My Capstone

Predictive Modeling
With
Financial data

Data Science through Predictive Modeling

- Using data provided by LendingClub.com I developed a model that will allow investors, of this peer-to-peer lending group, to make more informed decisions.

- Data analytics, financial data interpretation and a drive to come to a profitable conclusion.

- LendingClub.com - Peer to peer lending company operating in all 50 states.
  - Investors peers lending money to either portfolios provided by LendingClub.com or by choosing individual applicants to invest.

- LendingClub.com –
  - Investors can view credit report and 2 letter scores assigned by LendingClub.com
    - Grade values A, B, C, D, F, G

Data and Assumptions

- LendingClub.com loans totaling 421097 with 115 features.
  - Using Amazon Web Services during machine learning
- Features broken into subsects:
  - Applicant metrics
  - Credit reporting data
  - LendingClub.com loan data

- Assumptions:
  - LendingClub.com actual loan data deleting from project

Data and Assumptions

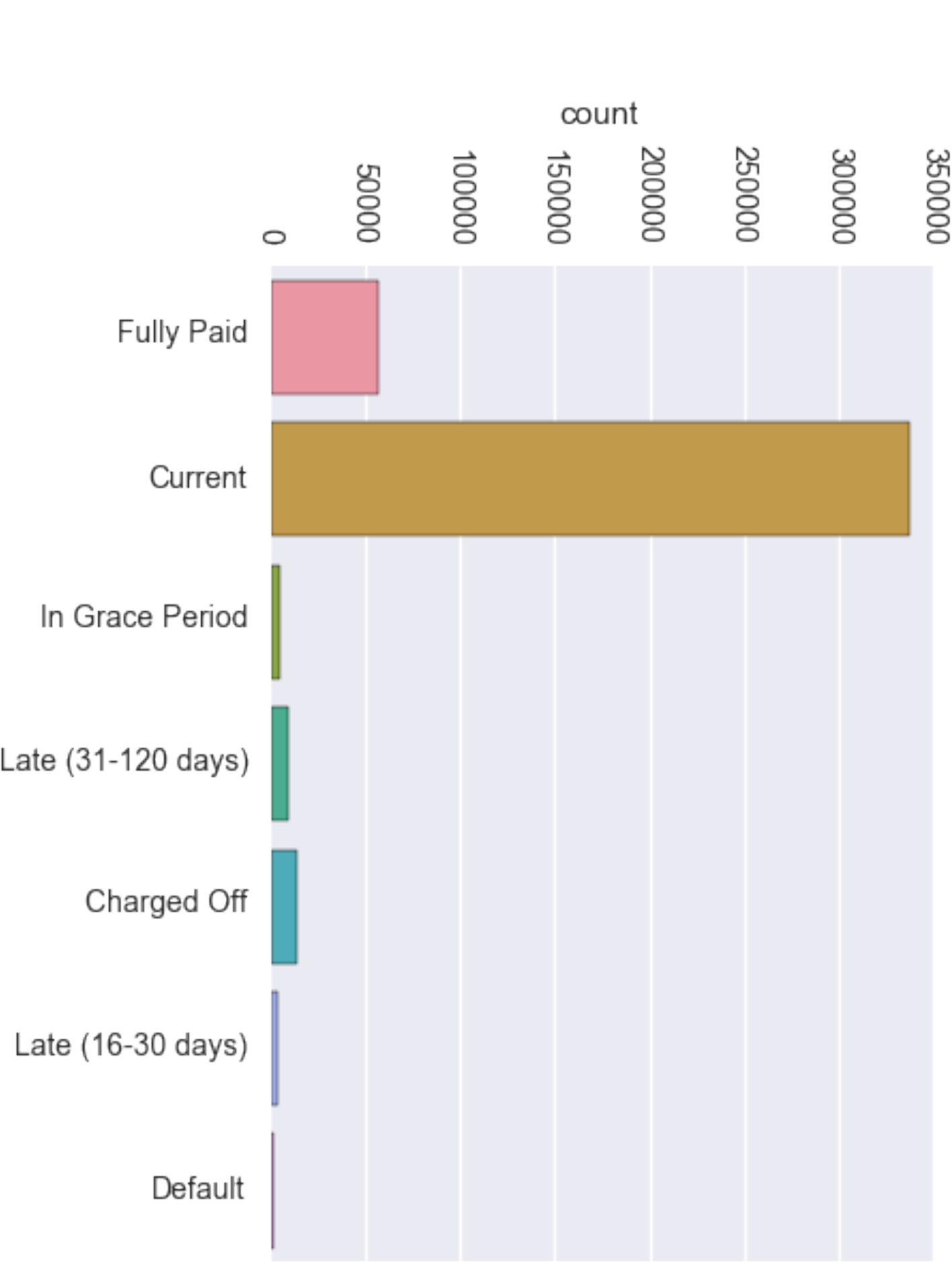
- Assumptions (cont.):
  - Missing values for credit reporting data filled with 0's
  - Variables or features are not dependent

Data Cleaning

- Data set: LendingClub.com customers, and the current status of their loan.
  - Target Loan Status
  - Current – loans current
  - Fully-paid – loans fully paid
  - Grace period – those less than 15 days late on current payment
  - Late (16-30 days)
  - Late (31-120 days)
  - Default – loans defaulted still in negotiations with applicant

Data Cleaning

- Target Column:
  - Setting Target to Good and Bad where
    - Good: Fully paid, Current and Grace (0)
    - Bad: Charged Off, Late, Default
    - Roughly 80% of all loans current, and 4% default , charge off or significantly late
  - Fillna(0) entired data set


count
0
50000
100000
150000
200000
250000
300000
350000
Fully Paid
Current
In Grace Period
Late (31-120 days)
Charged Off
Late (16-30 days)
Default

- Dropping and Why:

  -application_type, dti_joint, annual_inc_joint

    - Joint status dropped

  -next_pymnt, last_cr_pulled, int_rate, member_id etc.

    - Info for current loan not useful for this project

- Home_ownership – changed ‘Any’ to ‘Mortgage

Data Cleaning

- Features
  - Revol_bal, revol_util
    - Removed % and returned as decimals
  - Label Encode
    - sub-grade, grade, emp_length, verification_status
  - Dummy variables
    - purpose
    - home_ownership

Early Data Analysis

- What does my data look like?
- What is the reason for the loans, and how do if any effect loan status?
- Is it a better investment to chose an applicant who is a home owner vs. renter

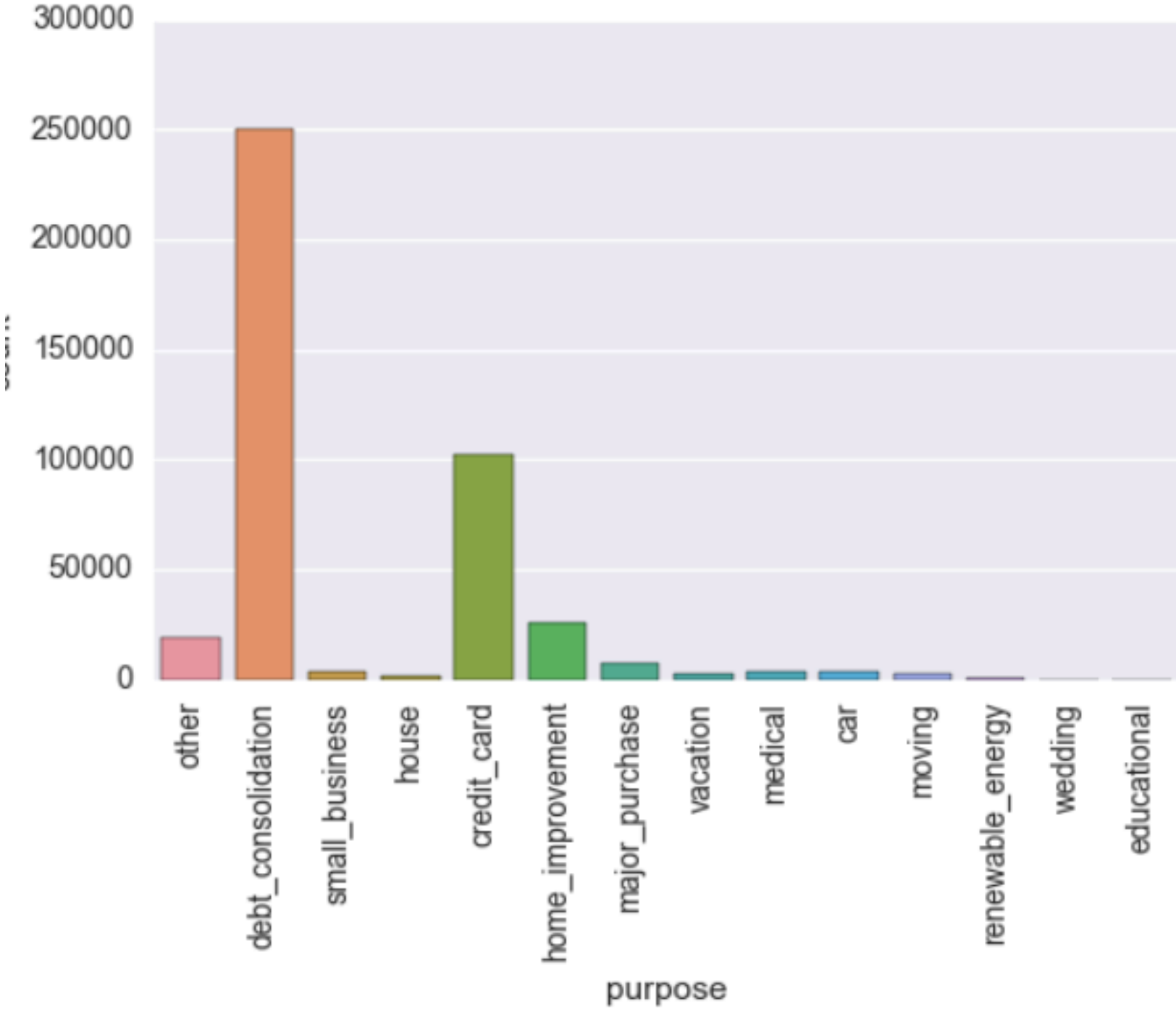
Early Data Analysis

- What role, if any, does an applicant's annual income have to play in loan status

- Since my 'Bad' target is a low frequency what can I do to make my models more accurate
  - 'balance' everything must be in balance
  - shuffle/split
  - stratify / cross-validate

Reason for loans

- Purpose column
  - 15 categories containing other
  - Whoa – rate of default higher:
    - Renewal energy
    - Home/mortgages
    - Moving
    - small business
  - Wedding/education – inconclusive
    - 4 wedding either current or fully paid
    - Education only 1 fully paid not much


300000
250000
200000
150000
100000
50000
0
other
debt_consolidation
small_business
house
credit_card
home_improvement
major_purchase
vacation
medical
car
moving
renewable_energy
wedding
educational
purpose

loan_status	Charged Off	Current	Default	Fully Paid	In Grace Period	Late (16-30 days)	Late (31-120 days)
purpose							
car	2.7%	78.2%	0.1%	15.8%	1.0%	0.6%	1.6%
credit_card	2.1%	84.3%	0.1%	10.8%	0.7%	0.5%	1.5%
debt_consolidation	3.4%	78.8%	0.1%	13.8%	1.0%	0.6%	2.2%
educational	nan%	nan%	nan%	100.0%	nan%	nan%	nan%
home_improvement	2.8%	78.5%	0.1%	15.2%	1.0%	0.6%	1.8%
house	5.6%	68.8%	0.3%	21.1%	0.8%	0.7%	2.6%
major_purchase	3.8%	78.1%	0.1%	14.4%	0.9%	0.6%	2.1%
medical	4.7%	76.7%	0.2%	14.4%	1.2%	0.5%	2.2%
moving	6.0%	72.9%	0.2%	16.6%	1.3%	0.8%	2.3%
other	3.9%	77.1%	0.2%	15.0%	1.1%	0.6%	2.3%
renewable_energy	4.9%	70.5%	nan%	18.3%	0.4%	0.9%	4.9%
small_business	6.4%	74.3%	0.2%	13.3%	1.2%	0.9%	3.6%
vacation	4.2%	75.4%	0.1%	16.8%	0.8%	0.4%	2.3%
wedding	nan%	75.0%	nan%	25.0%	nan%	nan%	nan%

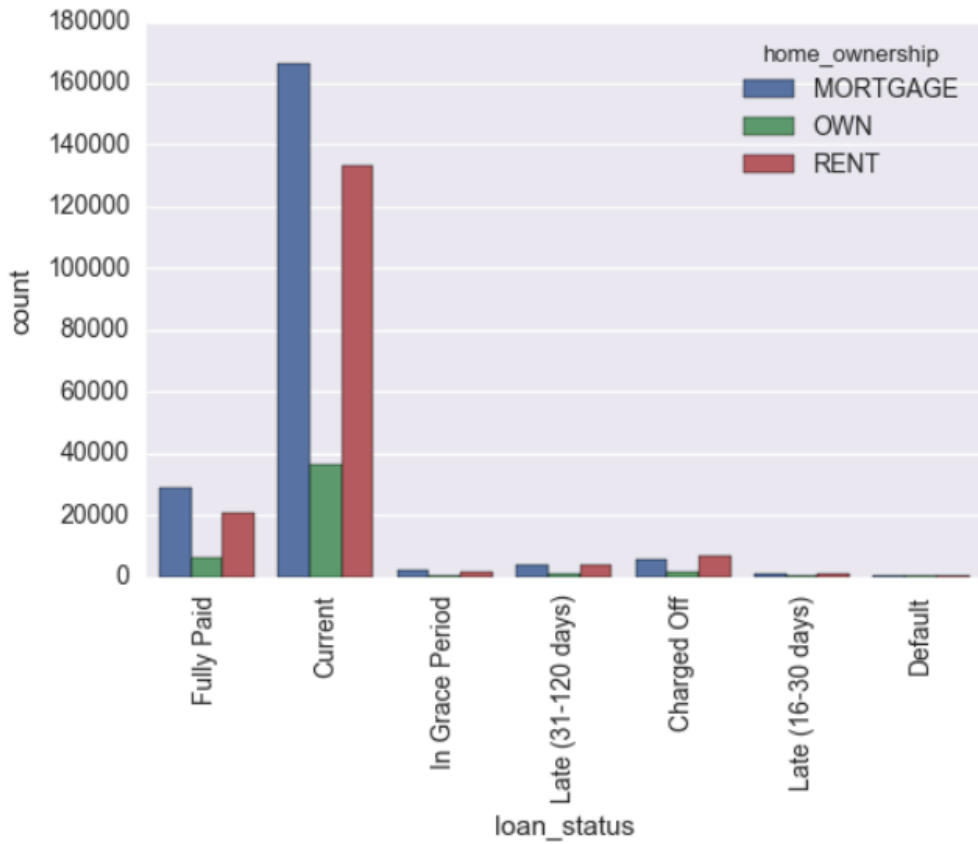
Early Data Analysis

Employment length

loan_status	Charged Off	Current	Default	Fully Paid	In Grace Period	Late (16-30 days)	Late (31-120 days)
emp_length							
1 year	3.6%	79.2%	0.1%	13.2%	1.0%	0.7%	2.2%
10+ years	2.7%	80.3%	0.1%	13.6%	0.9%	0.5%	1.8%
2 years	3.3%	79.3%	0.1%	13.6%	1.0%	0.6%	2.1%
3 years	3.2%	79.7%	0.2%	13.2%	1.0%	0.5%	2.2%
4 years	3.4%	79.7%	0.1%	13.2%	1.0%	0.7%	1.9%
5 years	3.3%	79.8%	0.1%	13.1%	1.0%	0.7%	2.1%
6 years	3.4%	79.5%	0.2%	13.4%	0.9%	0.6%	2.1%
7 years	3.2%	79.3%	0.1%	13.7%	1.0%	0.7%	1.9%
8 years	3.0%	79.7%	0.2%	13.4%	0.9%	0.6%	2.2%
9 years	3.1%	79.7%	0.1%	13.4%	1.0%	0.7%	1.9%
< 1 year	3.6%	79.1%	0.1%	13.2%	1.1%	0.7%	2.2%
n/a	4.0%	81.2%	0.2%	10.9%	0.8%	0.5%	2.4%

Early Data Analysis

Home Ownership



Early Data Analysis

- Annual Income
  - Didn't find much
- Grade
  - grades D, E, F, G above norm for default and late payments
  - sub-grades – From C4 below the rates of default, and late payments rise

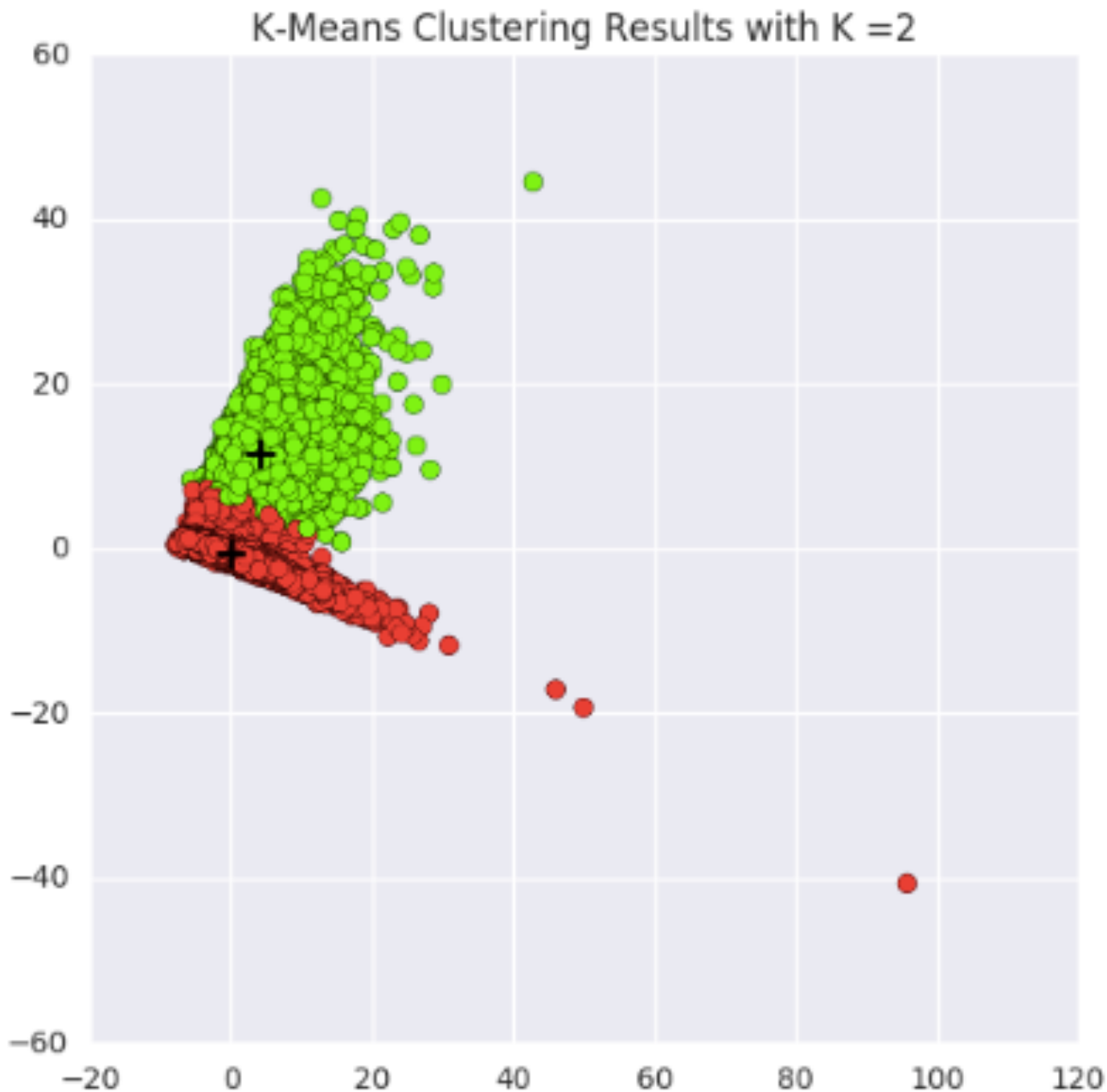
Grade and Subgrade Scores

loan_status	Charged Off	Current	Default	Fully Paid	In Grace Period	Late (16-30 days)	Late (31-120 days)
grade							
0	495	63032	28	9051	216	133	381
1	1766	98283	86	14898	767	439	1367
2	3729	95764	157	16259	1249	801	2608
3	3425	46797	133	8826	891	588	1994
4	2516	24954	88	4942	583	372	1493
5	1059	6256	34	1529	221	138	580
6	335	1181	13	400	47	44	147

Feature Selection and Model Building

- PCA
  - Why I began with Principal Component Analysis
  - Was it informative
  - Can I ascertain feature significance

- K – MEANS
  - Look at those boy's cluster!


K-Means Clustering Results with K =2
60
40
20
0
−20
−40
−60
−20
0
20
40
60
80
100
120

Feature Selection and Model Building

- LOGISTIC REGRESSION – LASSO PENALTY
  - Feature importance?
    - cut the following features 'last_fico_range_low, collection_12_mths_ex_med, debt_consolidation, num_actv_bc_tl
  - False positives and how I feel about them for my models.
  - Lasso – Least Absolute Shrinkage which involves penalizing the absolute size of the regression coefficients

```
Train/Test Metrics:
Accuracy:                0.9084
Precision:               0.3771
Recall:                  0.8385
False Positive Rate:     0.0872
Area Under ROC Curve:    0.9407
Area Under P-R Curve:    0.7413
```

	Predicted Good	Predicted Bad
Good	72319	6910
Bad	806	4184

	precision	recall	f1-score	support
0	0.91	0.99	0.95	73125
1	0.84	0.38	0.52	11094
avg / total	0.90	0.91	0.89	84219

Feature Selection and Model Building

- Grade and PCA models
  - Very low scores
  - The PCA produces a very high false positive rate
  - It appears the when combining the ‘grade’ and ‘sub-grade’ models to predict is not a good idea.

Feature Selection and Model Building

- Decision Tree
  - cross_val score of 0.945
  - Uncorrelated features? Yes and No

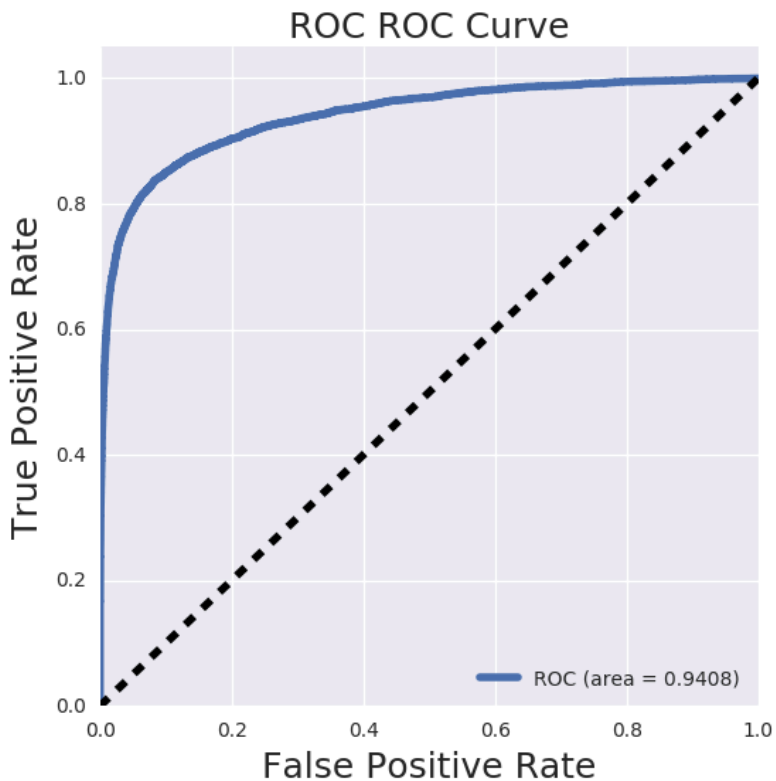
- BaggingTree/RandomForestClassifier
  - WTF!!!
  - using RandomForestClassifier adds boost and by adding ‘randomness’ to construction!

Feature Selection and Model Building

- What features were cut and why?
  - ‘gini’ less than 0.02

Feature Selection and Model Building

- Gridsearch Logistic Regression
  - Best parameters
  - cross_val score of 0.909



Train/Test Metrics:

Accuracy:	0.9081
Precision:	0.3766
Recall:	0.8401
False Positive Rate:	0.0876
Area Under ROC Curve:	0.9408
Area Under P-R Curve:	0.7409

	Predicted Good	Predicted Bad
Good	72291	6938
Bad	798	4192

	precision	recall	f1-score	support
0	0.91	0.99	0.95	73089
1	0.84	0.38	0.52	11130
avg / total	0.90	0.91	0.89	84219

Model!

	Accuracy	Area Under P-R Curve	Area Under Roc Curve	False Positve Rate	Precision	Recall	Cross-Val Score
Decision Tree Model	0.944917	0.549072	0.753088	0.029295	0.535149	0.535471	.945
Bagging Tree ModeId	0.951685	0.700043	0.917650	0.000114	0.990415	0.186373	.956
Lasso Model	0.908382	0.741254	0.940700	0.087216	0.377141	0.838477	.909
PCA Model	0.476377	0.072388	0.549929	0.531404	0.067242	0.597792	.479
LendingClub.com Model	0.653392	0.125441	0.705493	0.345846	0.104572	0.641283	.653
GridsearchLog	0.908144	0.740940	0.940827	0.087569	0.376640	0.840080	.909

- My best model
- Recall
- What does it mean to be True and predicted true
- Precision
- Predictive power
- False Positives
- How often does my model give us incorrect positive results? Good or bad?


I CAN'T
EVEN