DIGITAL MOSAIC

Prosper EDA & Sample Model Fit:

Overview of Data

Time Series Completed Loans:
1/2006 – 3/2014

Mean Estimated Return:
9.2%

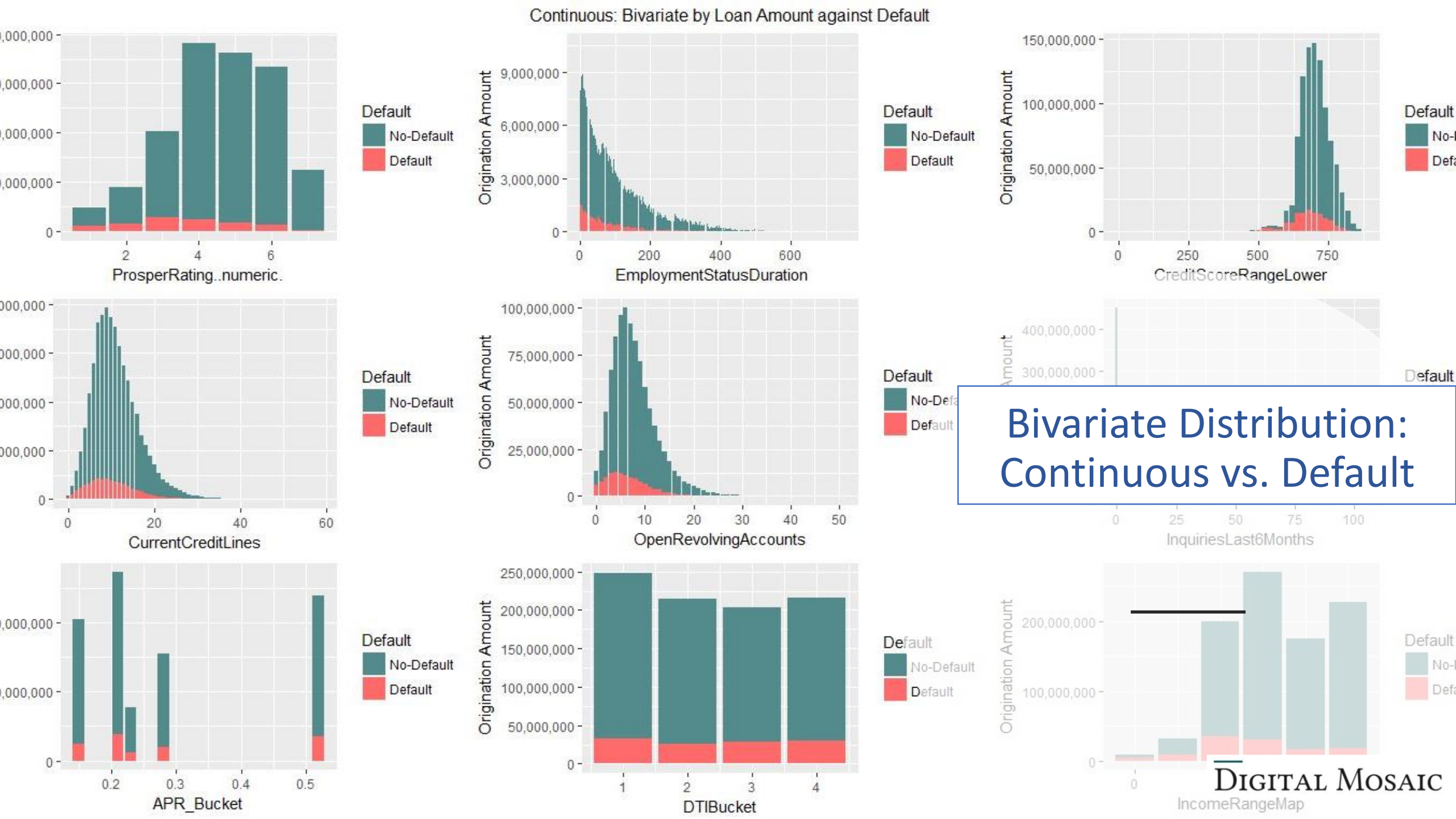
Default/Delinquency Rate:
16.93%

Sample Size:
~114,000 Loans

Filter	Visualize variable relationship to default[1] through bivariate distributions to make first selection of variables for model.
Correlations	Run correlations against other variables and against default to refine feature list.
Model	Run selected features through models to predict default.

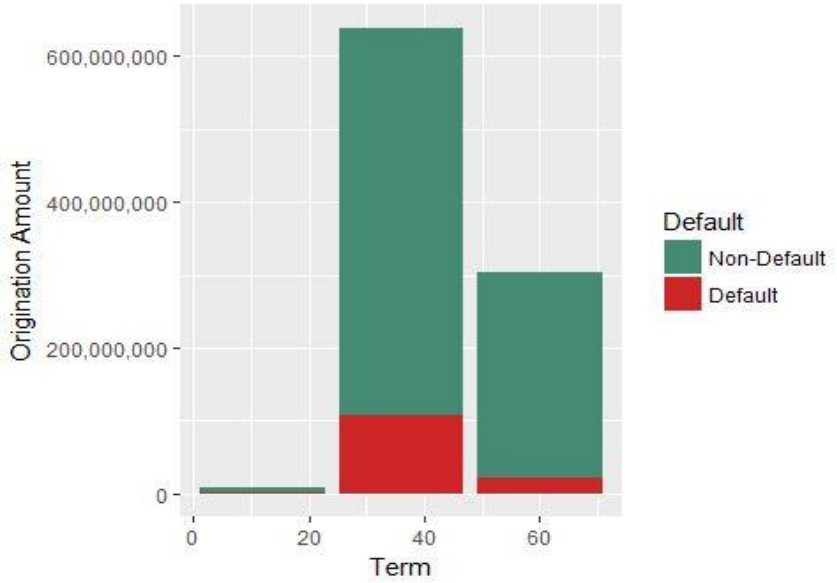
Exploration and Model Fit Overview

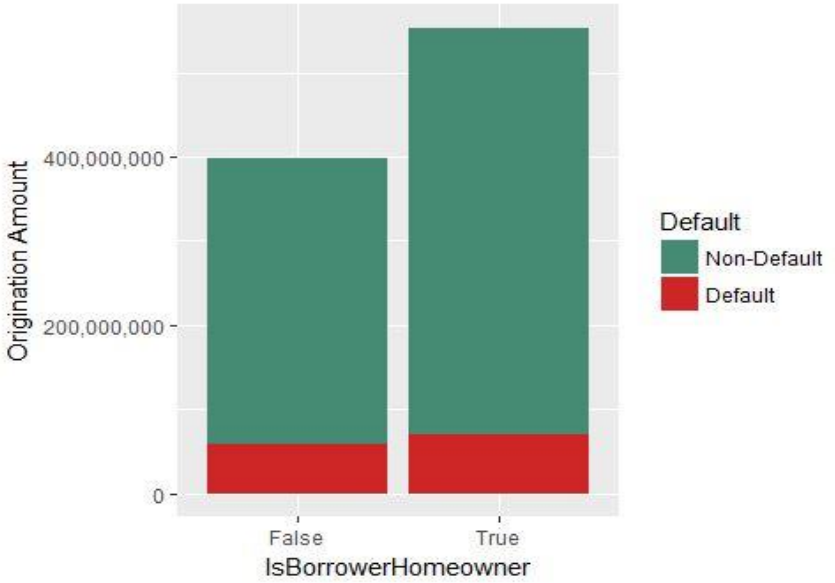
Default defined as any loan status that is **not** "Current," "Completed," or "Final Payment in Progress."

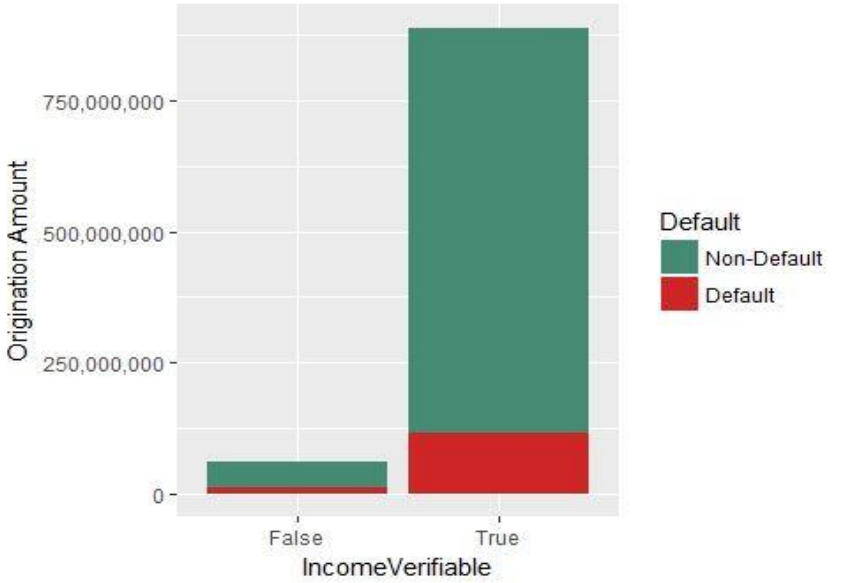


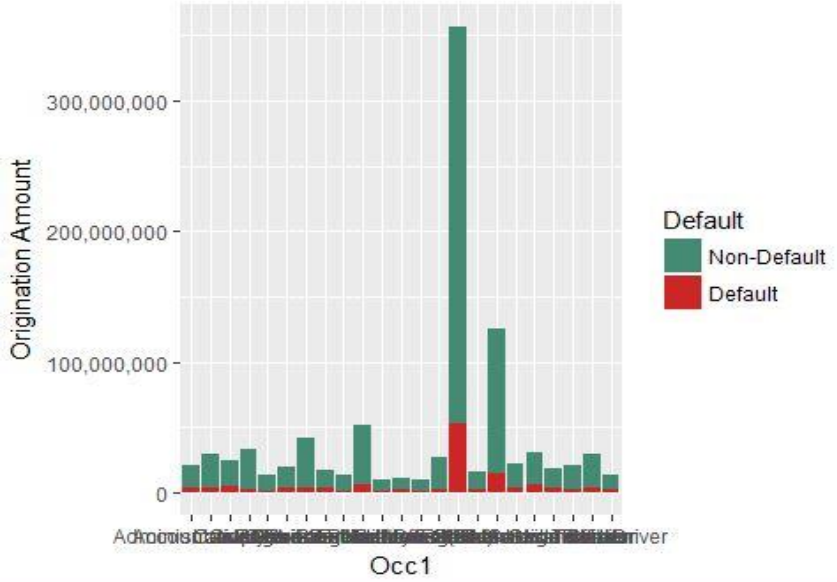
Bivariate Distribution: Continuous vs. Default

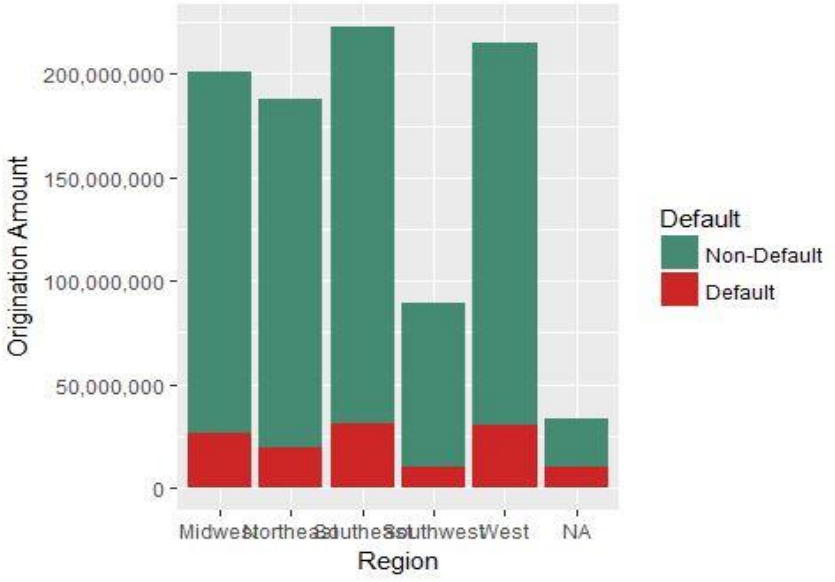
DIGITAL MOSAIC

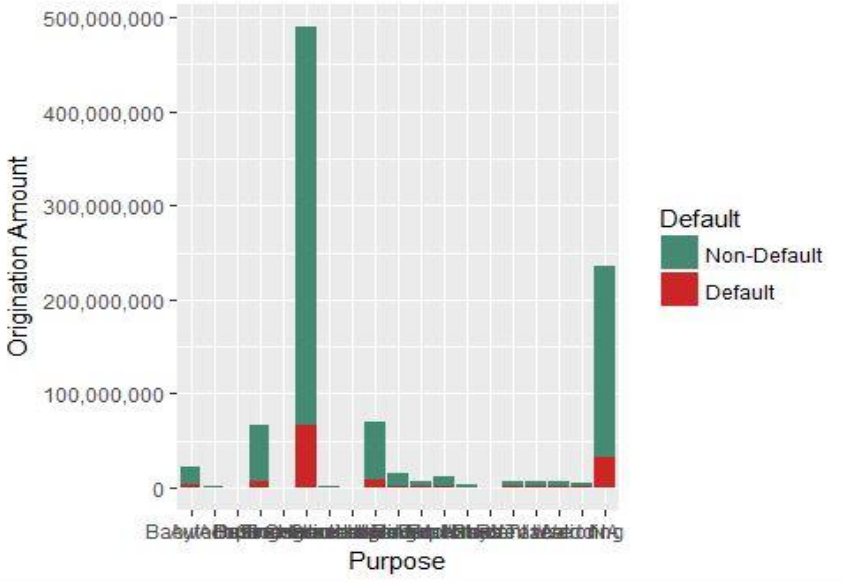












Categorical Distribution vs. Default

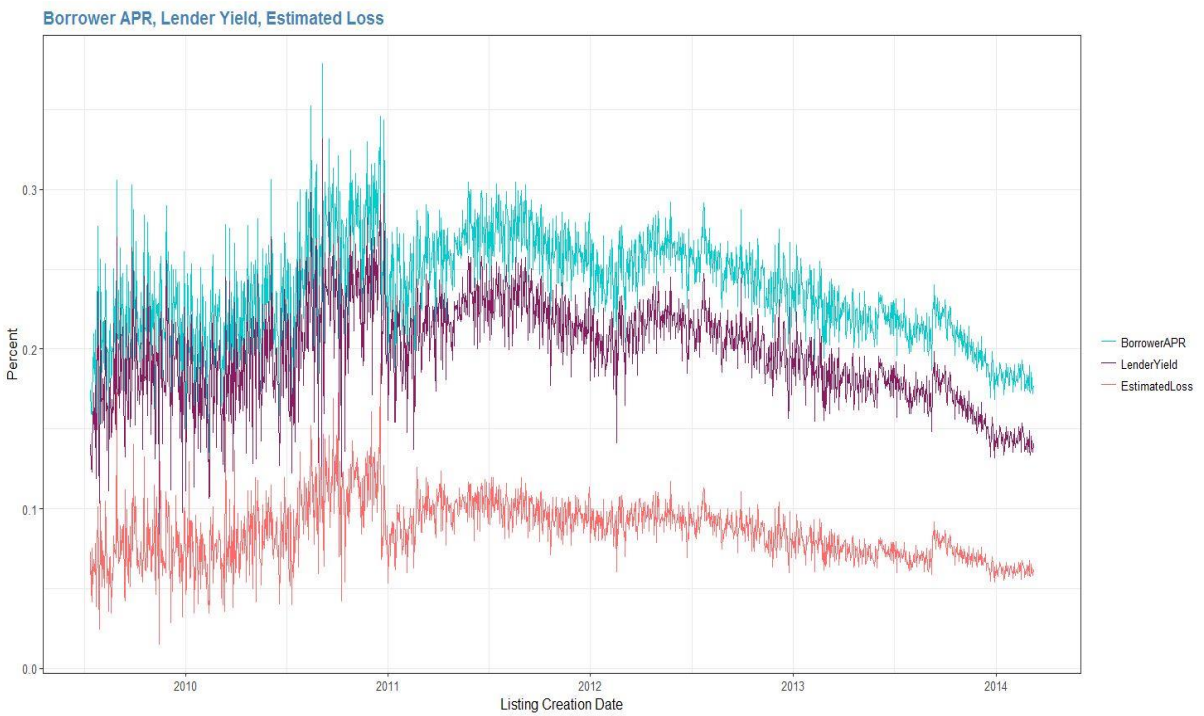
DIGITAL MOSAIC

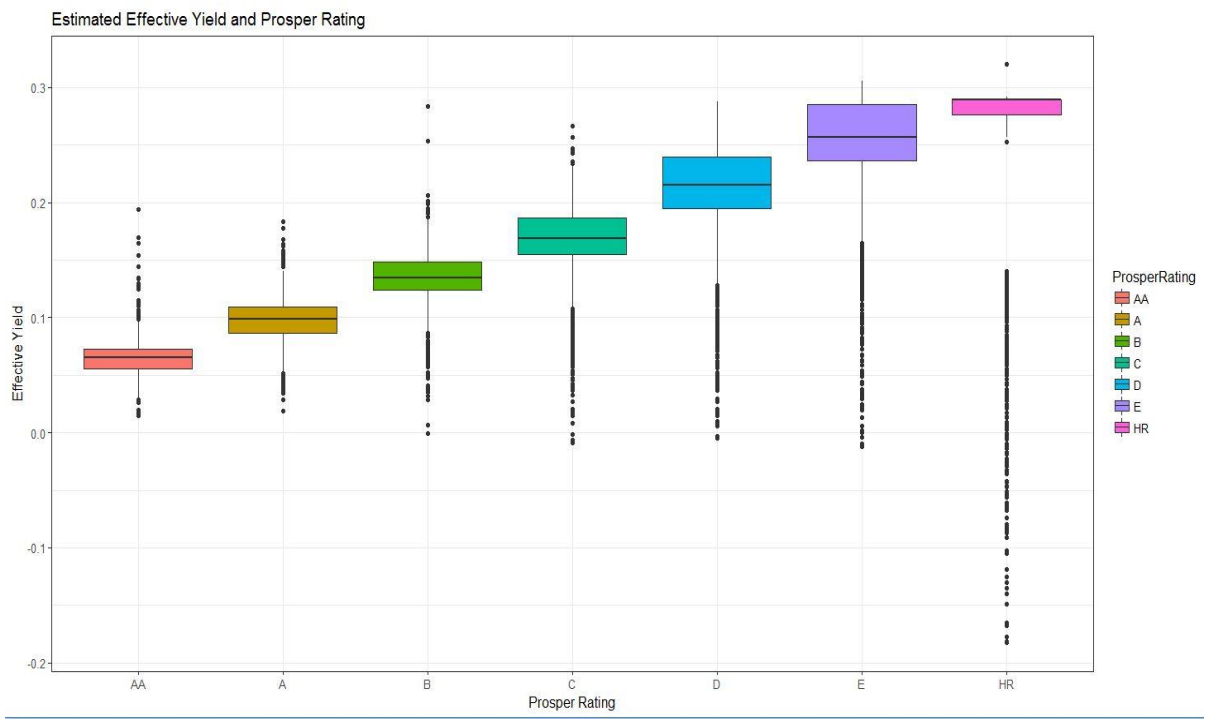
Examining the distributions

```
  value mean_default count  mean
2    36   0.19584634 87778 19.6%
3    60   0.08123854 24545  8.1%
1    12   0.06319703  1614  6.3%
[1] "EmploymentStatus"
          value mean_default count  mean
4 Not available   0.42453712  5347 42.5%
1                 0.36186253  2255 36.2%
8       Retired   0.30691824   795 30.7%
5  Not employed   0.29101796   835 29.1%
3     Full-time   0.28996395 26355 29.0%
7     Part-time   0.24816176  1088 24.8%
9 Self-employed   0.20639061  6134 20.6%
6         Other   0.12480294  3806 12.5%
2      Employed   0.09003001 67322  9.0%
[1] "IsBorrowerHomeowner"
  value mean_default count   mean
1 False    0.1871624 56459 18.72%
2  True    0.1517102 57478 15.17%
[1] "IncomeVerifiable"
  value mean_default  count   mean
1 False    0.2012920   8669 20.13%
2  True    0.1666413 105268 16.66%
[1] "Occ1"
                           value mean_default count  mean
6                       Clerical    0.2477876  3164 24.8%
1                                   0.2318841  3588 23.2%
19             Sales - Commission    0.2231573  3446 22.3%
12                        Laborer    0.2169279  1595 21.7%
20                Sales - Retail    0.2091527  2797 20.9%
3          Administrative Assistant    0.2060738  3688 20.6%
```

Continuous	Categorical
Borrower APR	Employment Status
Credit Range Score : Lower	Income Verified
Current Credit Lines	isHomeowner
Debt to Income	Occupation
Delinquencies - 7 years	Purpose
Delinquencies - Current	Region
Employment Duration	Term
Income	
Inquiries Last 6 Months	
Investment from Friends	
Number of Investors	
Open Revolving Accounts	
Percent Funded	
Prosper Rating	
Public Inquiries	

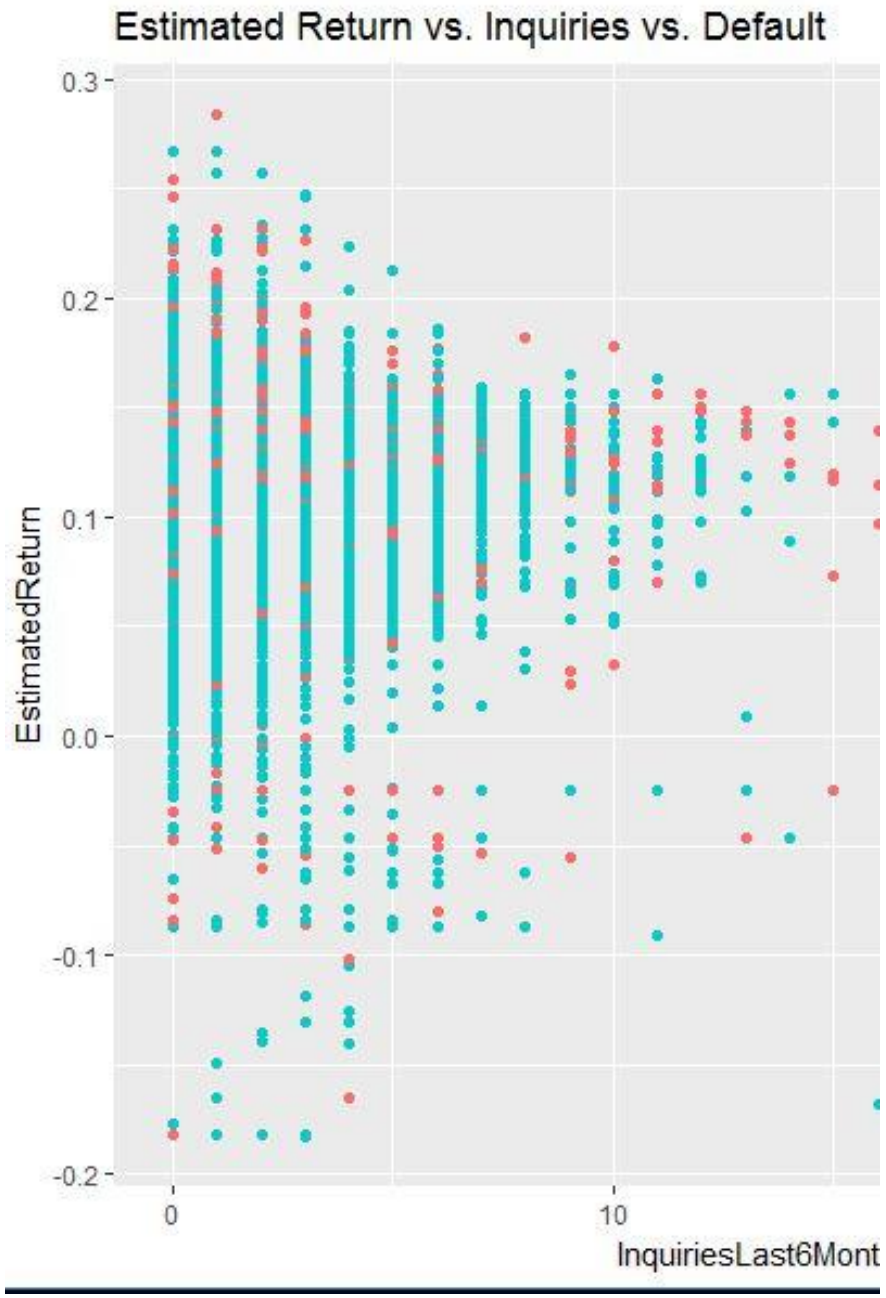
- Create table of defaults
- Select preliminary set of features
  - Split into Continuous and Categorical (Nominal) variables
  - Select based on table & bivariate distributions

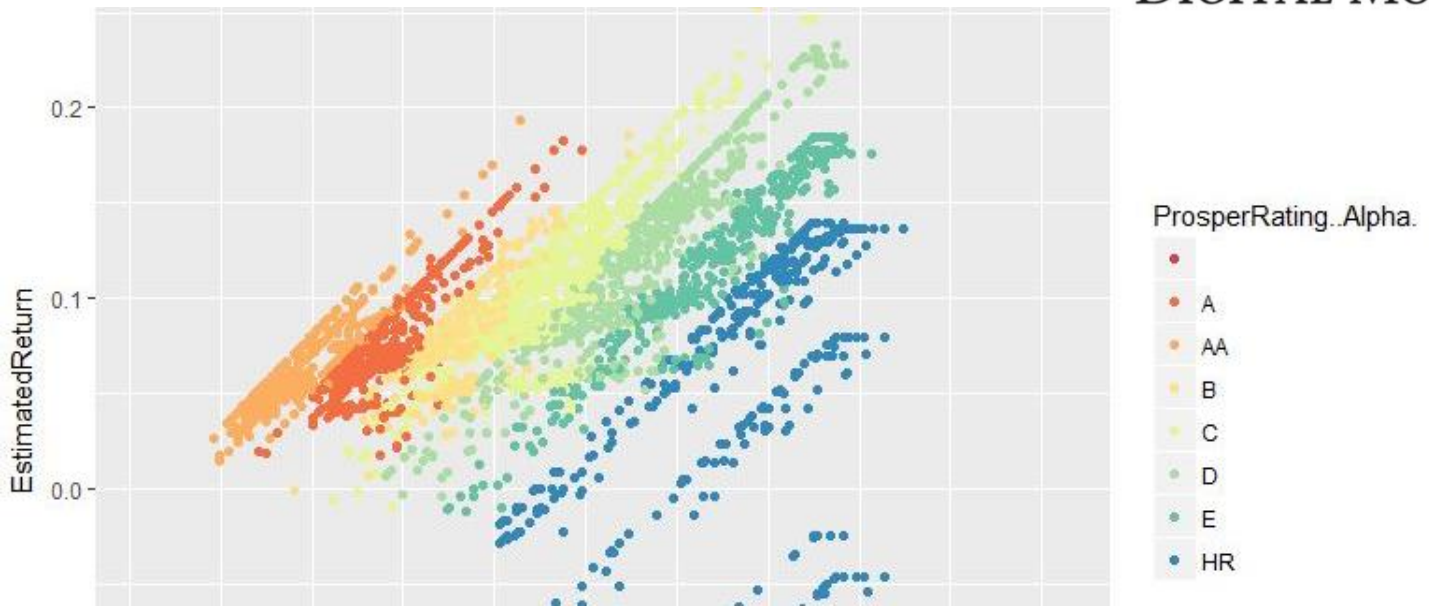




Yield Analysis I:

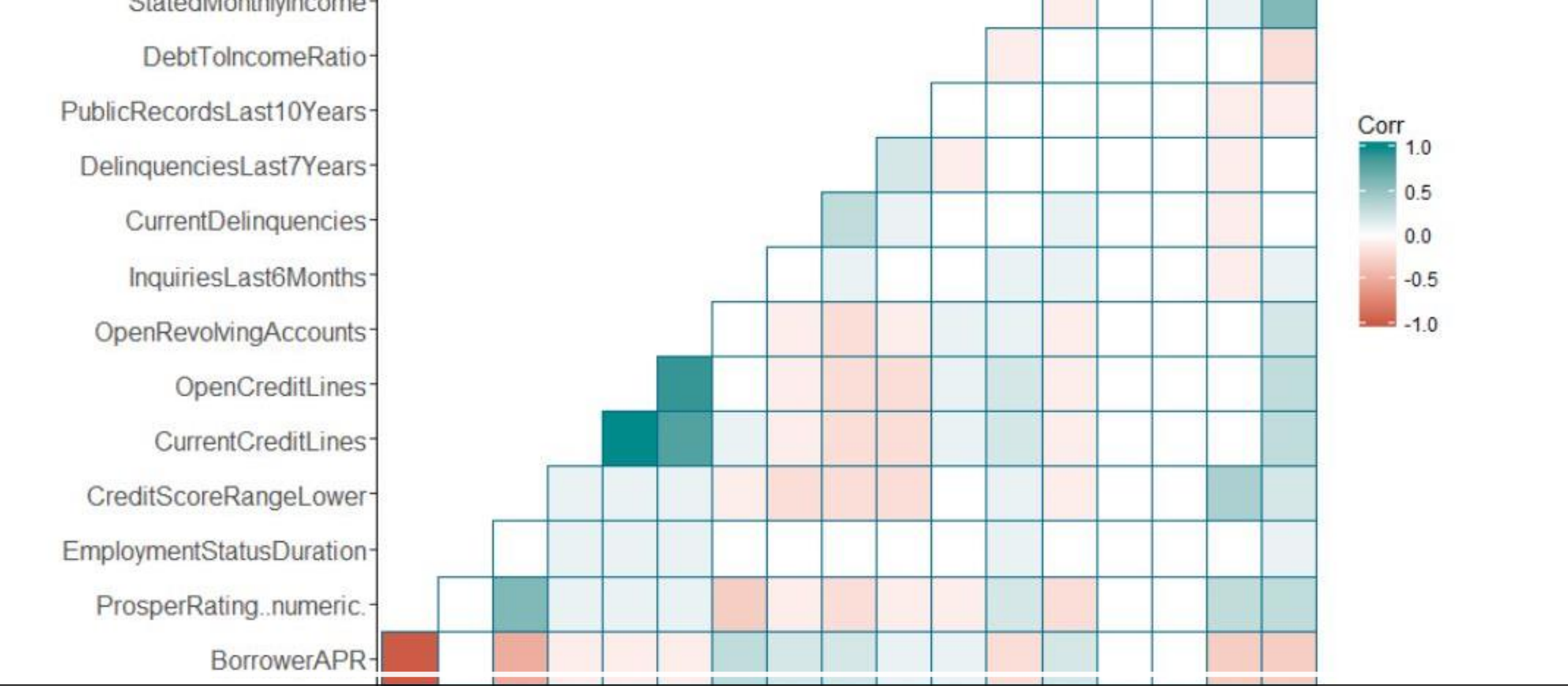
- Stabilizing returns post-2011
- Higher returns with lower rating, but much higher volatility





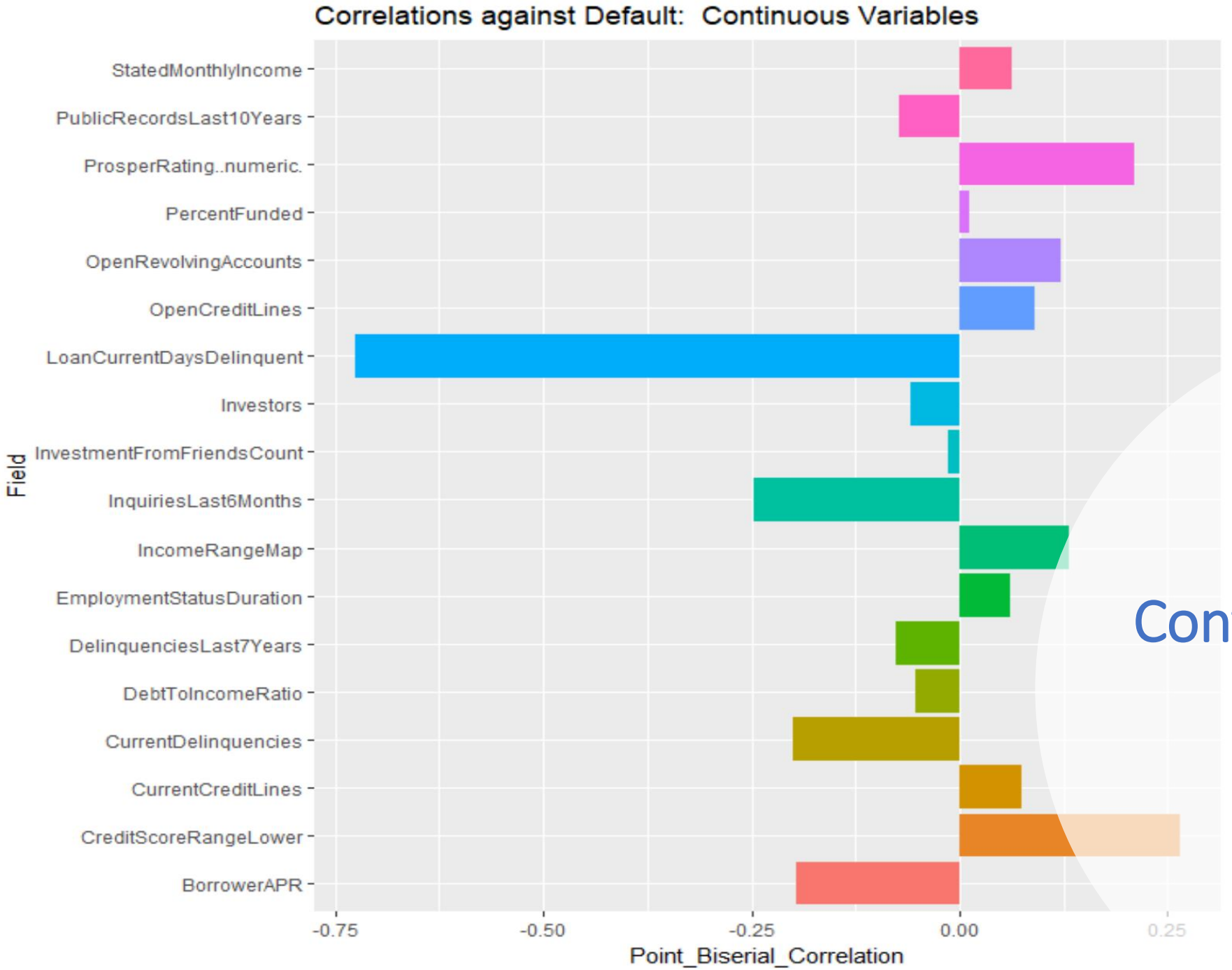
Yield Analysis II: Multivariate Examples

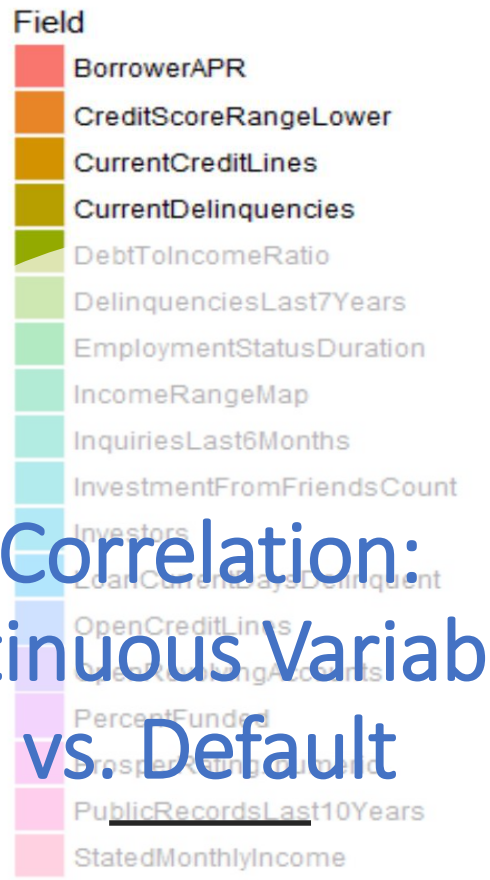
- Defaults increase with inquiries, but often still produce high returns.
- Highest loss-adjusted returns for B-E rated notes.



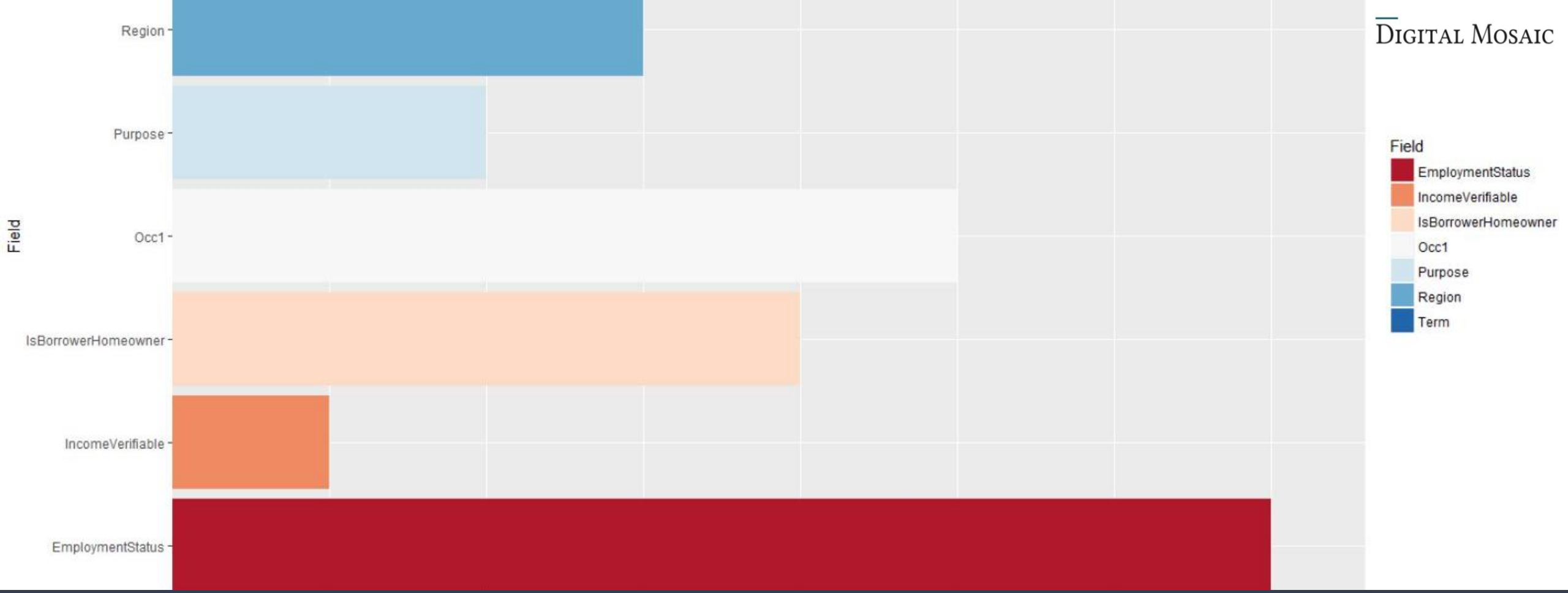
Correlation Matrix: Continuous Variables

DIGITAL MOSAIC





Correlation: Continuous Variables vs. Default



Correlation: Categorical Variables vs. Default

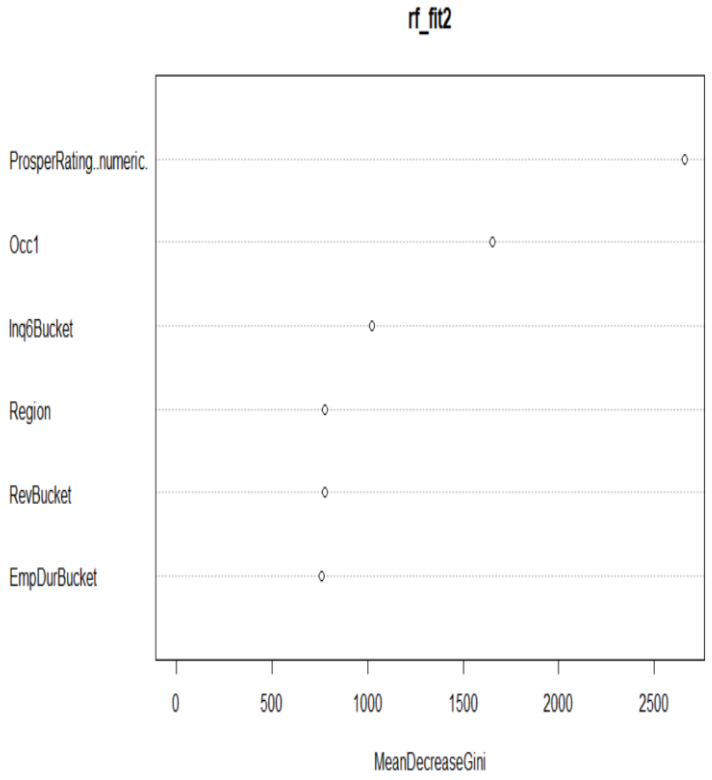
Correlation: Refining the Feature Set

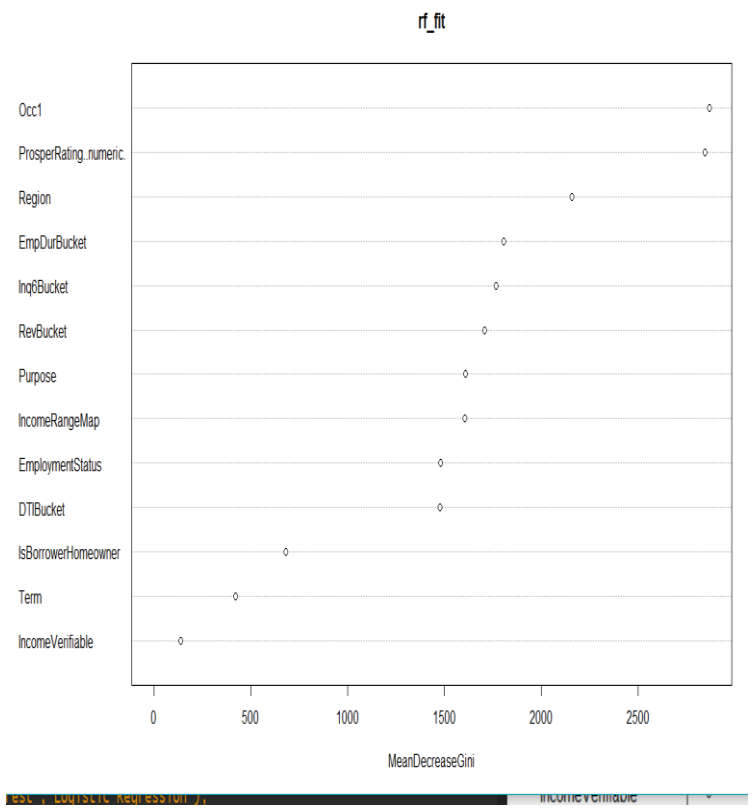
Default Correlation

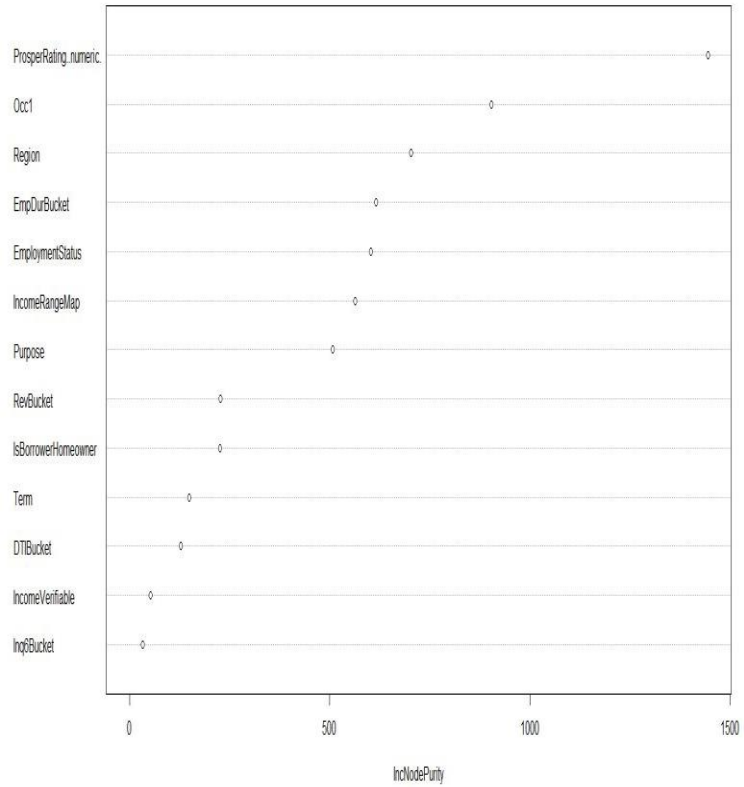
- Measures
  - Continuous Variables: Point Biserial
  - Categorical: Cramer's V (Chi-Square)
- Refine feature set by removing
  - Low correlation
  - Unclear relation to returns

Correlation Matrix

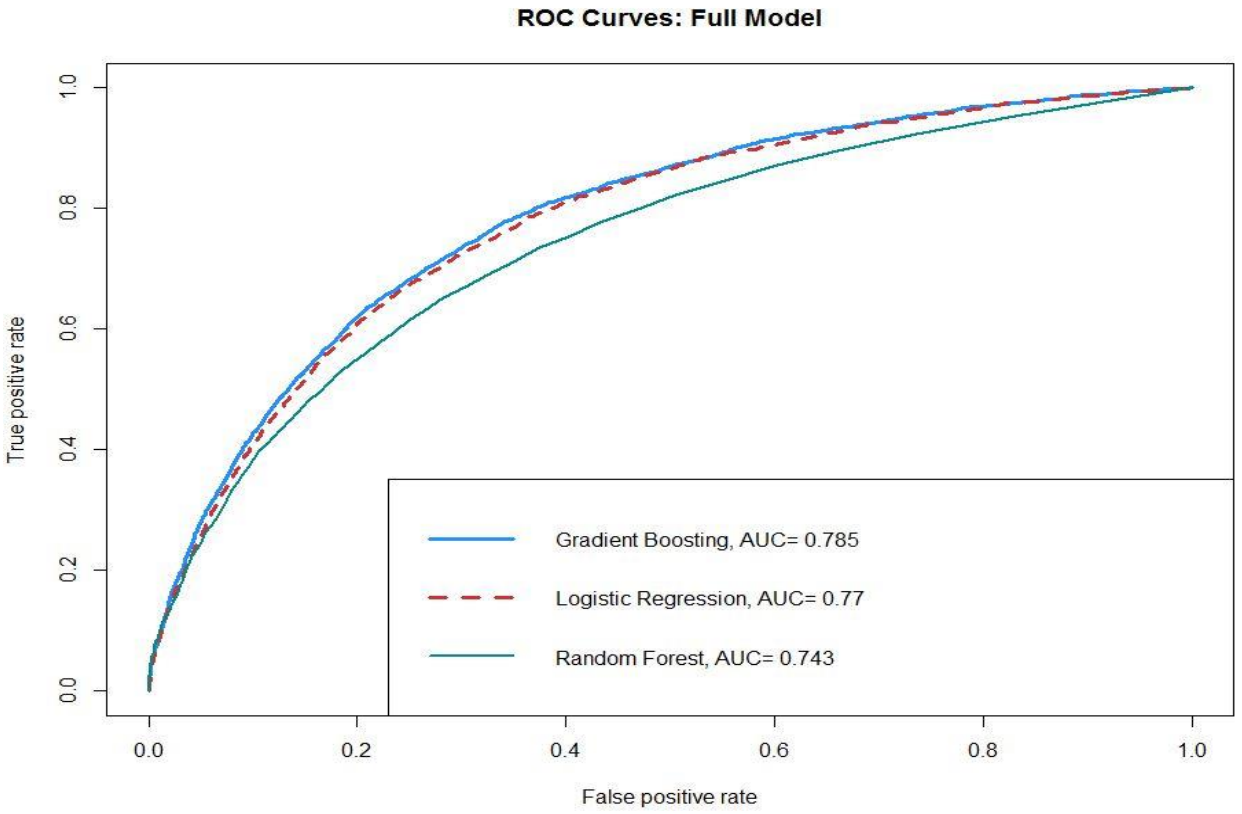
- Examine correlation between variables
- Remove features that are redundant.
  - Credit Lines (high correlation with Revolving Accounts)
  - Credit Range Lower / APR (Prosper Rating)

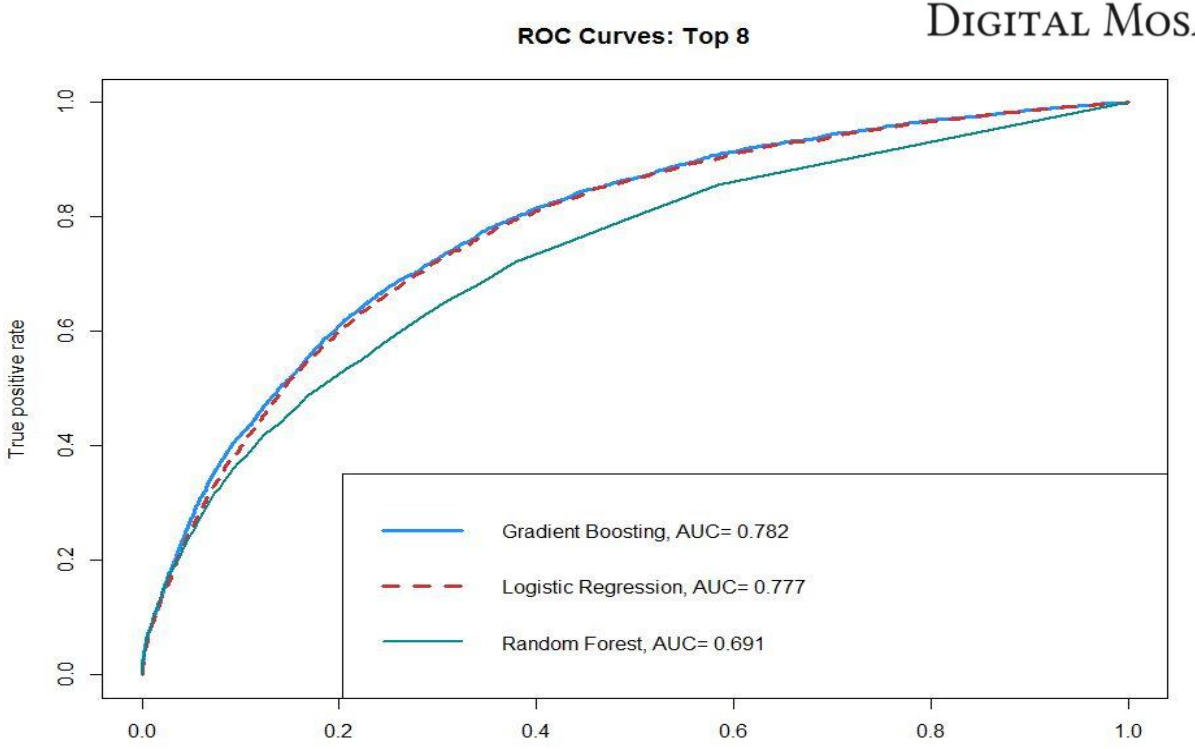






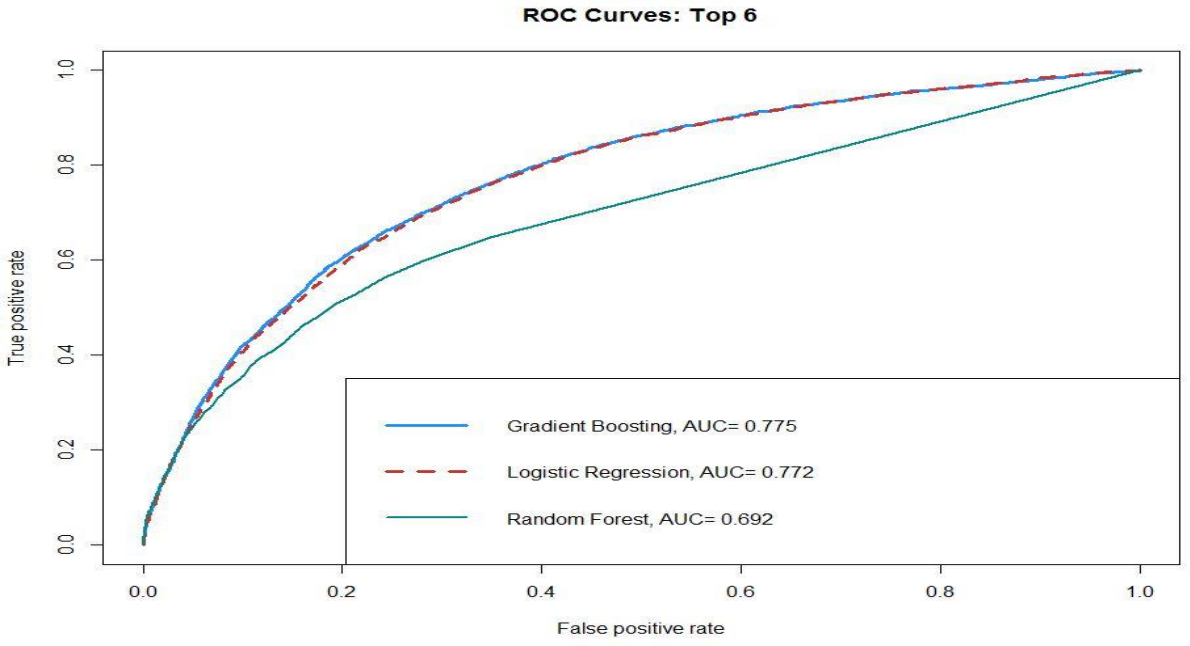
Variable Importance: Different seed & Feature Sets

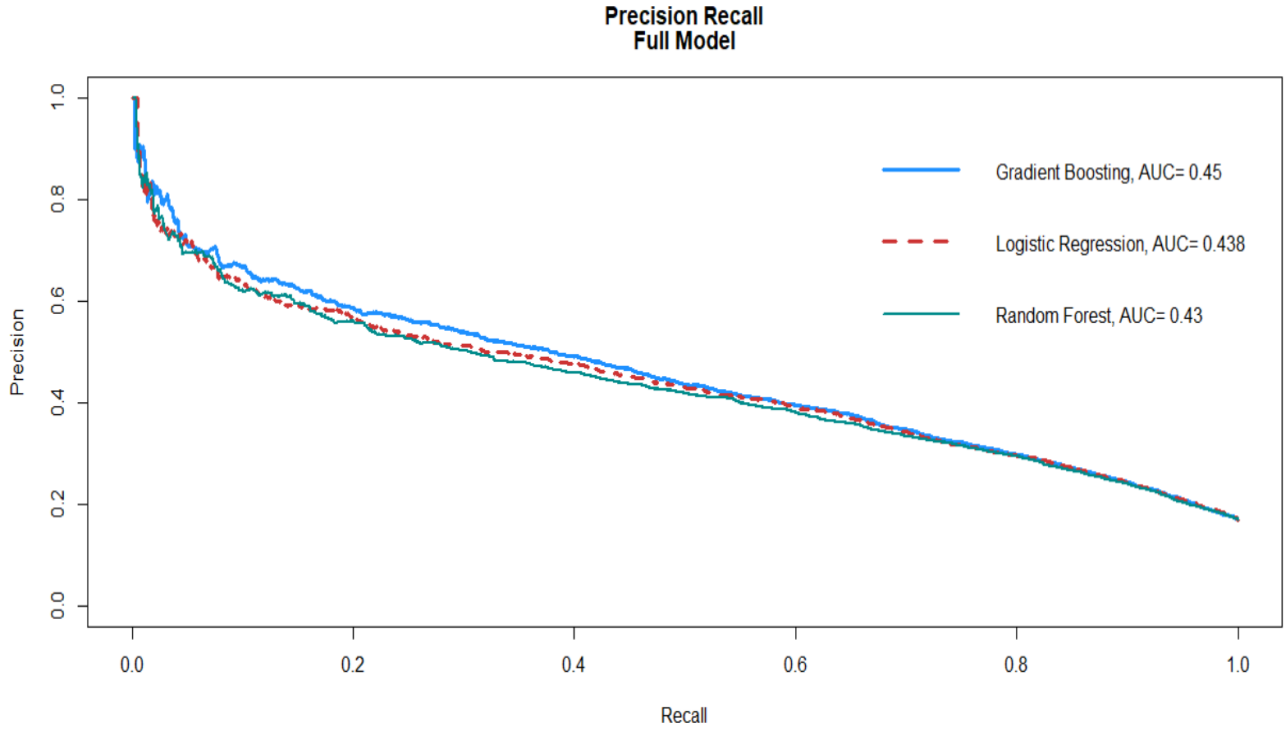


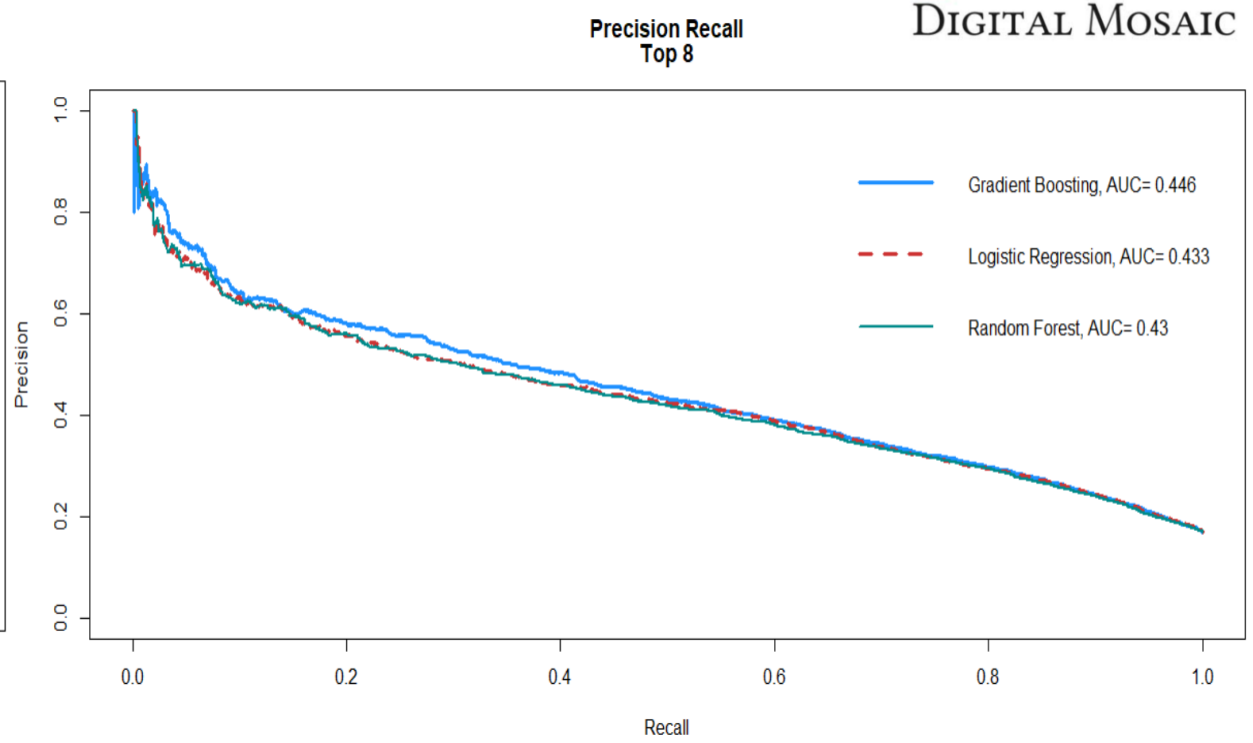


Model Fit & Comparison

- Ensemble vs. Logistic Regression
- Compare different seeds for randomizesd trees and change feature mix.
- Random Forest more sensitive to # of features
- Outperformance from Gradient Boosting







Adjusting for Default Frequency

- Defaults are low frequency events.
- We use Precision-recall to isolate only default outcome and predictions.
- Gradient Boosting still outperforms Random Forest and Logistic Regression, but the difference is smaller.