Milestone 1: Credit Risk Analysis in Social Lending Ecosystem

Aveek Choudhury - Harshita Ved - Sagar Singh - Sarang Pande

Peer-to-peer lending refers to the practice of lending money to individuals (or small businesses) via online services that match anonymous lenders with borrowers. Due to this anonymity, informed decisions are very critical for third parties working in between i.e. LendingClub here. One of the interesting features of the peer-to-peer lending market is the richness of the historical data available, which is 2.25M loans between 2007 & 2018 in our dataset.

One of the most important steps is to decide which loan characteristics tend to indicate a loan will be "good" (Fully Paid) or "bad" (Default or Charged Off), and another key step would be to identify the interest rate to be offered to individuals based on their application for the loan. Since the dataset is sparse and consists of 145 variables, there is a need for dimensionality reduction. As a first intuition we could leave out all rows consisting of NA values, but that leads to losing the entire dataset. As an alternative, we decided to first drop columns with more than 90% of missing data which resulted in dropping 38 variables. As the next step for this milestone, we decided to focus on cleaning records so we removed all rows with NA values.

Marching towards our first goal of inferring based on region-wise loan targets and answering to questions related to ideas for business expansion based on regions with a higher level of operating activity, we conducted our preliminary exploration of good and bad loan distribution concerning the state-wise distribution of average income, interest rates being offered and dti ratio. Based on our current observations, Idaho(ID) has the highest average dti ratio of 22%, and Washington(DC) had the highest average income of 91.48K with the lowest overall dti ratio of 15%. California(CA) alone accounted for 13% of the total loan shares, followed by New York(9%), Texas(8%) and Florida(8%). These top states along with many others (Fig 1.1) had only 4-5% default rate of Grade A loans. However, as we go down the Grade level especially for G (Fig 1.2), surprisingly more than 50% turned out to be of bad type i.e, were either defaulted or charged-off, thus indicative of the fact that Grade plays an important role in determining Loan Quality.

We then looked into the percentage variation in Interest rate based on each grade level with respect to different loan purposes (Fig 1.3). It was seen that moving from just Grade A to Grade B led to 3-4% increase in the interest rate and among all the loan purposes, Debt Consolidation and Credit Card had the highest share of bad loans among all categories and so Lending Club can ensure strict monitoring on these loan categories to curb revenue losses.

For our next goal of identifying patterns within young people and loan terms(Fig 1.4) we found a striking variation of 3-4% in interest rates across 36-month and 60-month loan tenure and so almost all preferred taking loans with lesser installment periods. Moreover, in contrast with our initial assumption of interest rate fluctuation with years of employment, we observed no significant difference in terms of interest rate across employment years, except for a slight decrease in the interest rate for 10+ experience level. Summarizing the results so obtained, we thus refute our initial claim that young people being new in the job will tend to take loans of higher installment periods and so will end up paying more interest on average. This way focusing on business expansion plans based on attracting young customers will not be much impactful for LendingClub.

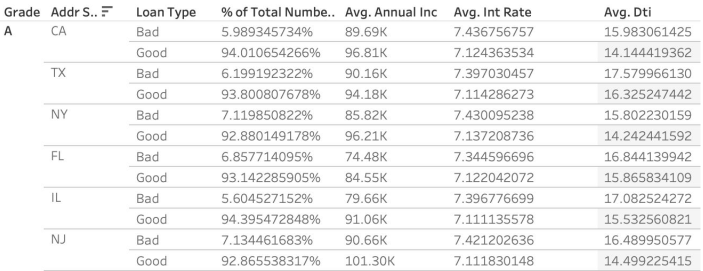
While performing a visualization between grades/subgrades and interest rates using the box plot(Fig 1.5), we can observe a linear increase in interest rates as the grades worsened. This strongly suggested that any model consisting of this feature should be considered a low-risk model, as there is a direct relationship that required no data transformation or feature engineering to conclude interest rates. However, the most interesting point about the plot is the extreme outliers in all of the grades. This made us wonder why would there be loans issued with interest rates way outside their respective whiskers. Could it be because of some internal contacts? Or could it just be a data anomaly?. So we started investigating these outliers, and interestingly we noticed that majority of the outliers were people with an experience more than 7 years and many open credit lines, this suggests that it might be possible that they might be in contact with authorities at Lending club which gave them an out of the box interest rates. But since the data is masked to maintain the privacy of the customers, to deduce any further relation among these outliers we will have to apply unsupervised techniques to generate concrete results about the correlation between these issued loans at uncharacteristic interest rates. However, we plan to estimate the amount of interest lost by Lending Club due to these loans.

The first phase to answer if borrower attributes can be used to satisfactorily estimate the interest rate at which a loan might be offered for a particular borrower, we kicked off by creating a subset of features by selecting borrower characteristics before a loan is applied as documented by Lending club. Post creation of the dataset, we applied forward selection with a max number of variables as 20 from possible 40 values to determine the best model to predict our response variable which is the interest rate. However, due to a huge number of factors in the employment title feature and various date columns like the earliest credit line, issued date, etc, we were running out of computational power. After dropping these columns from our subset, we were able to derive the model. The RSS and adjusted R^2 score generated for this model was not so significant as depicted in figure 2.1. Thus we decided to add new features to our subset which impacted interest rates significantly like grades and subgrades. The model generated after considering our new variables reported significantly better values as shown in figure 2.2. Thus, the next phase for our prediction will involve predicting grades of the loan application based on borrower characteristics.

For our second phase of building a classification system to distinguish between good and bad loans, we filter our dataset for completed loans only as we do not know the end state of current loans. Finally, after all the filtering, we are left with 8598 clean records with a distribution of 75% good vs. 25% bad loans for our classification engine. We created an 80% - 20% train - test partition of our dataset. Since we are faced with more than 100 variables, our next step was to attack the dimensionality reduction problem. Due to a large number of features, it wasn't feasible for us to use the best subset selection. With the existing variables and forward selection method, the best features obtained were dummy variables pertaining to categorical features with high variability like zip codes, employment titles, earliest credit lines, etc which weren't very informative. Since these were variables with very high variability (employment titles had roughly 5000 unique values), we decided to drop them from our dataset from a classification point of view. Also, the methods of the best subset, forward and backward selection are computationally heavy and tedious. Therefore, as a next step, we decided to use Lasso regularization with logistic regression to select features. As seen in fig. 3.1, the left dashed vertical line indicates that the log of the optimal value of lambda is approximately -11, which is the one that minimizes the prediction error. The exact value of lambda is observed as $2*10^{-5}$. The lasso model using this optimal lambda value generates coefficients for 114 out of 213 features (after one-hot encoding of categorical variables). The model attains an overall accuracy of 99% and specificity of 99% in our test set (observed from figure 3.2) making it a highly flexible and seemingly overfit model. Since the model is still built on a high number of predictors, there is a need to further investigate the reduction of features to generate a less complex and more interpretable model.

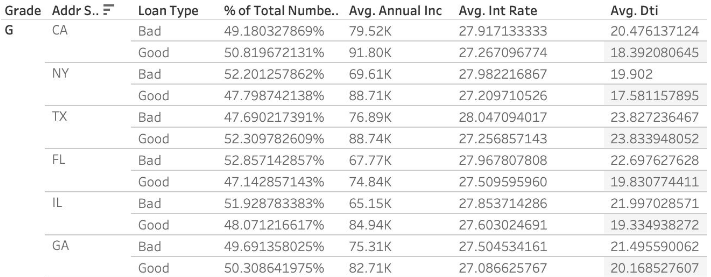
As future steps, we plan to use Principal Component Analysis (PCA) for dimensionality reduction. Currently, we have a class imbalance of 75:25 and so for model reliability, we intend to use bootstrap techniques for upsampling. We would also like to explore different classification techniques like logistic regression, decision trees, random forest, and support vector machines in conjunction with the resampling techniques to conclude on the best approach to a model with an optimal bias-variance tradeoff in our next milestone.

Our observation of the strong relationship between the grades/subgrades and the interest rate also motivates us to analyze a new perspective of using borrower characteristics to assign grades/subgrades to loan applicants and then use their conjunction to make predictions of interest rates in our next milestone.



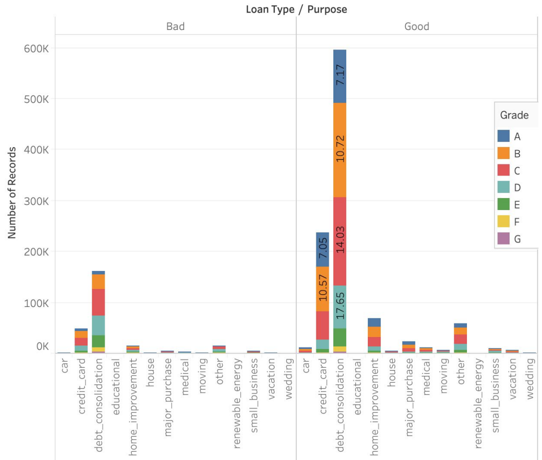
Grade	Addr S..	Loan Type	% of Total Numbe..	Avg. Annual Inc	Avg. Int Rate	Avg. Dti
A	CA	Bad	5.989345734%	89.69K	7.436756757	15.983061425
		Good	94.010654266%	96.81K	7.124363534	14.144419362
	TX	Bad	6.199192322%	90.16K	7.397030457	17.579966130
		Good	93.800807678%	94.18K	7.114286273	16.325247442
	NY	Bad	7.119850822%	85.82K	7.430095238	15.802230159
		Good	92.880149178%	96.21K	7.137208736	14.242441592
	FL	Bad	6.857714095%	74.48K	7.344596696	16.844139942
		Good	93.142285905%	84.55K	7.122042072	15.865834109
	IL	Bad	5.604527152%	79.66K	7.396776699	17.082524272
		Good	94.395472848%	91.06K	7.111135578	15.532560821
	NJ	Bad	7.134461683%	90.66K	7.421202636	16.489950577
		Good	92.865538317%	101.30K	7.111830148	14.499225415

<u>Fig 1.1:</u> Grade A loan Distribution wrt to bad and good loan

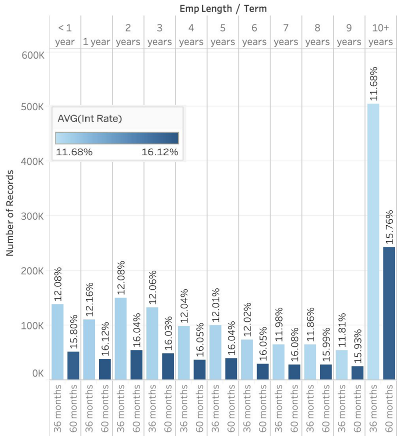


Grade	Addr S..	Loan Type	% of Total Numbe..	Avg. Annual Inc	Avg. Int Rate	Avg. Dti
G	CA	Bad	49.180327869%	79.52K	27.917133333	20.476137124
		Good	50.819672131%	91.80K	27.267096774	18.392080645
	NY	Bad	52.201257862%	69.61K	27.982216867	19.902
		Good	47.798742138%	88.71K	27.209710526	17.581157895
	TX	Bad	47.690217391%	76.89K	28.047094017	23.827236467
		Good	52.309782609%	88.74K	27.256857143	23.833948052
	FL	Bad	52.857142857%	67.77K	27.967807808	22.697627628
		Good	47.142857143%	74.84K	27.509595960	19.830774411
	IL	Bad	51.928783383%	65.15K	27.853714286	21.997028571
		Good	48.071216617%	84.94K	27.603024691	19.334938272
	GA	Bad	49.691358025%	75.31K	27.504534161	21.495590062
		Good	50.308641975%	82.71K	27.086625767	20.168527607

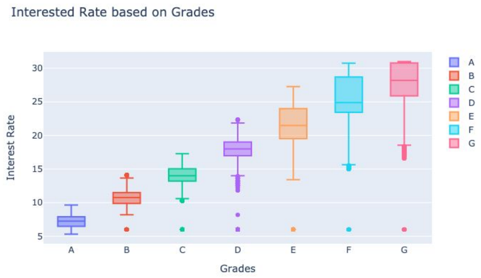
<u>Fig 1.2:</u> Grade G loan Distribution wrt to bad and good loan



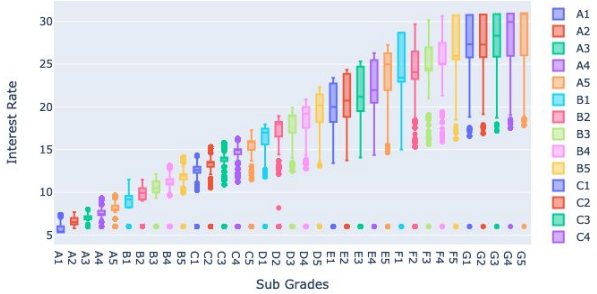
<u>Fig 1.3:</u> Loan Purpose vs Grade distribution and interest rate



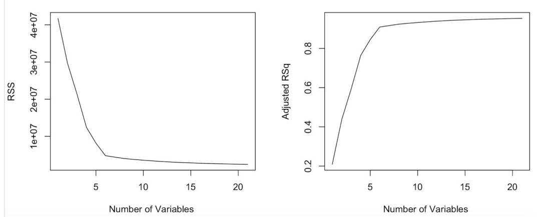
<u>Fig 1.4:</u> Term vs Employment years and interest rates



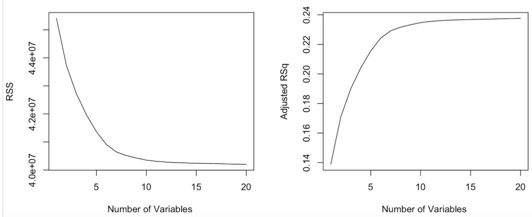
<u>Fig 1.5:</u> Distribution of Interest rates with respect to Grades



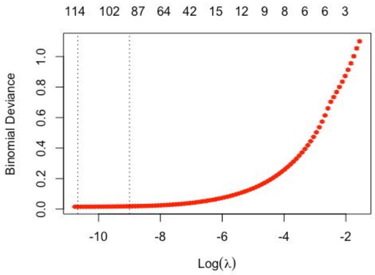
<u>Fig 1.6:</u> Distribution of Interest rates with respect to Sub-Grades



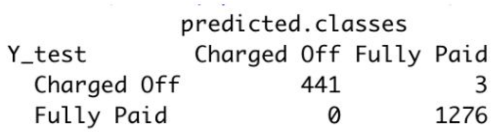
<u>Fig 2.1:</u> RSS and Adjusted R2 without grade and subgrade



<u>Fig 2.2:</u> RSS and Adjusted R2 with grade and subgrade



<u>Fig 3.1:</u> Lasso Regression for Good loan vs. Bad loan



```
              predicted.classes
Y_test         Charged Off Fully Paid
  Charged Off          441          3
  Fully Paid             0       1276
```

<u>Fig 3.2:</u> Confusion Matrix