Hi Srdjan,

This is just the rough final draft of the Capstone Report.

I’m a little concerned about the results of the predictive model. It looks like for some reason the data for the ‘bad’ loans is not complete, it has zero variance. This result in a trivial fit. I discuss this in the last section of the report.

Any feedback would be welcome.

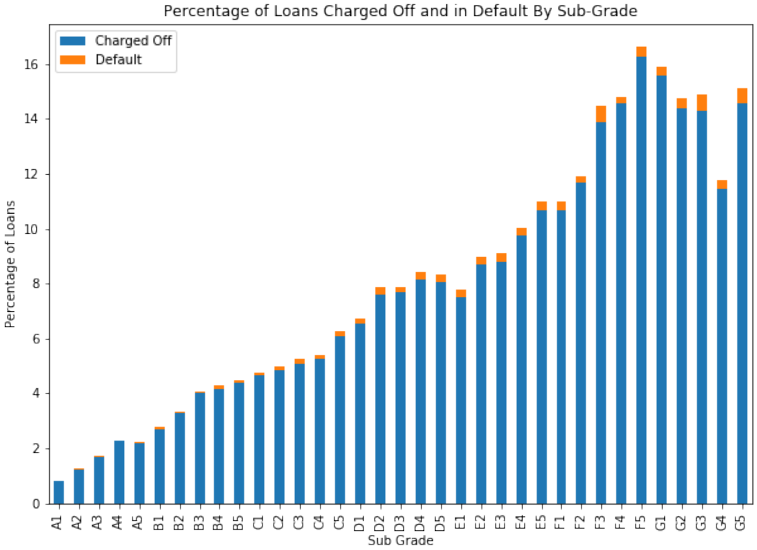
Best regards,
David Ermer

Inferential Statistics and Supervised Learning Applied to LendingClub Data

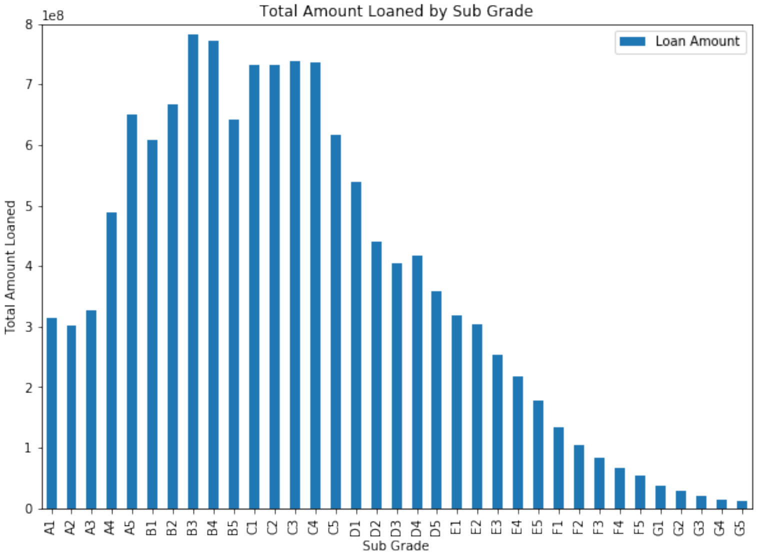
David Ermer

LendingClub Grading, Performance and Data Quality:

The LendingClub (LC) data has information about 887379 loans with 74 columns of data. The columns with FICO scores are missing and might contain the primary information used in the loan approval process. The LC process involved the ranking of the loans into 7 grades and 35 sub-grades. Calculating the percentage of charged off/in default loans in each of the 35 sub-grades shows that what ever process used by LC is reasonable and predictive of the performance of the loan.



LC also used their grading system to proportion the total amount loaned to each sub-grade.



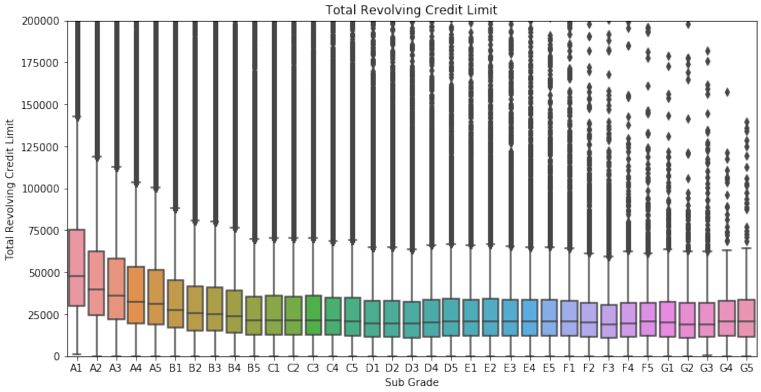
The provided data were processed and 38 columns were selected as being the most likely to contain information needed. As can be seen below there are a number of columns with 21372 rows of non-NaN data.

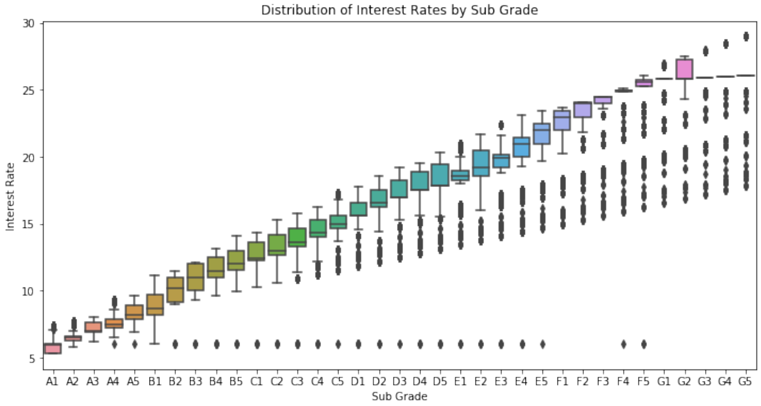
```
Sub Grade                      887379
Loan Amount                    887379
Interest Rage                  887379
Annual Income                  887375
Annual Income - Joint             511
Income Verification            887379
Income Verification - Joint    887379
Data Issued                    887379
Debt Income Ratio              887379
Debt Income Ratio - Joint         509
Revolving Credit Balance       887379
Revolving Credit Ratio         886877
All Credit Ratio                21372
Current Deliquent Accounts     887379
Collections (12mths)           887379
Delquient (2yrs)               887379
Years Employed                 887379
Home Ownership                 887379
Inquiries (6mths)              887379
Personal Finance Inq           887379
Credt Inq 12mts                 17389
Inst Credit Ratio              887379
Earliest Credit                887379
Max Bal Revol                   21372
Total Coll E Owed              817103
Mths Last Delq                 433067
Mths Last Derog                221703
Mths Last Rec                  137053
Mths Last Inst                  20810
Num Open Acc                   887350
Num Open Acc 6mths              21372
Inst Open 12mths                21372
Inst Open 24mths                21372
Inst Open 6mths                 21372
Open Revolve 12m                21372
Open Revolve 24m                21372
Total Revolve Limit            817103
Loan Status                    887379
```

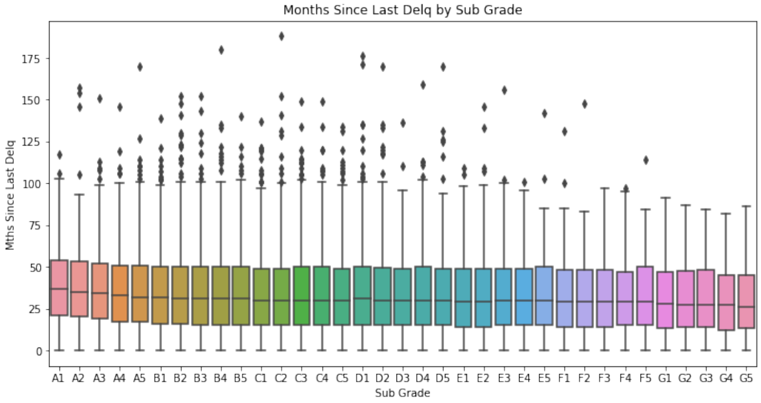
The data that is provided suffers from selection bias, i.e. it only includes data from loans that were approved. The data has also been edited, for example FICO scores have been removed. We may find that other data that has been obfuscated for legal, liability or other reasons.

<u>Statistical Difference of Data Across LC Sub-Grades:</u>

One question that needs to be determined is ‘are the differences in the factors across the different sub-grades statistically significant?’ Box graphs of most of our 38 columns grouped by sub-grade were made. For example:

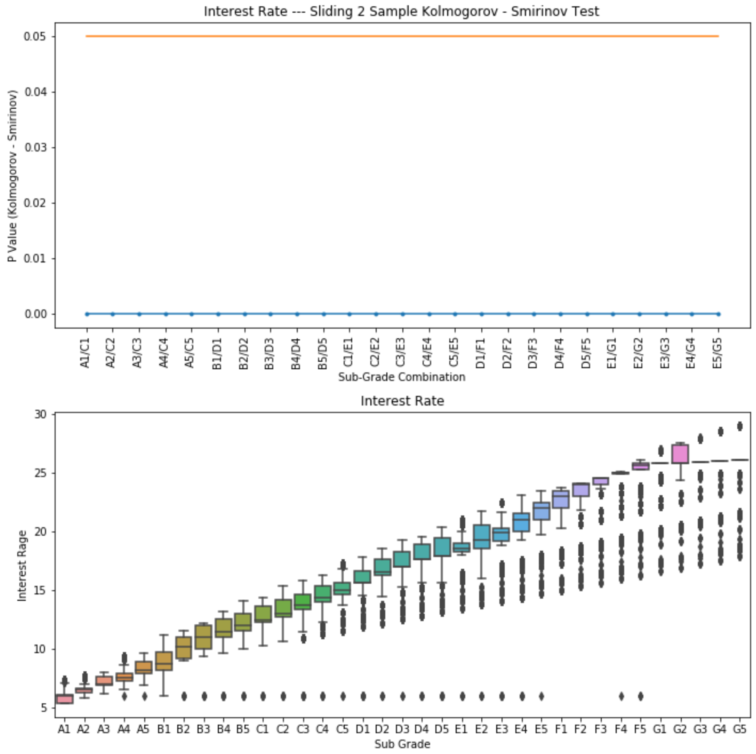






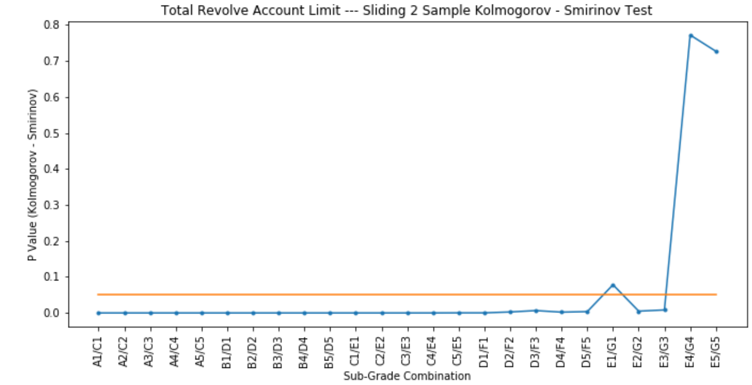
The interest rate is obviously assigned during and not a factor used in the loan approval process and it has as a result a strong correlation with the assigned sub-grade. The *Total Revolve Limit* and *Mths Last Delq* are both important factors in determining the sub-grade. However from the graphs it is not obvious that there is a significant difference in these factors among the different sub-grades. The mean of the credit limit changes by a factor of ~2, but the mean of the months since last delinquency does not vary much more than about 18%. However in both cases there is quite a bit of overlap between the distributions across all of the sub-grades. Most of the data in this data set have quite a few outliers.

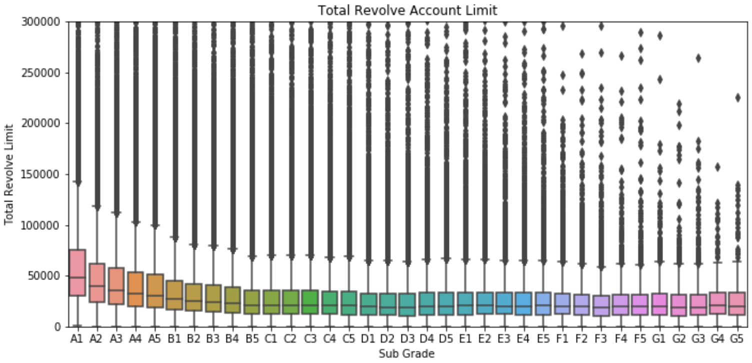
The Kolmogorov-Smirnov test was selected to examine the statistical independence of the the factors across the sub-grades. The distribution of sub-grades spaced by 10 grades (i.e. two main grades) was tested and the p-value graphed for most of the columns. Other intervals were tried but similar results were obtained.



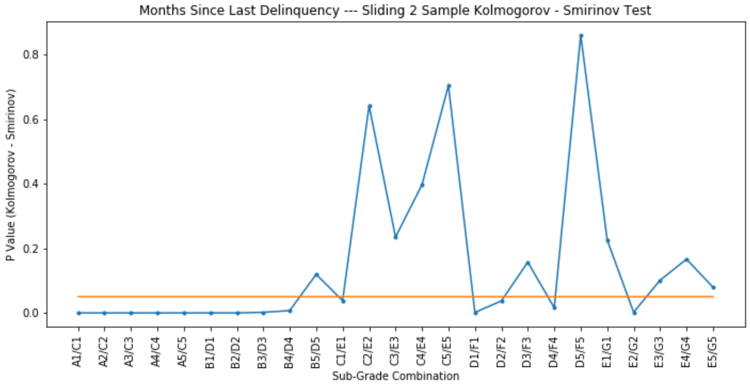
The interest rate obviously shows independence of the distributions across the sub-grades.

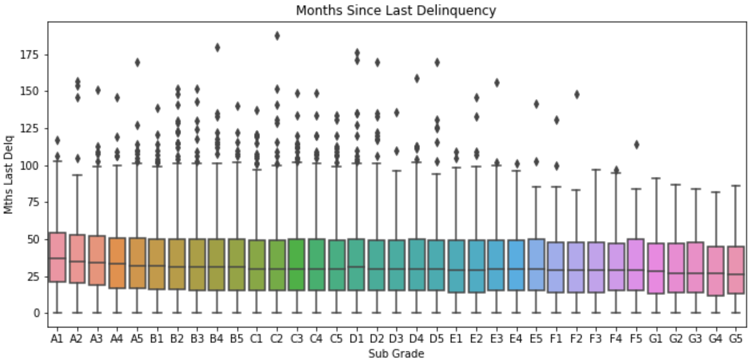
The distribution of the total revolving credit limit data is by the KS test significantly different across the sub-grades. There are many fewer instances at the lower sub-grades which may effect the test between those grades.





The *Mths Last Delq* data shows much less variance across the sub-grades and KS test results in a p-value above 0.05 across a range where there are actually a large number of instances in the sub-grade.





Most of the columns with numeric data were tested this way with various results. Some of the data are quite homogenized across the sub-grades while others are not.

Predictive Model Using Supervised Learning:

Fifteen columns of the data were selected to be used to train a Support Vector Machine (SVM) to categorized the loans into ‘good’ or ‘bad’ based on the *Loan Status* column. All rows with NaN were dropped leaving 201165 loans, 153929 ‘good’ and 47228 ‘bad’.

```
Debt Income Ratio             201165
Revolving Credit Balance      201165
Current Deliquent Accounts    201165
Collections (12mths)          201165
Delquient (2yrs)              201165
Years Employed                201165
Inquiries (6mths)             201165
Personal Finance Inq          201165
Inst Credit Ratio             201165
Earliest Credit               201165
Annual Income                 201165
Revolving Credit Ratio        201165
Num Open Acc                  201165
Total Coll E Owed             201165
IVCode                        201165
```

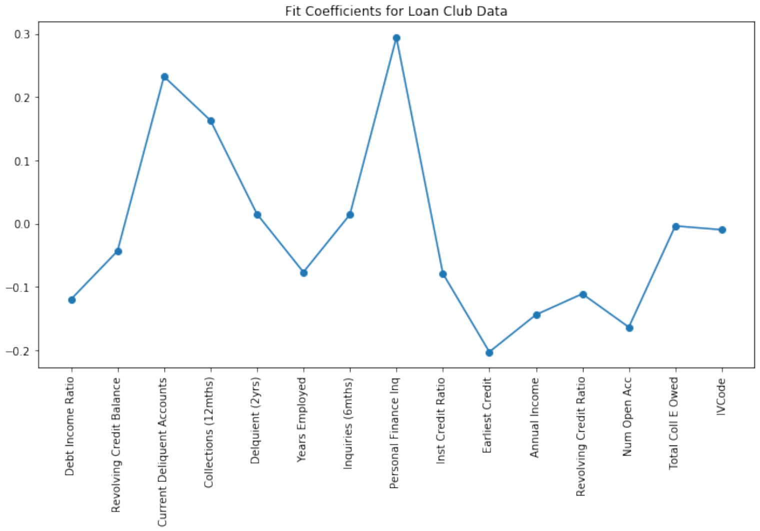
Both ‘linear’ and ‘rbf’ kernels were used and the penalty ‘C’ and kernel coefficient ‘gamma’ hyper-parameters were tuned using a grid search. The data was split into a training set and a test set (75%/25%). The final model using a ‘liner’ kernel had an accuracy score of 0.999946975271 on the training data and an accuracy score of 1.0 on the test data. This will be discussed completely in the next section. The confusion matrix for the training data was,

Training	Predicted: Good	Predicted: Bad
Actual: Good	115498	8
Actual: Bad	0	35367

And on the training data,

Test	Predicted: Good	Predicted: Bad
Actual: Good	38431	0
Actual: Bad	0	11861

These results vary depending on the kernel, random_state, and hyper-parameters used but in all cases the fit is very good. A graph of the fit coefficients for a typical ‘linear’ kernel fit is show below.



Fit and Data Quality:

A fit with a precision score of 1.0 is indicative of an issue with either the fitting method or with the data. This doesn't appear to be a problem of over fitting to the training data or with the fitting method. To explore the reason for such a good fit the variance of the various columns for the 'good' and 'bad' loans.

Variance of Data in Columns	Good	Bad
Debt Income Ratio	6.081339e+01	0.0
Revolving Credit Balance	4.101893e+08	0.0
Current Deliquent Accounts	4.544634e-03	0.0
Collections (12mths)	1.034459e-02	0.0
Delquient (2yrs)	6.226807e-01	0.0
Years Employed	1.492368e+01	0.0
Inquiries (6mths)	1.123380e+00	0.0
Personal Finance Inq	4.461511e-03	0.0
Inst Credit Ratio	4.869610e+00	0.0
Earliest Credit	6.572376e+06	0.0
Annual Income	3.488163e+09	0.0
Revolving Credit Ratio	5.737877e+02	0.0
Num Open Acc	2.465460e+01	0.0
Total Coll E Owed	5.470880e+08	0.0
IVCode	6.977742e-01	0.0

It would appear that the data for the ‘bad’ loans has at some point and for an unknown reason been completely homogenized. This means that for the columns selected, all ‘bad’ loans are mapped to a single point in the 16 dimensional space, making this a trivial fitting problem. And it was not evident until after the SVM fit was done.

Other columns were examined to see if they contain non-trivial data on the ‘bad’ loans. The eight columns with 21372 valid values contain no data for ‘bad’ loans. *Mths Last Delq*, *Mths Last Derog* and *Mths Last Rec* have rows with ‘bad’ loans but in low numbers so there use would greatly reduce the number of training/test instances. The rest of the columns have little valid data and for the most part have no rows for the ‘bad’ loans.