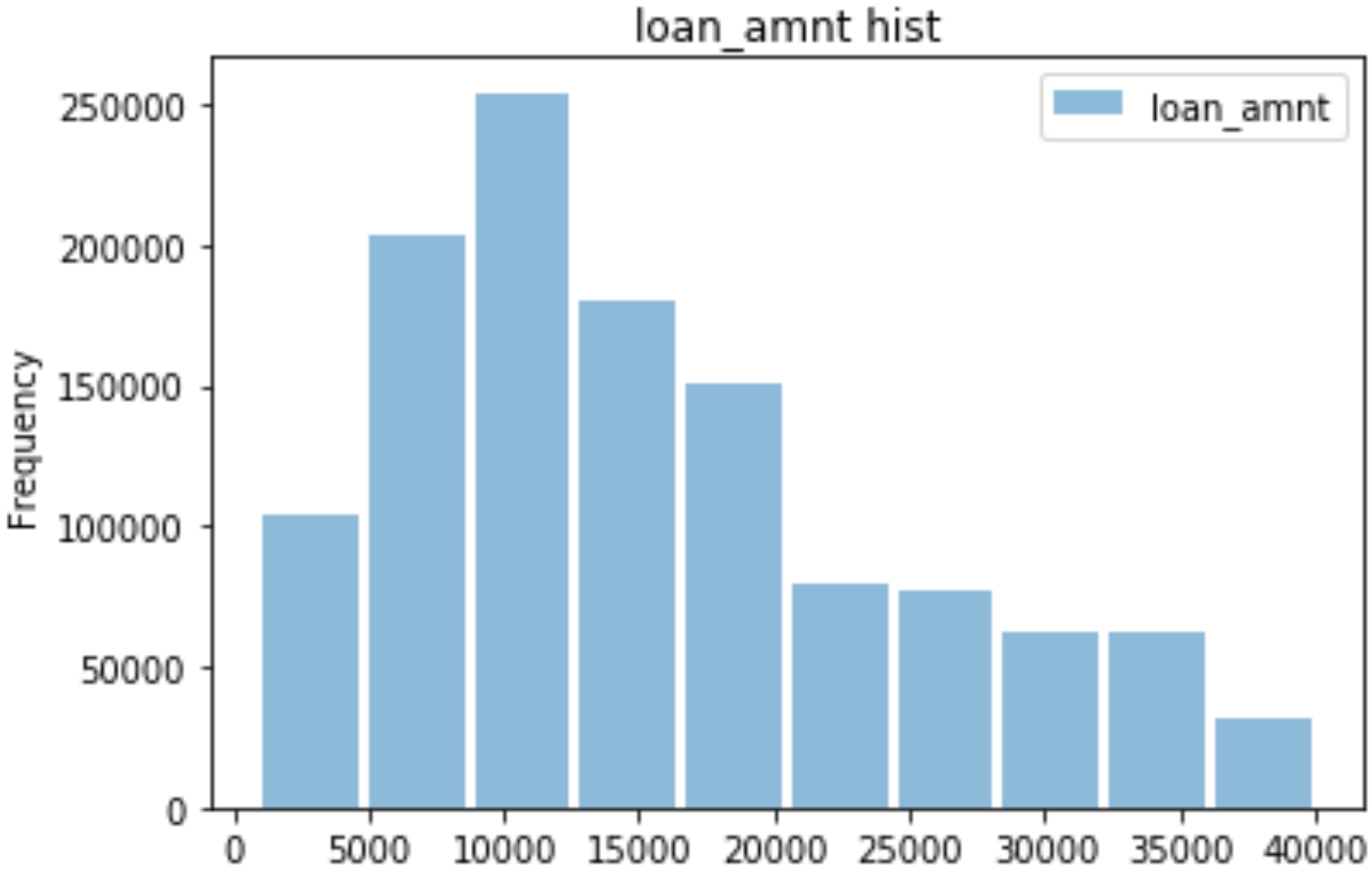
Annual Income Distribution

annual_inc	
count	1.20695e+06
mean	78306.2
std	139121
min	0
25%	45200
50%	65000
75%	93600
max	1.1e+08

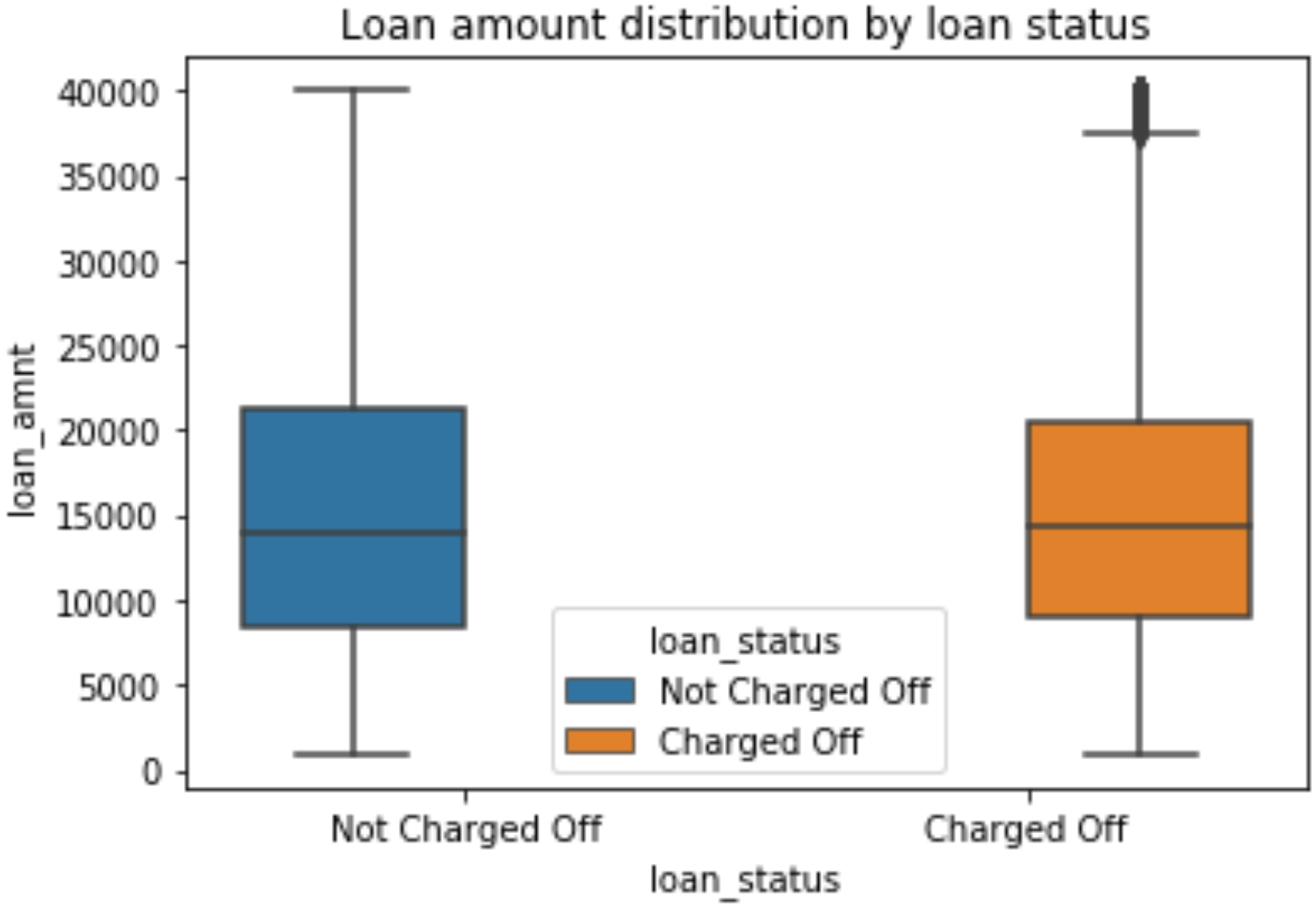


Loan Amount Distribution

loan_amnt	
count	1.20695e+06
mean	15871.5
std	9555.97
min	900
25%	8500
50%	14000
75%	21000
max	40000



Loan Amount Comparison



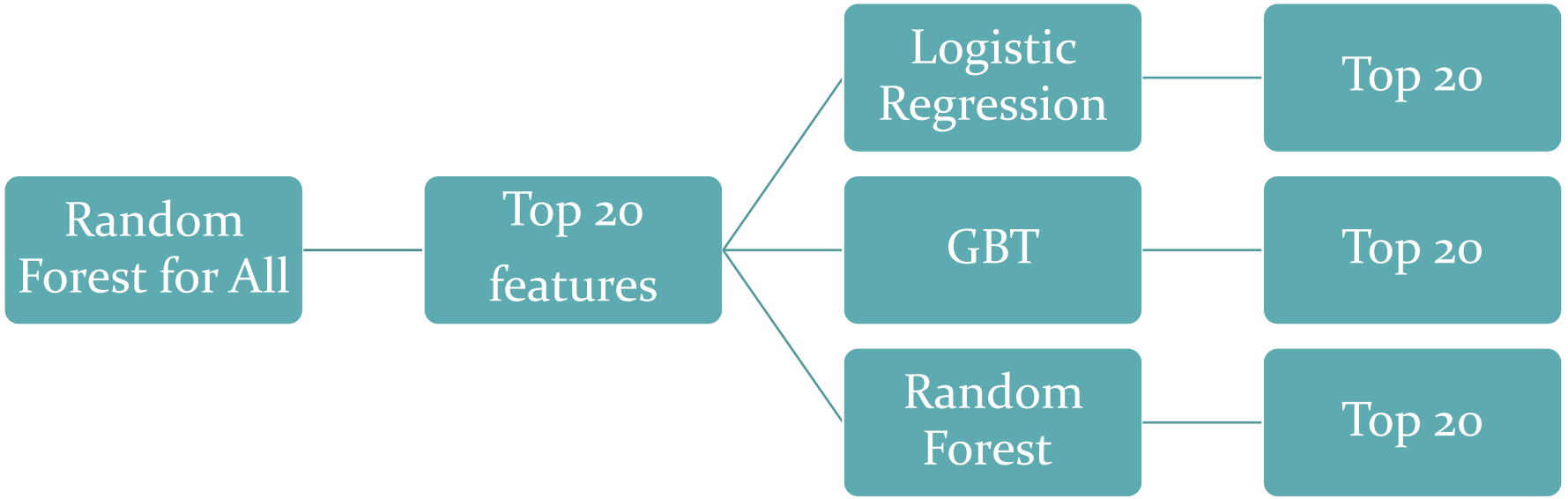
Modelling

4

Model Flow

- Random Forest
  - chooses features that have little-to-no correlation
  - Handles big number of data fields well

- Classification
  - Not Charged Off
  - Charged Off



Random Forest (Top 20)

- Performed a Random Forest model on all features to extract the important ones
- Used the Top 20 features for further models

- But Why?
  - Training future models faster
  - Reduces over-fitting further
  - Keep it simple

Look Out For..!!

Total Credit Revolving Balance – 19%

Number of Derogatory Public Records – 15%

Number of Open Trades – 12%

Total Current Balance – 9%

Outstanding Principal for funded amount – 8%

Logistic Regression

- Performed a Logistic Regression model with penalty factors on the Top 20

➢ Logistic Regression
  - Widely used in industry because of ability to interpret feature importance using probability (log odds)

  ➢ Classification
    - Not Charged Off
    - Charged Off

Gradiant Boosting Tree Classifier

- Performed a GBT model on the Top 20 features

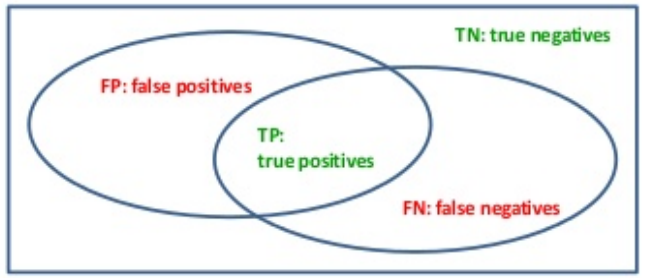
➢ Gradient Boosting Tree
  - Predictions of the final ensemble model is the weighted sum of the predictions made by the previous tree models.

  ➢ Classification
    - Not Charged Off
    - Charged Off

Evaluate Models

Definitions

Precision, Recall, and F-measure



Precision:

$$p = \frac{TP}{TP + FP}$$

Recall:

$$r = \frac{TP}{TP + FN}$$

F-measure: Harmonic mean of precision and recall

$$F = \frac{1}{\frac{1}{2}\left(\frac{1}{p} + \frac{1}{r}\right)} = \frac{2pr}{p + r}$$

- **Area under the AUC curve**
- **Precision**
- **Recall**
- **F-1 Score**

Random Forest (Top 20)

AUC: 0.86

F1-Score:
0.91 “Not Charged Off”
0.65 “Charged Off”

Recall:
0.93 “Not Charged Off”
0.61 “Charged Off”

Precision:
0.90 “Not Charged Off”
0.71 “Charged Off”

	column	weight
14	revol_bal	0.189613
13	pub_rec	0.148968
28	open_acc_6m	0.124547
27	tot_cur_bal	0.094344
18	out_prncp	0.089755
33	total_bal_il	0.035971
25	collections_12_mths_ex_med	0.033773
32	mths_since_rcnt_il	0.029548
24	last_pymnt_amnt	0.027444
31	open_il_24m	0.025129
16	total_acc	0.021584
30	open_il_12m	0.015044
23	collection_recovery_fee	0.013587
15	revol_util	0.013378
42	bc_open_to_buy	0.011513
26	tot_coll_amt	0.011428
45	mo_sin_old_il_acct	0.010851
36	open_rv_24m	0.010317
3	grade	0.010057
29	open_act_il	0.006203

Logistic Regression

AUC: 0.84

F1-Score:
0.90 “Not Charged Off”
0.52 “Charged Off”

Recall:
0.95 “Not Charged Off”
0.42 “Charged Off”

Precision:
0.86 “Not Charged Off”
0.69 “Charged Off”

	word	weight
19	open_act_il	1.954890e-01
1	pub_rec	6.268813e-02
10	total_acc	2.068996e-02
9	open_il_24m	1.352888e-02
13	revol_util	7.271916e-03
12	collection_recovery_fee	8.966213e-04
17	open_rv_24m	3.587220e-04
8	last_pymnt_amnt	4.681214e-05
5	total_bal_il	3.801108e-06
0	revol_bal	3.107899e-06
3	tot_cur_bal	9.584614e-08
15	tot_coll_amt	-3.918885e-06
14	bc_open_to_buy	-1.627264e-05
4	out_prncp	-1.634688e-04
11	open_il_12m	-2.789346e-04

Gradient Boosting Tree Classifier

AUC: 0.90

F1-Score:
0.92 “Not Charged Off”
0.68 “Charged Off”

Recall:
0.93 “Not Charged Off”
0.65 “Charged Off”

Precision:
0.91 “Not Charged Off”
0.71 “Charged Off”

	column	weight
14	bc_open_to_buy	0.189613
13	revol_util	0.148968
18	grade	0.089755
16	mo_sin_old_il_acct	0.021584
15	tot_coll_amt	0.013378
3	tot_cur_bal	0.010057
0	revol_bal	0.005227
1	pub_rec	0.002999
2	open_acc_6m	0.002800
7	mths_since_rcnt_il	0.001561
10	total_acc	0.001070
11	open_il_12m	0.000776
19	open_act_il	0.000506
4	out_prncp	0.000430
12	collection_recovery_fee	0.000401

Model Comparison

Random Forest

Recall

0.93 NCO

0.61 CO

AUC

0.86

Logistic Regression

Recall

0.95 NCO

0.42 CO

AUC

0.84

Gradiant Boosting Tree

Recall

0.93 NCO

0.66 CO

AUC

0.90

Problems Encountered

5

Problems Encountered

Feature Engineer

StringIndex , Label Encode , Vector Assemble, Parse string data, Correlation problems – for 145columns

Wrong Values

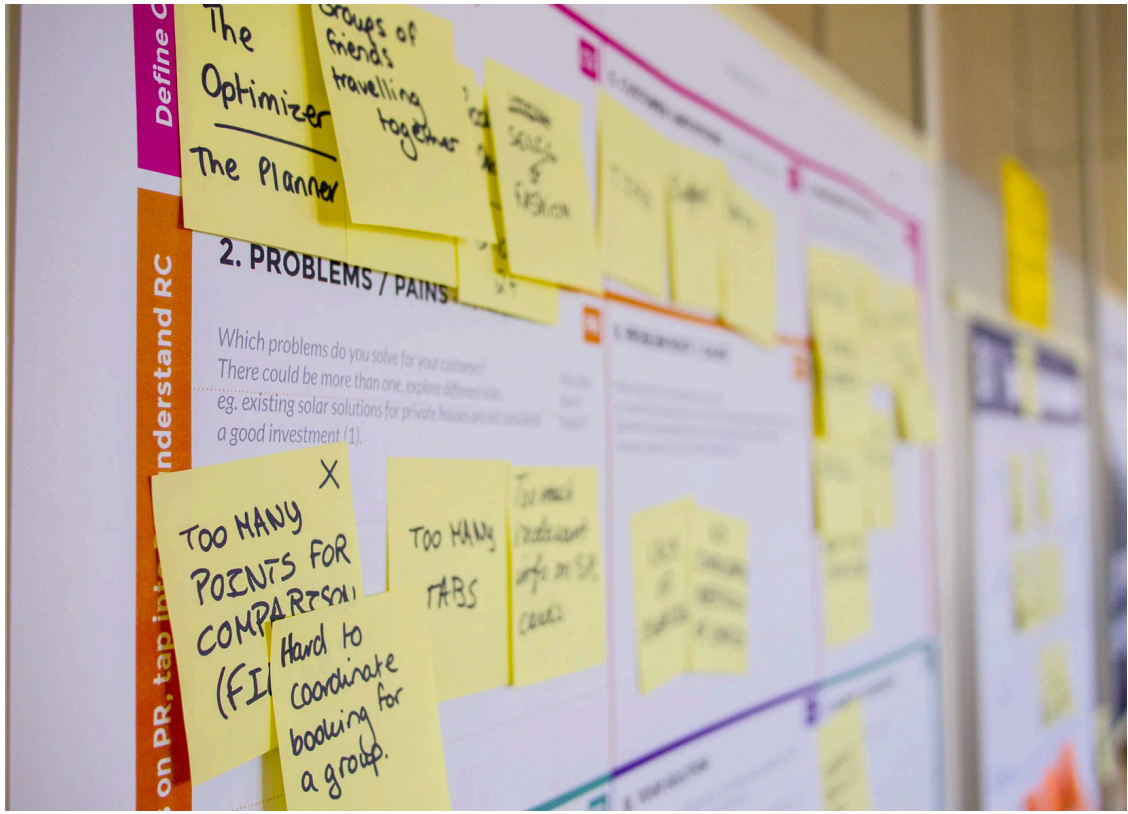
Another major challenge we faced is parsing the correct data. 1 instead of "One" makes all the difference

Missing Values

Every type of feature – numerical/Categorical?string requires its own type of NA interpolation

Knowing Data

Anyone can model, But the main role of data scientists is to interpret it in a business sense



Workload Divided

6

Work Divided

Abhiram

- Initial Cleaning, NA interpolation, winsorizing of about 50 columns & feature engineering
- Evaluation metrics, Random Forest model & feature extraction

Harper

- Cleaning, NA interpolation, winsorizing of about 25 columns & feature engineering
- Correlation matrix
- Descriptive analysis
- Visualizations, presentation format & work

Benjamin

- Setting up a Git repo
- Cleaning NA, interpolation, winsorizing of about 25 columns & feature engineering
- Combining teams parts into a single pipeline
- GBT & Logistic Regression

Q & A

7


THANK YOU
24Slides