Springboard Capstone Project #1

Will the loan be paid-off ?

Loan

Investment Vehicle

Borrowers

LendingClub

PEER-TO-PEER LENDING COMPANY

Lenders

Origination Fee

Service Fee

kaggle

LendingClub

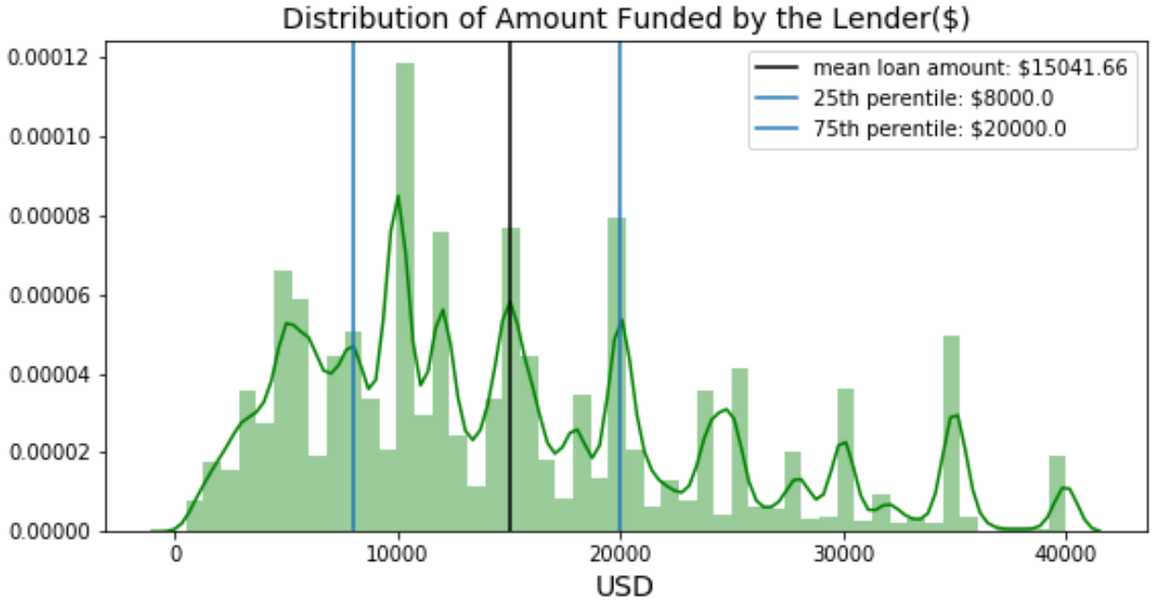
- each and every loan
- from 2007-2019

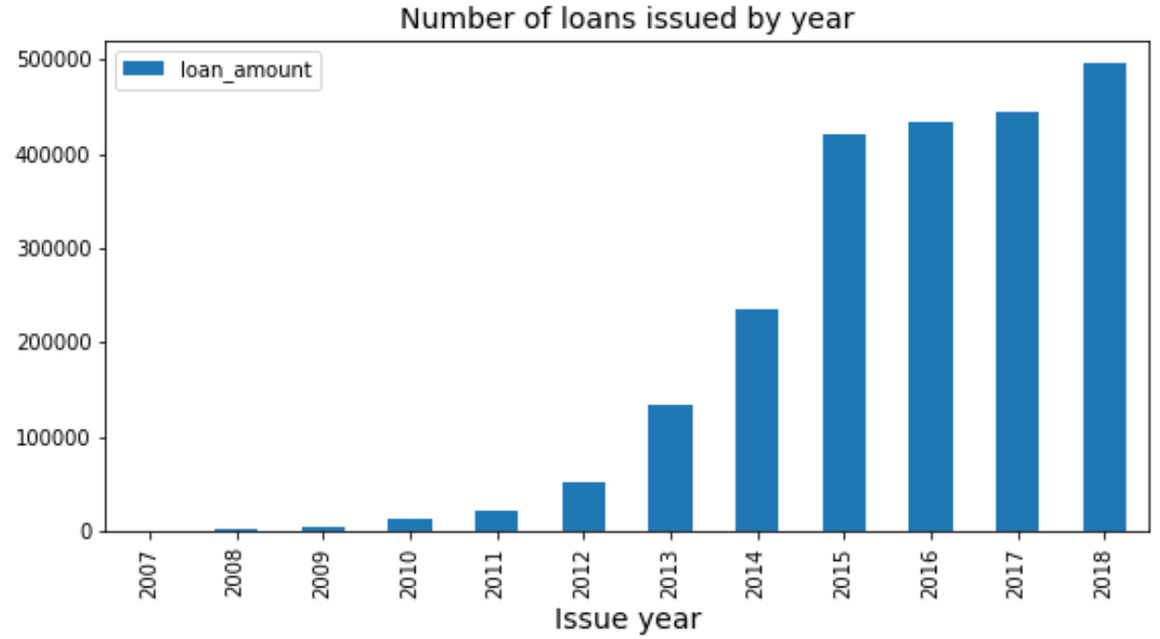
- 1.19 GB
- 2.26 mil. rows
- 145 features

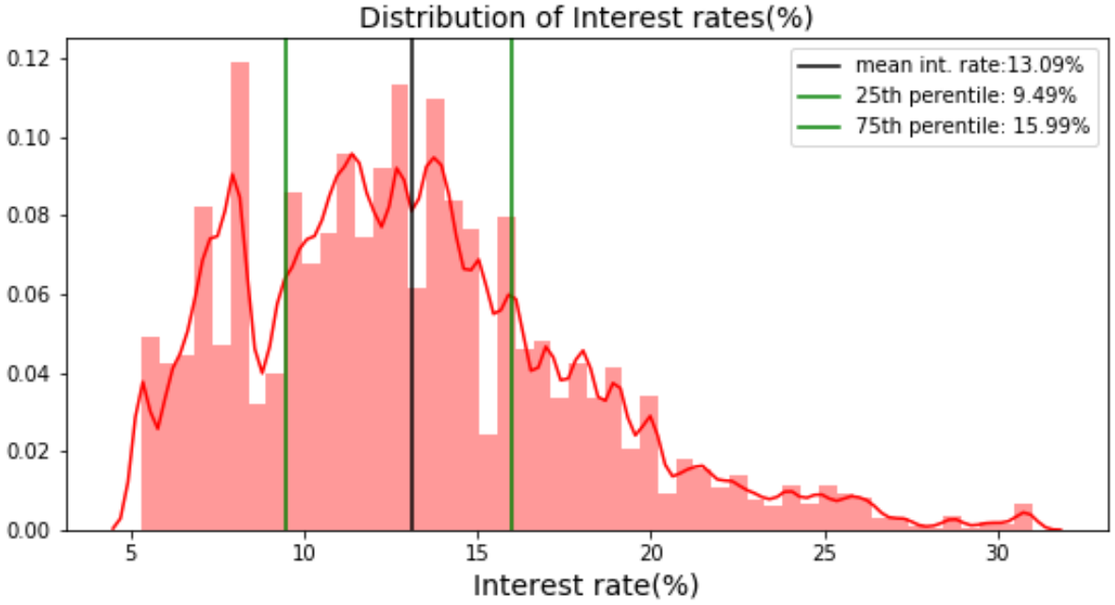
```
1 df.head()
```

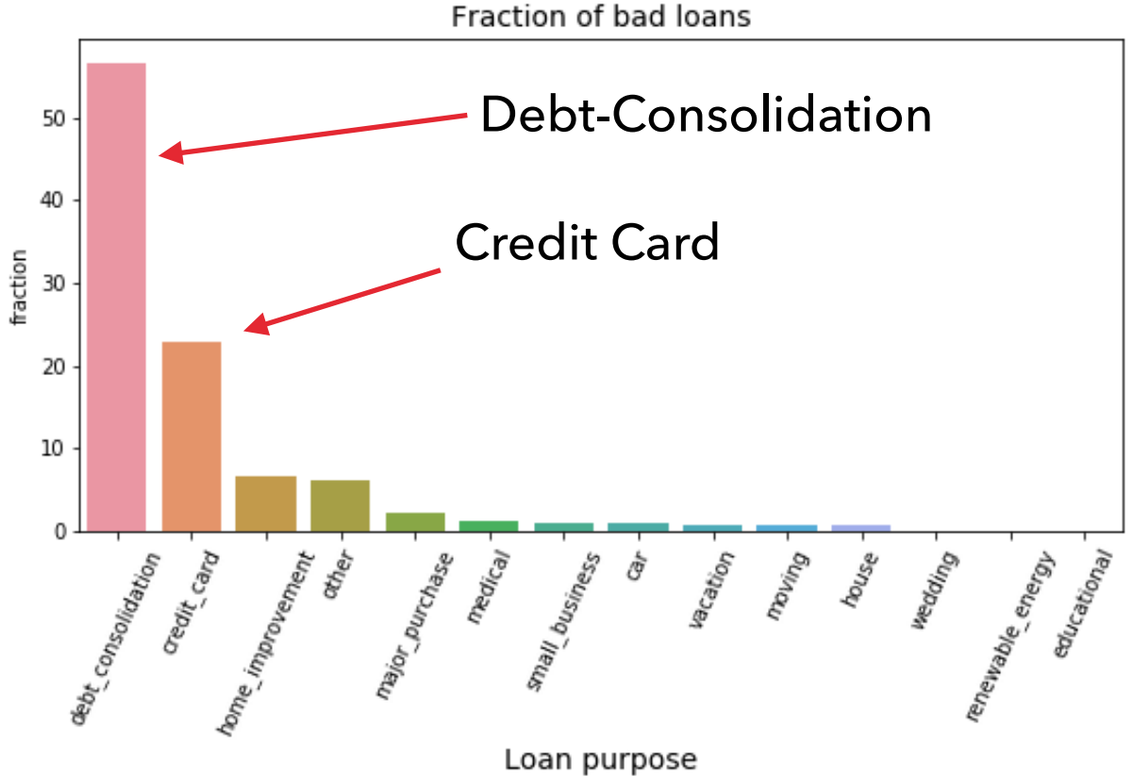
	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	grade	sub_grade	emp_title	emp_length
0	2500	2500	2500.0000	36 months	13.5600	84.9200	C	C1	Chef	10+ years
1	30000	30000	30000.0000	60 months	18.9400	777.2300	D	D2	Postmaster	10+ years
2	5000	5000	5000.0000	36 months	17.9700	180.6900	D	D1	Administrative	6 years
3	4000	4000	4000.0000	36 months	18.9400	146.5100	D	D2	IT Supervisor	10+ years
4	30000	30000	30000.0000	60 months	16.1400	731.7800	C	C4	Mechanic	10+ years

Key Metrics


Distribution of Amount Funded by the Lender($)
mean loan amount: $15041.66
25th perentile: $8000.0
75th perentile: $20000.0
0.00012
0.00010
0.00008
0.00006
0.00004
0.00002
0.00000
0
10000
20000
30000
40000
USD

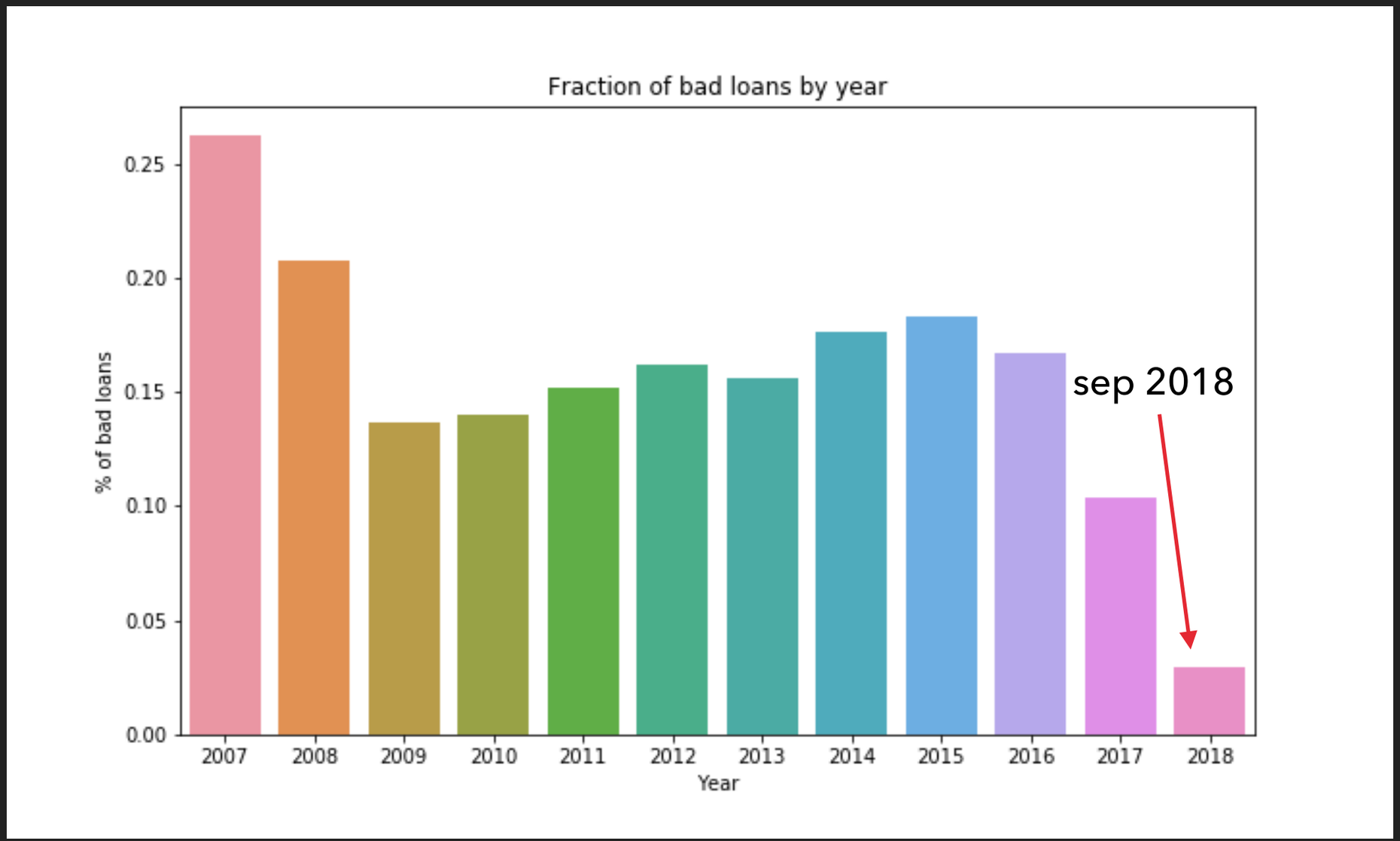

Number of loans issued by year
loan_amount
500000
400000
300000
200000
100000
0
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
Issue year


Distribution of Interest rates(%)
mean int. rate:13.09%
25th perentile: 9.49%
75th perentile: 15.99%
0.12
0.10
0.08
0.06
0.04
0.02
0.00
5
10
15
20
25
30
Interest rate(%)


Fraction of bad loans
Debt-Consolidation
Credit Card
fraction
50
40
30
20
10
0
debt_consolidation
credit_card
home_improvement
other
major_purchase
medical
small_business
car
vacation
moving
house
wedding
renewable_energy
educational
Loan purpose

Problem: **Minimize the Fraction of Bad Loans**

Proposal: **To build a predictive ML model to identify future bad loans at the time of application.**



Which features have predictive power?

sklearn.feature_selection.chi2

`sklearn.feature_selection.` **`chi2`**`(X, y)` [source]

Compute chi-squared stats between each non-negative feature and class.

This score can be used to select the n_features features with the highest values for the test chi-squared statistic from X, which must contain only non-negative features such as booleans or frequencies (e.g., term counts in document classification), relative to the classes.

Recall that the chi-square test measures dependence between stochastic variables, so using this function "weeds out" the features that are the most likely to be independent of class and therefore irrelevant for classification.

```
1 chi2_table.sort_values('statistic',ascending=False)
```

	feature	statistic	p-value
22	last_pymnt_amnt	62822.9637	0.0000
17	total_rec_prncp	42694.9679	0.0000
15	total_pymnt	22972.1608	0.0000
16	total_pymnt_inv	22906.4785	0.0000
79	grade_A	20865.6040	0.0000
83	grade_E	14766.6788	0.0000
97	sub_grade_other	13072.8918	0.0000
3	int_rate	10572.7377	0.0000
82	grade_D	9809.4045	0.0000
80	grade_B	7833.3168	0.0000
78	term_ 36 months	7356.5606	0.0000

Hard-Learnt Lesson/Rookie Mistake

- KNOW YOUR FEATURES !!!

Preprocessing & making sure the data doesn't leak

! Binary Classification

! Feature Engine

! Pipeline

Numerical	Categorical
IQR outlier treatment	Rare Label Encoding
Discretization KBins/CAIMD	OneHot Encoding
Standard Scaler	

Modeling

	AUC	Precision Score
Logistic Regression	0.9963	0.9990
Random Forest Classifier	0.9946	0.9982

```
chi2_table.sort_values('statistic',ascending=False)
```

	feature	statistic	p-value
22	last_pymnt_amnt	62822.9637	0.0000
17	total_rec_prncp	42694.9679	0.0000
15	total_pymnt	22972.1608	0.0000
16	total_pymnt_inv	22906.4785	0.0000
79	grade_A	20865.6040	0.0000
83	grade_E	14766.6788	0.0000
97	sub_grade_other	13072.8918	0.0000
3	int_rate	1057	
82	grade_D	980	
80	grade_B	783	
78	term_ 36 months	735	

Hard-Learnt Lesson/Rookie Mistake

- KNOW YOUR FEATURES !!!

Analysis uses data that would have not been known under real conditions.

Look-Ahead Bias

By WILL KENTON | Updated Feb 16, 2020

What Is the Look-Ahead Bias?

Look-ahead bias occurs by using information or data in a study or simulation that would not have been known or available during the period being analyzed. This can lead to inaccurate results in the study or simulation. More importantly, a look-ahead bias can unintentionally sway simulation results closer into line with the desired outcome of the test. This leads to economists and analysts putting too much confidence in their models and the ability of the model to predict and mitigate future events. Investors also need to be aware of the potential for look-ahead bias when evaluating particular trading strategies using past data.

Results after the biased features were removed:

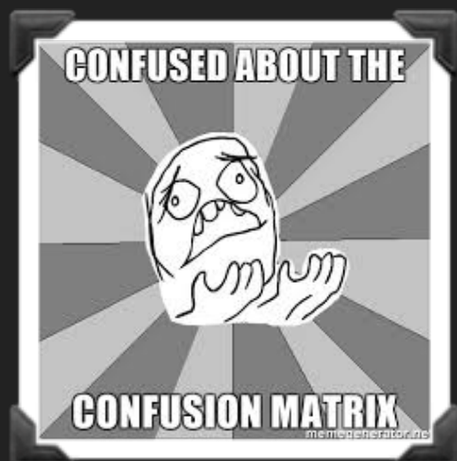
	AUC	Precision Score
Logistic Regression	0.7510	0.9146
Gradient Boosting Classifier	0.7361	0.9098
Gaussian NB	0.7274	0.9026

Classification report

Logistic Regression (Accuracy 0.79)	Precision	Recall	f1-score
Charged-off	0.47	0.44	0.45
Paid-off	0.87	0.88	0.87

Adjusting predict_proba threshold

Logistic Regression (Accuracy 0.81)	Precision	Recall	f1-score
Charged-off	0.21	0.98	0.35
Paid-off	0.96	0.11	0.19



Final steps

GridsearchCV - optimal C is 15

Model was picked and saved

Future steps

Log transformation

Deep learning

Conclusion

Proposal: **To build a predictive ML model to identify future bad loans at the time of application.**

Not great, not terrible

What did I learn?

- Skepticism, maybe.

- Getting and preparing data is 85% of the job.

- Experimenting is FUNdamental.

- There are lots of libraries that make your life easier and save you a lot of time.

- Existing pre-processing and modeling routine code that that generalizes easily.

- Classification reports are awesome !