CAPSTONE:
PREDICTING LENDINGCLUB'S CHARGED-OFF LOANS

Jean GOH / Jan 2018

MOTIVATION

- Professional background in finance
- Strong interest in FinTech, especially businesses which have disrupted traditional banking activities
- Publicly available data

CONTENTS

INTRODUCTION

- Lending Club (LC) is a peer-to-peer lending company with over $15 billion originated loans since 2007.
- Investors lend money directly to borrowers via the online platform
- Offers loans up to $40,000 across 2 loan terms (36/60 months).
- Loans are issued according to various grades, corresponding to different interest rates

HYPOTHESIS AND MODELING

- Dataset includes information for every loan issued by LC in 2015.
- Hypothesis: Machine learning techniques can be used to identify if a LC loan will be charged-off.
- Train-test random split (80/20) was done to identify the best performing model.

LOAN STATUSES

STATUS	DEFINITION
CURRENT	Up to date on all outstanding payments
FULLY PAID	Fully repaid either at expiry or result of a prepayment
IN GRACE PERIOD	Loan is past due but within 15-day grace period
LATE (16-30 DAYS)	Not been current for 16-30 days
LATE (31-120 DAYS)	Not been current for 31 to 120 days
DEFAULT	In general, when note is 121+days past due
CHARGED OFF	**When there is no longer a reasonable expectation of further payments. Charge off typically occurs when a loan is 150 days past due (ie 30 days after the "Default" status is reached)**

Source: LendingClub's website

RAW DATA

- 421,097 rows, 111 features

	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	grade	sub_grade	...	num_tl_90g_dpd_24m
0	68009401	72868139.0	16000.0	16000.0	16000.0	60 months	14.85%	379.39	C	C5	...	0.0
1	68354783	73244544.0	9600.0	9600.0	9600.0	36 months	7.49%	298.58	A	A4	...	0.0
2	68466916	73356753.0	25000.0	25000.0	25000.0	36 months	7.49%	777.55	A	A4	...	0.0
3	68466961	73356799.0	28000.0	28000.0	28000.0	36 months	6.49%	858.05	A	A2	...	0.0
4	68495092	73384866.0	8650.0	8650.0	8650.0	36 months	19.89%	320.99	E	E3	...	0.0

num_tl_op_past_12m	pct_tl_nvr_dlq	percent_bc_gt_75	pub_rec_bankruptcies	tax_liens	tot_hi_cred_lim	total_bal_ex_mort	total_bc_limit	total_il_high_credit_limit
2.0	78.9	0.0	0.0	2.0	298100.0	31329.0	281300.0	13400.0
2.0	100.0	66.7	0.0	0.0	88635.0	55387.0	12500.0	75635.0
0.0	100.0	20.0	0.0	0.0	373572.0	68056.0	38400.0	82117.0
0.0	91.7	22.2	0.0	0.0	304003.0	74920.0	41500.0	42503.0
12.0	100.0	50.0	1.0	0.0	38998.0	18926.0	2750.0	18248.0

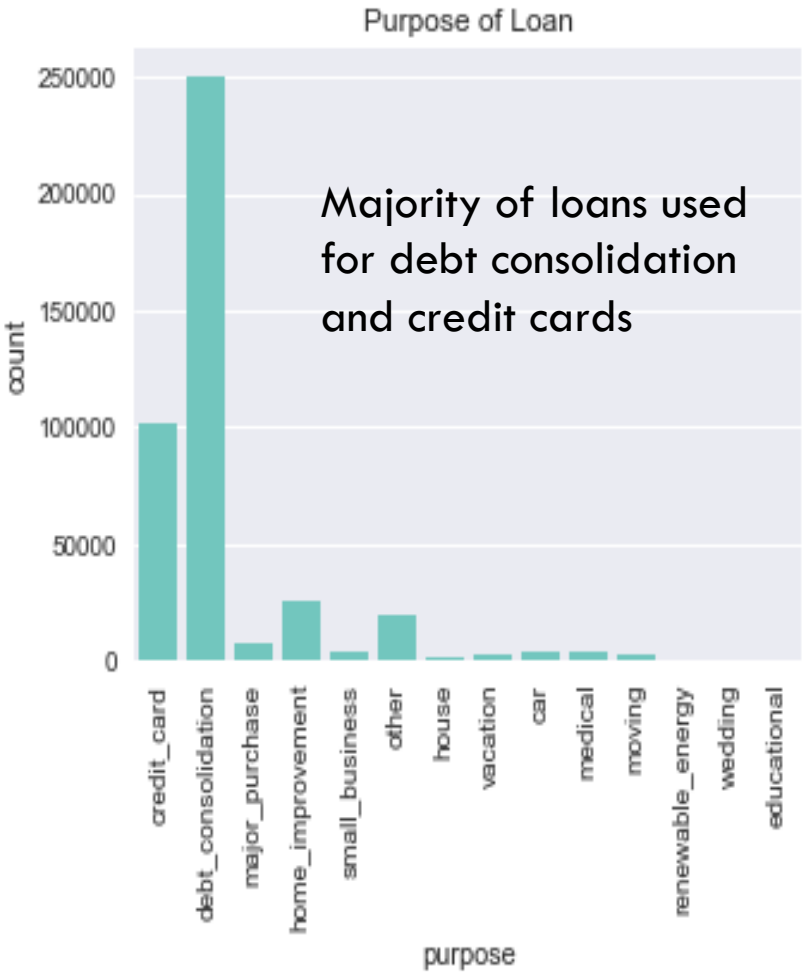
DATA CLEANING

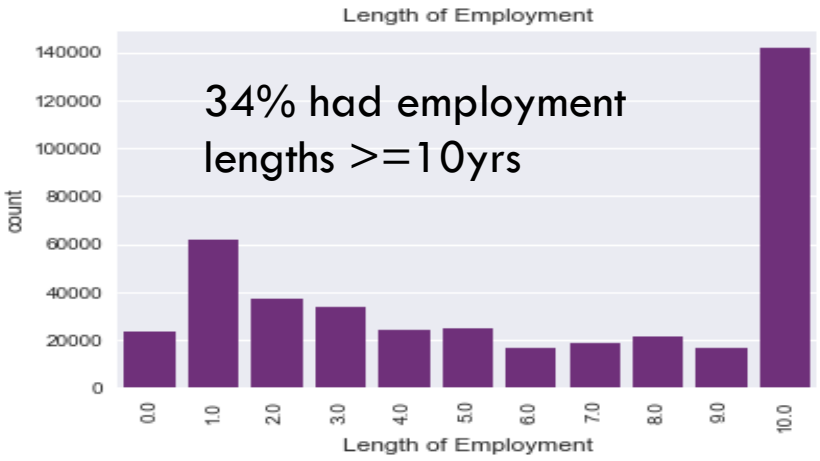
- Data Removal:
  - Variables with >40% null values
  - Rows which did not contain any data
  - Objects with >15 unique categories
  - Features which constituted less than 10% of total data
  - Features which were >80% correlated
  - Repeated variables (eg. loan_amt, funded_amnt, funded_amnt_inv had the same data)
  - Used only data which were either Charged Off or Fully Paid to predict chance of loan being Charged Off

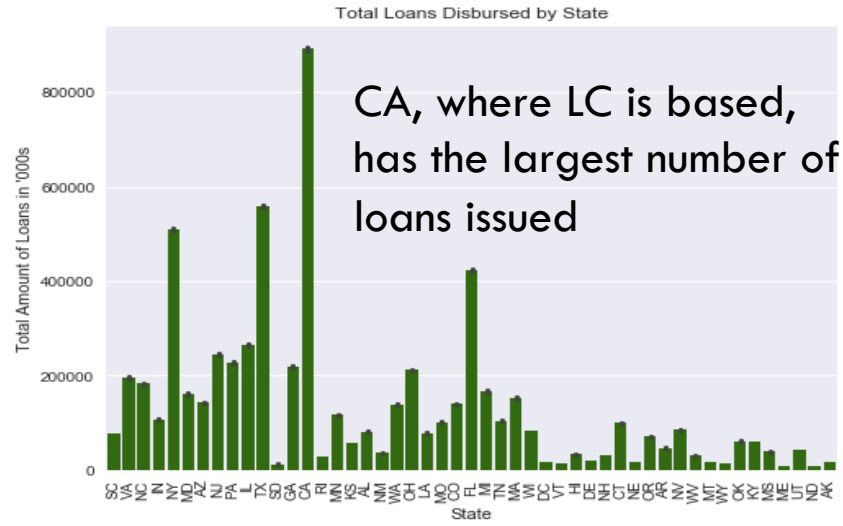
DATA CLEANING

- Data Transformation:
  - Perform get_dummies for categorical data (home_ownership, initial_list_status, purpose, sub_grade, verification_status)
  - Convert to boolean (loan_status, delinq_2yrs, inq_last_6mths)
  - Symbols such as '%', '+' and '<' were removed and objects converted to float

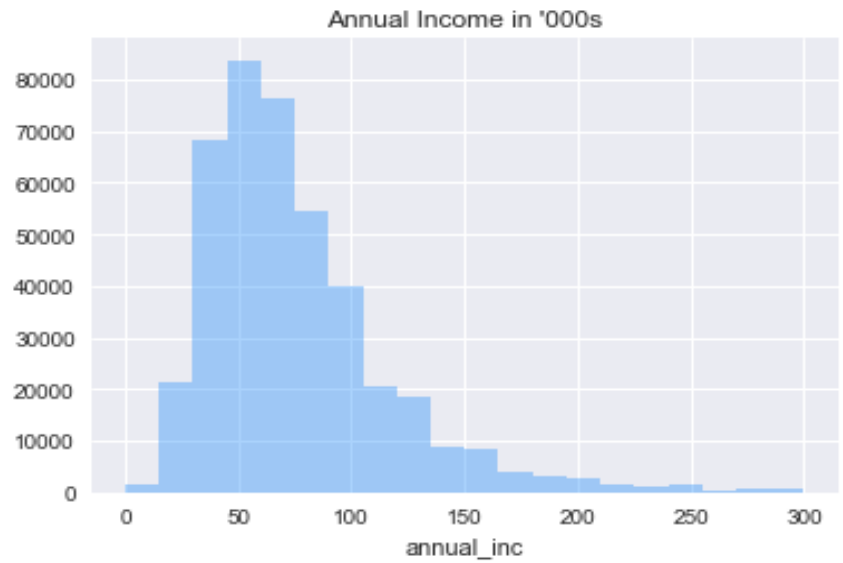
BORROWER PROFILES

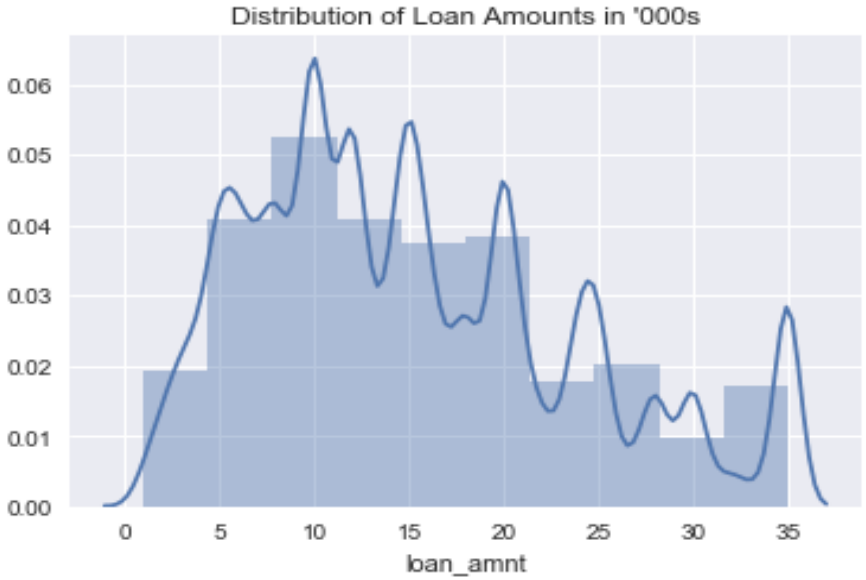


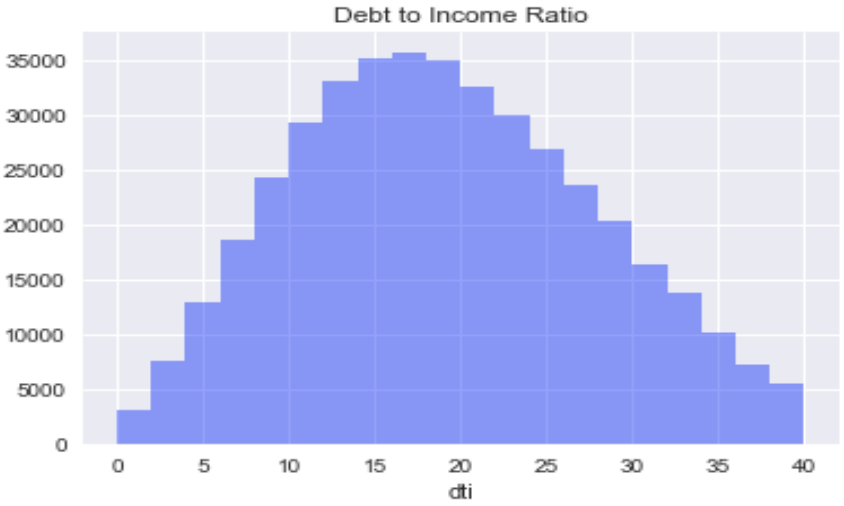


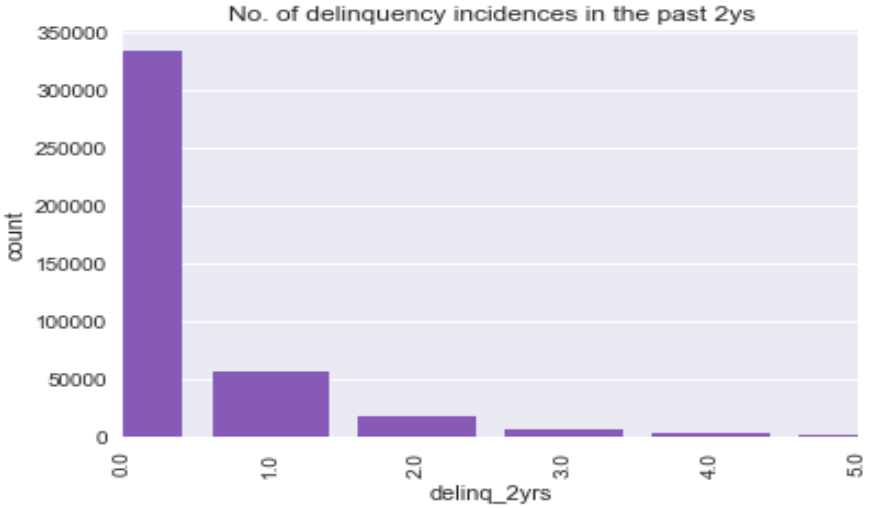


BORROWER PROFILES

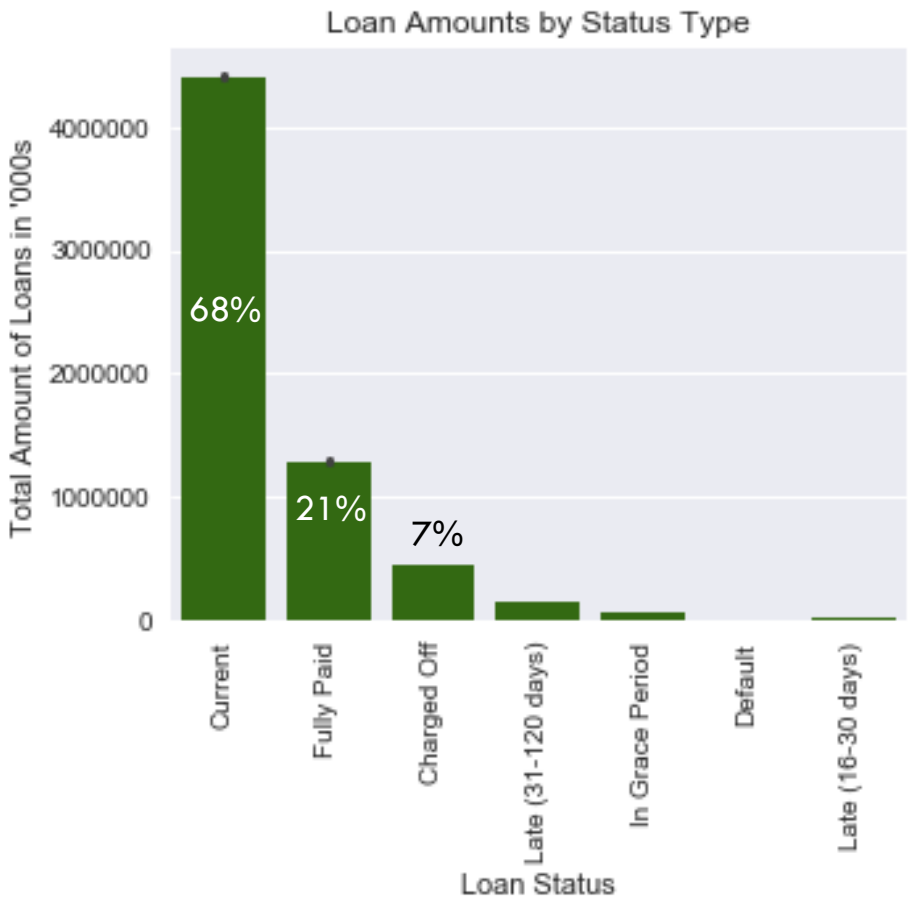


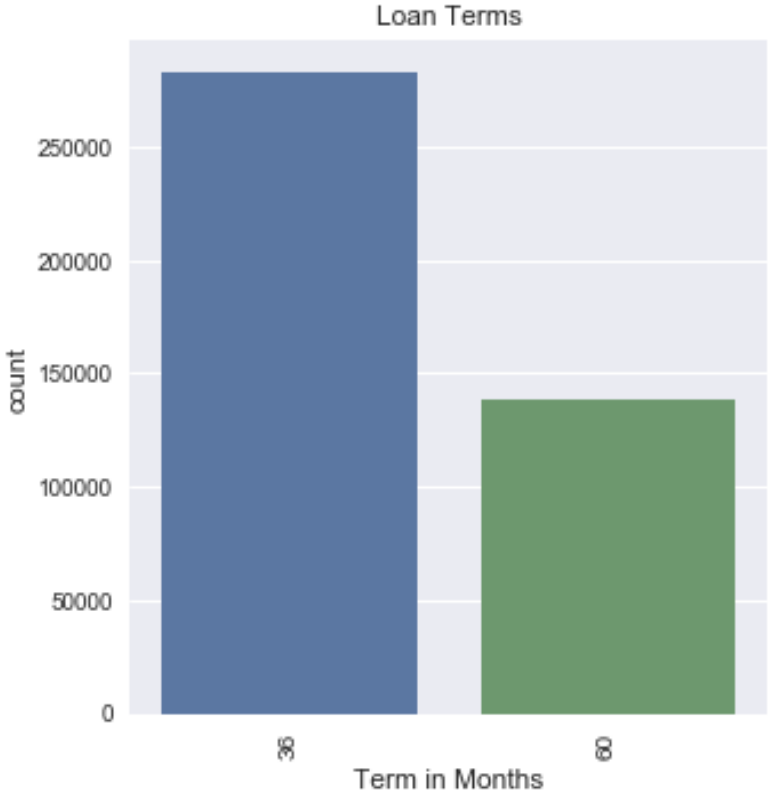






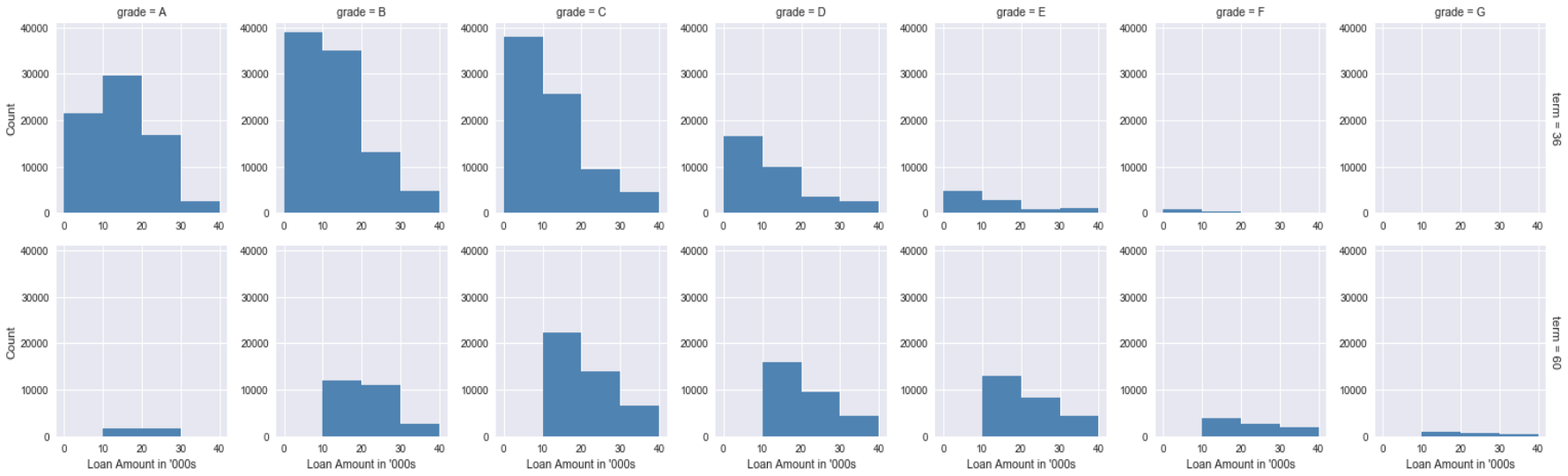
LOAN DETAILS





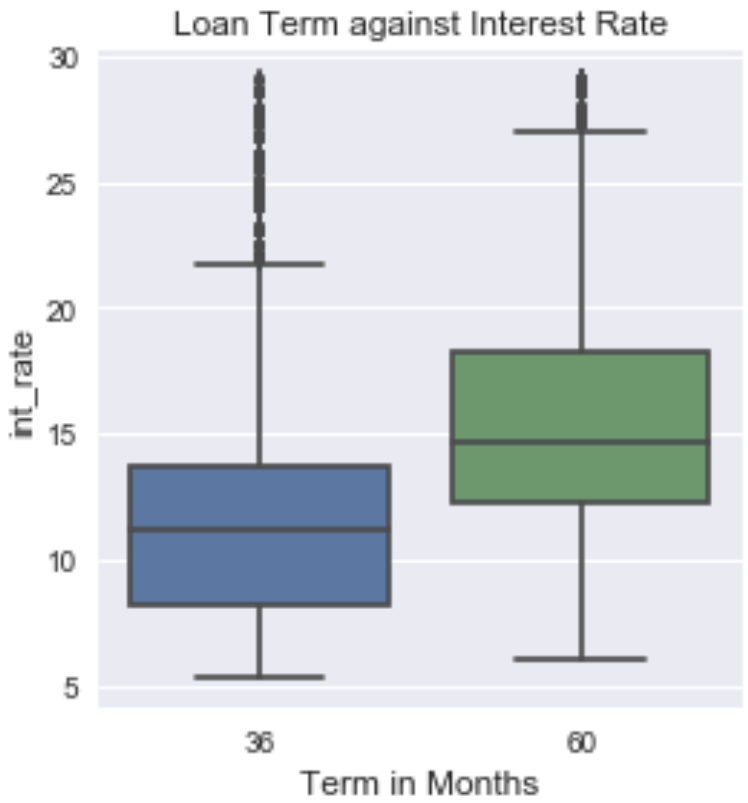
LOAN DETAILS

- Breakdown of the loans according to Grades
- Majority were within A-D grades



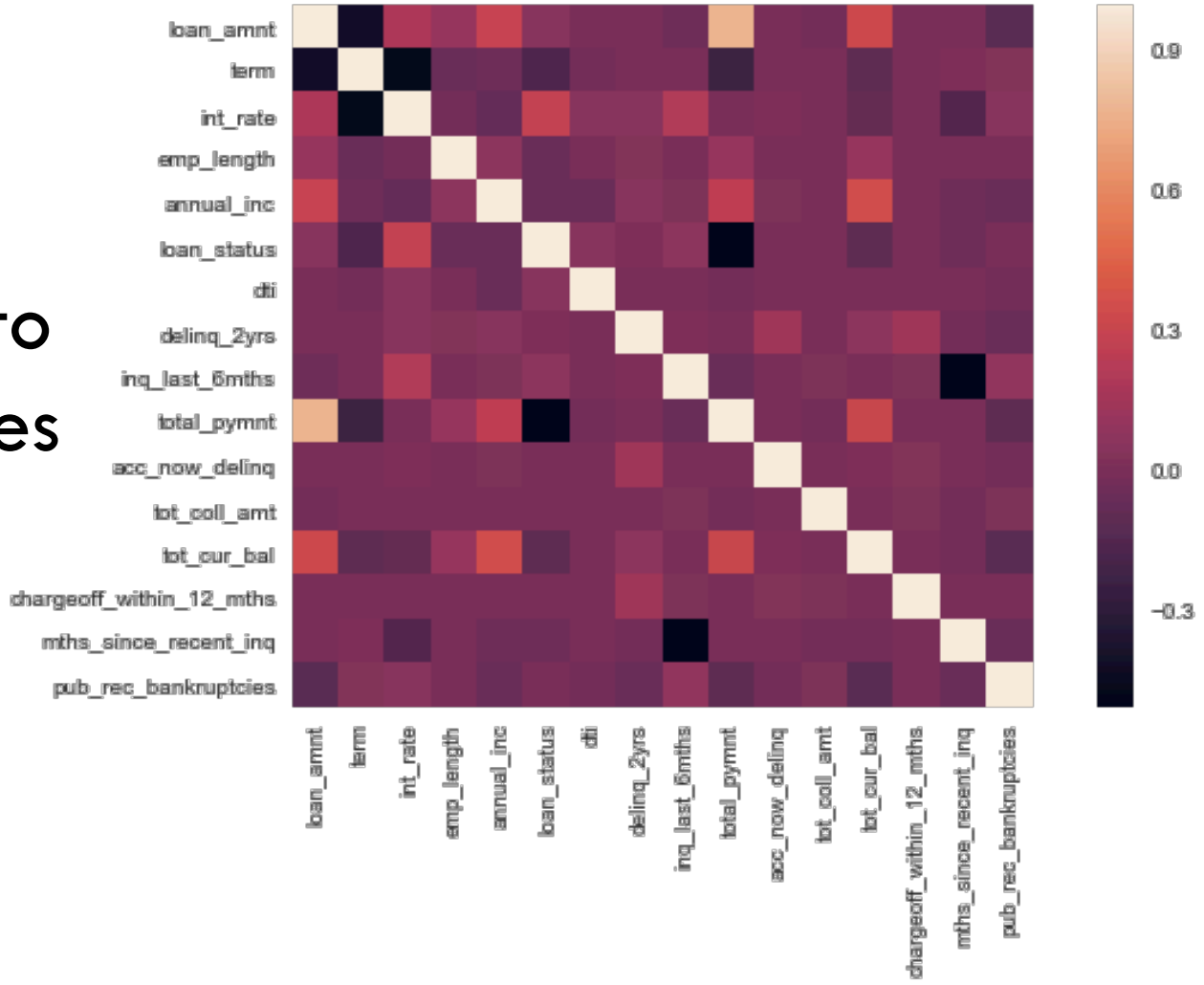
INTEREST RATES

GRADE	INTEREST RATE (AVG)
A	6.941684
B	10.044497
C	13.299105
D	16.725335
E	19.29209
F	23.620010
G	26.835187



FEATURE CORRELATION

Reduced to 21 features



PRINCIPAL COMPONENT ANALYSIS

- PCA was performed to identify the top 10 principal components explaining the variance in the data.
- It seems ~99% of the variance can be explained by just 1 component!

The percentage of total variance in the dataset explained by each component from Sklearn PCA.

[9.96253250e-01 3.51489052e-03 2.31607802e-04 2.02515001e-07
4.52474222e-08 1.41626178e-09 1.22100438e-09 6.25210970e-10
5.43010387e-10 1.91687052e-11]

RANDOM FOREST CLASSIFIER

- Leverage strength of decision trees, handle non-linear relationships, determine predictor importance, less prone to over fitting, and robust to data outliers

Accuracy of Random Forest Classifier: 0.98340
[0.99027055 0.9912947 0.98549117 0.98762482 0.97712725
0.97900486 0.96014338 0.962021 0.96218846 0.93615023]
Mean cross validation score is: 0.97313

Confusion Matrix	
70339	2
56	23336

0.06% of the data were False
(True Negative or False Positive)

GRADIENT BOOSTING

- Works for classification and regression models.
- Each tree helps to correct errors made by previous tree, hence improving the model.

Accuracy of Gradient Boosting Classifier on Test: 0.99509

[0.99846377 0.99863446 0.99871981 0.99854912 0.99769566
0.997781 0.99786635 0.99436716 0.98873336 0.98326931]

Mean cross validation score is: 0.99541

Confusion Matrix	
70341	1
386	23006

0.4% of the data were False
(True Negative or False Positive)

LOGISTIC REGRESSION

- Used as target variable is binary
- Can handle various types of relationships as it applies a non-linear log transformation

Accuracy of logistic regression classifier on test set: 0.97691
[0.96842195 0.98694205 0.99863446 0.99752496 0.99675685
0.98276009 0.98438167 0.96825126 0.97046774 0.92283397]
Mean cross validation score is: 0.97770

Confusion Matrix	
69591	750
1338	22054

2.2% of the data were False
(True Negative or False Positive)

NAÏVE BAYES CLASSIFIER

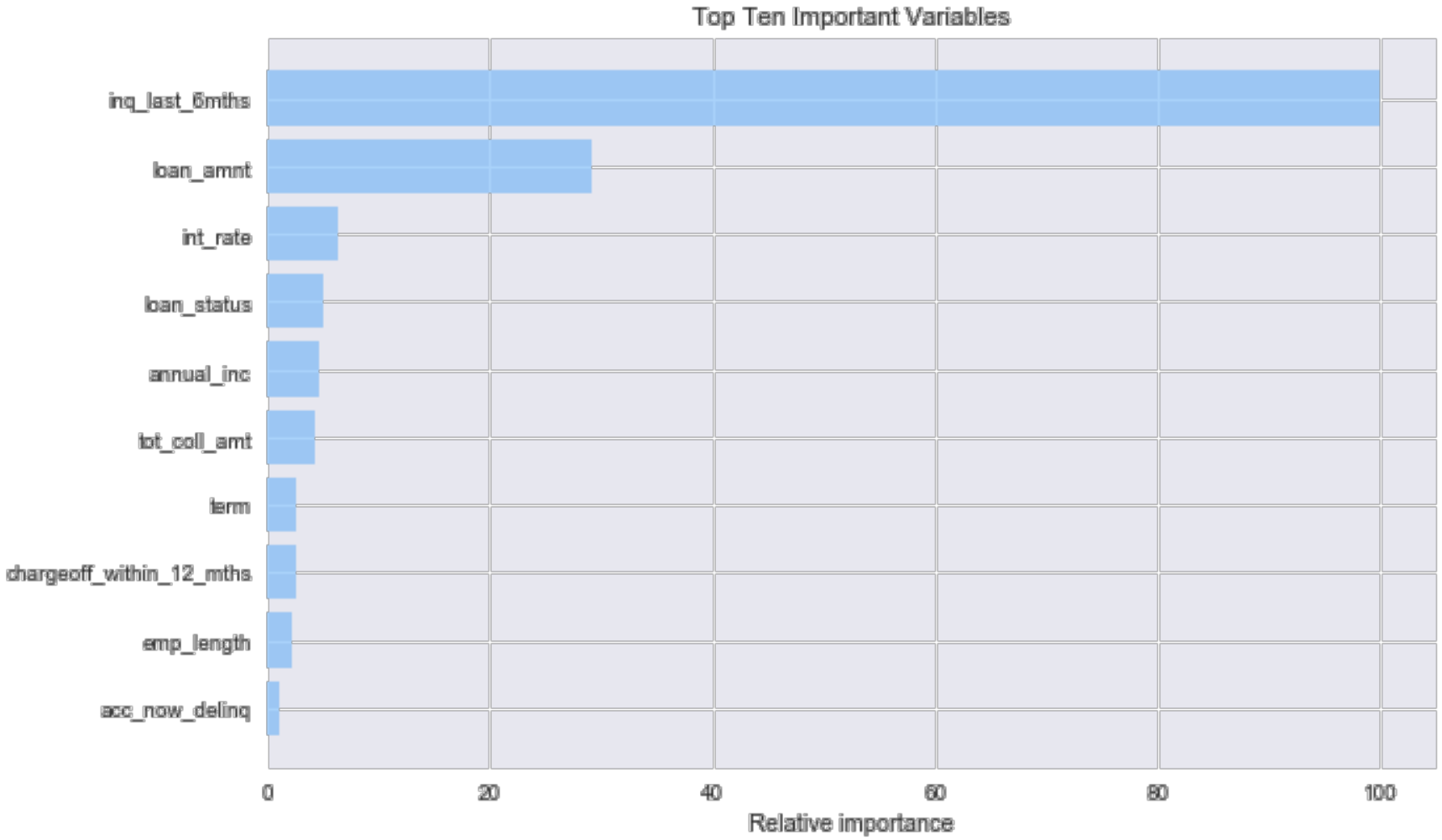
- Used for binary/boolean features
- Assumes probability of each attribute is independent of all others

Accuracy of Bernoulli classifier: 0.80631
[0.7451566 0.7356832 0.75010668 0.73764616 0.74328 0.732184
0.73115985 0.73286677 0.73694094 0.73154076]
Mean cross validation score is: 0.73765

Confusion Matrix	
62281	7212
9510	13730

19% of the data were False
(True Negative or False Positive)

FEATURE IMPORTANCE



SUMMARY

Model	Accuracy	10-Fold CV Score
RandomForestClassifier	0.98340	0.97313
Gradient Boosting Classifier	**0.99509**	**0.99504**
Logistic Regression	0.97691	0.97770
Naive Bayes	0.80268	0.73765

Gradient Boosting Classifier was the best model used for this data set. Each tree built learns from the previous tree, and hence model becomes more expressive. Modified features such as learning rate and depth of tree are kept low to allow for slow learning and better generalization.

LIMITATIONS

- Only 2015's data was used, hence I was unable to model time series data.
- Data has not been tested during periods of financial crisis, as interest rates and default rates would likely increase during those periods. RF can only return values within a range it has seen before.
- Credit scores were not given for 2015's data. This would likely be another feature that could affect the model