Predicting Loan Default

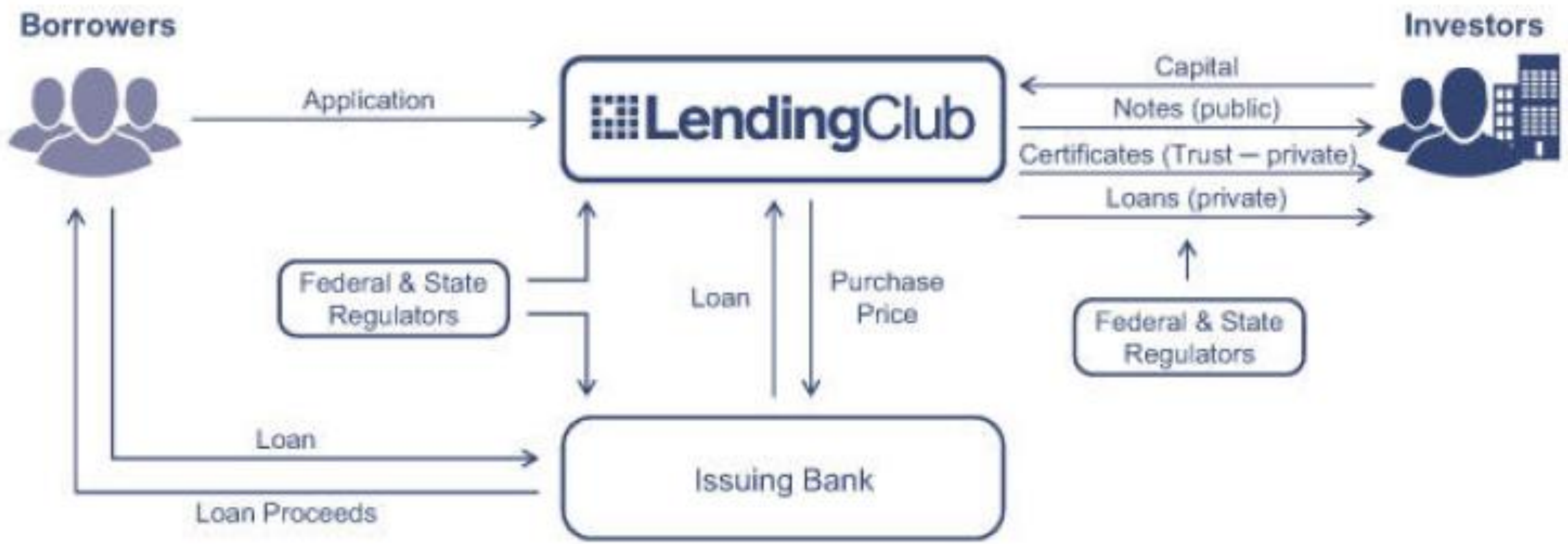
- Yang (Stefan) Lyu

Project Outline

- Company Introduction
- Data Preprocessing
- Exploratory Data Analysis
- Feature Engineering
- Model Building/Evaluation
- Recommendation

- Largest p2p lending platform
- Head-quartered in San Francisco
- Issued more than 10 billion by 2015
- Loans between $1,000 to $40,000

Business Model



Source: https://www.hbs.edu/openforum/openforum.hbs.org/challenge/understand-digital-transformation-of-business/business-model/lending-club-opening-the-floodgates-on-credit/comments.html

Objective

- **Problem:** Potential default risk
- **Solution:** Machine learning models to predict default
- **Data:**
  - All transactions issued between 2012 and 2013
  - 188,183 rows and 145 features
  - 108 numerical variables and 37 categorical variables

Data Description

Credit History

Number of open accounts,
Credit inquiries, delinquency,
total credit revolving balance,
total credit limit, etc

User Info

State, employment length,
employment title,
annual income, dti, zip code,
home ownership, member id, etc

Loan Info

Application type, description, purpose,
grade, interest rate, term, issue date,
loan amount, funded amount, etc

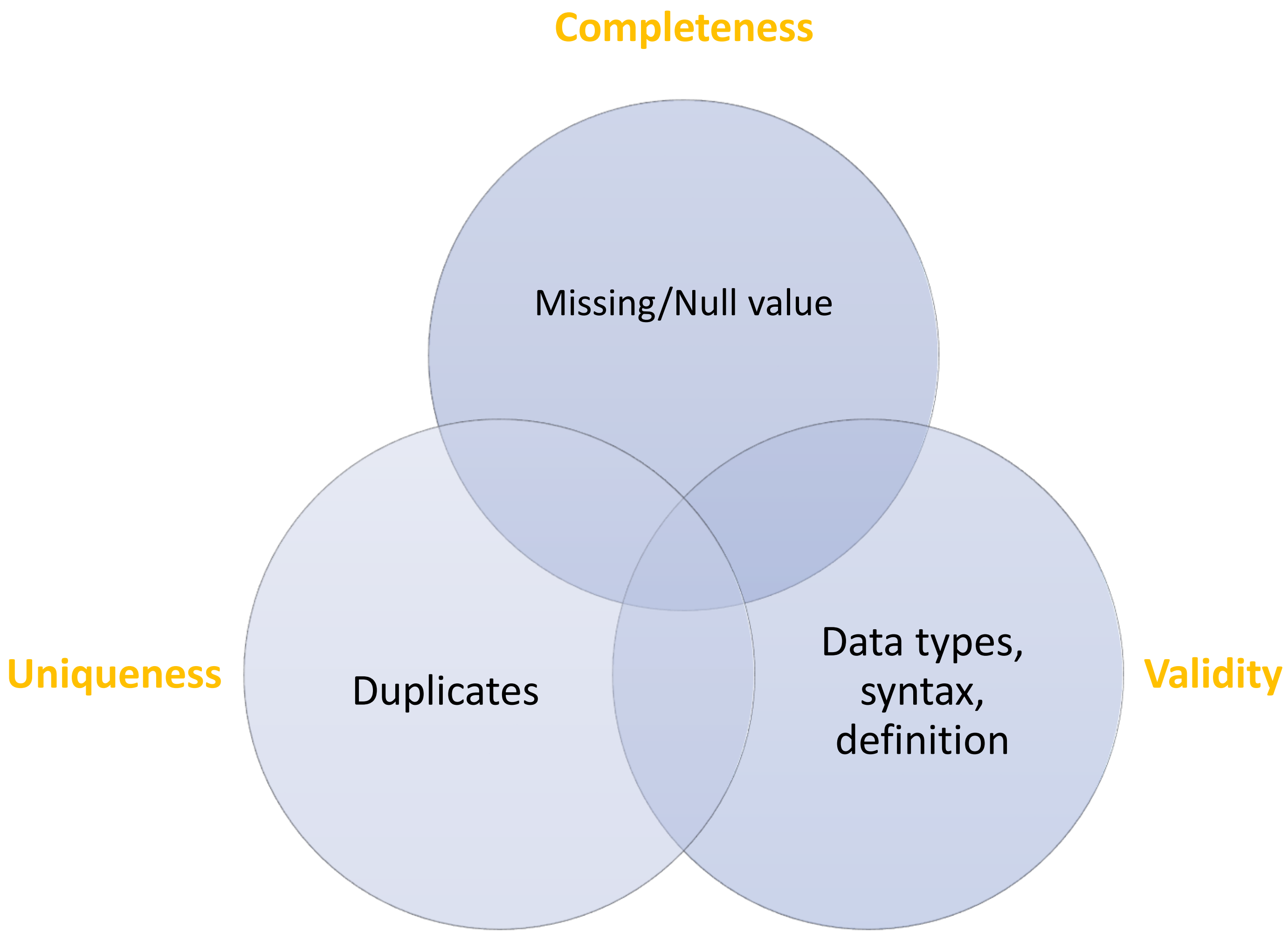
Payment Info

Last payment date, last payment amount,
interest, late fee, principal received to date,
etc.

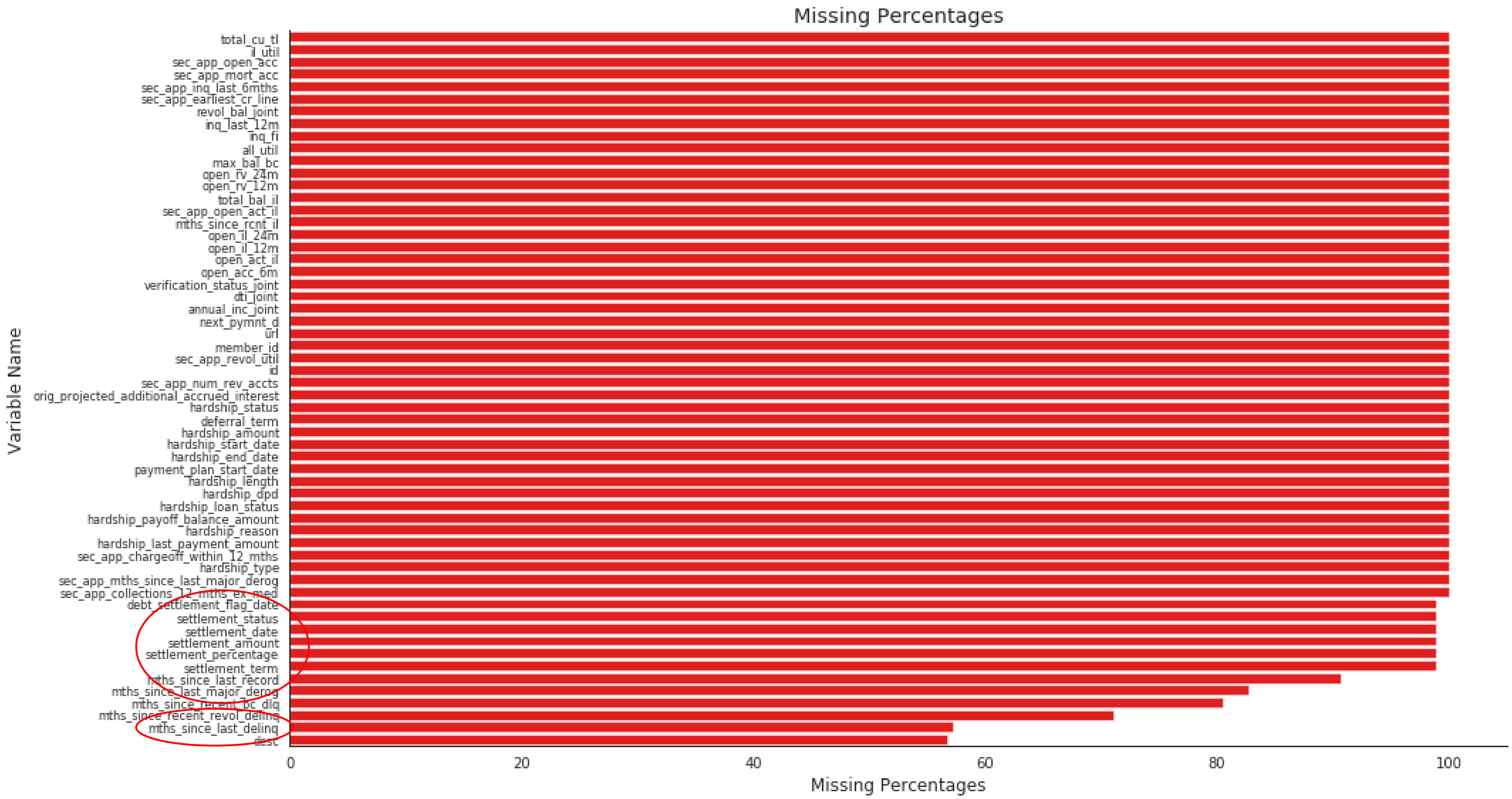
A peak at the data

	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	grade	sub_grade	emp_title	emp_length	home_ownership
0	NaN	NaN	12000.0	12000.0	12000.0	36 months	6.62%	368.45	A	A2	MANAGER INFORMATION DELIVERY	10+ years	MORTGAGE
1	NaN	NaN	28000.0	28000.0	28000.0	36 months	7.62%	872.52	A	A3	Area Sales Manager	5 years	MORTGAGE
2	NaN	NaN	27050.0	27050.0	27050.0	36 months	10.99%	885.46	B	B2	Team Leadern Customer Ops & Systems	10+ years	OWN
3	NaN	NaN	12000.0	12000.0	12000.0	36 months	11.99%	398.52	B	B3	LTC	10+ years	MORTGAGE
4	NaN	NaN	12000.0	12000.0	12000.0	36 months	7.62%	373.94	A	A3	Systems Engineer	3 years	MORTGAGE

Data Preprocessing

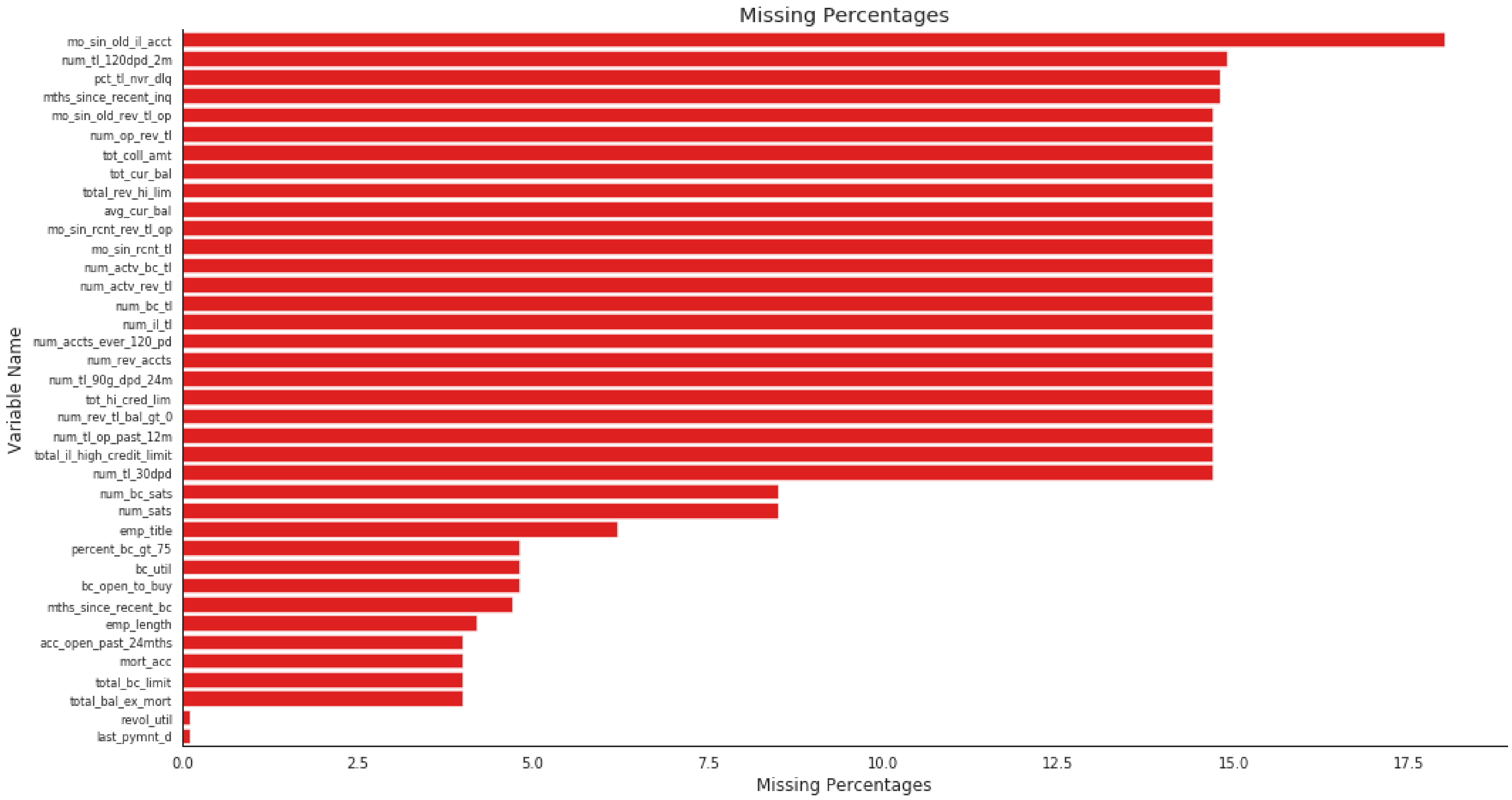


Completeness – Missing > 50%

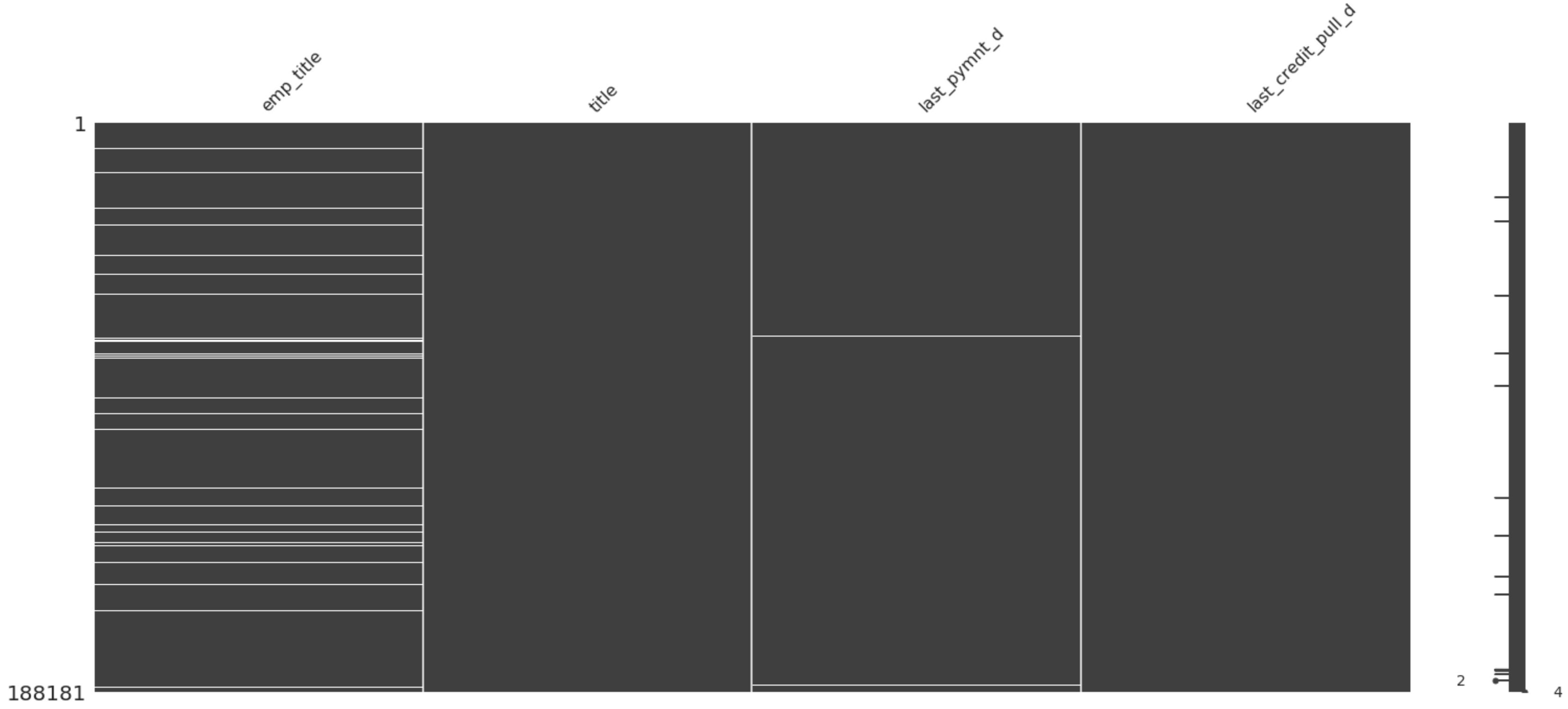


Missing suggest no delinquency records, convert to categorical variable

Missing < 50%


Missing Percentages
Variable Name
mo_sin_old_il_acct
num_tl_120dpd_2m
pct_tl_nvr_dlq
mths_since_recent_inq
mo_sin_old_rev_tl_op
num_op_rev_tl
tot_coll_amt
tot_cur_bal
total_rev_hi_lim
avg_cur_bal
mo_sin_rcnt_rev_tl_op
mo_sin_rcnt_tl
num_actv_bc_tl
num_actv_rev_tl
num_bc_tl
num_il_tl
num_accts_ever_120_pd
num_rev_accts
num_tl_90g_dpd_24m
tot_hi_cred_lim
num_rev_tl_bal_gt_0
num_tl_op_past_12m
total_il_high_credit_limit
num_tl_30dpd
num_bc_sats
num_sats
emp_title
percent_bc_gt_75
bc_util
bc_open_to_buy
mths_since_recent_bc
emp_length
acc_open_past_24mths
mort_acc
total_bc_limit
total_bal_ex_mort
revol_util
last_pymnt_d
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Missing Percentages

Missing - Categorical



- Few observations missing – remove rows with missing in last payment date and last credit pull
- Ignore missing in employment title and title

Missing - Numerical

	count	mean	std	min	25%	50%	75%	max
annual_inc	188021.0	72240.624961	51833.633269	4800.0	45000.0	62000.0	87000.0	7141778.0
revol_bal	188021.0	16322.667085	19287.939650	0.0	7136.0	12440.0	20674.0	2568995.0
tot_cur_bal	160311.0	137372.156558	150765.340446	0.0	27490.0	80839.0	208229.0	8000078.0
tot_hi_cred_lim	160311.0	165600.050938	167267.220360	0.0	44820.5	108628.0	243804.5	9999999.0

- Remove outliers > 4 standard deviation of median
- Impute ordinal variables with median
- Impute numerical variables with mean

Uniqueness

- Columns to check
  - Loan amount
  - Term
  - Interest Rate
  - Grade
  - Employment length
  - Home ownership
  - Issue date
  - Purpose
  - Zip code
- There are no duplicated records

Validity

- Convert data types

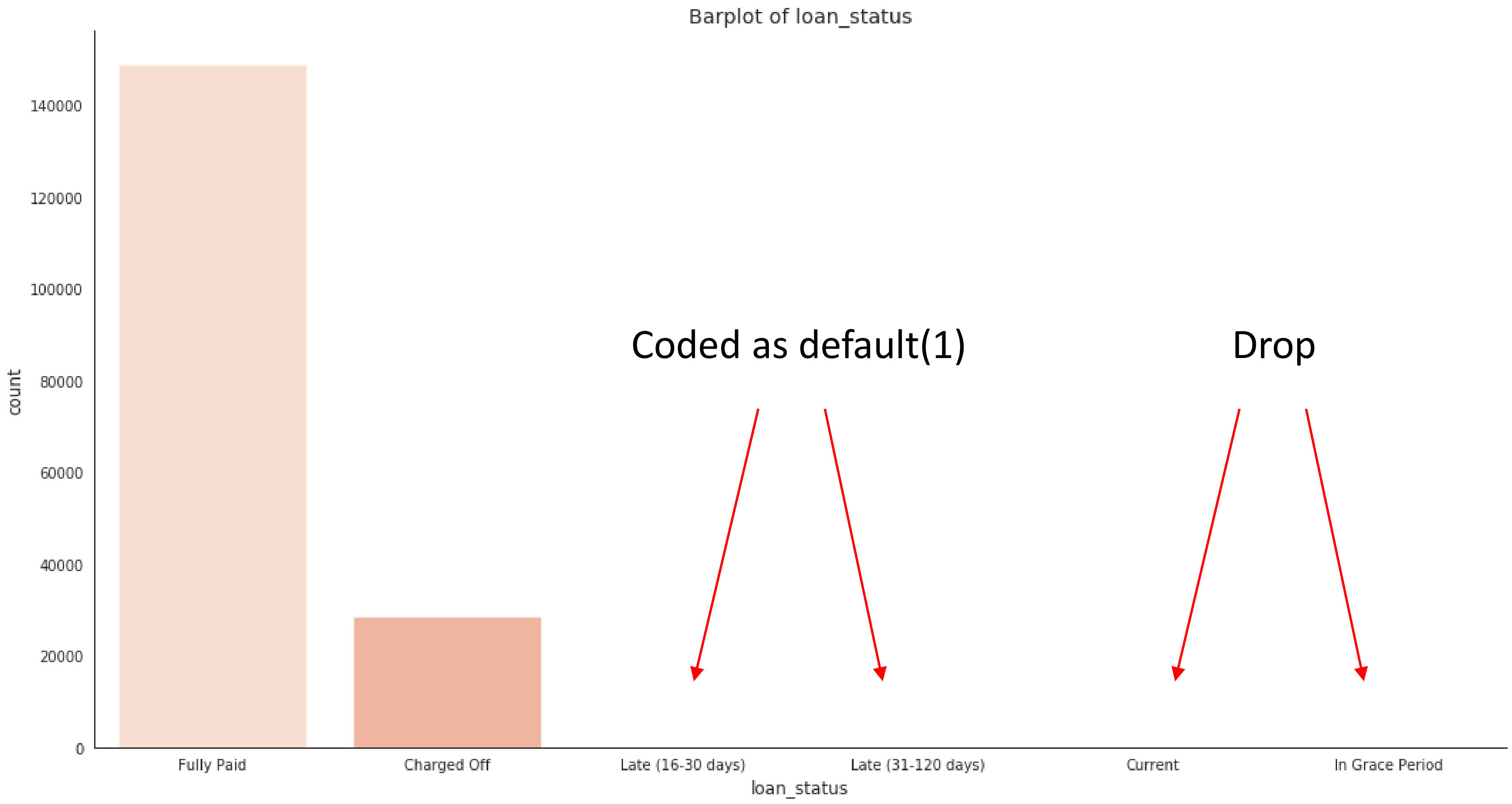
int_rate	revol_util	emp_length
6.62%	21.6%	10+ years
7.62%	54.6%	5 years

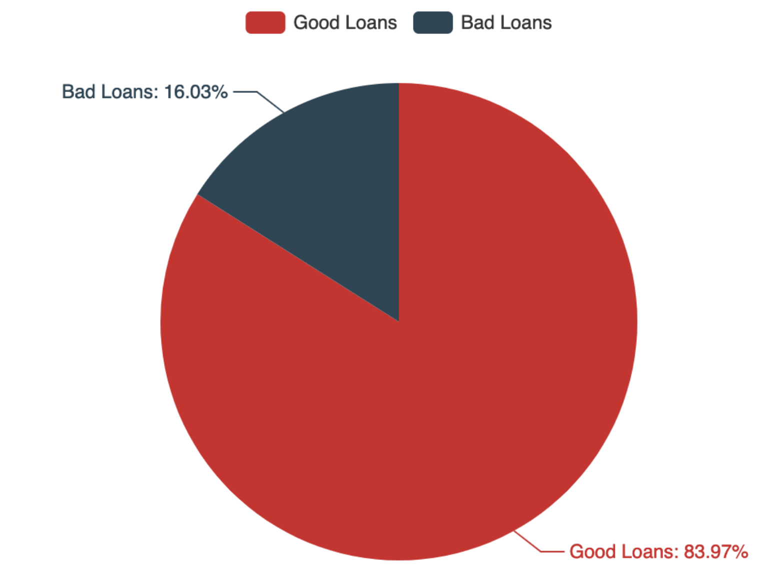
- Convert to datetime object
  - Issue date, earliest credit line, last payment date, last credit pull date
- Drop features
  - Title, employment title, zip code, debt settlement flag

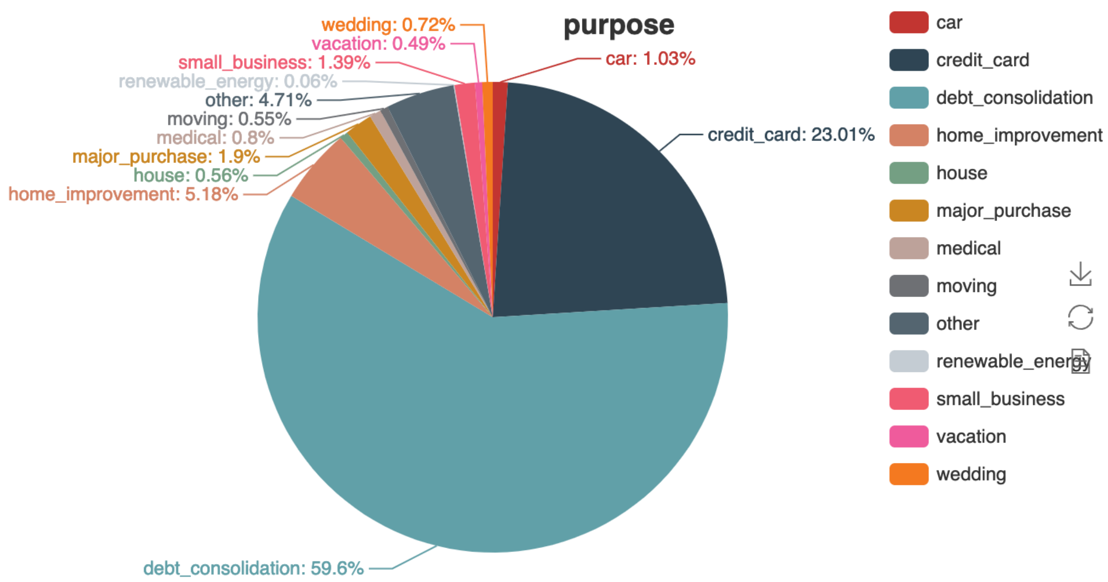
Exploratory Data Analysis

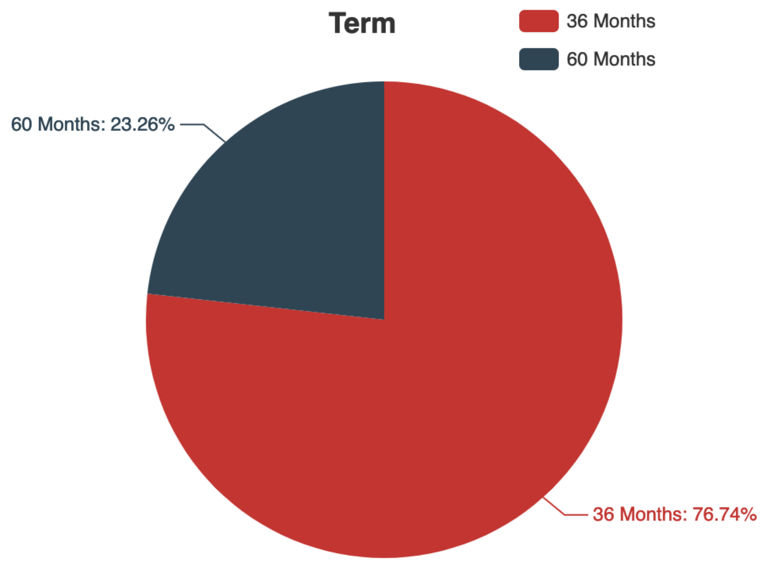
Univariate	Multivariate
Analyze a single variable	Analyze relationships between variable of interest

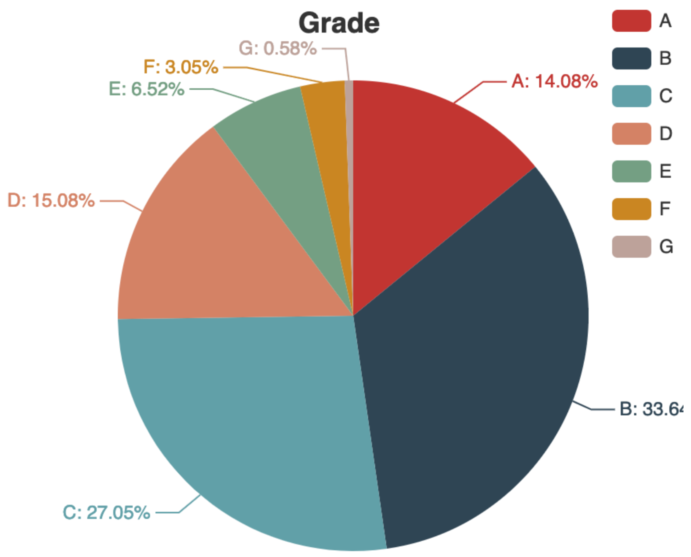
Univariate – Loan Status



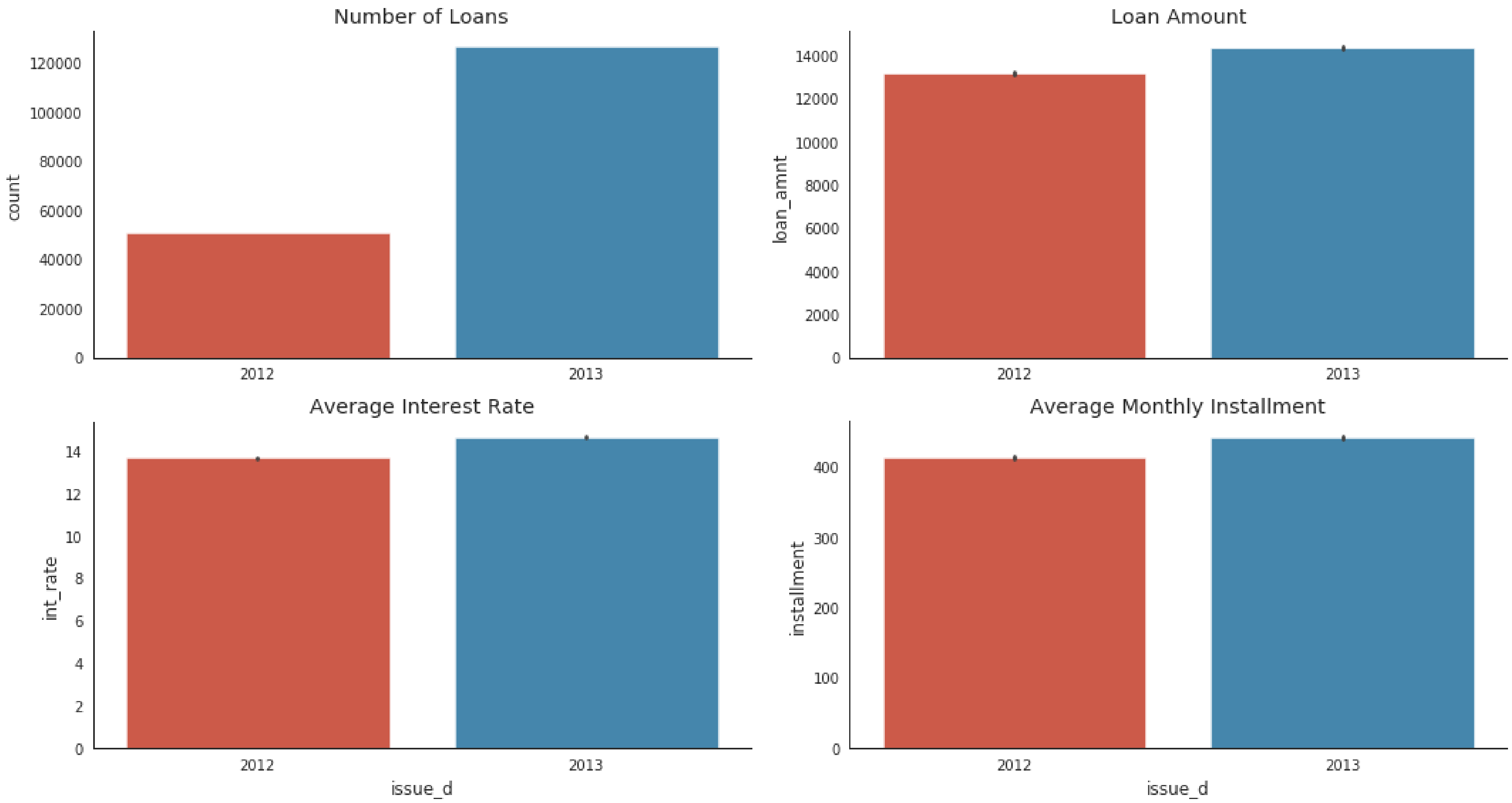

Good Loans
Bad Loans
Bad Loans: 16.03%
Good Loans: 83.97%


purpose
wedding: 0.72%
vacation: 0.49%
small_business: 1.39%
renewable_energy: 0.06%
other: 4.71%
moving: 0.55%
medical: 0.8%
major_purchase: 1.9%
house: 0.56%
home_improvement: 5.18%
car: 1.03%
credit_card: 23.01%
debt_consolidation: 59.6%
car
credit_card
debt_consolidation
home_improvement
house
major_purchase
medical
moving
other
renewable_energy
small_business
vacation
wedding

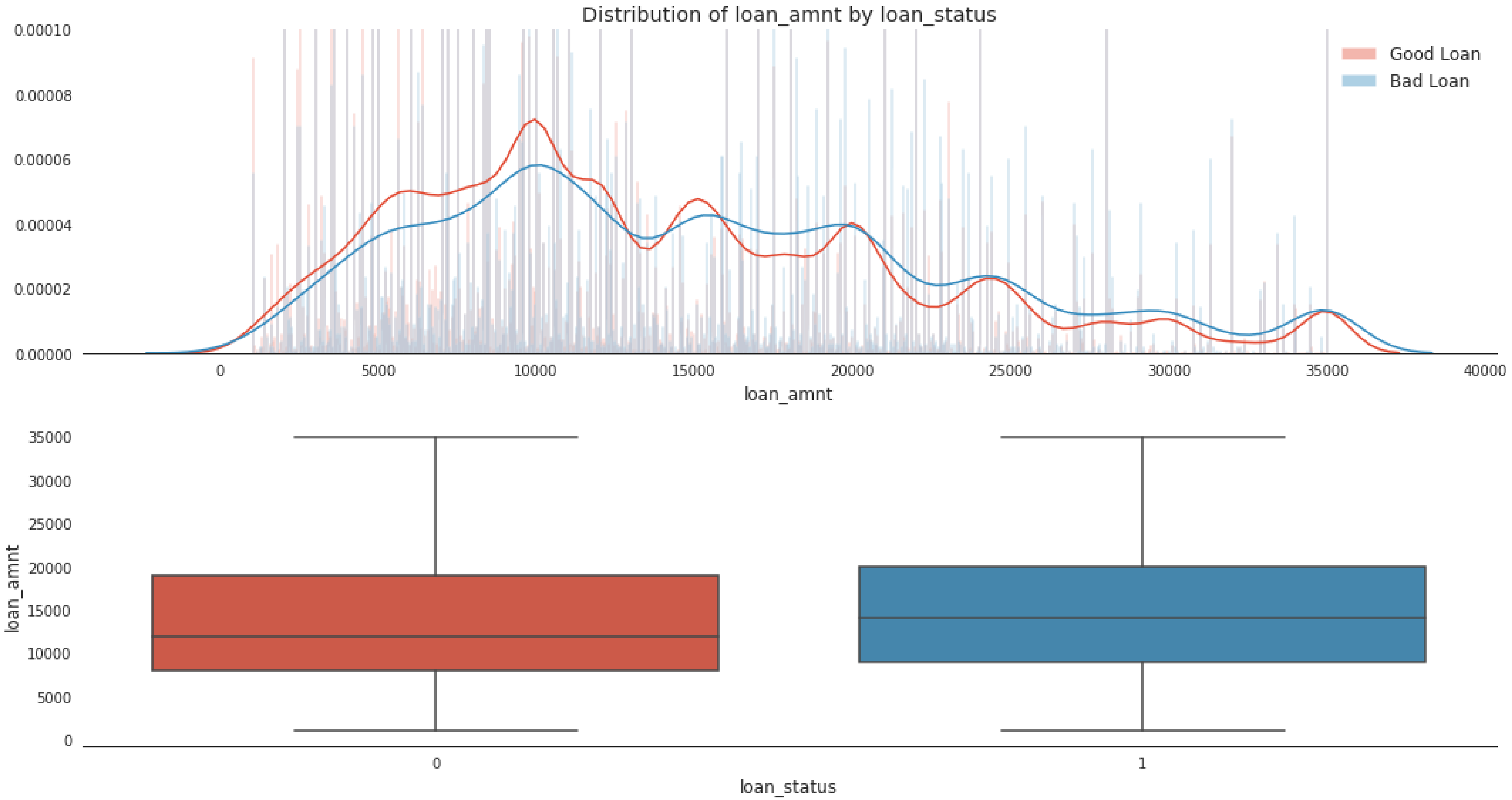

Term
36 Months
60 Months
60 Months: 23.26%
36 Months: 76.74%


Grade
G: 0.58%
F: 3.05%
E: 6.52%
A: 14.08%
D: 15.08%
B: 33.6
C: 27.05%
A
B
C
D
E
F
G

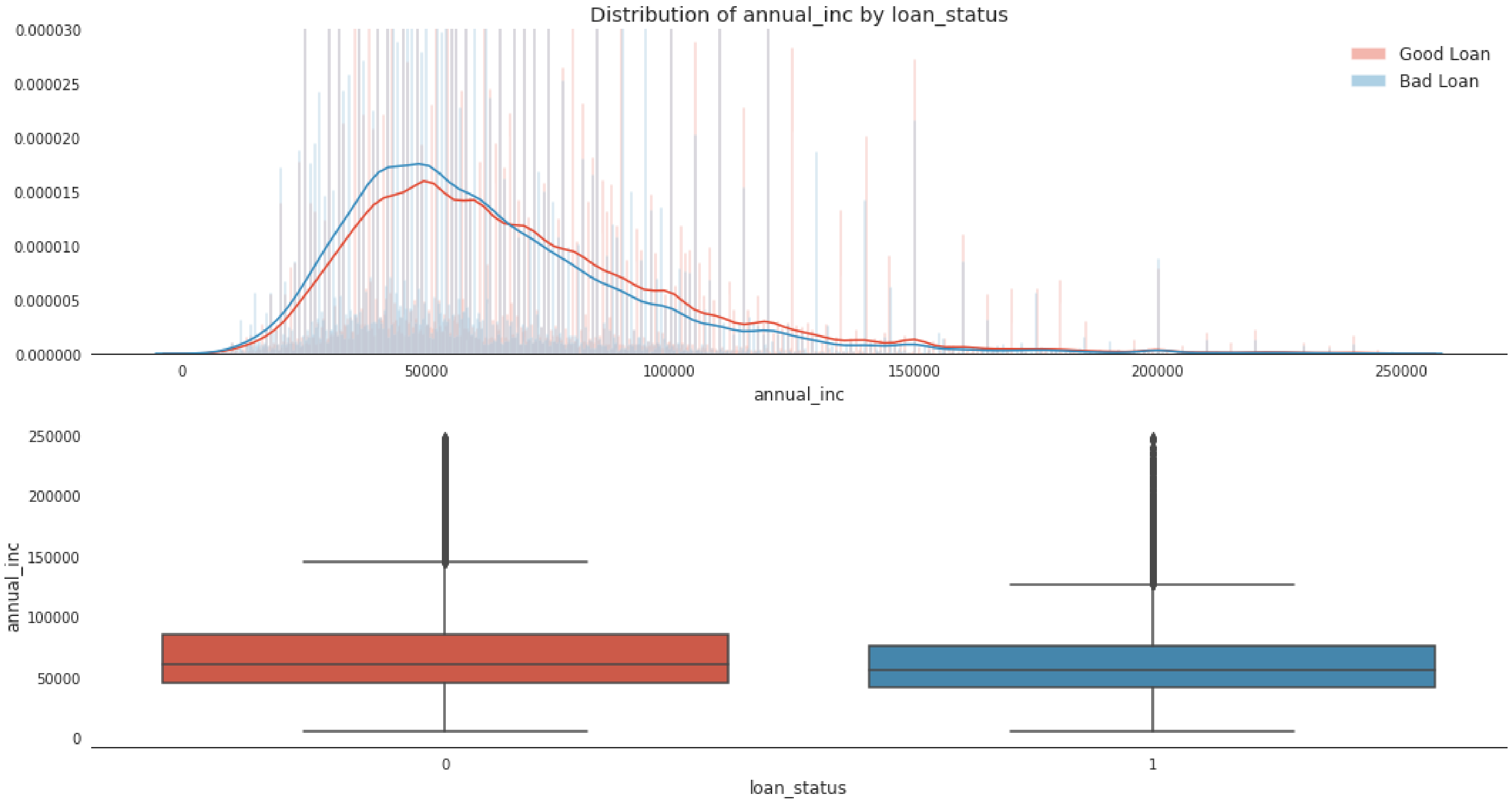
Multivariate



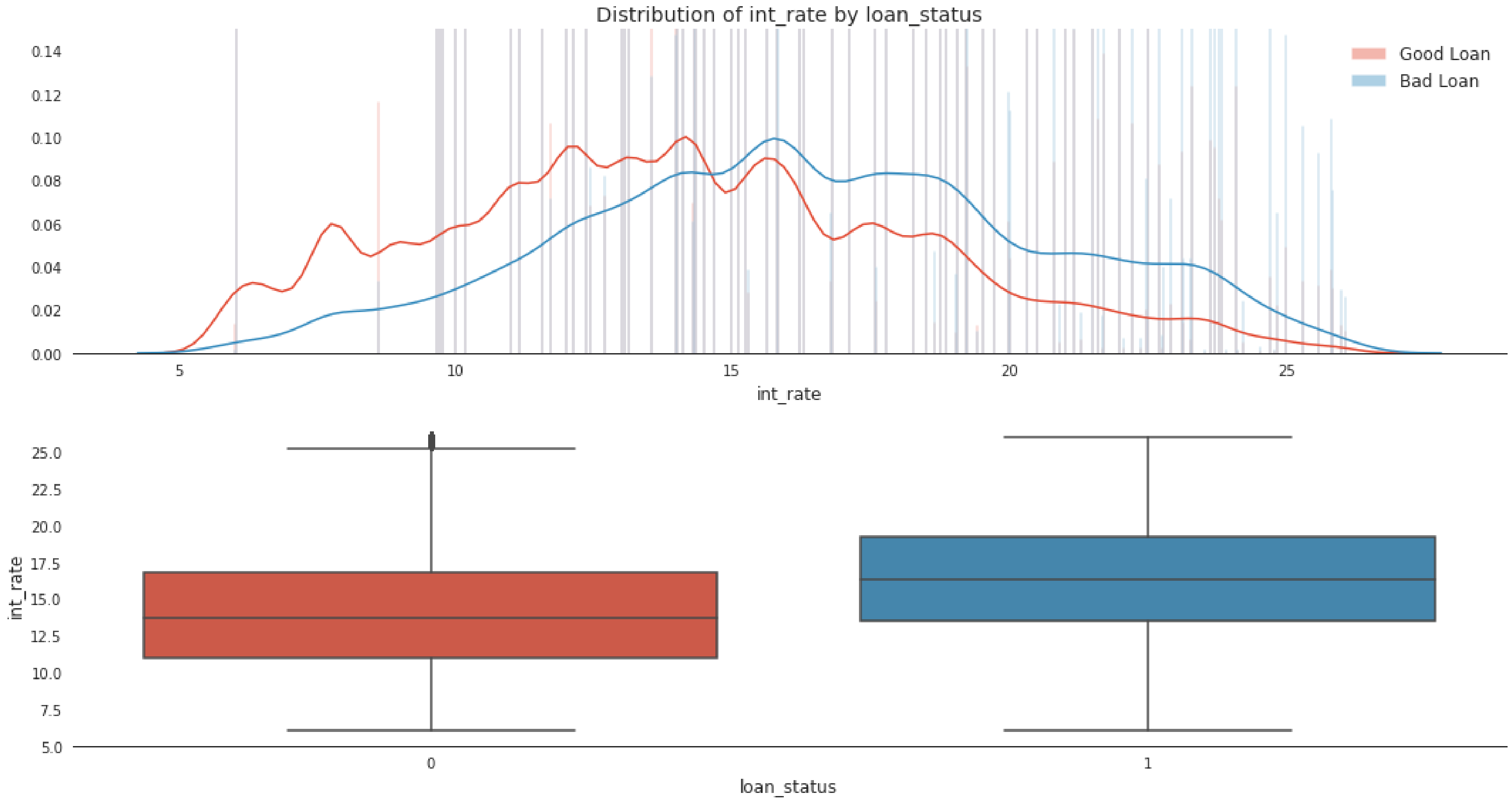
Loan amount by Loan Status



Annual Income by Loan Status



Interest Rate by Loan Status



Two-sample Z-test

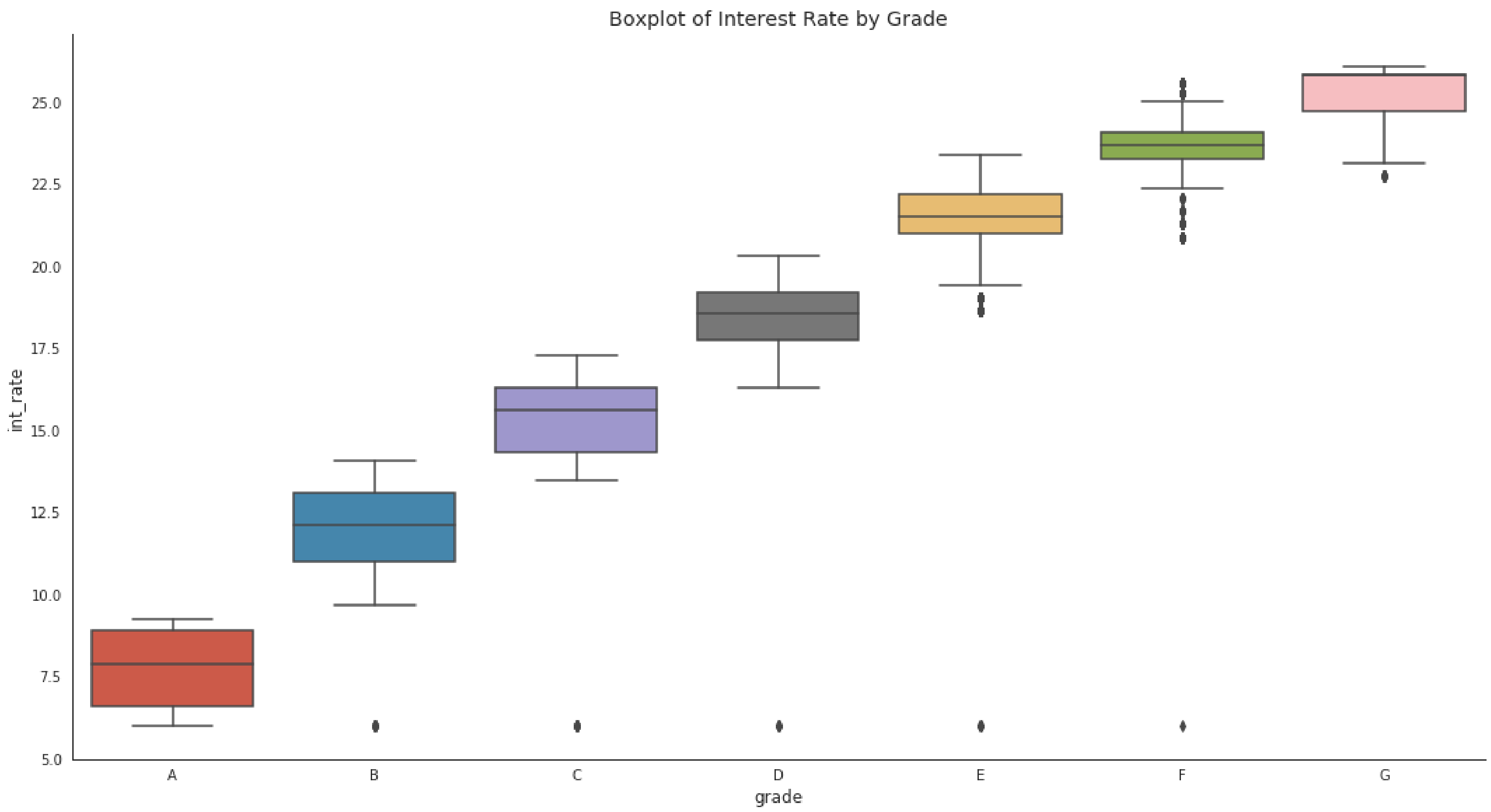
- Test for difference in mean

$H_0\colon \mu_1 = \mu_2$

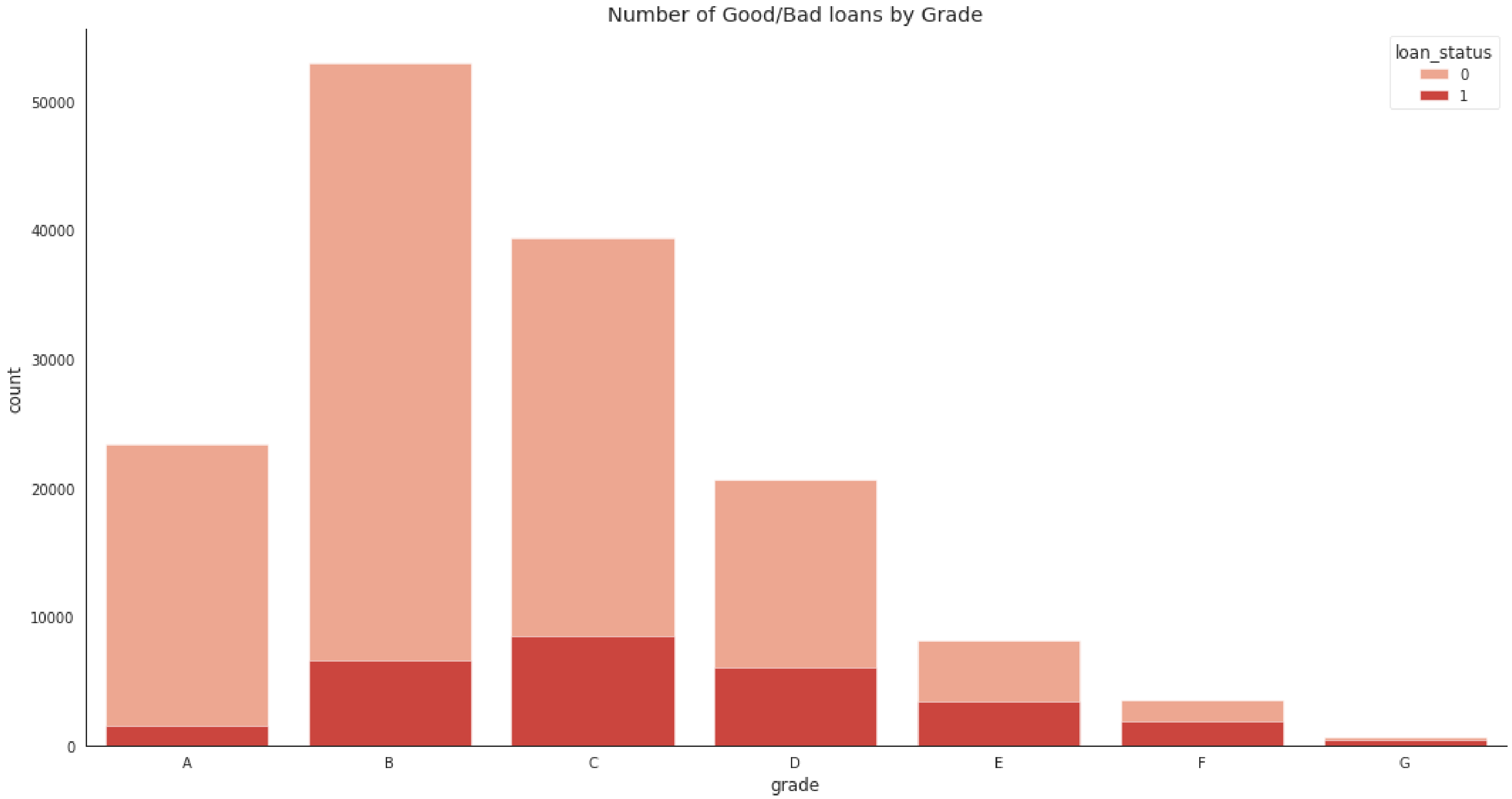
$H_1\colon \mu_1 \neq \mu_2$

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}}$$

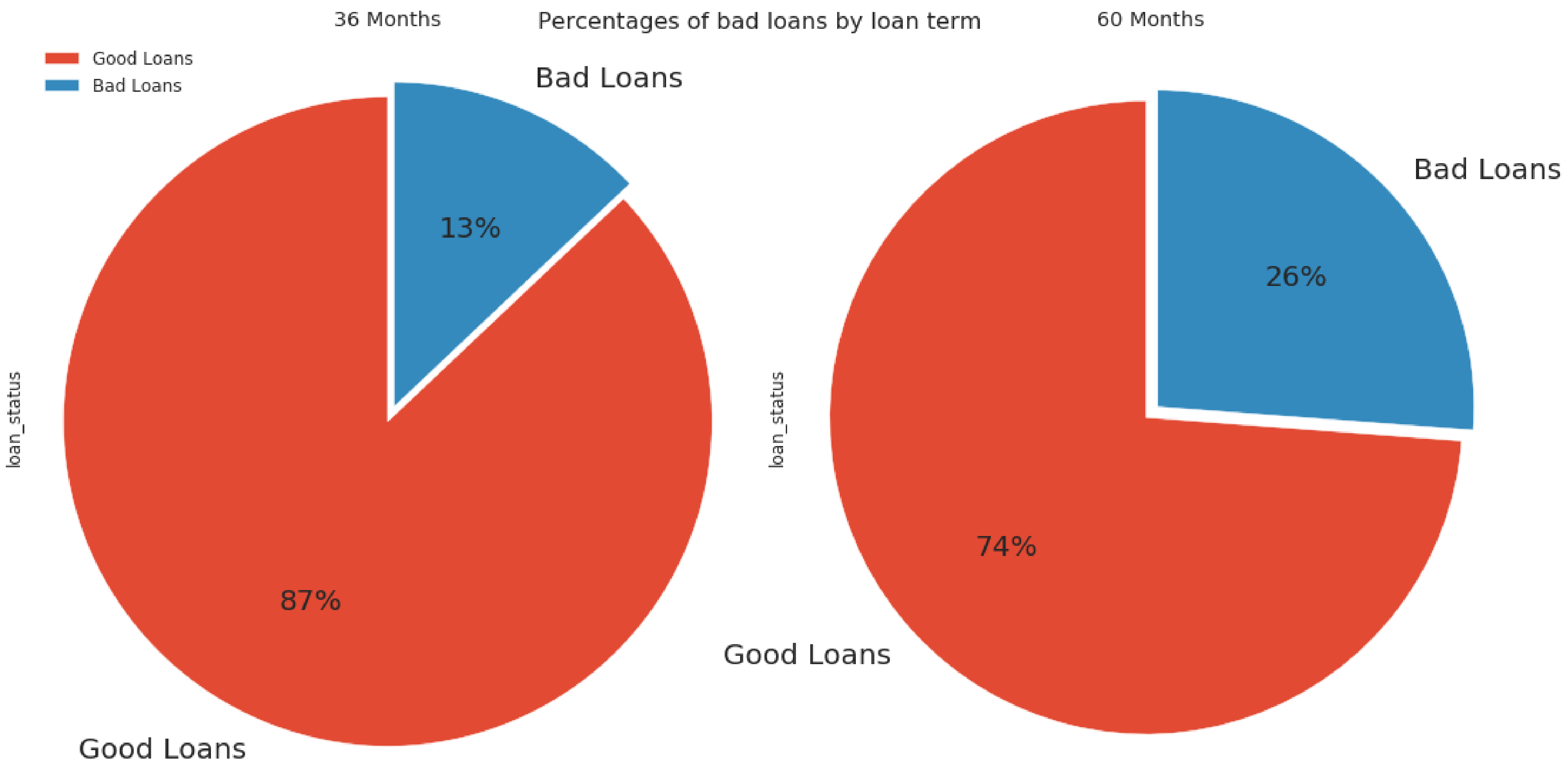
Interest Rate by Grade



Grade by Loan Status



Term by Loan Status



Chi-square test

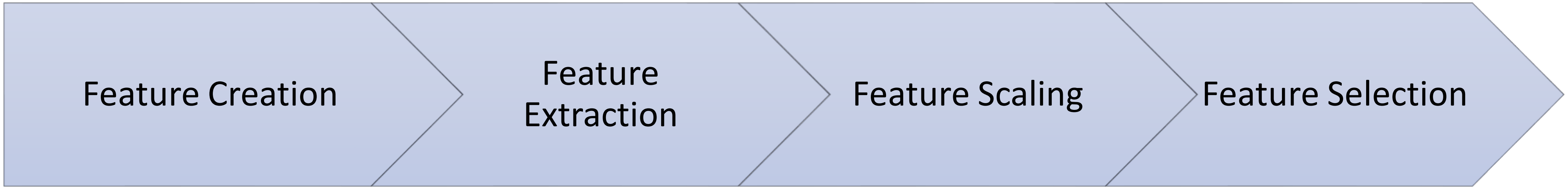
- Test for independence between categorical groups

$H_0: Two\ categorical\ variables\ are\ independent$

$H_1: Two\ categorical\ variables\ are\ not\ independent$

$$x^2 = \sum \frac{(Observed\ \ - Expected)^2}{Expected}$$

Feature Engineering



Feature Creation

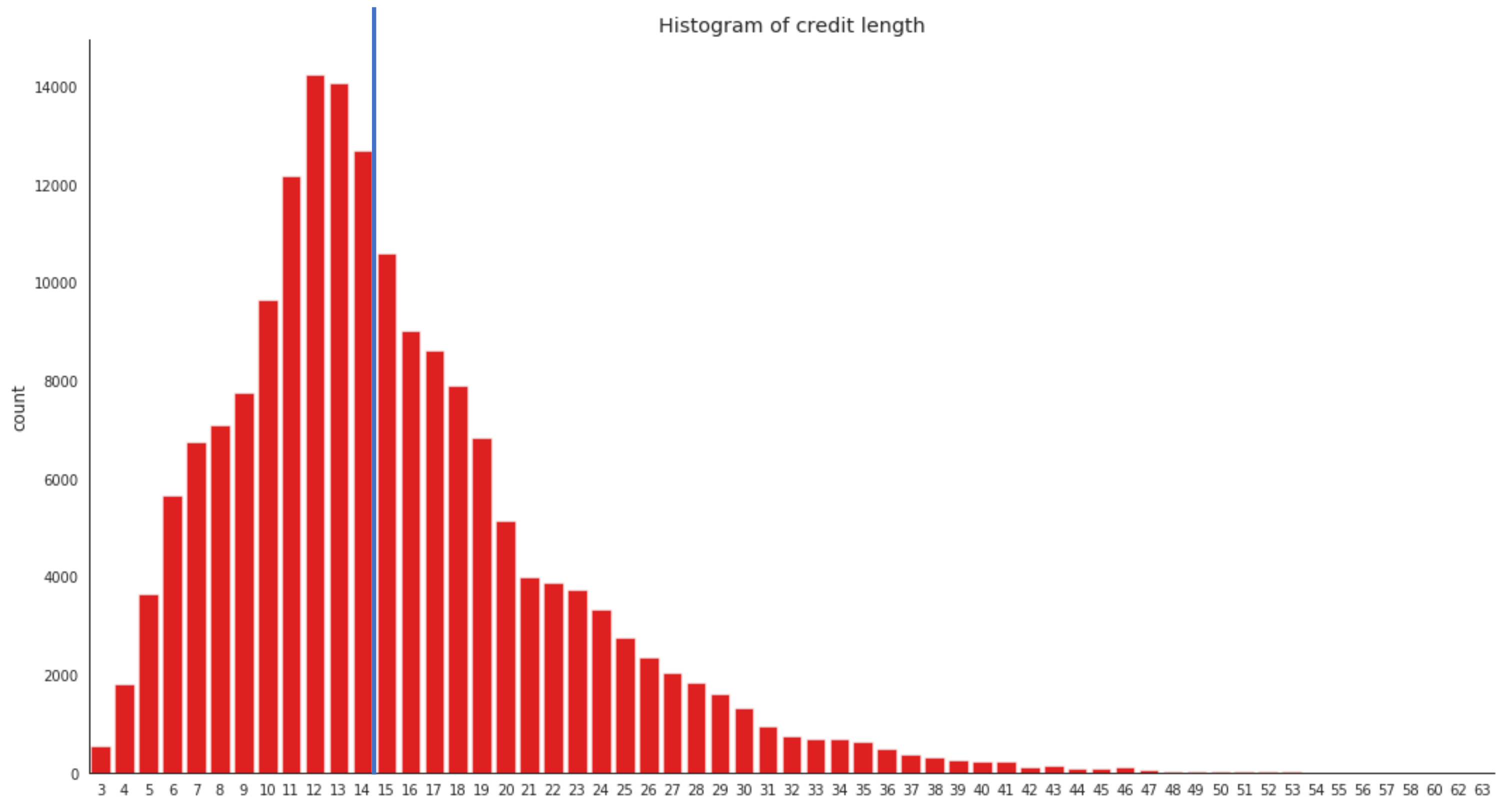
- Credit length = issue year – earliest credit line year
- Installment Feat = monthly installment/monthly income

Feature Extraction - Binning

- Binning (create buckets for variables)
    - Employment length
    - Delinquency in last 2 years
    - Num of derogatory record
    - Inquiry in last 6 months
    - Number of open accounts
    - Etc.

Median Line

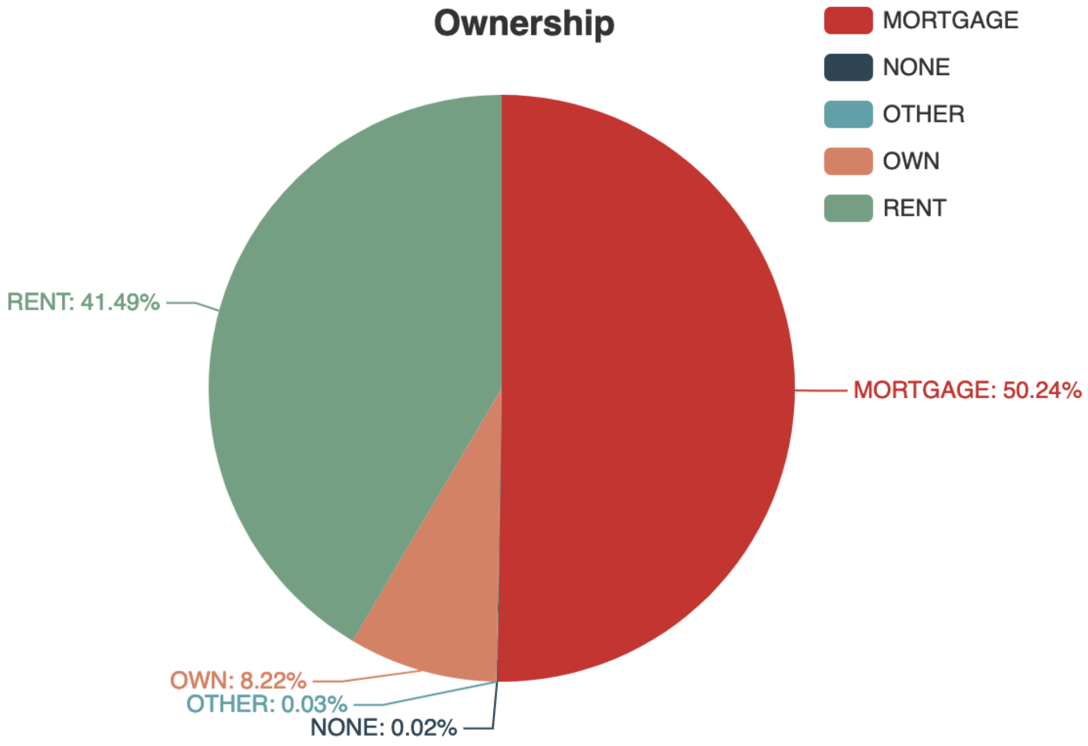
<= 14: coded as 0
> 14: coded as 1



Feature Extraction - Grouping

- Group some categories into larger group

```
map_list = {
    'purpose':{
        'renewable_energy':'other',
        'moving':'home_improvement',
        'house':'home_improvement',
        'vacation':'other',
        'wedding': 'other'},
    'home_ownership':{
        'OTHER':'MORTGAGE',
        'NONE':'MORTGAGE'
    }
}
```



Feature Scaling

- Numerical: Standardization

	loan_amnt	funded_amnt	funded_amnt_inv	int_rate	installment	annual_inc
0	-0.254098	-0.253850	-0.252597	-1.769860	-0.277778	1.143209
2	1.653496	1.654310	1.656842	-0.771640	1.921421	-0.379746
3	-0.254098	-0.253850	-0.252597	-0.543215	-0.149869	1.904686
4	-0.254098	-0.253850	-0.252597	-1.541435	-0.254425	0.884306
5	-0.818138	-0.818057	-0.817182	0.427594	-0.712122	-1.202141

- Categorical: Dummy coding

grade_A	grade_B	grade_C	grade_D	grade_E	grade_F	grade_G
1	0	0	0	0	0	0
0	1	0	0	0	0	0
0	1	0	0	0	0	0
1	0	0	0	0	0	0
0	0	1	0	0	0	0

Feature Selection

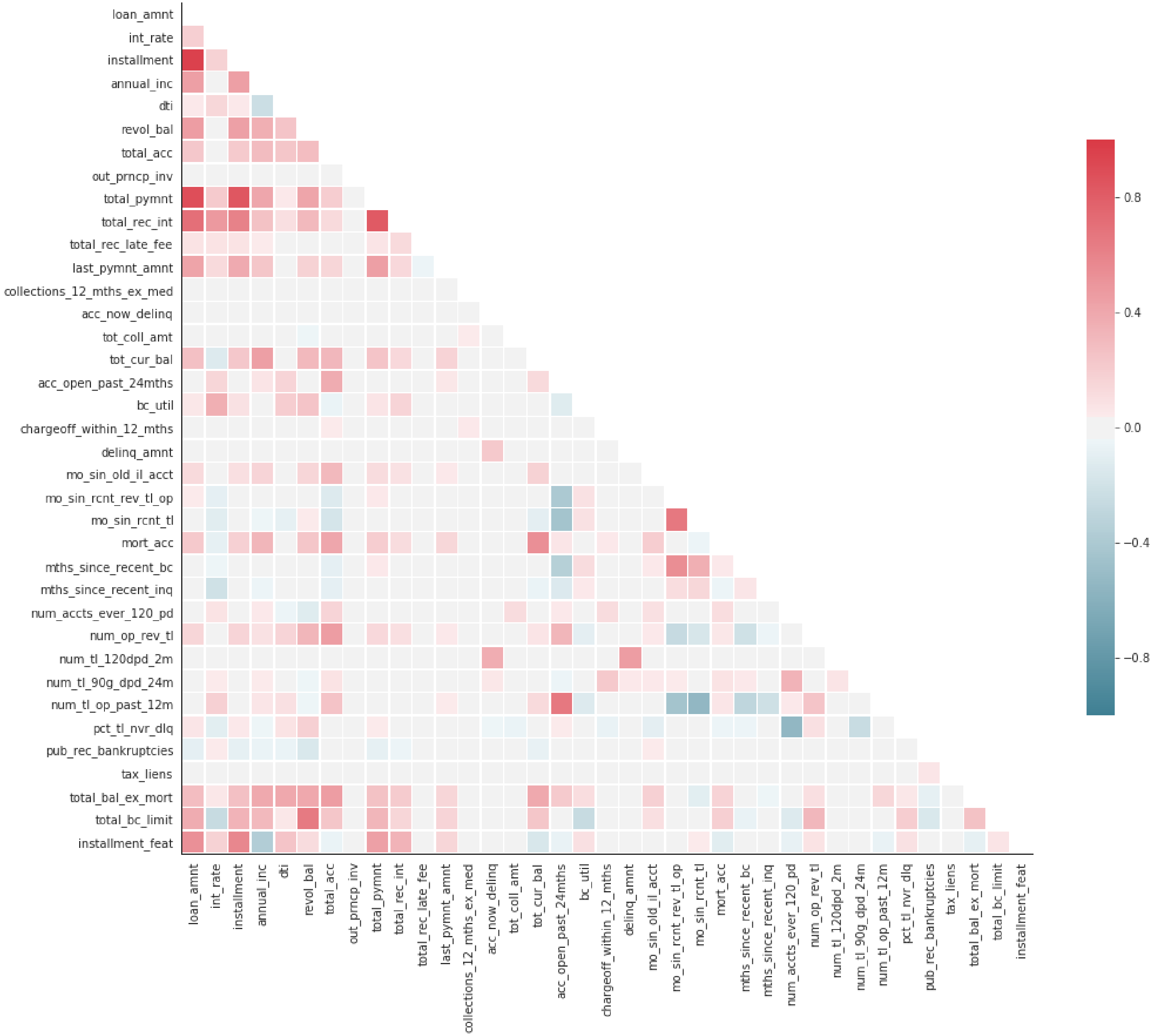
- Filtering Method: Correlation based
- Wrapper Method : Recursive Feature Elimination
- Embedded: Model based variable selection (Lasso, Random Forest)

Filtering

- Funded Amount, funded Amount investors, loan amount almost have correlation of 1
- Number of revolving account, open accounts, satisfactory account, bank account are similar variables

Filtering

- Removed 29 variables



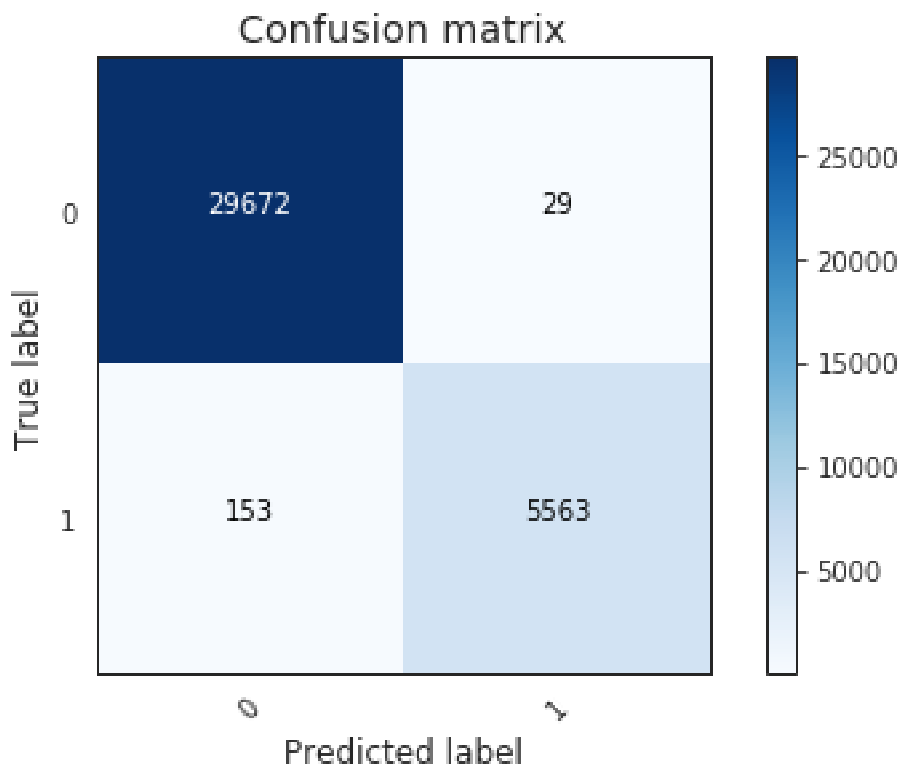
Recursive Feature Elimination

- Exhaustive
- Backward elimination using logistic regression
- Eliminate least important features until 30 variables left
- Metric: accuracy
- Cross-validation to automatically select number of features

Model Building

- Train-test split: 80% Training/20% Testing
- Model:
  - Logistic Regression
  - Decision Tree
  - Random Forest
- Evaluation: ROC Curve/Confusion Matrix

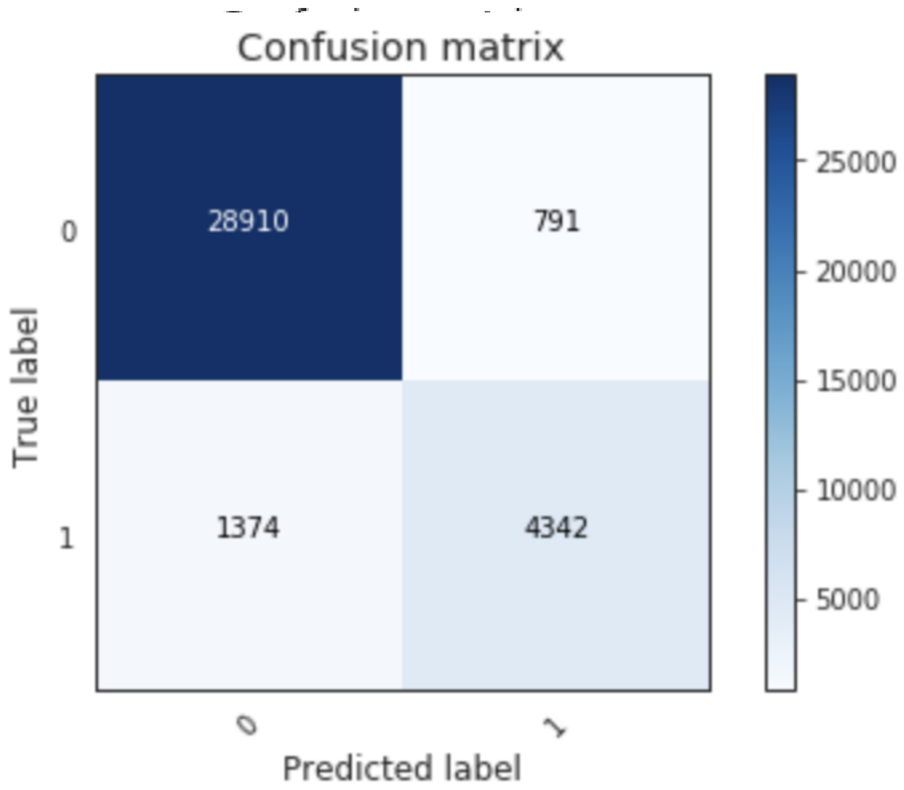
Logistic Regression




```
Precision =      0.995
Recall (TPR) =   0.973
Fallout (FPR) = 0.001
```

Threshold: 0.2

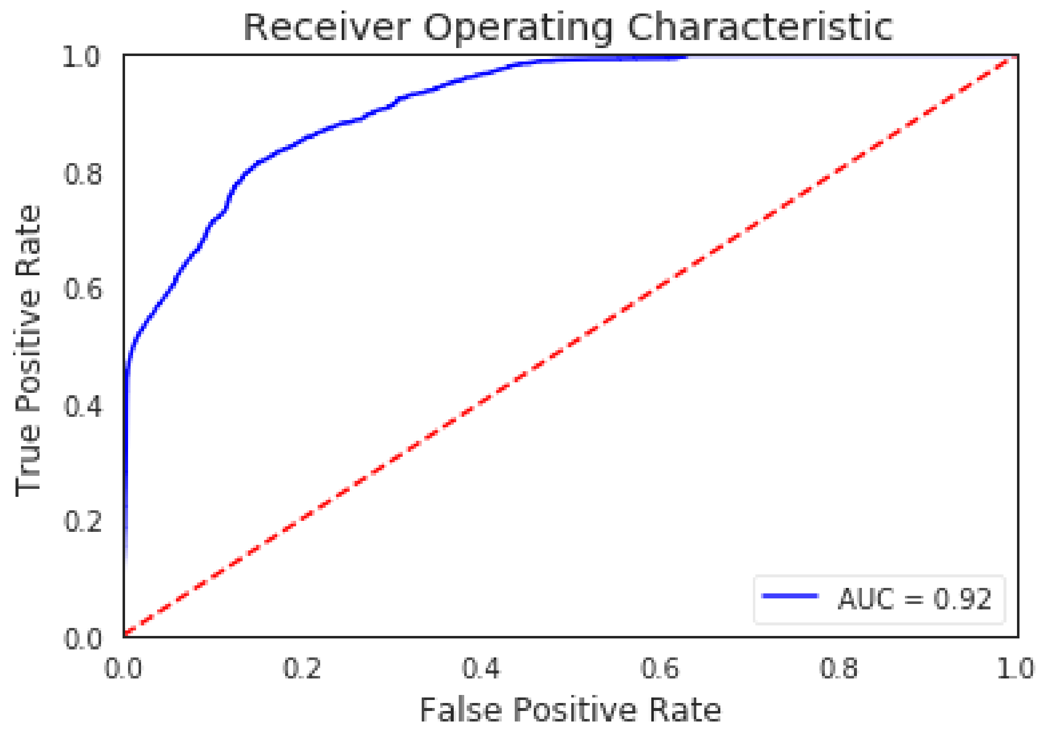
Decision Tree

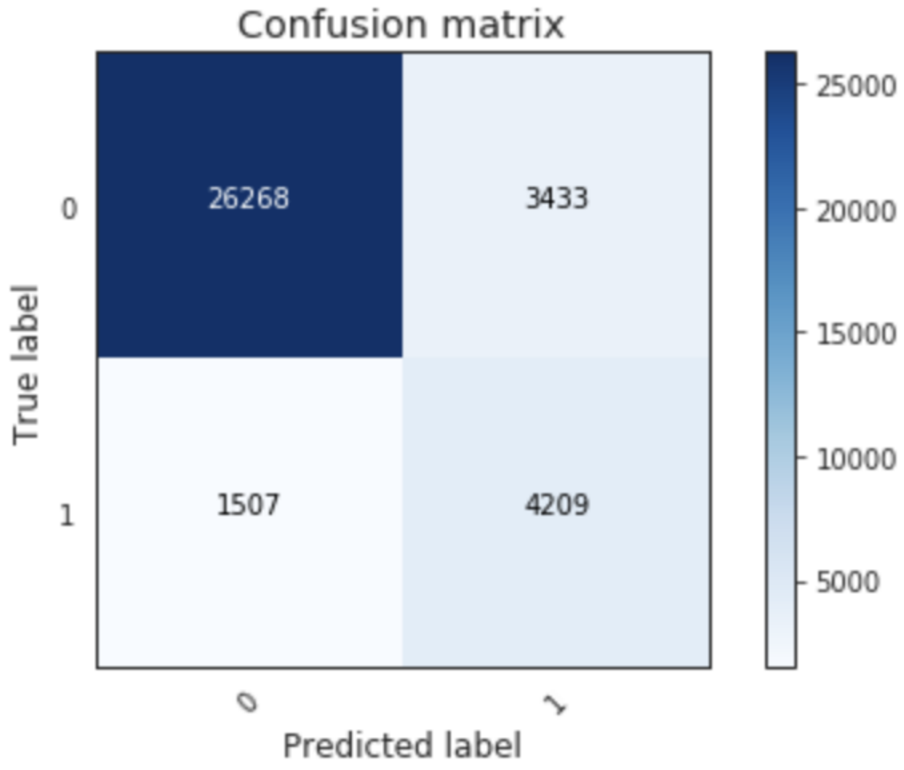


Threshold: 0.2

```
Precision =      0.846
Recall (TPR) =   0.760
Fallout (FPR) =  0.027
```

Random Forest

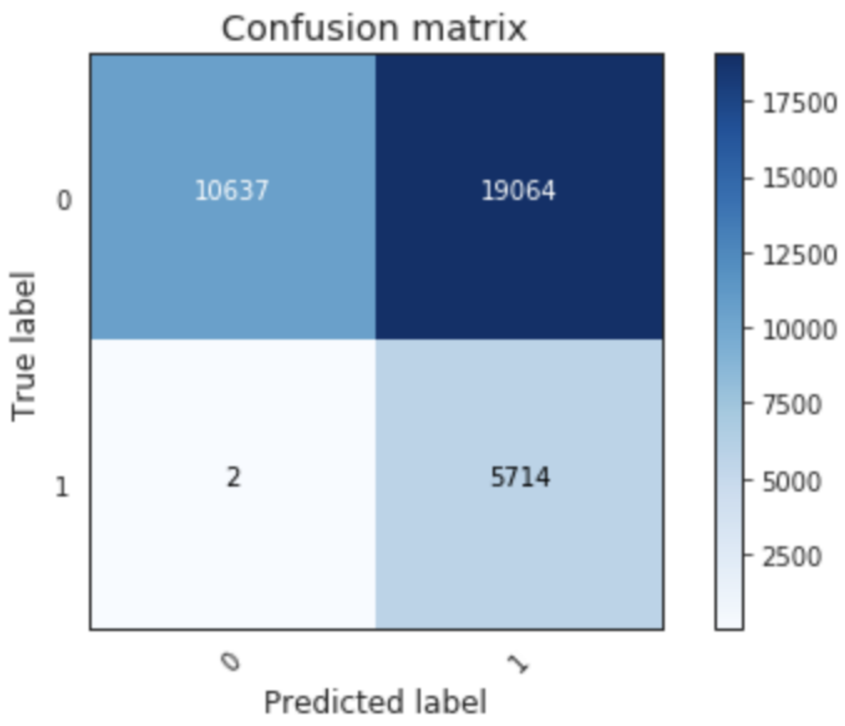




Threshold: 0.2

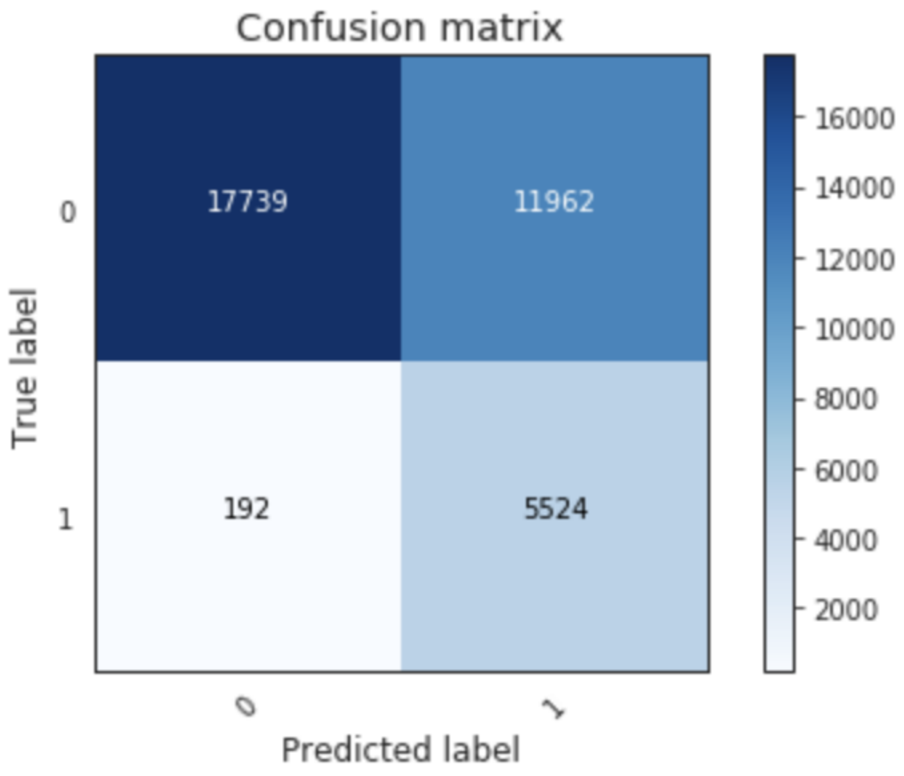
```
Precision =      0.551
Recall (TPR) =   0.736
Fallout (FPR) = 0.116
```

Random Forest




```
Precision =      0.231
Recall (TPR) =   1.000
Fallout (FPR) = 0.642
```

Threshold: 0.1



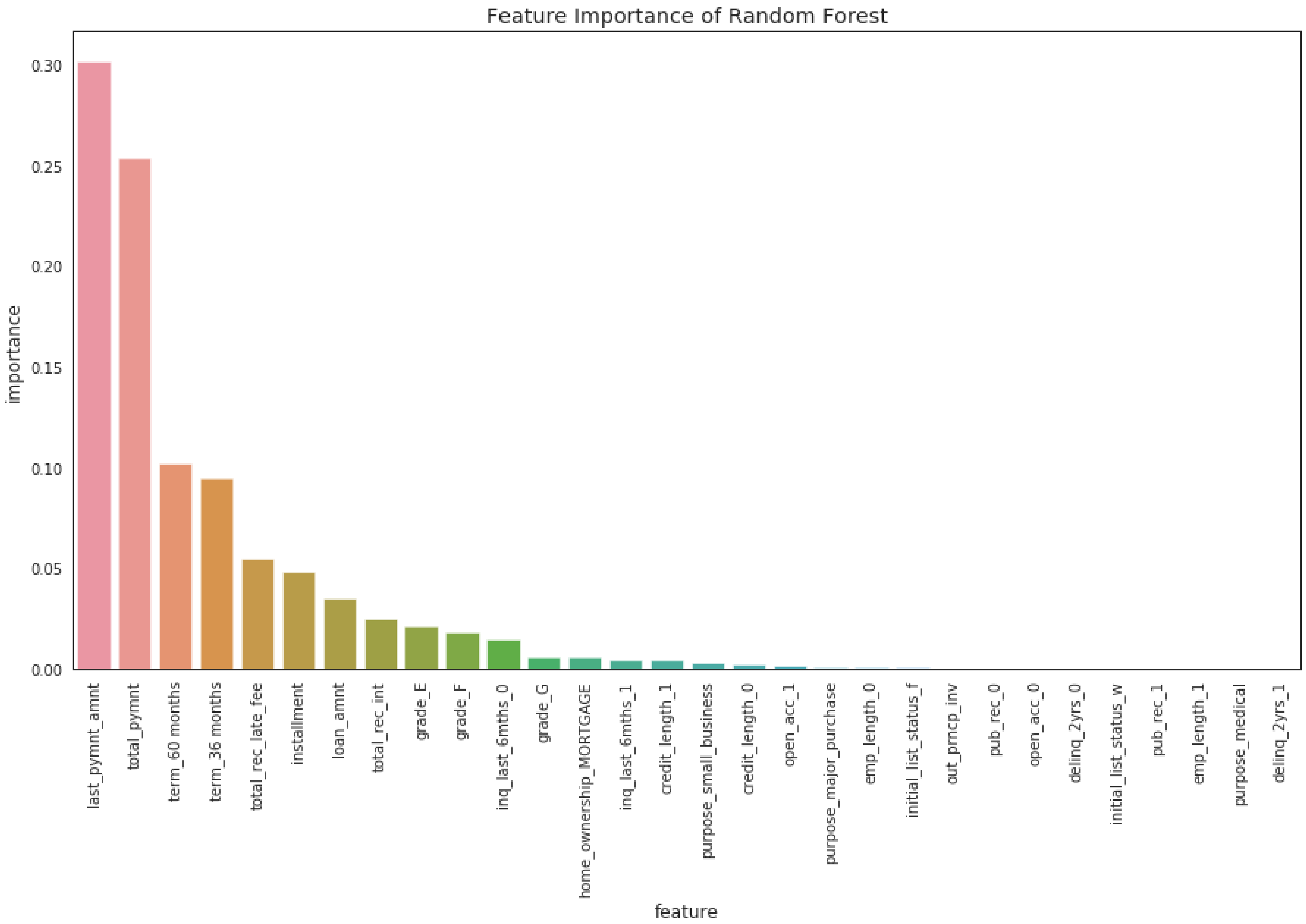

```
Precision =      0.316
Recall (TPR) =   0.966
Fallout (FPR) = 0.403
```

Threshold: 0.15

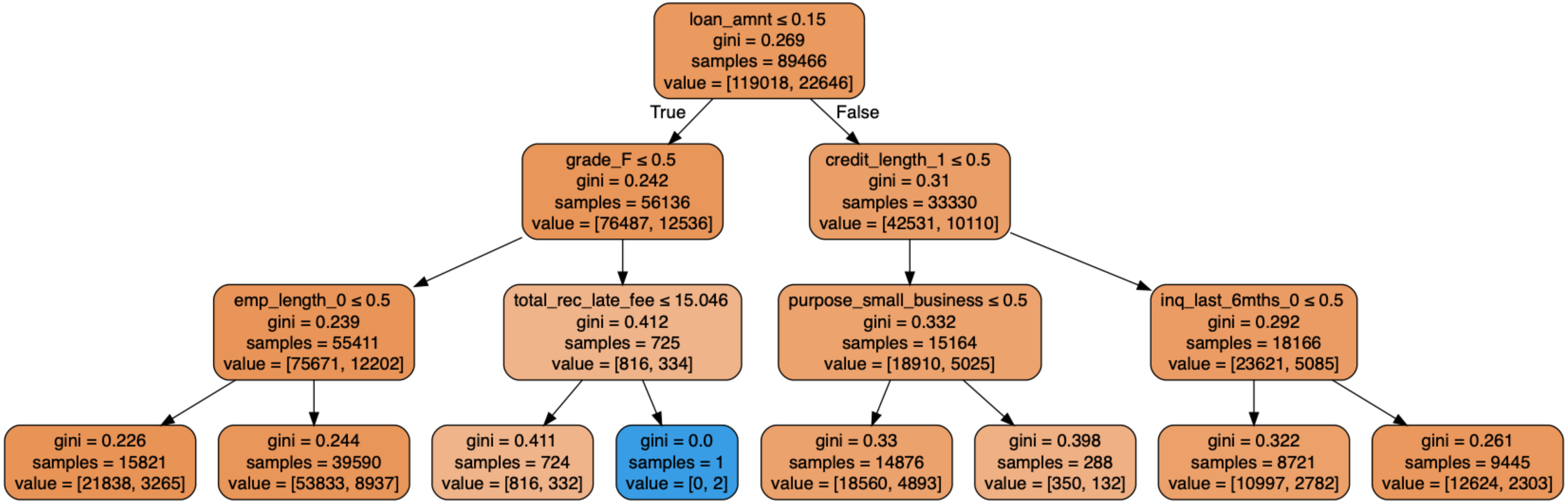
Feature Importance – Gini Index

Important Features:

- Last payment amount
- Total payment
- Term
- Total late fees paid
- Total Interest paid
- Loan amount
- Total principal paid
- Grade



Visualizing Random Forest



Conclusion

- In this dataset, all default/non-default cases are completely separated, which leads to good classification result using variables such as ‘total late fees received’.
- In reality, most of the customers are ‘current’, the predicting power are likely to decrease
- The best result is achieved by Logistic Regression, but needs to be cross-validated before put into production

Recommendation

- Recursive Feature Elimination using cross-validation
- SMOTE (Oversampling) to deal with unbalanced dataset
- Hyperparameter tuning and cross validation
- Gather more data, up-to-2018Q4
- Analysis on geographic dimension
    - https://public.tableau.com/views/Book2_15528449246150/Dashboard1?:embed=y&:display_count=yes

Thank you!