Introduction

Lenders are in the business of taking risks. They make loans to debtors with a probability of not collecting the expected total payments and perhaps even losing some of the principle. The interest rate is the expected gross profit reaped by a lender. The greater the perceived risk of a debtor defaulting on a loan, the higher the interest rate. Risk is widely and crudely measured by a credit grade, which is akin to a credit score. The perceived risk implied by a credit grade is a partial reflection of the interest rate the debtor is charged. Therefore, lenders are incentivized to take risks of varying levels. However, the lender has to carefully assess a pool of loan applicants to avoid harmful losses. Lending entities have to estimate the probability of default, and based off their estimation and perhaps other factors, decide as to whom loans are made. Here, we obtain data, explore the data, and build and evaluate predictive models which can be relied upon to make estimated predictions and assist in the decision making of lenders.

The data used includes over 42,536 observations, or loans, and was obtained from the Lending Club website. It was wrangled, cleaned, and formed into a dataset catered for exploration, inference, and to build predictive and classification models. The models were then tuned, tested, and evaluated with the goal of reducing defaults while doing our best at not decreasing fully repaid loans.

Data Wrangling

The dataset needed to be modified for our purposes. Columns were added and deleted and missing data and outliers were addressed. The first column created is called member id. It was made to track each loan as an account. A unique integer between 1 and 42536 was assigned to each loan. Columns were then deleted.

The columns which were first dropped were title, employment title, and description of the loan. These columns were removed because the data within these columns were of text data types that were informal and not uniform. They couldn’t serve as categories. Luckily, the dataset includes other columns which provide data concerning the purpose of each loan and employment. Columns that had to do with settlements and collections were also dropped because settlements and collections take place after a loan is made. Lenders and their agents would not know anything about settlements or collections of a loan prior to making that loan. The same reason led to deleting columns related to payments made during the life of a loan. Other columns were deleted for various reasons.

Columns that contained datetime data types were omitted. These include earliest credit line, last credit pull date, and last payment date. These columns were removed because using them adds complications which include establishing a date as a benchmark and calculating the time since that benchmark date. This may have been interesting to have done for last credit pull date because that may serve as a significant predictor. But it was decided against to reserve time and simplicity. Columns related to addresses and location of residence were also dropped. Months since last delinquency and record were also removed due to the amount of the data missing within those columns. Both columns were missing well over half of the data. The remaining columns included data that had to do with many of the deleted columns. Below is a list of the columns used and their data types.

Data columns (total 23 columns):

loan\_amnt 42535 non-null int64  
 funded\_amnt 42535 non-null int64  
 funded\_amnt\_inv 42535 non-null float64  
 term 42535 non-null category  
int\_rate 42535 non-null float64  
 installment 42535 non-null float64  
 grade 42535 non-null category  
 emp\_length 42535 non-null category  
 home\_ownership 42535 non-null category  
 annual\_inc 42535 non-null float64  
 verification\_status 42535 non-null category  
 loan\_status 42535 non-null category  
 purpose 42535 non-null category

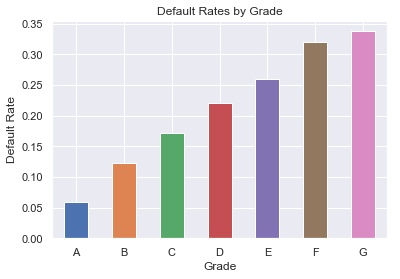
dti 42535 non-null float64  
 delinq\_2yrs 42535 non-null float64  
 inq\_last\_6mths 42535 non-null object  
 open\_acc 42535 non-null float64  
 pub\_rec 42535 non-null float64  
 revol\_bal 42535 non-null int64  
 revol\_util 42535 non-null float64  
 total\_acc 42535 non-null float64  
 total\_pymnt 42535 non-null float64  
 pub\_rec\_bankruptcies 42535 non-null object

The data that were of categorical type were converted into dummy variables for the purpose of building the predictive models. The loan status column, which contains categorical data, includes the values fully paid and charged off, defaulted. These values were of course converted into a dummy variables and the fully paid dummy variable was dropped. The charged off dummy variable is the variable we are predicting for. A value of 0 indicates a loan that was fully paid, and 1 identifies the defaulted loan.

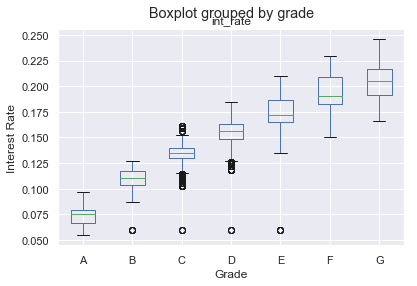
Columns that contain profit or loss margins and total profits or losses were added for exploratory purposes. These columns were not used for prediction. With a modified dataset, exploratory data analysis was performed to better understand the data.

Exploratory Analysis

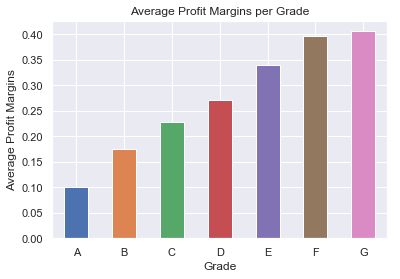
Lenders are most interested in participating in loans that will be fully repaid and avoiding those which may be defaulted on. The uncertainty associated with lending, or risk, influences the reward. The reward for a creditor in making a loan is the interest received. Generally, the greater the preconceived risk, the more interest is potentially earned. Therefore, creditors are incentivized to make loans to those who are perceived as more risky. Let’s see if this is the case within the data.



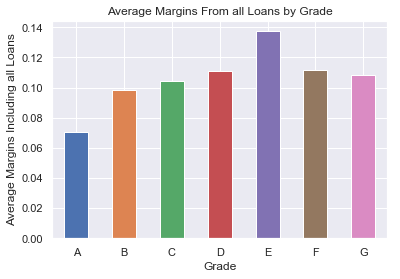
As mentioned in the introduction, the credit grade of a borrower is in part an estimation of risk. The graph above groups each grade and displays the proportion of those that default per each grade. It shows that A graded creditors are the safest and that the default rate per grade increases subsequently as we go from A to G. We can explore interest rates per grade to see if the assumption about the relationship between potential reward and risk is correct.



The plot above shows that the distributions of interest rates increase as perceived risk, as reflected by the grades, increases. It must follow that loans made to lower graded creditors that are fully repaid must be associated with higher profit margins due to the differences in interest rates. Let’s see if this is in fact the case in the data.



As seen from the illustration, the average of profit margins from fully repaid loans are in fact higher for lower grades, as explained by the distributions in interest rates for each grade. It is also important to explore how losses from defaults influence total profit margins.



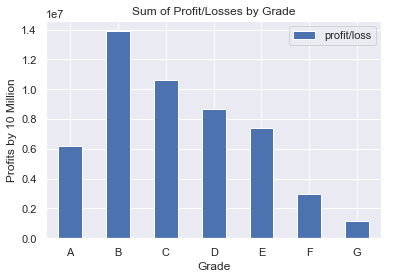
The differences in average profit margins per grade differ less when the losses are accounted for. The average margins from fully paid loans are greater for G than it is for F, and greater for F than it is for E. The differences in margins for those grades are reversed when losses are recognized. Also, the range in average margins when losses are considered is roughly from 0.7 to 0.14, while for only fully paid loans it is roughly from 0.10 to 0.40. It is ideal for the margins when losses are accounted for to be as close to the gross profit margins from fully repaid loans as possible. Let’s compare the margins, or percentages, in gains and losses respectively to gain context in this regard.



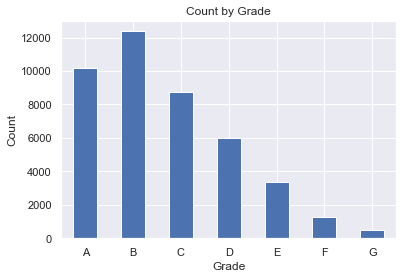
The visual just above shows that the average percentage losses from defaults have very little variation between grades. The average percentage loss among all grades is roughly four to five percent. That is larger than the average profit margin from fully paid loans from any single grade. The reason why lenders do not solely provide credit to lower graded borrowers to take advantage of those higher interest rates is because of their greater likelihood of defaulting. Looking at gains and losses in terms of dollars provides context in this regard. But before looking at absolute figures, let’s take a look at averages per grade.



The average gain from fully paid loans to lower grades is greater than that of higher grades. The losses increase also. However, the difference between average losses and average gains decrease as grade lowers. This again elaborates on the incentives for making loans to lower graded creditors, and also motivates lowering defaults across all grades. The later is because of the difference between average lose and average gain for higher grades and because of higher default rates for lower grades. Also, the loan sizes on average, tend to increase for lower grades. It is now worth looking at absolute gains and loses.



The plot above shows the sum of all returns and losses. If a loan resulted in a return, the amount of the return was added, if a loan resulted in a loss, the amount of the loss was subtracted. The plot tells us that absolute profits peak at grade B and decrease from there. It moves inversely compared to margins, or percentages. We can take a look at the count of borrowers per grade to see if that may contribute to explaining this fact.



Lenders in this dataset appear to be cautious. They made a lot less loans to lower grades. This is reasonable, given that the aim of a lender is to make loans that provide returns and that higher graded debtors are more likely to repay fully. However, there clearly is incentive for making loans to those perceived as more risky. There is also motive to decrease defaults among all grades for reasons aforementioned. With these things in mind, a lender would be very interested in decreasing defaults while taking on enough risk to reap the benefits from well informed lending decisions.

In the next section, we perform statistical inference to investigate if the differences in defaults rates, interest rates, and the percentages of profits and losses between grades are statistically significant. If the differences are significant, predictive models that are segmented by grade may prove to be more reliable. This is discussed later.

Statistical Inference

Given the differences shown in the exploratory section between grades, it is worth investigating the statistical significances of these differences. If there are statistical differences in several regards then it may be reasonable to construct predictive models for each grade and compare them to a model that contains all of the grades. The default rate between the grades appear to be different. The default rate of each grade is also a measure of risk, which is the underpin in terms of potential reward and weather or not it is reasonable to continue with a segmented approach. The alpha in these analyses is 0.05.

A function was written to calculate the p value for the differences in default rates between grades given the null hypothesis that the default rates come from the same distribution. As it turns out, the p values are all less than 0.0 except for that between grades F and G, the two lowest grades. The p value between F and G is 0.24, which suggests that the default rates likely come from the same distribution. However, as mentioned, there are statistical differences between all of the other grades in terms of risk, or default rates. Let’s take a look at interest rates to see if these differences in default rates translates to greater potential rewards.

When examining the statistical differences between the grades in terms of interest rates, we find what may be expected. The p values between the grades are all less than 0.0. This suggests that none of the distributions of interest rates of the grades share the same distribution. They are all significantly different. It is now known, that each of the grades are different in terms of risk and potential reward. We now consider percentages of losses and gains.

The p values for the differences of the distributions of the percentages of the returns and losses are less than 0.0 between most of the grades. There are exceptions. The p value between B and C is 0.014, which still suggests statistical significance in their differences. B and G, and C and D have p values of 0.02 and 0.04, respectively. Interestingly, the p value between C and G is 0.17, which suggest no difference between these two grades. Also, the three lowest grades, D, E, and F all share the same distribution in this regard, except between D and E. Below, are the p values aside the grades being compared.

D F 0.0668  
 D G 0.4011  
 E F 0.2596  
 E G 0.1261  
 F G 0.2978

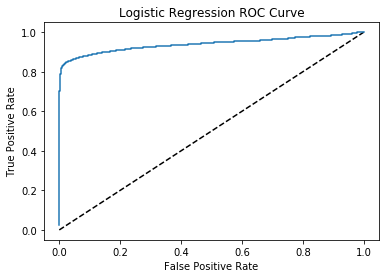
The statistical inferences suggests that there are significant differences, for the most part, in terms of risk and potential reward. However, the differences in reward are not fully exploited given the fact that C and G likely share the same distribution in terms of margins. A lender would want to have the distributions of margins of C and G to come from different distributions, given the differences in their risks. Now that statistical significance is granted in terms of their differences, it is pragmatic to approach a segmentation of grades when constructing predictive models. This is especially true given the practical differences as well. In the next section, in depth analysis is done on a segmented basis using logistic regression. Comparisons between an aggregate model (a model containing all of the grades) and the segmented models are made. A random forest model is built and evaluated as well.

In Depth Analysis

The in depth analysis consists of machine learning techniques with the aim of predicting the future status of a loan.The status of a loan will either be fully paid or charged off, defaulted. As aforementioned, this predictive analysis was performed using logistic regression and random forest algorithms. The data however, needed an additional minor preparation before commencing in the predictive task.

To prepare for the in depth analysis using machine learning techniques the data set was split into two, the features set and their corresponding labels. The collective features are called X, while the corresponding labels are called y. The features are all the variables except for the loan status, while the labels contain the status of each loan. Once this preparation was completed, the logistic regression containing all grades was built, followed by segmented models which were compared to the aggregate model. This is followed by the construction and evaluation of a random forest model.

To insure that the evaluation of the models are not influenced by chance, cross validation was utilized using five folds. Cross validation was also used to optimize the models, as is discussed later. The data was split into five folds and each fold was used as the test data as the remaining folds were used to train, or fit, the model. This was done for each fold. The scores of the folds were then averaged. The first scoring metric that was used was the area under the receiver operating characteristic curve to evaluate the aggregated model (the model containing all of the grades.) The average of the five scores for this model is 0.9435 without tuning, which implies that the true positive rate to false positive rate ratio is relatively high as shown by the graph.



This model was then optimized also using cross validation.

To tune the model, the best C value and penalty was sought after. These are both regularizations that combat the potential of overfitting the model to the training set due to coefficients that are relatively high. And as aforementioned, cross validation was used in this process. The best C value for this particular model is 8.483428982440725e-05. The penalty which best minimizes overfitting is l1, which sums the square of all the coefficients. This sum is then multiplied by C, and this product is added to the error function. The score measures the extent to which this modified error function is minimized. With these hyper parameters, the model was again tested using cross validation, which lead the an average area under the ROC curve of 0.9403, slightly below the untuned model. However, when examining the precision, recall, and f1 scores of the models, we get a different result.

The precision, recall, and f1 scores were obtained for the test features dataset and the predictions that resulted from them. These metric figures are shown below. As we can see, the hyper parameters improved the recall of defaults by a point. The f1 score also improved from 0.87 to 0.89. Optimization makes the mode more reliable. It is also worth noting, that the model had a recall of 1.0 for loans that were fully paid. This is important because lenders are interested in making as many successful loans as possible and the model does maximize that potential. The model of course, misclassified some loans as fully paid by overestimating the amount, but that may be worth the cost given the fact that it correctly identifies all of the fully paid loans in the test set. A lender that is too cautious may leave money on the table.

Untuned Model

precision recall f1-score support  
 0 0.96 1.00 0.98 9063  
 1 0.98 0.79 0.87 1571

avg / total 0.97 0.97 0.96 10634  
  
 Tuned Model

precision recall f1-score support

0 0.97 1.00 0.98 9063  
 1 0.99 0.80 0.89 1571  
 avg / total 0.97 0.97 0.97 10634

The dataset was divided among the grades to construct the segmented models. Cross validation was again used for each grade to evaluate the performances of the models. Interestingly, the average area under the ROC curve between all of the untuned segmented models is 0.9617, compared for 0.9435 for the untuned aggregate model, which had the larger area under the curve between the versions of the aggregate model, tuned and untuned. What is more of interest however, are the precision, recall, and f1 scores for the segmented models after optimization.

For the segmented models, the best C value and penalty was searched for. This was also done with cross validation and the grid search. What is interesting is that for all the segmented models, l1 was the best penalty expect for the A grade model, which is l2. The tuning improved the estimated models, but especially the lower grades. Below are the precision, recall, and f1 score metric figured for each of the tuned models.

A precision recall f1-score support  
  
 0 0.99 1.00 1.00 2387  
 1 0.99 0.91 0.95 159

avg / total 0.99 0.99 0.99 2546

B precision recall f1-score support  
  
 0 0.98 1.00 0.99 2709  
 1 1.00 0.84 0.91 389

avg / total 0.98 0.98 0.98 3098

C precision recall f1-score support  
  
 0 0.96 1.00 0.98 1790  
 1 1.00 0.83 0.90 395  
  
avg / total 0.97 0.97 0.97 2185  
  
D precision recall f1-score support  
  
 0 0.95 0.99 0.97 1172  
 1 0.97 0.83 0.89 332  
  
avg / total 0.96 0.96 0.95 1504  
  
E precision recall f1-score support  
  
 0 0.92 0.98 0.95 608  
 1 0.94 0.80 0.86 241  
  
avg / total 0.93 0.93 0.93 849

F precision recall f1-score support  
  
 0 0.91 1.00 0.95 221  
 1 1.00 0.79 0.88 105  
  
 avg / total 0.94 0.93 0.93 326  
  
 G precision recall f1-score support  
  
 0 0.97 1.00 0.98 83  
 1 1.00 0.93 0.97 45  
  
 avg / total 0.98 0.98 0.98 128

Interestly, the A grade model estimation has the greatest disproportion between fully paid and defaulted loans with 5% being defaults. Yet, the differences in the metric figures between fully paid loans and defaults are the most minimal for the A grade compared to all of the other segmented models, further relaxing the need for upsampling. The segmented models are then compared to the tuned aggregate model, which has higher f1 scores than the untuned model, to investigate if there is justification in their use.

There are mixed results when making the comparisons just mentioned. The A, B, and G grade model estimations perform better than the aggregate estimated model in terms of f1 scores by a mere point or two. The A and G grade models identify the defaults in their respective test set noticeably better with a recall of 0.91 and 0.93 respectively, compared to 0.80. The average f1 score for the C grade model is identical to that of the aggregate model, with a recall of defaults of 0.83. Let’s take a close look at the lower graded model estimations.

The D, E, and F graded estimated models seemingly did not perform as well as the aggregate model. However, it is still possible that they may be useful. To find this out, the aggregate model can be tested against all D graded loans from which metric figures can be obtained and compared to those of the segmented estimated model. This is not done here, but can be done in a subsequent project. A lender, or agent, would like to keep the recall of fully paid loans as close to 1.0 as possible at the expense of precision. But would also like to simultaneously reduce the recall of defaults are much as possible. These two goals must be balanced along with risk and potential reward. To get further assistance from predictive models we turn to a non parametric one, the random forest.

A random forest model was first built without hyper parameters. The data was split into the training set and the testing set, with 25% of the data being reserved for testing. The following figures make up the classification report of the untuned random forest model.

precision recall f1-score support  
  
 0 0.93 1.00 0.96 9063  
 1 0.99 0.57 0.72 1571  
  
avg / total 0.94 0.94 0.93 10634

As seen, the recall is much lower than the logistic regression models, however, the random forest model still performs relatively well given its tendency to overfit. The model was tuned in order address this overfitting.

The random forest model was also optimized with cross validation and a grid search for five hyper parameters. The first hyper parameter tuned was the amount of estimators, or trees, in the forest. It was discovered that 400 trees was optimal for the random forest model. The max depth was then tuned to none. The trees in the forest had no limits and were grown to purity or until two samples were used for a split. The maximum amount of features considered for each split is eight, and the better criterion is entropy. With tuned hyper parameters, the optimized random forest model has a recall of 0.63, compared to 0.57 without tuning. The following is the classification report of the tuned random forest model.

precision recall f1-score support  
  
 0 0.94 1.00 0.97 9063  
 1 1.00 0.63 0.78 1571  
  
avg / total 0.95 0.95 0.94 10634

The logistic regression models have higher recall rates for defaults, but the non parametric model in the random forest it still effective in its predictions as it would have reduced defaults by 63% if the estimated model existed and was relied upon when making the loans in the dataset. These model estimations would serve a lender well for it would reduce defaults significantly if all factors not accounted for in the dataset remained equal. These models would also maximize loans which are fully paid, according to their evaluations.

Conclusion

The estimated models give good foresight to lenders when making lending decisions by estimating the probabilities of defaults and using those probabilities to predict weather a loan will be fully paid or defaulted on. In working towards the construction of the model the data was prepared, explored, and used to make inferences that led to motivations behind the in depth analysis. Successively, the predictive models were built then optimized.

The dataset contained many irrelevant features that needed to be discarded, and missing values that prompted to be filled. For example, features related to settlements and collections contained information that would not be known by a lender prior to making a loan, therefore any effect from those features would be of no use. Missing data was filled with methods that made sense on a case to case basis. For example, missing annual income data was filled with the mode, to avoid influences by outliers, the amount of time an applicant has been at their current employer was filled with the forward fill method, and so on. The cleaned data set was then explored to provide insights leading to the in depth analysis.

As we saw earlier, the grades had average defaults that increased as the grades became lower. The A grades were the safest and the G grades were the riskiest. But with greater risk comes greater potential reward, as the lower grades paid higher interest rates. Lenders in this data set however, were weary of lending to higher risk grades. Yet, lower grades provided greater opportunity for higher profits. The profit margins from lower grades were from distributions that significantly and practically differed those of higher grades. This can serve as motivation for finding those with low default probability relative to their grades. These differences inspired the decision to build estimated models by grade and compare those with a model estimation containing all of the grades.

In depth analysis was then performed to estimate predictions concerning the result of loans. The segmented estimated models, on average, had greater areas under the ROC curve than the aggregate model estimation, which suggest that the segmented models may serve well for ranking each applicant by grade. For A and B grades, the ability to find potential defaults improved when compared to the aggregate model. However, for most of the lower grades with higher interest rates, the aggregate model seemingly performs better, except for the lowest grade, G. But, that does not mean that those models are not useful.This will be the topic of a future project. The random forest algorithm was also used for prediction with interesting and decent results. The random forest model performed well given the algorithm’s tendency to overfit.

It may be worth noting however, that these models are not perfect. They assume that factors that are not accounted for in the dataset will remain unchanged. Also, the dataset only contains data on loans that were initially accepted. There is no data available from loans that were rejected. However, a lender can still rely on these estimated models, as they will almost certainly reduce defaults significantly while maintaining a high recall on fully paid loans.