Statistical Analysis and Supervised Learning Applied to LendingClub Data

Intermediate Data Science Course Capstone Project
SpringBoard

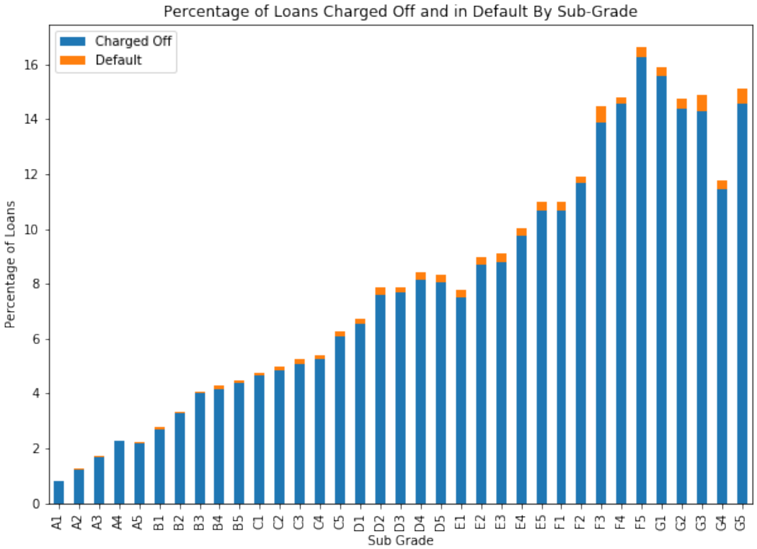
by
David Ermer

Objectives:

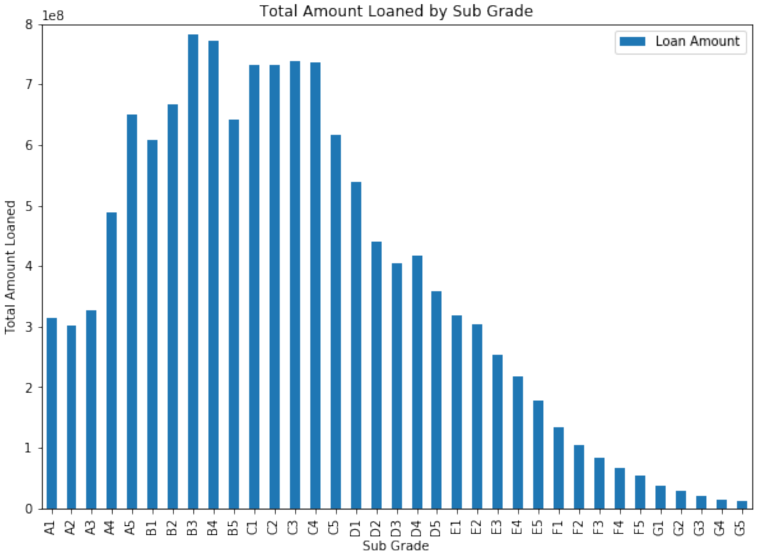
Using the data provided, examine the performance of the selection process used by LendingClub, analyze the statistical differences in the data and see if it can be used to refine the slection process through predictive modeling.

LendingClub Grading, Performance and Data Quality:

The LendingClub (LC) data (obtained from kaggle.com) 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.



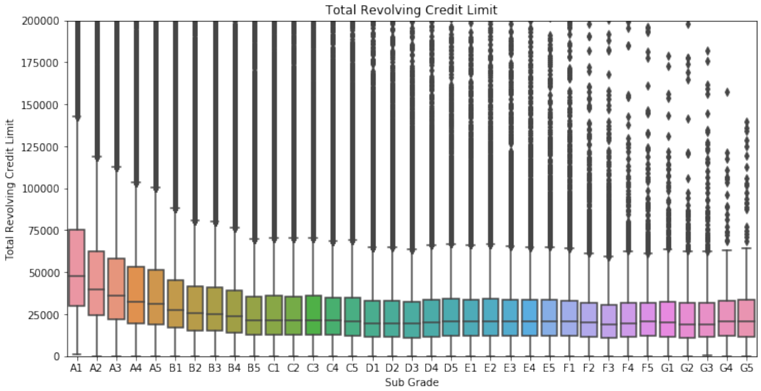
The provided data were processed and 38 columns were selected as being the most likely to contain the information needed. As can be seen below there are a number of columns with 21372 rows of non-NaN data.

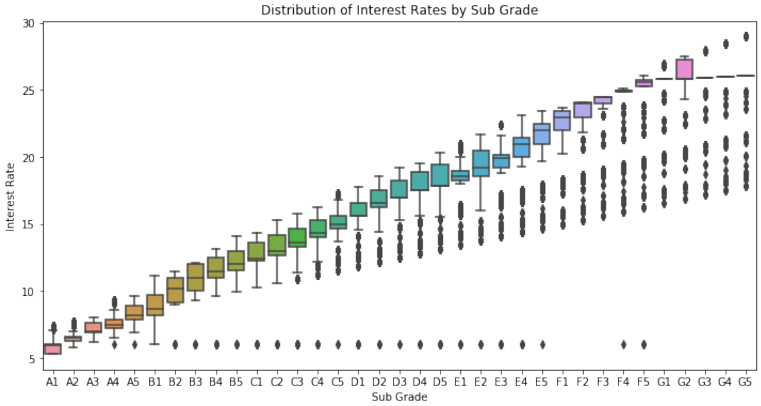
```
Sub Grade                       887379
Loan Amount                     887379
Interest Rage                   887379
Annual Income                   887375
Annual Income - Joint              511
Income Verification             887379
Income Verification - Joint     887379
Data Issued                     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
```

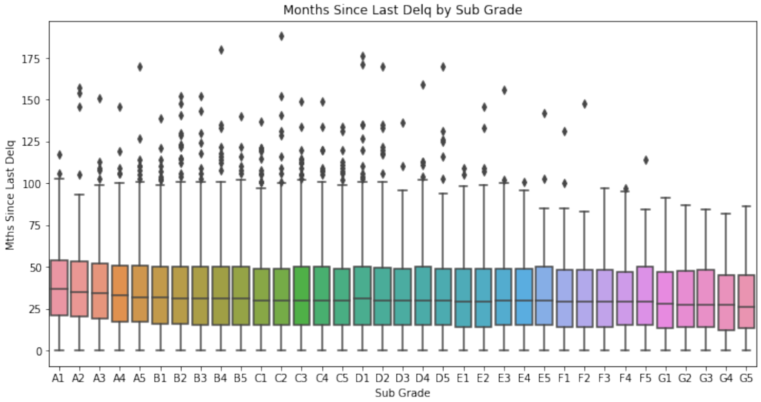
The data that is provided suffers from selection bias, i.e. it only includes data from loans that were approved. This will probably have some effect on the results of predictive modeling which will be discussed in the relevant section below.

Statistical Difference of Data Across LC Sub-Grades:

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:

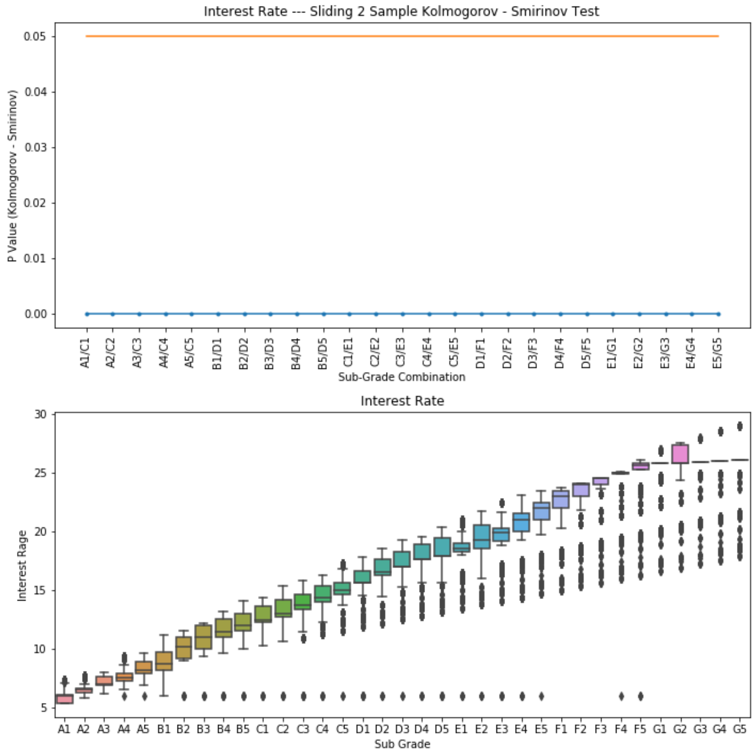






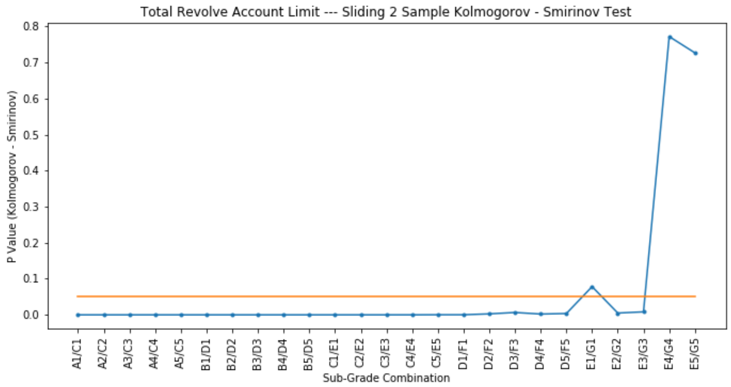
The interest rate is obviously assigned during and not a factor used in the loan approval process and it has as a result a strong correlation with the assigned sub-grade. The *Total Revolve Limit* and *Mths Last Delq* are both important factors in determining the sub-grade. However from the graphs it 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.

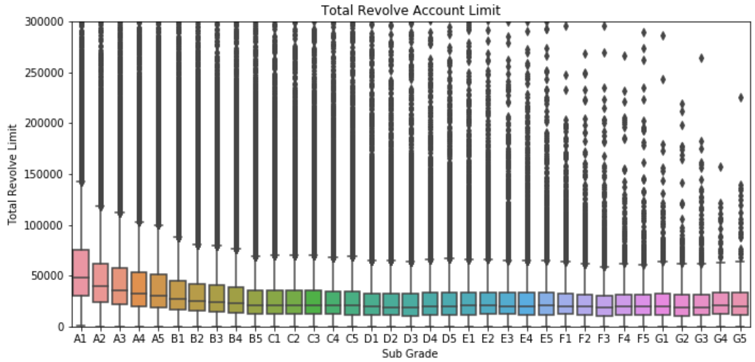
The Kolmogorov-Smirnov test was selected to examine the statistical independence of the factors across the sub-grades. The distribution of sub-grades spaced by 10 grades (i.e. two main grades) was tested and the p-value graphed for most of the columns. Other intervals were tried but similar results were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



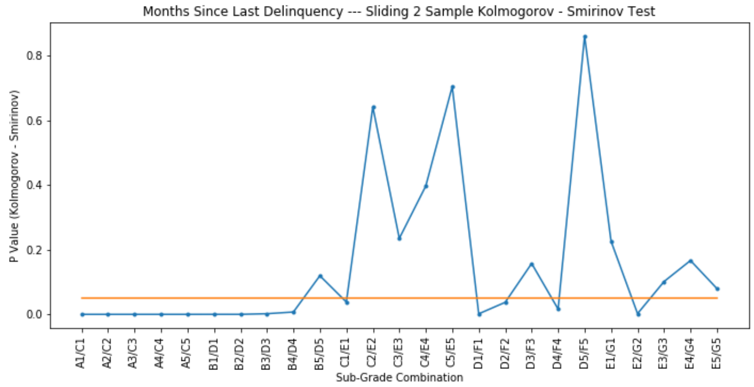
The interest rate obviously shows independence of the distributions across the sub-grades.

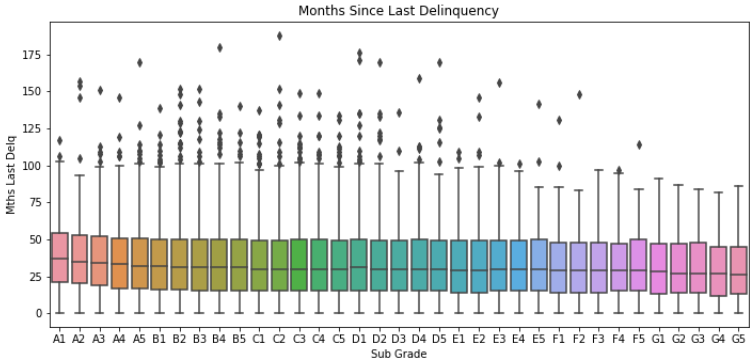
The distribution of the total revolving credit limit data is by the KS test significantly different across the sub-grades. There are many fewer instances at the lower sub-grades which may effect the test between those grades.





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. Although the distributions of every column relevant to loan performance has a high degree of overlap across all sub-grades. This seems odd given the default/charge off rate difference of a factor of about 16 from the lowest to highest (1% and 16%). This may indicate a problem with the data and may effect the quality of the predictive model developed in the next section.

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 246593 loans, 202048 'good' and 44545 'bad'.

```
Debt to Income Ratio            246593
Revolving Credit Balance        246593
Current Delinquent Accounts     246593
Collections (12mths)            246593
Delinquent (2yrs)               246593
Inquiries (6mths)               246593
Annual Income                   246593
Revolving Credit Ratio          246593
Open Accounts                   246593
Income Verified  (code)         246593
Employment Length (code)        246593
Earliest Credit (days)          246593
```

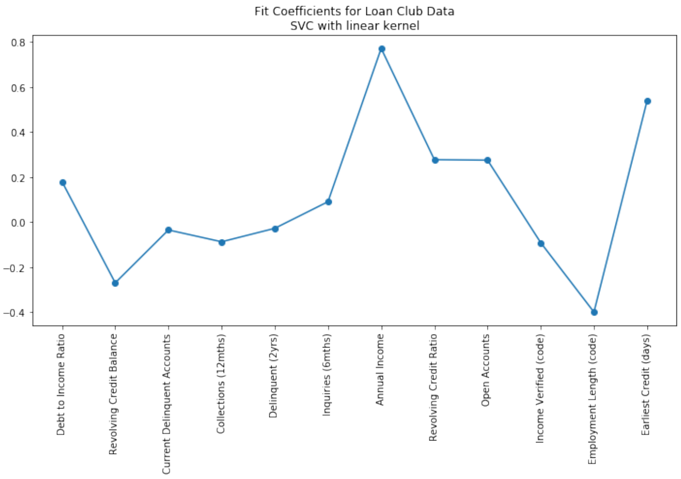
Both 'linear' and 'rbf' kernels were tried but the run time for the 'rbf' case was prohibitively long. The data was split into a training set and a test set (75%/25%). The final model using a 'linear' kernel had an accuracy score of 0.50386 on the training data and an accuracy score of 0.5034 on the test data. The confusion matrix for the training data was,

Training	Predicted: Good	Predicted: Bad
Actual: Good	76941	74426
Actual: Bad	17332	16245

And on the training data,

Test	Predicted: Good	Predicted: Bad
Actual: Good	25713	24968
Actual: Bad	5645	5322

A graph of the fit coefficients for a typical 'linear' kernel fit is show below.



A KNeighborsClassifier with n_neighbors set to 12 was tried. The accuracy score on the training data was 0.8352 and 0.7896 on the test data. The calculated confusion matrices are as follows:

Training	Predicted: Good	Predicted: Bad
Actual: Good	147816	3718
Actual: Bad	26746	6664

Testing	Predicted: Good	Predicted: Bad
Actual: Good	47821	2693
Actual: Bad	10278	856

The accuracy is reasonably high but the confusion matrices show that the predictive modeling of this data is not really good enough to make decisions on loan approval. For instance throwing out loans that are predicted as bad would potentially throw out more good loans than bad.

Most likely, one of two conditions apply, the first is that the training data is unbalanced i.e there are many more good examples. The second is that the data is simply not differentiated enough to make a good prediction.

To test for the first case the data was scales and then split up 75%/25% into training and testing data. The training data was then randomly split into 10 sets such that in each set there are an equal number of good and bad examples. All of the bad loan data was in each training set. Ten classifiers were trained and used to predict the test data. The final prediction was determined by letting the ten classifiers ‘vote’ and then applying all possible thresholds to the vote counts.

For SVM prediction (C = 2.0, linear kernel) we obtained an accuracy of 0.7488 with a threshold of 10 i.e. all classifiers had to predict a bad loan to result in a bad classification. The confusion matrix obtained:

	Predicted: Good	Predicted: Bad
Actual: Good	44971	5541
Actual: Bad	9945	1191

For KNeighborsClassifier (n_neighbors = 12, algorithm = ‘auto’) an accuracy of 0.7652 was seen. The highest accuracy was again obtained by requiring all classifiers to vote for a bad classification. The confusion matrix was:

	Predicted: Good	Predicted: Bad
Actual: Good	45338	5174
Actual: Bad	9298	1838

Again these results have reasonable accuracy but are really sufficient to make decisions regarding loan approval.

A final technique was tried by having two thresholds and adding an ‘uncertain’ classification. Vote tallies equal to or below the first threshold were classified as ‘good,’equal to or above the second as ‘bad.’ In between the loans were classified as ‘uncertain.’ Training/testing results from the KNeighborsClassifer run from above were used. The thresholds were 8 and 10 and the highest accuracy was 0.7663 (uncertain classification does not count as a prediction). The confusion matrix was as follows:

	Predicted: Good	Predicted: Bad	Uncertain
Actual: Good	39833	5174	5505
Actual: Bad	7536	1838	1762

And it appears that this result also suffers from the same problems as seen in the other results.

Conclusions:

This data has very little variation across all of the LC assigned sub-grades even though many of the tests performed show that there is a statistically significant difference in a number of the data columns. It is also unbalanced in that there are more ‘good’ loans than ‘bad.’ In the predictive modeling we tried to address the problem of the unbalanced classes and used the two learning algorithms best suited for the data set.

Looking at the accuracy and confusion matrices results and statistical analysis, it appears that there is not enough information in the data to provide a usable predictive model.