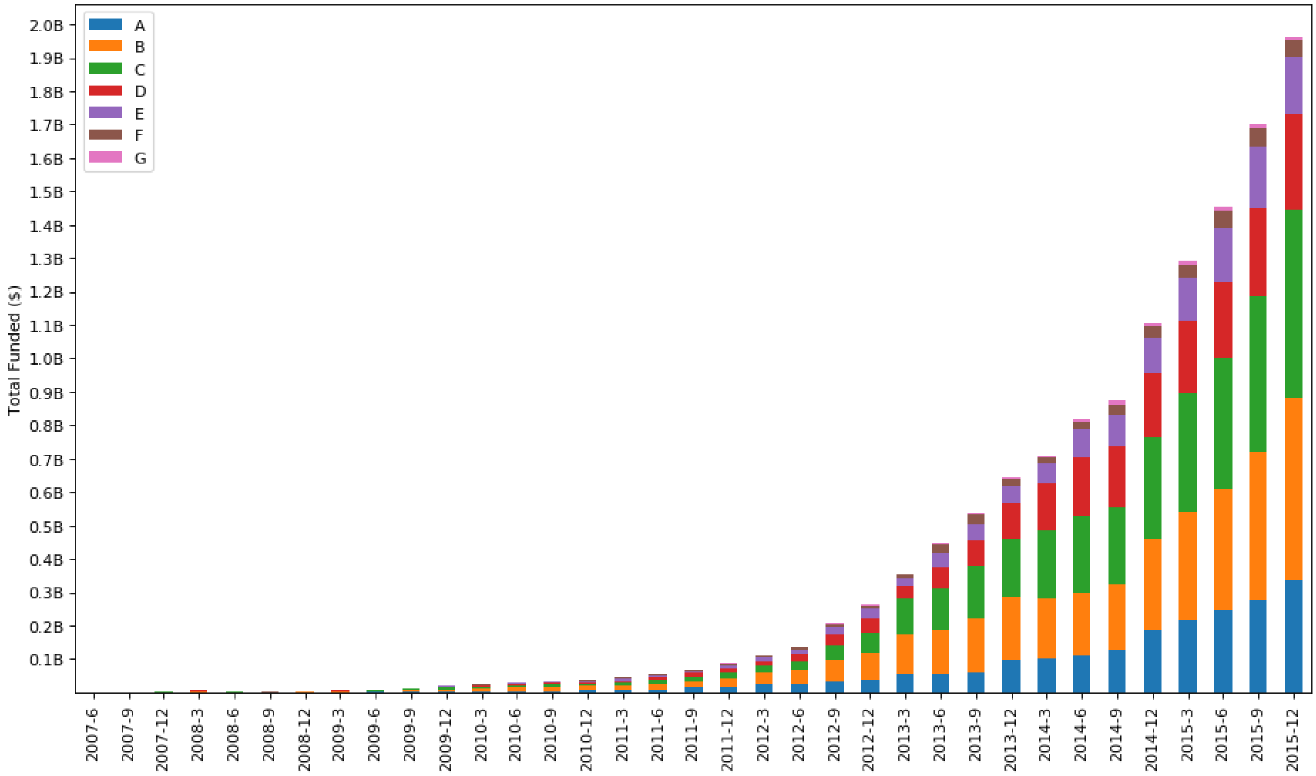
Bx Data Take Home Assignment
Peter Will

Part 1 – Cleaning and Exploration

The total dollar amount of loans issued by Lending Club has increased exponentially with time, and the composition by loan-grade stays more or less steady.

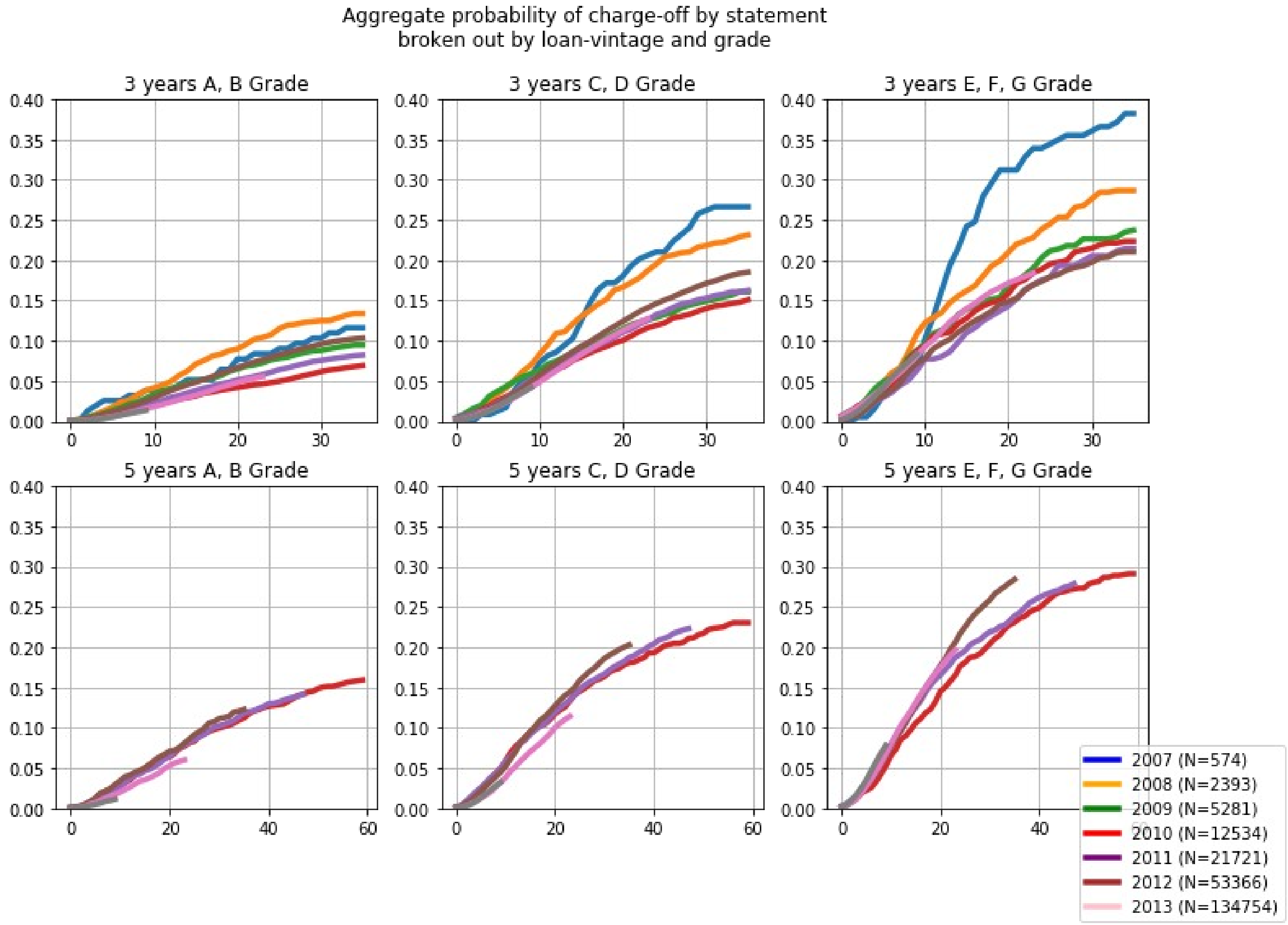


Overall, there are about 3 times as many active loans as there are closed loans in the dataset, among closed loans about 4 loans are closed satisfactorily for each one that defaults, although this changes significantly by year.

Overall Counts	
Open	631141
closed_paid	209682
closed_bad	46006

category	closed_bad	closed_paid	open
issue_y			
2007	27.00%	73.00%	0.00%
2008	20.73%	79.27%	0.00%
2009	13.69%	86.31%	0.00%
2010	13.99%	85.90%	0.11%
2011	14.76%	75.83%	9.42%
2012	15.15%	77.73%	7.13%
2013	11.01%	41.85%	47.14%
2014	5.93%	23.22%	70.85%
2015	0.66%	5.46%	93.88%

There were significant differences in the expected default-rate for each vintage and loan-type, this is especially pronounced for 2007 and 2008 vintages and is a likely result of the financial recession.



Part 2 – Business Analysis

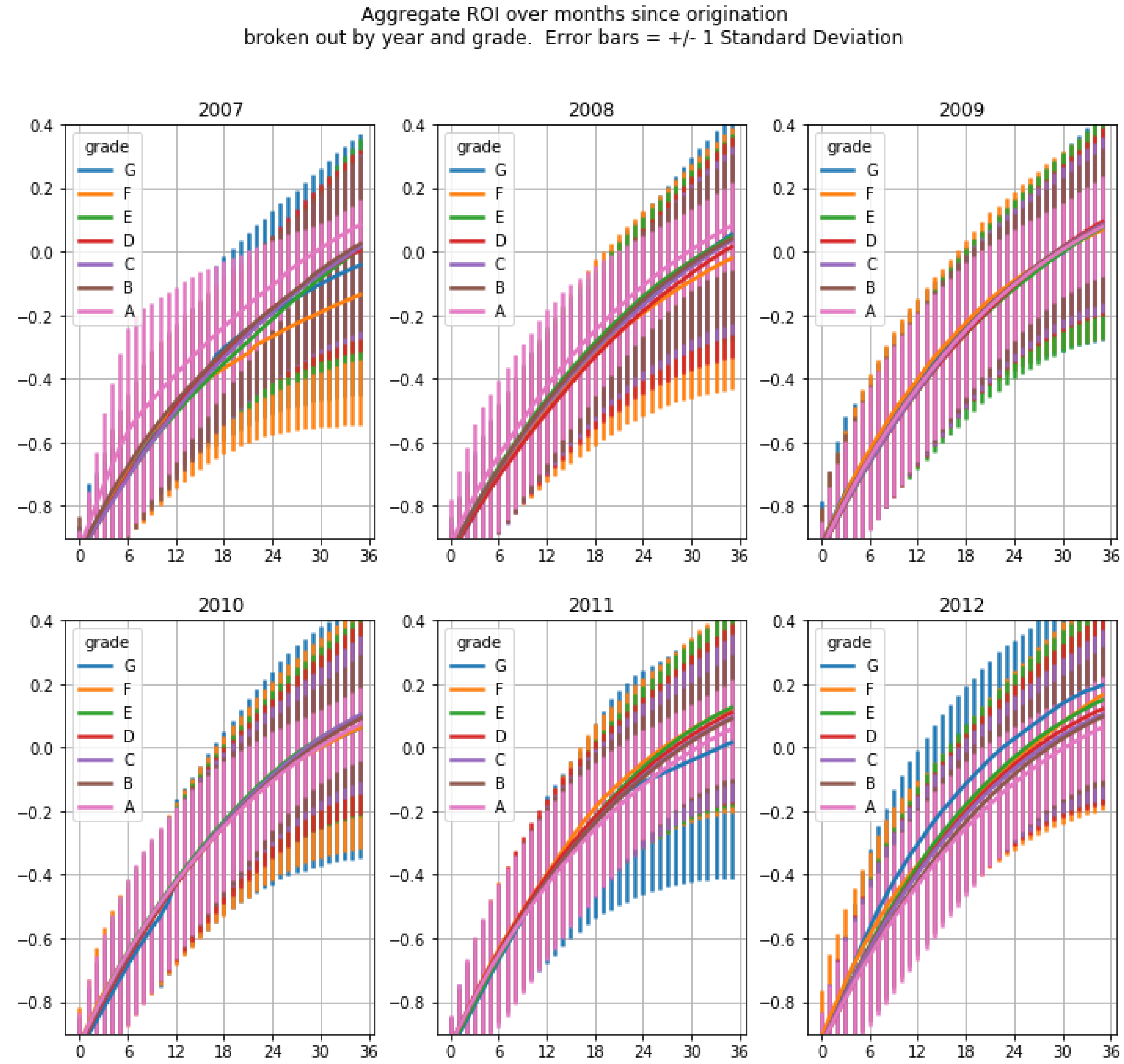
My approach to determining the ROI of investing in Lending Club loans was to construct payment histories for each loan, and then to compare the sum of the payment-flow at points in time to the upfront cost of funding the loan. I believe this approach ultimately presents a cleaner picture of loan performance. It provides a basis where ROI is compared across regular time-intervals -- an important quality because the loans exhibit different payment behavior and duration. Additionally, I found that it handled edge-cases such as early prepayment or continued servicing of loans after their intended maturity dates better than other ROI calculation methods.

I did make a number of simplifying assumptions in constructing the payment histories.

- Borrowers who repaid the loans early were assumed to do so in equal allotments.

- Collections for loans that had defaulted were assumed to be disbursed 5-months after the last payment date (this aligns with information found on Lending Club’s FAQ) .
- This is not a NPV analysis, there is no discounting of future payments
- The ROI calculated assumes that loans can not be re-sold.
- The ROI calculated ignores any fees Lending Club may impose on the lender.
- Finally, in the results that follow I focused exclusively on loans issued before 2013. The dataset contains data current to 2016-01-01, thus loans issued after 2013 didn’t have sufficient payment history to determine returns.

A Chart of the average ROI for payment periods 1 through 36 is shown below for 36 month term loans originated between 2007-01-01 and 2012-12-31.



Aggregate ROI over months since origination
broken out by year and grade. Error bars = +/- 1 Standard Deviation

2007 2008 2009 2010 2011 2012

Error bars are shown in the chart above reflecting the standard deviation of the ROI for each population. Two general trends are apparent. First, there’s much less volatility for higher-grade loans, a reflection of their relatively low-probability of default. Second, from 2007 to 2012 there’s a general trend wherein the ROI of more risky loans surpasses that of safer loans. This is likely due to the effects of the financial recession and it’s macro-level impact on the lending environment.

The tables below summarizes this information and presents overall expected returns and volatility, along with the standard error on these statistics.

Table 1: Expected ROI over 3 years (5 years) for 36 (60) month loans

		2007	2008	2009	2010	2011	2012	AVG	+/-
grade	**term**								
A	**36**	8.58	8.16	7.91	7.39	5.9	6.35	**7.38**	**0.96**
B	**36**	2.56	4.04	8.14	9.21	9.11	9.76	**7.14**	**2.79**
C	**36**	1.65	3.57	8.61	10.43	9.47	10.44	**7.36**	**3.46**
D	**36**	0.53	1.58	9.51	9.33	11.03	12.18	**7.36**	**4.57**
E	**36**	0.99	4.74	7.66	9.97	12.59	14.95	**8.48**	**4.69**
F	**36**	-13.44	-2.02	6.93	6.24	12.29	16.59	**4.43**	**9.83**
G	**36**	-4.06	5.55	7.66	6.21	1.65	19.76	**6.13**	**7.22**
A	**60**	NaN	NaN	NaN	10.45	7.17	5.14	**7.59**	**2.19**
B	**60**	NaN	NaN	NaN	11.66	5.53	1.9	**6.36**	**4.03**
C	**60**	NaN	NaN	NaN	14.01	5.18	1.66	**6.95**	**5.19**
D	**60**	NaN	NaN	NaN	15.51	4.6	-0.39	**6.57**	**6.64**
E	**60**	NaN	NaN	NaN	16.72	7.16	-0.58	**7.77**	**7.08**
F	**60**	NaN	NaN	NaN	18.43	5.64	0.92	**8.33**	**7.4**
G	**60**	NaN	NaN	NaN	18.95	5.59	-3.36	**7.06**	**9.17**

Table 2: Expected Volatility over 3 years (5 years) for 36 (60) month loans

		2007	2008	2009	2010	2011	2012	AVG	+/-
grade	**term**								
A	**36**	8.01	13.67	15.87	11.69	15.34	16.29	**13.48**	**2.89**
B	**36**	27.71	26.56	24.38	19.99	20.36	21.87	**23.48**	**2.96**
C	**36**	28.91	29.72	27.77	24.95	26.43	26.73	**27.42**	**1.59**
D	**36**	31.94	34.84	29.47	30.18	28.34	30.45	**30.87**	**2.08**
E	**36**	34.76	32.51	35	31.68	31.26	32.37	**32.93**	**1.44**
F	**36**	41.14	41.29	34.61	38.06	32.58	36.31	**37.33**	**3.21**
G	**36**	41.2	38.35	35.8	40.98	42.8	36.58	**39.28**	**2.56**
A	**60**	NaN	NaN	NaN	18.43	21.3	22.06	**20.6**	**1.56**
B	**60**	NaN	NaN	NaN	30.05	30.94	32.84	**31.28**	**1.16**
C	**60**	NaN	NaN	NaN	35.24	37.88	37.33	**36.82**	**1.14**
D	**60**	NaN	NaN	NaN	39.09	41.38	41.56	**40.68**	**1.12**
E	**60**	NaN	NaN	NaN	40.99	43.44	42.89	**42.44**	**1.05**
F	**60**	NaN	NaN	NaN	40.11	47.03	44.95	**44.03**	**2.9**
G	**60**	NaN	NaN	NaN	47.56	49.36	46.7	**47.87**	**1.11**

In presenting these final results it’s important to acknowledge that they are averages of the average statistics observed for each year. In other words, the population of loans originated in 2007 is given equal weight to the population of loans originated in 2012, despite the fact that the number of loans originated increased exponentially with time. Consequently, the final averages displayed are heavily impacted by the relatively bad performance of 2007 and 2008 vintages. I favor this approach as it underscores temporal differences in loan performance. Averaging ROI without stratification by year would produce misleadingly high results, and overweight the 2011 and 2012 vintages.

My overall recommendation is that lending club hasn't been operating long enough to generate enough payment history data to be confident about expected returns. For the dataset available, it's only possible to get full payment histories for 36-month term loans issued between 2007 and 2013. **In this time period, 36-month E-grade loans had the highest average return at 8.48% over 3 years, with an average volatility of 32.93%** While these results are certainly underwhelming compared to other investment areas, it's worth noting that loan performance generally trended upwards with loan-vintage. Thus, there's reason to think that future returns will be higher than the averages displayed here. Additionally, none of the 5-year loans included in this analysis

Part 3 – Modeling

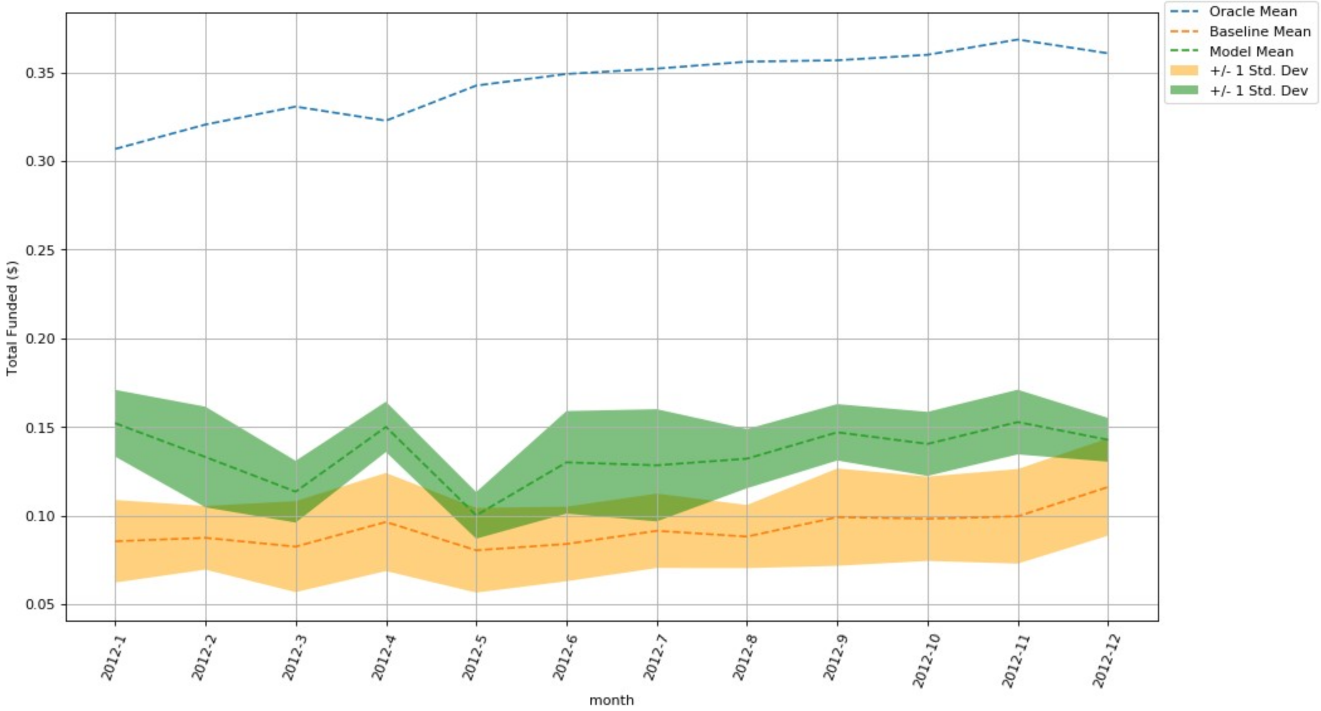
The first step in any modeling task is to first formulate a well-posed definition of the problem. To that end, the modeling problem I set out to solve can be described as follows:

- Assume access to a set of historical loans for which the full payment history is known. Because, of the relative scarcity of 60-month term loans in the data set and their long duration, I decided to focus exclusively on 36-month term loans.
- The quantity of interest is the 36-month ROI obtained by buying loans at issuance and holding them for 36 months. The dataset contains loans issued between 2007-06-01 and 2016-01-01. Since 36 months of payment history are needed to determine the target variable, only loans issued before 2013-01-01 were used for modeling.
- Because training involves using data that would not be available until 36 months after loan-issuance, **the test set of loans must originate 36 months after the issuance of any loan included in the training set**. With this in mind, I decided that my test-set would be composed of loans issued between 2012-01-01, 2013-01-01.
- Modeling performance is evaluated via a sequence of simulations over 1-month increments. Over the first time period, the model is trained on loans issued between 2007-06-01 and 2009-01-01. It is used to predict the 36-month ROI of loans originating between 2012-01-01 and 2012-02-01. For the second time period, these date ranges are incremented by 1 month, for the third time period by 2 months, and so forth. The final time period uses loans issued between 2007-06-01 and 2009-12-01 to predict ROI of loans originating between 2012-12-01 and 2013-01-01.
- The simulations can be thought of as a discrete series of points in time, where the model makes predictions as to what loans a lender should invest in given all available historical data up until that point in time. To gauge model performance, the simulations assume that a lender has a certain amount of money that he must invest each month. In practical terms, each simulation rank orders the loans by predicted ROI, and the lender then selects the top loans that fit within his monthly budget.
- The ROI of the portfolio of loans selected by the model is then compared to two other hypothetical portfolios. The first is a so-called "oracle" in which the loans with the highest actual ROIs are selected. This represents an ideal investment strategy with perfect knowledge and establishes an upper bound for model performance. The second is a so-called "baseline" portfolio in which loans are randomly selected randomly. This represents a benchmark that we hope to see the modeled-portfolio out-perform. All portfolios use the same monthly budget.

The final results of my model are presented in the chart below for a monthly investment budget of 1 million dollars. Here, the dashed blue line represents the oracle's results. Roughly, it achieves a 36-month ROI of around 35%, or close to 12% a year. The dashed orange line shows what you could expect from randomly selecting from loans issued in 2012. It returns close to 9% on average, or a yearly ROI of about 3% a year. The final model I used outperformed the baseline by about 4.2% percent on average, translating into an increased yearly ROI of roughly 4.1%.

Since both the model and the baseline incorporated random elements into their results, I ran the full simulated back test 10 times to produce one-standard deviation confidence intervals. These are shown as the shaded regions in the graph. As can be seen, the model out-performing the baseline is a fairly robust finding.

Chart: Oracle, Model, and Baseline results for $1-million monthly budget



Overall, I think the model can be declared modestly successful. It regularly produces real alpha in excess of 4% for the 36-month ROI relative to the baseline approach of investing randomly. For the 12-million dollars invested over the full duration of the simulation, this would result in more than $480,000 over 3 years.

That said, one thing I wondered about but did not explore due to time-constraints, was the performance of better baseline heuristics. Randomly selecting loans may be considered a weak benchmark. No investor does this in practice. Even if they aren't using a model *per* se, they are likely investing according to some guiding principles or rules-of-thumb. Investigating out-perform relative to baselines constructed from these rules would be interesting.

Another point worth mentioning is that there really isn't a lot of historical data to train on. As explained above training takes place over 2007, 2008, and some 2009 vintage loans. Not only do these years account for a very small portion of the total loans issued, they're also some what anomalous. Referring back to the plot of Aggregate ROI Over Months Since Origination, it's clear that the ROI curves for 2007 and 2008 are fundamentally different from those of 2012. More training data, and particularly training data not as impacted by the recession, would likely help the model generate more lift.

It is also interesting to examine the impact of the “monthly budget” on the modeling results. Intuitively one would expect smaller monthly budgets to produce lower-bias, higher-variance overall model results. And, this is exactly what the results show.   The following charts show results for a 100k and 10-million monthly budgets. The average alpha of the model across all months simulated is was %5.2 percent for the 100k budget and 2.37% for the 10-million budgets, but the volatility was much more severe for the former.

Chart: Oracle, Model, and Baseline results for $100k monthly budget: Avg. Alpha 5.20%

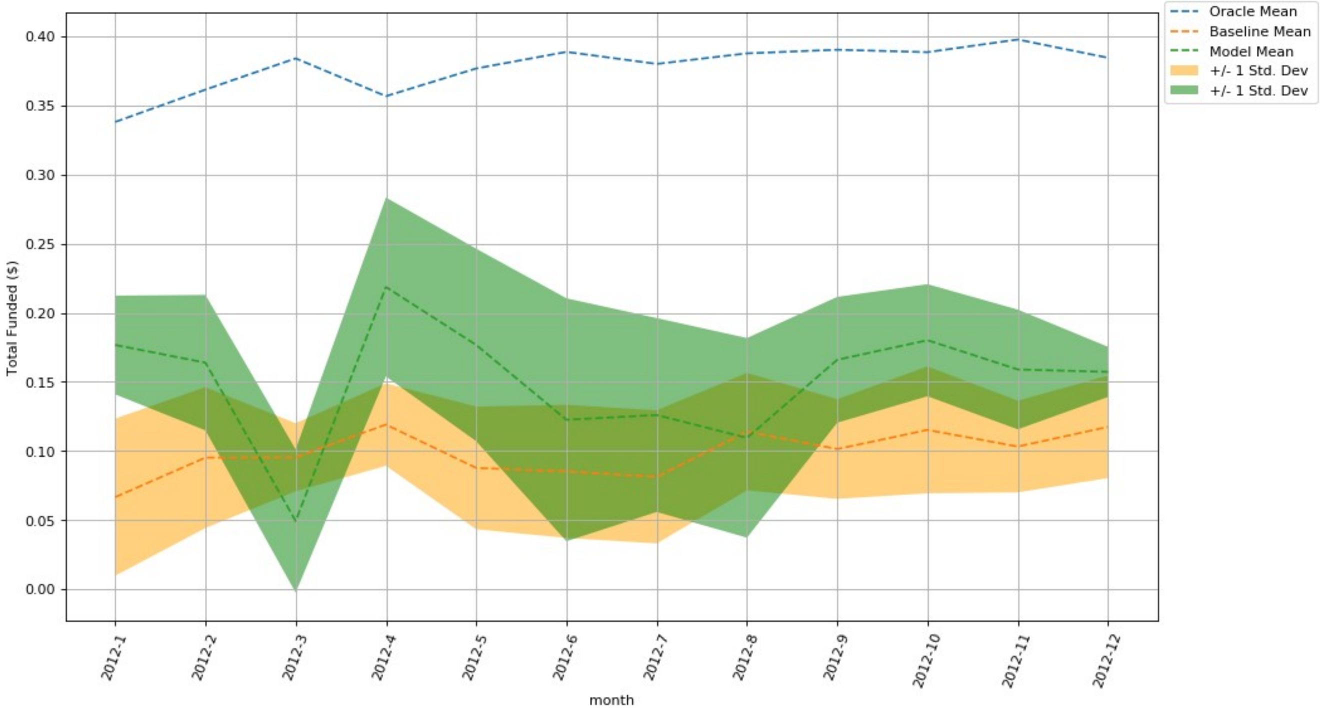
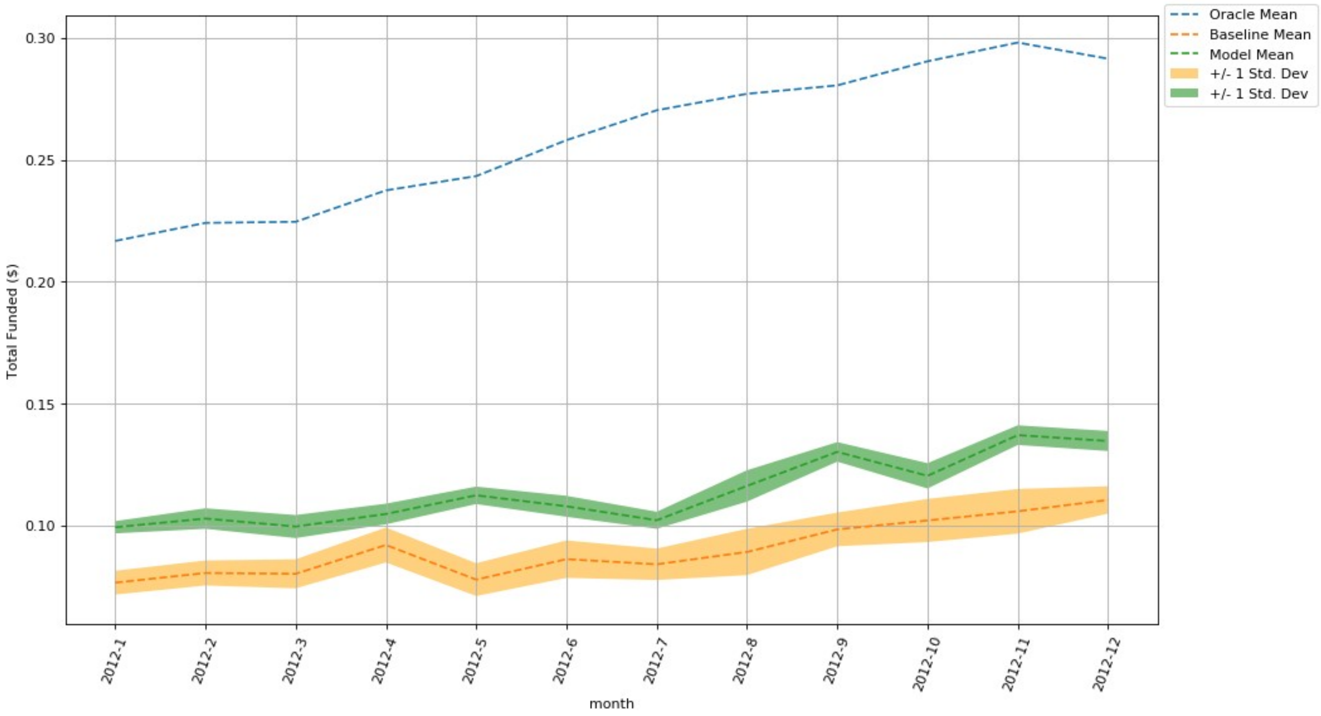


Chart: Oracle, Model, and Baseline results for $10-million monthly budget: AVG Alpha 2.37%

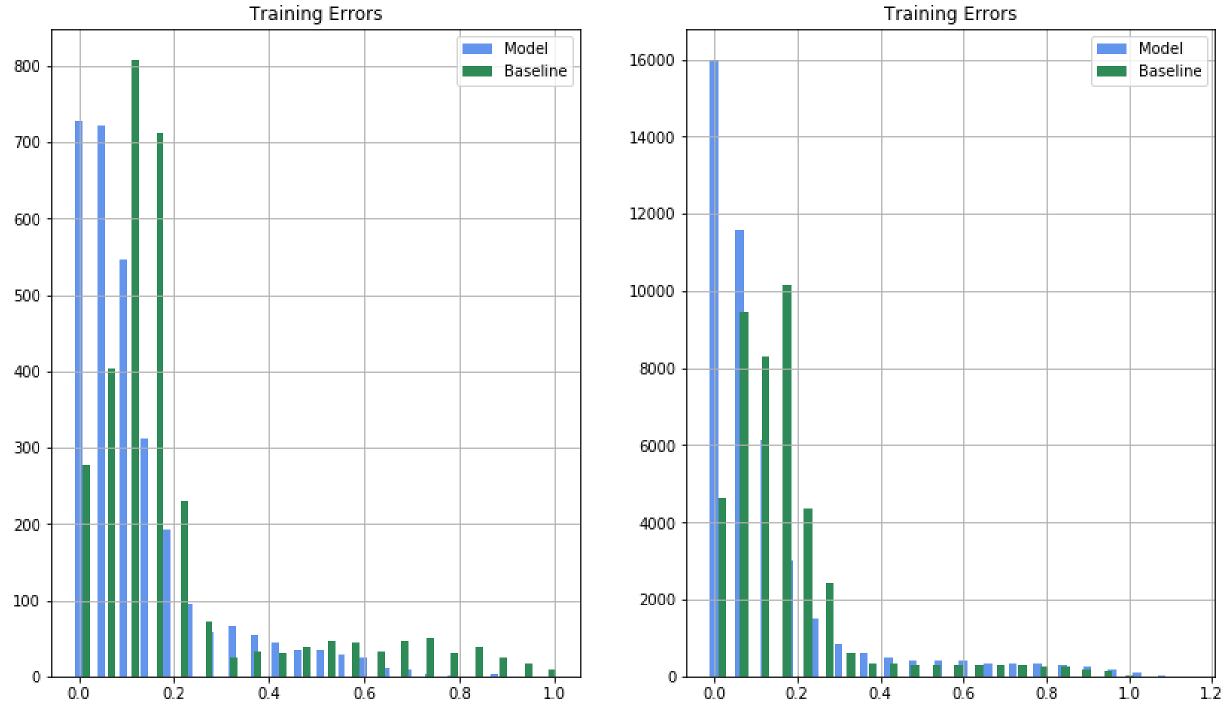


Finally, I'd like to conclude with a brief discussion of model selection and evaluation. Because, the full back-testing simulation described above is computationally costly, I used a modified version of the problem for initial model selection. Here I simply used all historical loans from 2007-06-01 to 2009-01-01 to predict ROIs for all loans issued in 2012, rather than stepping through month-by-month.

I first used a linear model with L2-normalization to do a quick assessment of model feasibility.   Since the results showed some promise, I quickly moved onto a model that would include non-linear interaction terms. My first go to was a support vector regression model with a radial basis function kernel. I've had success with this in the past, and spent a fair amount of scaling features and tuning. Ultimately, however, it's results paled in comparison with scikit-learn's Gradient Boosted random forest implementation. I believe this is because of the large number of categorical features in the dataset and their relatively unbalanced representation. Random forests typically handle these sorts of features well. Additional the random forest model had the upside of being much quicker to train.

While performing initial model selection, I plotted and considered the the dist of errors, predictions, and resulting aggregate ROIs. Representative plots for the final model selected are shown below.

ROI Regression Err. Training (2007–06-01 – 2009-01-01) Testing (2012-01-01 – 2013-01-01)



ROI predictions Training (2007–06-01 – 2009-01-01) Testing (2012-01-01 – 2013-01-01)

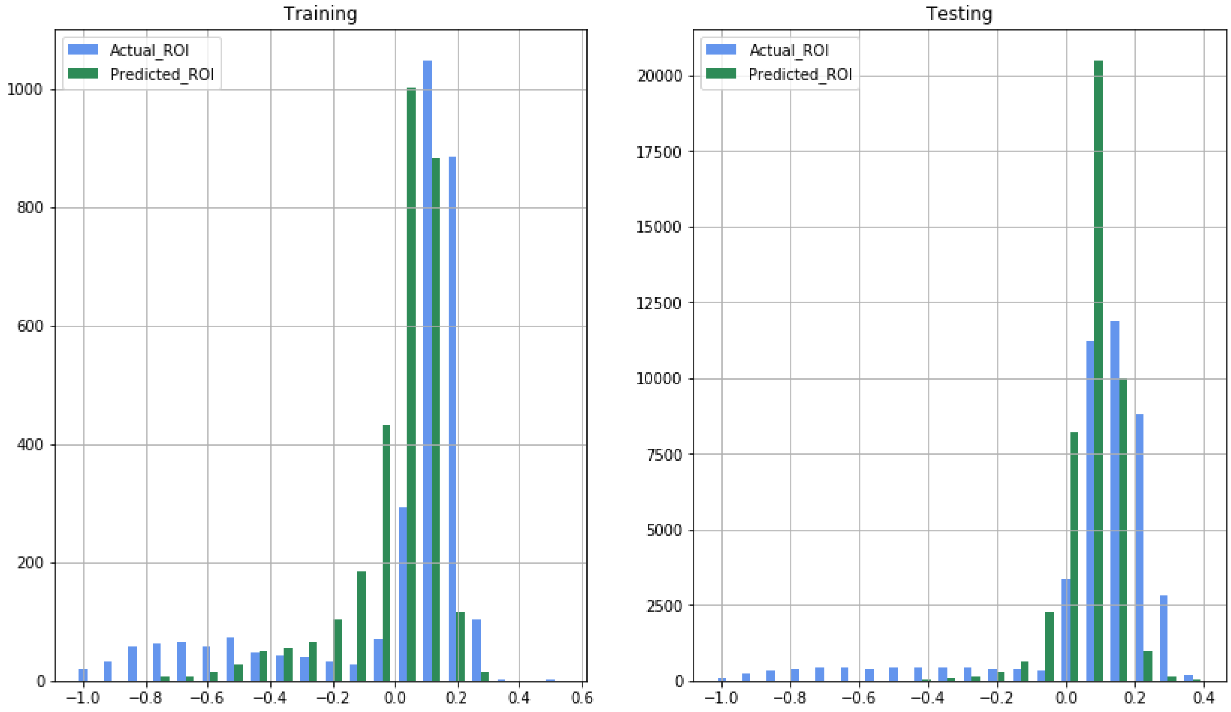
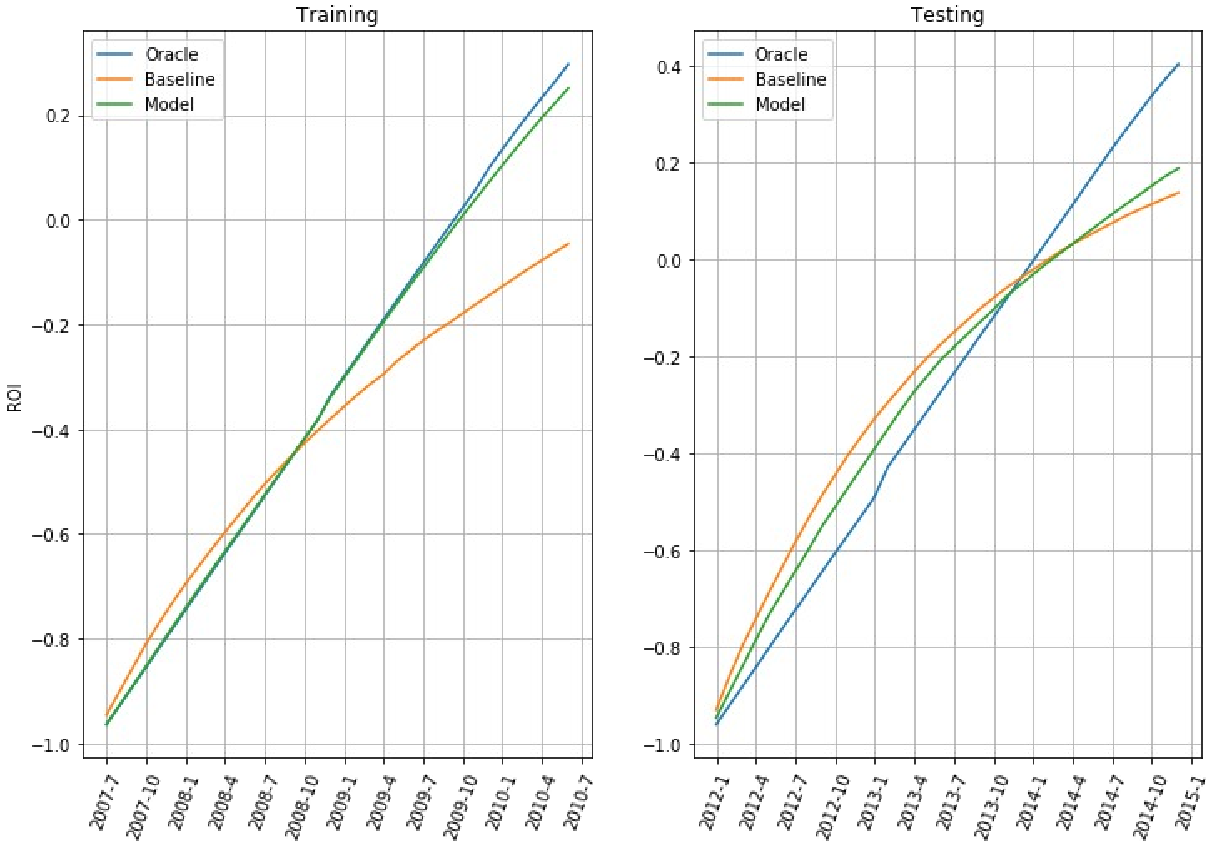


Chart: Aggregate ROI Training (2007–06-01 – 2009-01-01) Testing (2012-01-01 – 2013-01-01)


Training
Oracle
Baseline
Model
ROI
Testing
Oracle
Baseline
Model