Overview

The Lending Club originations file which forms the basis of this project provides a treasure trove of insights into how borrower credit characteristics and loan attributes correlate with credit and prepayment risk. I have attempted to ETL and summarize the data provided, in addition to conducting some basic statistical analyses to investigate how selected credit characteristics (borrower income, borrower DTI, and funded amount) influence annualized returns on a subset of loans. I have also attempted to validate that the results hold across the time period for which data is provided.

Selected Descriptive Statistics

I performed a basic aggregation on interest rates charged and amount funded by term and grade. I've also included the average values for two important credit attributes of the borrower, Annual Income and DTI.

Term	Grade	Int. Rate (avg)	Funded Amnt (avg)	Funded Amnt (sum)	DTI (avg)	Annual Inc. (avg)
36 months	A	7.22%	$13,868	$1,983,393,550	15.58%	$86,568
36 months	B	10.86%	$12,379	$2,624,402,425	17.24%	$72,568
36 months	C	13.92%	$11,847	$1,924,171,150	18.70%	$66,407
36 months	D	17.16%	$11,772	$902,738,250	19.29%	$63,487
36 months	E	19.76%	$11,914	$260,215,900	19.43%	$61,424
36 months	F	23.33%	$10,033	$44,807,475	18.66%	$57,475
36 months	G	23.98%	$12,642	$8,672,700	19.87%	$69,742
60 months	A	8.00%	$18,294	$94,891,050	15.53%	$97,284
60 months	B	10.66%	$19,869	$844,959,275	17.64%	$90,043
60 months	C	14.09%	$19,541	$1,630,529,425	18.97%	$81,522
60 months	D	17.19%	$19,939	$1,253,255,375	20.67%	$76,102
60 months	E	19.96%	$20,678	$1,010,411,775	20.75%	$76,900
60 months	F	23.64%	$21,278	$395,343,675	20.12%	$77,725
60 months	G	25.86%	$21,620	$103,840,600	19.08%	$81,356

Especially at the lower grades, there's not a clear relationship between annual income or dti and grade / rate. Intuitively we would expect these two characteristics to be very tightly correlated with risk, so this result is surprising. I will revisit these characteristics below.

I also present two summary visualizations of the evolution of origination characteristics over time. Given that "Grade" is a central characteristic provided by lending club for portfolio selection, it has been used to stratify both plots.

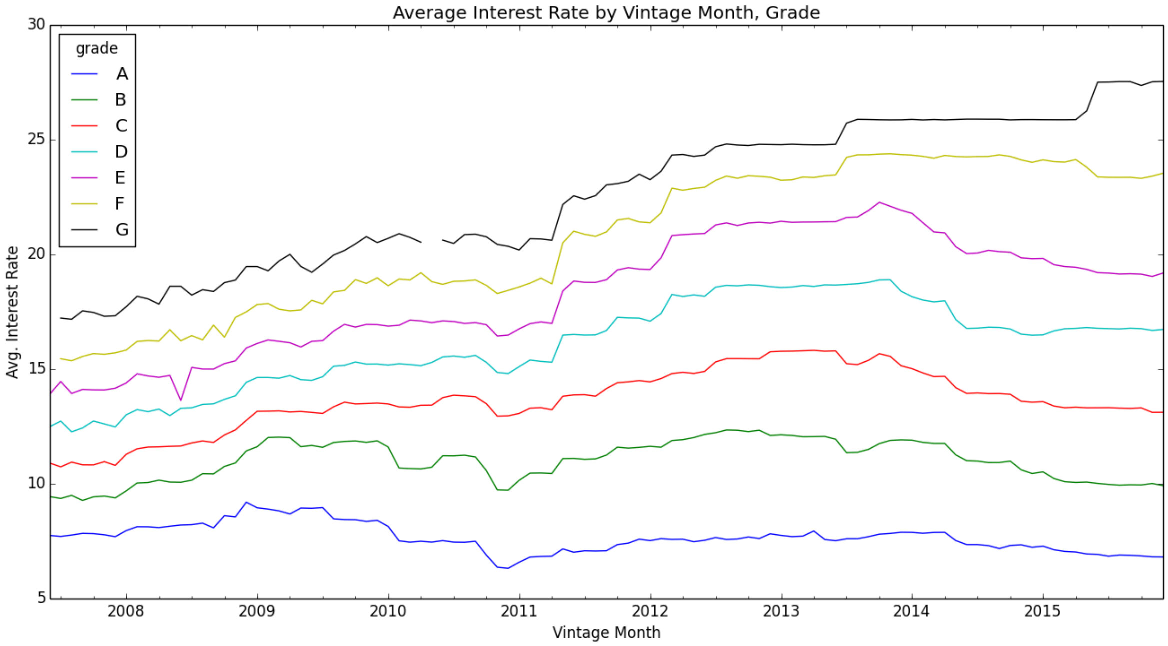


Figure 1 (Average Interest Rate by Vintage Month, Grade)

Two interesting trends emerge in figure (1) above: the overall increase over the first six years of originations, potentially in response to sub-market returns, as well as the subsequent tightening of spreads for the higher-quality grades since 2012.

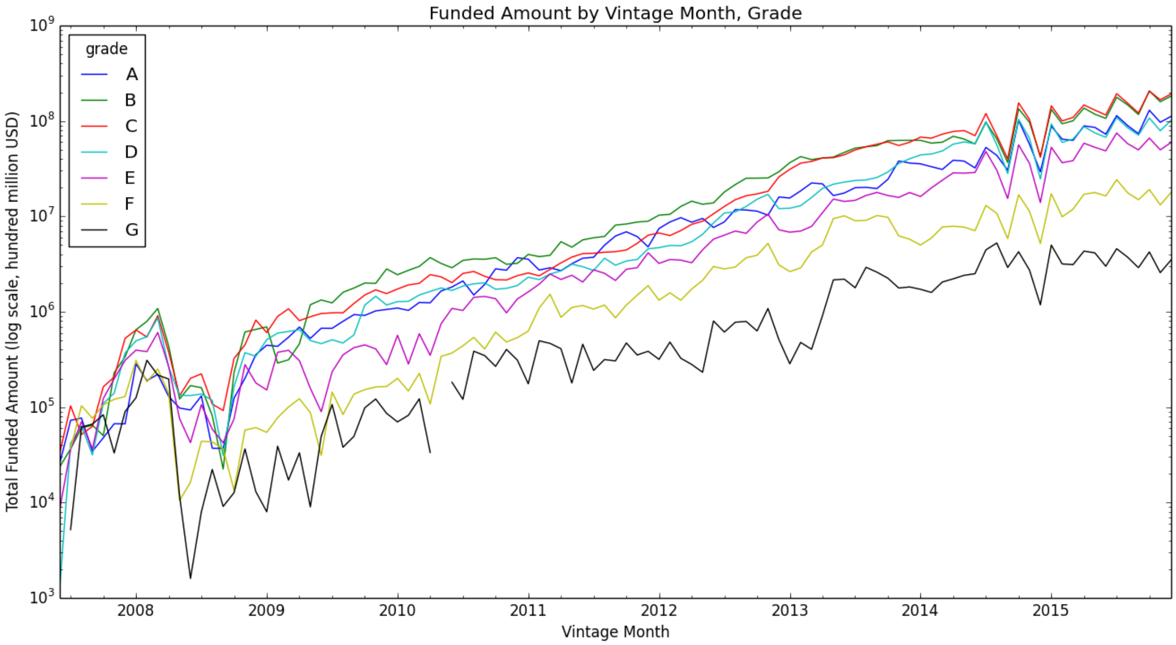


Figure 2 (Funded Amount by Vintage Month, Grade)

Figure (2) presents log-scale originations by grade over time. All grades appear to grow in lockstep over time, apart from the lowest quality F and G grade. Growth rates slow in recent years, likely in reflection of increasing market saturation.

Returns on Matured 36 month loans

Returns were calculated for 36 month loans who had a full history in the data. The return calculation naively assumes that payments received during the life of the loan are held as cash throughout the remainder of the loan's life, when these proceeds could be continually re-invested in similar assets. Returns are calculated on an aggregate level and are therefore weighted by the original loan amount.

The annualized return on all qualifying loans during the period was **3.30%**. The cohort with the highest returns was the **G grade loans from 2007**, however this was on notably low origination volume (less than $100k total). A full breakout of loan returns by grade and vintage year is included at the end of this document.

Grade	Total Amount Funded	Annualized Return (weighted)
A	$182,881,949.0	2.28%
B	$247,783,130.9	3.28%
C	$149,870,777.7	3.70%
D	$98,290,235.6	4.16%
E	$26,235,381.2	4.84%
F	$5,781,745.9	3.41%
G	$2,486,001.6	3.94%

Model

Given the time limitations to this task, the goal was to create an intuitive model of factors that might drive abnormal returns for Lending Club loans. Various machine learning models were rejected because of complexity, perils resulting from over-fitting, and inability to coherently interpret the model parameters.
I ran three simple linear ("Ordinary Least Squares") regressions on several credit characteristics of the borrowers, using my calculated annual return as the dependent variable. The coefficients produced by this model, as well as the measures of statistical fit, are easily understood in terms of independent variables.

The origination characteristics chosen for the independent variables were annual income, DTI , and funded amount. These characteristics were chosen as they displayed relatively muted covariance in the dataset -- DTI specifically is an attempt to isolate the existing debt load of a borrower in proportion to their ability to service this debt, whereas annual income indicates the absolute flow of funds – much in the way that the GDP and Debt-to-GDP ratio of a country give us very different and uncorrelated information about the health of a country's economy.

The first model used the entire set of 36 month loans which have a full 36 months of performance history. The second and third models simply split the population between loans originated before December, 2010 and after. My hypothesis was that, given the public availability of this dataset and the growing sophistication of investors in the space, any clear relationships existing in the data early on will have been noticed by researchers, and the excess returns will have been arbitraged out of the market in the latter half of the performance history.

The model results mildly support this hypothesis, as two of three independent variables displayed reversing relationships with returns between the early and late vintages. Funded amount provided higher returns in the early period (with each additional $1000 borrowed adding 1.7bp of annualized return) but reversed to providing sub-normal returns in the later vintage (with each additional $1000 borrowed detracting 25bp of annualized return). DTI displayed a less statistically important trend – higher DTI borrowers yielded higher returns in the early vintage, with high statistical significance but a low absolute coefficient of relation of 5.5e-05 in the early vintage, flipping signs to -8e04 in the latter vintage, indicating a reversing relationship with returns. Annual income in both cases was *negatively* correlated with returns, with an extra dollar of income decreasing annual return by 2.4e-08 in the early cohort to -1.9e-08 in the later cohort – both results significant to > 99% confidence.

Full model outputs provided in the appendix

If I had the resources to create a more sophisticated model of performance -- for example, one encompassing more credit attributes, or using more complex machine learning algorithms -- I would similarly parameterize the model using this earlier set of performance history and validate whether these relationships hold in the second half of this history (i.e. the training set would be the early loan vintages and the testing set would be the more recent vintages).

Further work

Conceptual

The performance of loans in 2007-2009 is likely highly informative to the performance of this asset class at the end of a credit cycle, and could likely be very useful for approximating the default behavior of the current cohort of loans as we potentially enter a period with similar default and credit market characteristics. Perhaps when assessing an appropriate risk position a further investigation of this period, despite the low volume of originations, could be useful.

It appears there are differences between funded_amt and funded_amt_inv (principal coming in on all loans vs. those held by external investors). On their website, they exclusively use the total amounts issued and recovered, so it would be interesting to investigate performance differences between these two sets of loans.

Loan purpose very data-rich field. I found some interesting research on using NLP to tie borrower word choice and sentiment to loan performance, i.e. : https://papers.ssrn.com/sol3/Papers.cfm?abstract_id=2865327

Technical

This dataset is about 30 months out of date. Retrieving an updated dataset would provide larger sample pool for the model, and further validation of patterns displayed in early vintages. Additionally, scripts could be placed into a more accessible python package and the code could be optimized for speed / memory footprint, as well as generalized for re-use on similar datasets.

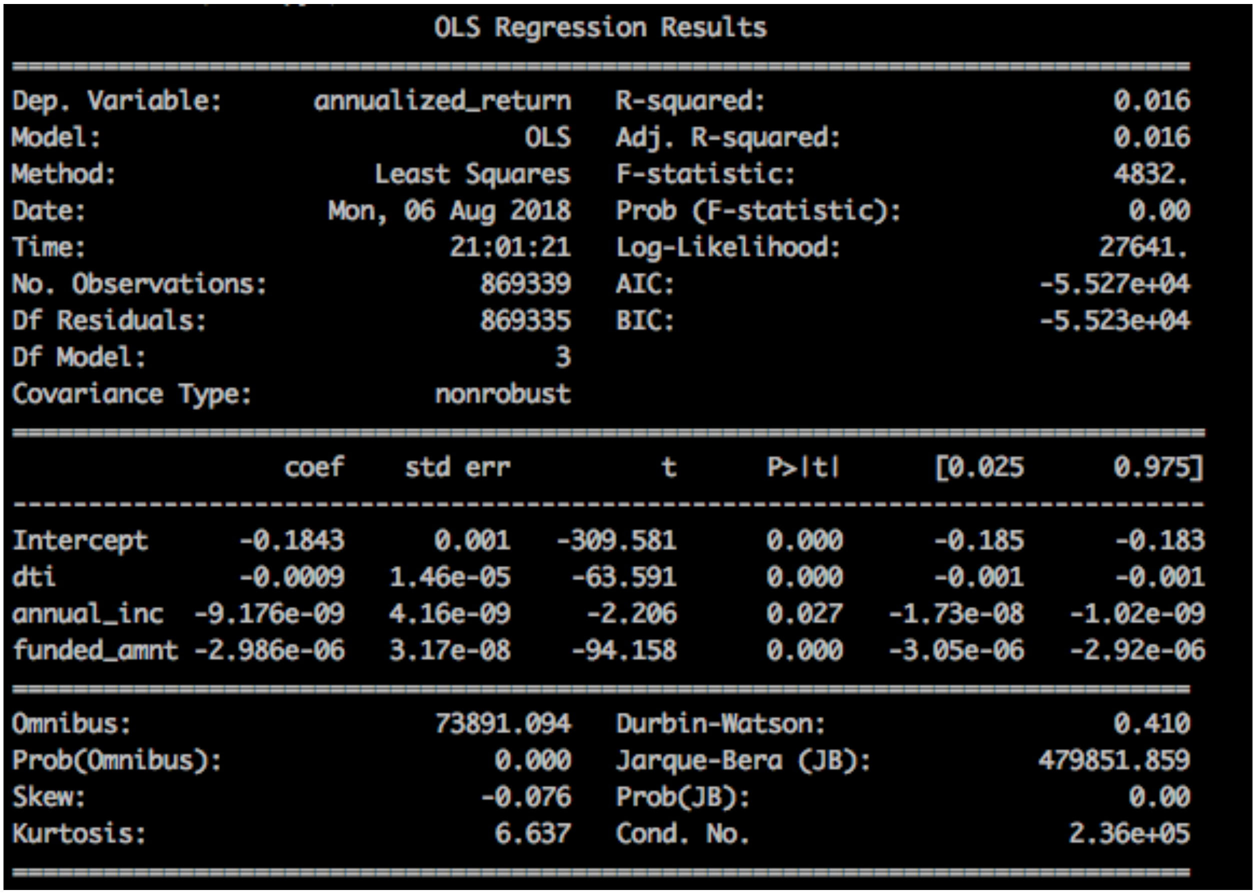
Annualized Returns by Grade, Vintage Year

grade	vintage_year	int_rate	funded_amnt_inv	annualized_return_wt
A	2007	7.75%	$127,163	3.07%
A	2008	8.39%	$953,971	2.41%
A	2009	8.64%	$8,395,865	2.70%
A	2010	7.15%	$19,945,784	2.54%
A	2011	7.08%	$49,494,913	2.21%
A	2012	7.58%	$103,964,252	2.23%
B	2007	9.41%	$120,889	3.23%
B	2008	10.38%	$1,891,955	2.48%
B	2009	11.79%	$14,199,848	2.61%
B	2010	10.71%	$27,646,090	3.20%
B	2011	11.01%	$44,738,443	3.32%
B	2012	12.14%	$159,185,906	3.35%
C	2007	10.84%	$130,248	5.00%
C	2008	11.85%	$1,534,127	2.39%
C	2009	13.32%	$11,076,035	3.02%
C	2010	13.52%	$18,630,670	3.73%
C	2011	13.93%	$19,778,722	3.64%
C	2012	15.19%	$98,720,975	3.79%

D	2007	12.55%	$122,032	1.56%
D	2008	13.36%	$1,210,806	1.27%
D	2009	14.93%	$7,354,666	3.48%
D	2010	15.20%	$13,333,025	3.77%
D	2011	16.19%	$13,285,828	4.37%
D	2012	18.21%	$62,983,879	4.34%
E	2007	14.13%	$142,439	1.92%
E	2008	14.83%	$1,029,790	2.96%
E	2009	16.49%	$3,046,311	3.12%
E	2010	16.85%	$3,832,720	3.98%
E	2011	18.19%	$3,599,859	5.42%
E	2012	20.76%	$14,584,262	5.42%
F	2007	15.58%	$98,138	-2.69%
F	2008	16.30%	$468,154	1.17%
F	2009	18.24%	$1,117,086	2.38%
F	2010	18.78%	$1,454,674	1.98%
F	2011	19.94%	$721,019	3.94%
F	2012	22.84%	$1,922,675	5.64%
G	2007	17.36%	$58,475	11.90%
G	2008	18.05%	$521,384	1.57%
G	2009	20.01%	$588,067	2.64%
G	2010	20.78%	$601,278	4.81%
G	2011	21.44%	$169,697	-1.11%
G	2012	24.59%	$547,100	7.03%

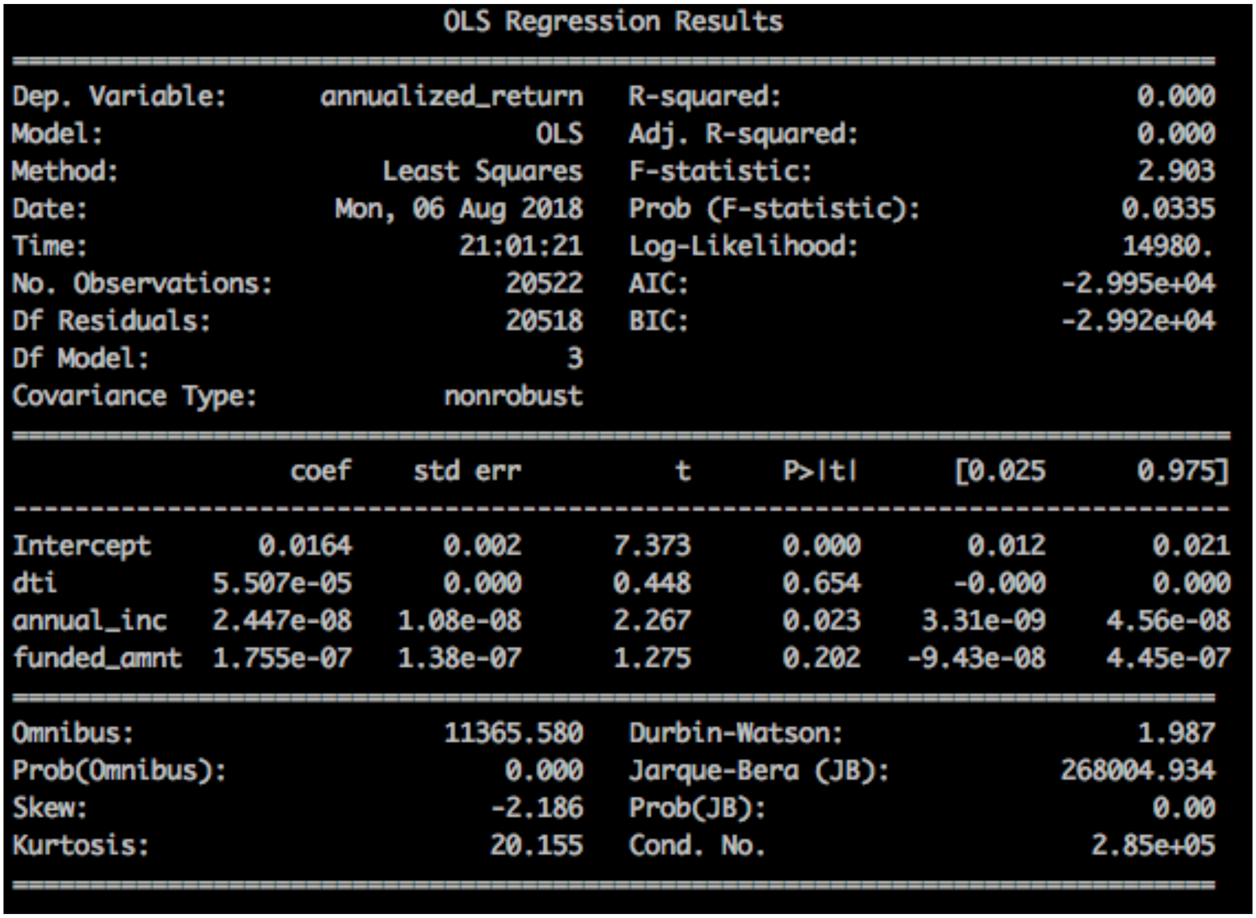
Full Model Outputs

Full Model [2007-2012]



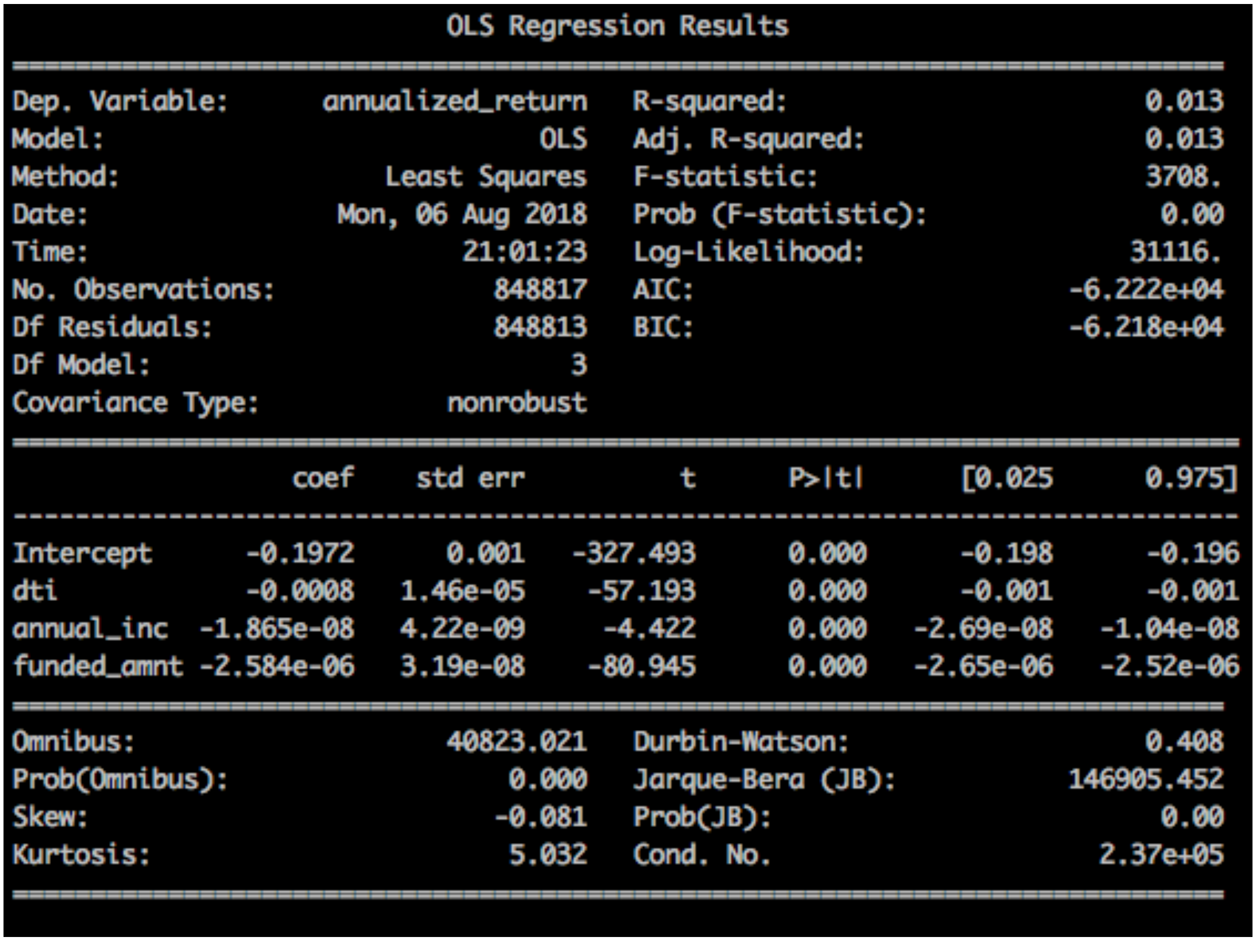
```
                          OLS Regression Results
==============================================================================
Dep. Variable:      annualized_return   R-squared:                       0.016
Model:                            OLS   Adj. R-squared:                  0.016
Method:                 Least Squares   F-statistic:                     4832.
Date:                Mon, 06 Aug 2018   Prob (F-statistic):               0.00
Time:                        21:01:21   Log-Likelihood:                 27641.
No. Observations:              869339   AIC:                        -5.527e+04
Df Residuals:                  869335   BIC:                        -5.523e+04
Df Model:                           3
Covariance Type:            nonrobust
===============================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------
Intercept      -0.1843      0.001   -309.581      0.000      -0.185      -0.183
dti            -0.0009   1.46e-05    -63.591      0.000      -0.001      -0.001
annual_inc  -9.176e-09   4.16e-09     -2.206      0.027   -1.73e-08   -1.02e-09
funded_amnt -2.986e-06   3.17e-08    -94.158      0.000   -3.05e-06   -2.92e-06
==============================================================================
Omnibus:                    73891.094   Durbin-Watson:                   0.410
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           479851.859
Skew:                          -0.076   Prob(JB):                         0.00
Kurtosis:                       6.637   Cond. No.                     2.36e+05
==============================================================================
```

Early Period Model [2007-2010]



```
                            OLS Regression Results
==============================================================================
Dep. Variable:      annualized_return   R-squared:                       0.000
Model:                            OLS   Adj. R-squared:                  0.000
Method:                 Least Squares   F-statistic:                     2.903
Date:                Mon, 06 Aug 2018   Prob (F-statistic):             0.0335
Time:                        21:01:21   Log-Likelihood:                 14980.
No. Observations:               20522   AIC:                        -2.995e+04
Df Residuals:                   20518   BIC:                        -2.992e+04
Df Model:                           3
Covariance Type:            nonrobust
================================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
Intercept      0.0164      0.002      7.373      0.000       0.012       0.021
dti         5.507e-05      0.000      0.448      0.654      -0.000       0.000
annual_inc  2.447e-08   1.08e-08      2.267      0.023    3.31e-09    4.56e-08
funded_amnt 1.755e-07   1.38e-07      1.275      0.202   -9.43e-08    4.45e-07
==============================================================================
Omnibus:                    11365.580   Durbin-Watson:                   1.987
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           268004.934
Skew:                          -2.186   Prob(JB):                         0.00
Kurtosis:                      20.155   Cond. No.                     2.85e+05
==============================================================================
```

Late Period Model [2011-2012]



```
                            OLS Regression Results
==============================================================================
Dep. Variable:      annualized_return   R-squared:                       0.013
Model:                            OLS   Adj. R-squared:                  0.013
Method:                 Least Squares   F-statistic:                     3708.
Date:                Mon, 06 Aug 2018   Prob (F-statistic):               0.00
Time:                        21:01:23   Log-Likelihood:                 31116.
No. Observations:              848817   AIC:                        -6.222e+04
Df Residuals:                  848813   BIC:                        -6.218e+04
Df Model:                           3
Covariance Type:            nonrobust
===============================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------
Intercept      -0.1972      0.001   -327.493      0.000      -0.198      -0.196
dti            -0.0008   1.46e-05    -57.193      0.000      -0.001      -0.001
annual_inc  -1.865e-08   4.22e-09     -4.422      0.000   -2.69e-08   -1.04e-08
funded_amnt -2.584e-06   3.19e-08    -80.945      0.000   -2.65e-06   -2.52e-06
==============================================================================
Omnibus:                    40823.021   Durbin-Watson:                   0.408
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           146905.452
Skew:                          -0.081   Prob(JB):                         0.00
Kurtosis:                       5.032   Cond. No.                     2.37e+05
==============================================================================
```