Mira Kasari, mkasari
Sairathan Rajuladevi, srajulad
Kelly McManus, kellymcm
Cole Thomas, nhthomas

April 7, 2022

Case Study - Phase 2
95-828: Machine Learning for Problem Solving

Throughout this phase, we will be cleaning and preparing the data for modeling.

Step 1
The zipped data files were unzipped and merged into one data frame named ‘final_data’. This will be our core dataset for the analysis.

Step 2
Many different removal techniques were used to reduce our dataset to hold only clean records.

- Payback date: Remove loans where there is missing information on the payback period. This will affect our calculations for returns so we dropped any record where it was less than or equal to zero.
- Loan status: Remove all loans that are still current. Keeping only loans with statuses 'Fully Paid', 'Charged Off', and 'Default'. While we recognize that according to the first phase documentation, the only final statuses are ‘Fully Paid’ and ‘Charged Off’ (see pg. 4 of assignment 1), we included ‘Default’ per phase two’s instruction set.
- Issue date: Remove loans that were issued before 2010.
- Null values: Remove loans that have null values in any of the numeric features. These features are required for prediction.

Step 3

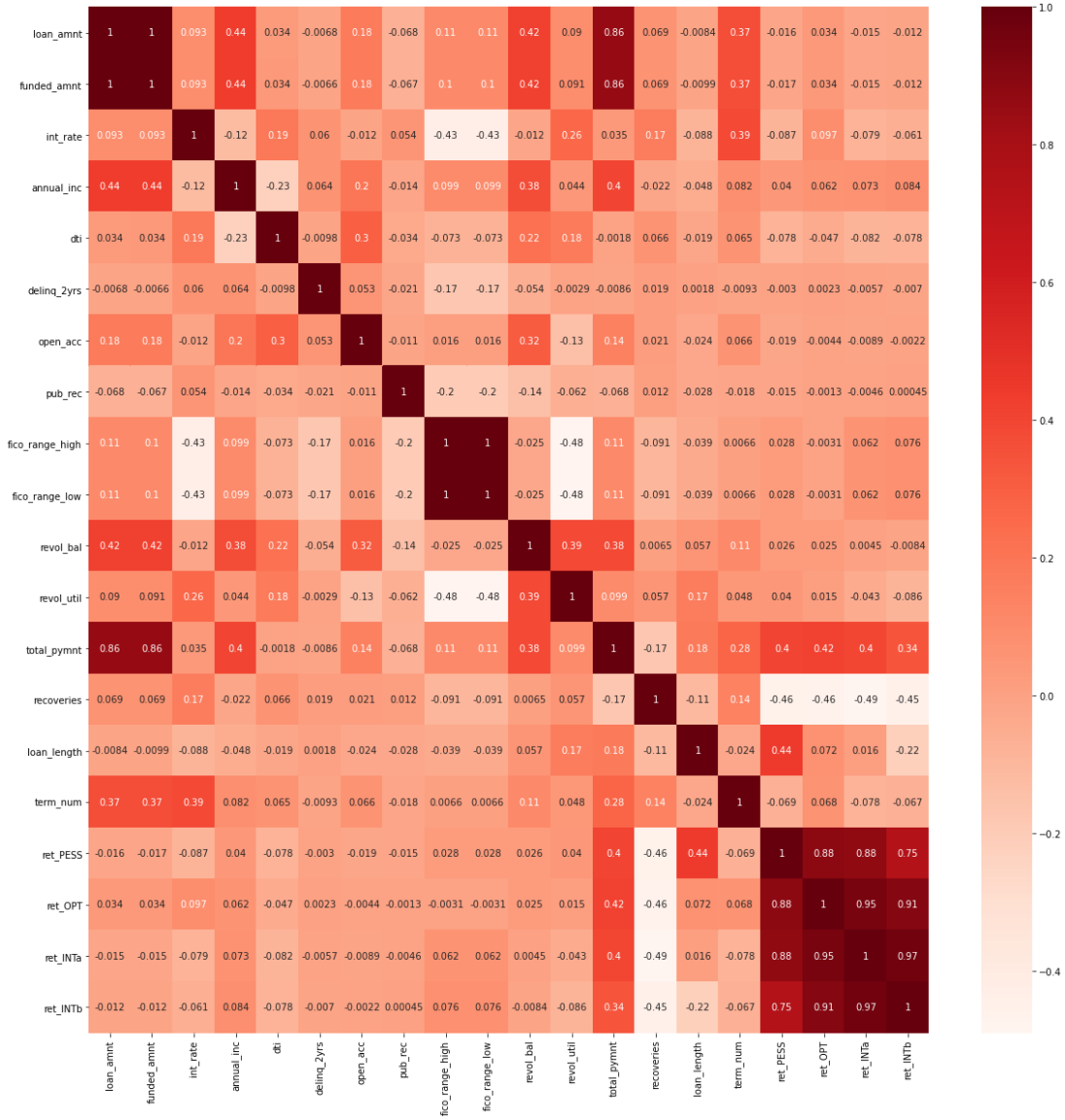
We employed different visualization techniques to analyze the features in the file. Namely, we plotted each float, categorical, and date column individually. After visually inspecting the features, we found that a subset of these features needed imputation and as well as outlier removal.

For imputation, we performed a business logic test on the variables. Two variables that caught our eye, and are interrelated, are DTI and annual income. There were only a few data points that had values of zero for these variables, however, we imputed the median income for the dataset. After all, does it make sense that a loan would be issued to someone without any income? This

risk strategy reflects the days of old with NINJA loans (no income, no job) and is largely not used by US financial institutions after the financial crisis of 2008.

This imputation not only serves to help us potentially better reflect the true nature of the underlying borrower, but also aids us when performing outlier detection. By shifting these ultra-low values to the median, we are able to ever so slightly change the distribution of the data and perform outlier elimination. The target columns we decided for outlier elimination were 'annual_inc', 'dti', 'open_acc', 'total_pymnt', 'recoveries', 'revol_bal', and 'revol_util'. We picked these columns by analyzing the visualization mentioned earlier. We removed any data point that was three standard deviations above or below the mean value from our dataset and reinspected our charts. The charts look much less skewed now.

Step 4



The above heatmap shows correlation values between all features in our dataset. Dark red means features are highly correlated, while white means features are not correlated. Some observations to note are as follows:

- There is a high correlation between ‘loan_amount’, ‘funded_amount’, and ‘total_pymnt’. This is likely due to the fact that we are only looking at non-current loans that have completed their loan life-cycle. We would expect that non-current loans were fully funded and were mostly paid off and we can see this by the high correlation values between these variables.
- There is a strong correlation between our calculated return rates because they have similar formulations and use many of the same features in their calculations. Similarly, ‘total_pymnt’ shows signs of correlation between the return rates. This value is used in the calculation of the return rates, but is more correlated with the value than other features we used, such as ‘funded_amnt’ and ‘loan_length’. When it comes time for model construction, we most likely cannot include the ‘total_pymnt’ column as a feature as this would result in data leakage, especially when we are trying to choose which loans to invest in.
- Loan amount, funded amount, and annual income show a strong positive correlation. Logically, it makes sense that individuals with high annual incomes will get approved for larger loans.
- On the other hand, there are some values that show a strong negative correlation, in particular, FICO scores and interest rates. Generally speaking, this makes sense as the higher the FICO score, you can expect to get a lower interest rate on your loan as you present less risk to the financial institution
- There is also a strong negative correlation between recoveries and the return rates. This makes sense logically because recoveries are only typically generated when there has been some sort of credit event on a security, such as a charge off or default.
- Lastly, the two FICO score features are perfectly correlated with one another because they are showing the low and high values of the same FICO score. We can consider removing one of these variables for modeling or splitting the difference between the FICO scores for modeling.

Overall, only a few features in our dataset show strong correlations that we should consider before modeling. Depending on the modeling technique we choose, the model itself may identify these correlated variables and remove them (ie. Lasso Regression feature selection). We will need to perform additional data preparation steps such as feature transformations and/or ingesting external complementary data sets to help us with our prediction task.

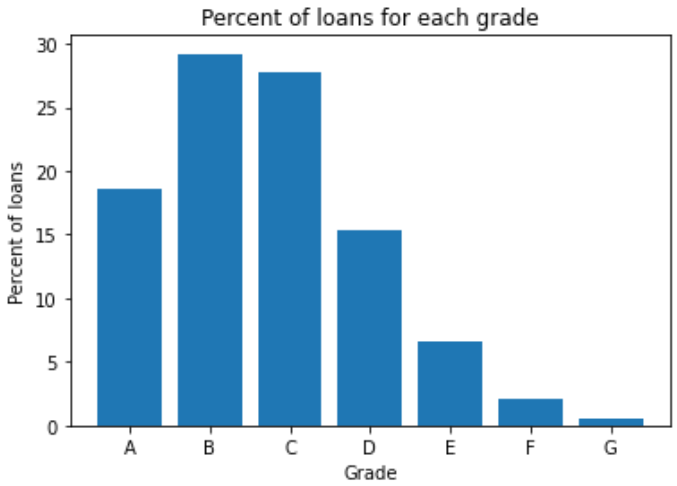
Step 5 & 6

The dataset was cut down to only include columns that are important to modeling our outcome. The data will be saved in a pickle titled 'clean_data.pickle' in the same folder where the Jupyter notebook is.

Step 7

(i) What percentage of loans are in each grade?

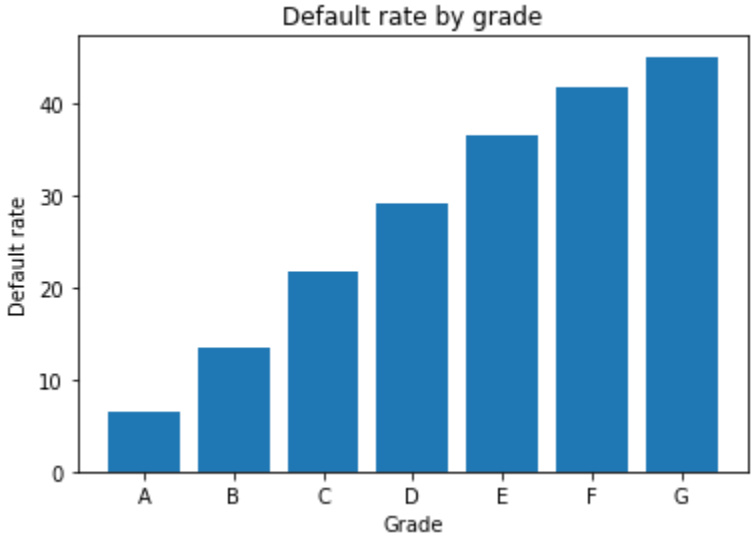
	perc_of_loans
A	18.539689
B	29.205654
C	27.763913
D	15.337104
E	6.531939
F	2.095674
G	0.526027



The above distributions show the percent of loans for each grade. We can see that the majority of loans are grades B or C, while very few loans are in grades E, F, and G.

(ii) What is the default rate in each grade? How do you interpret those numbers?

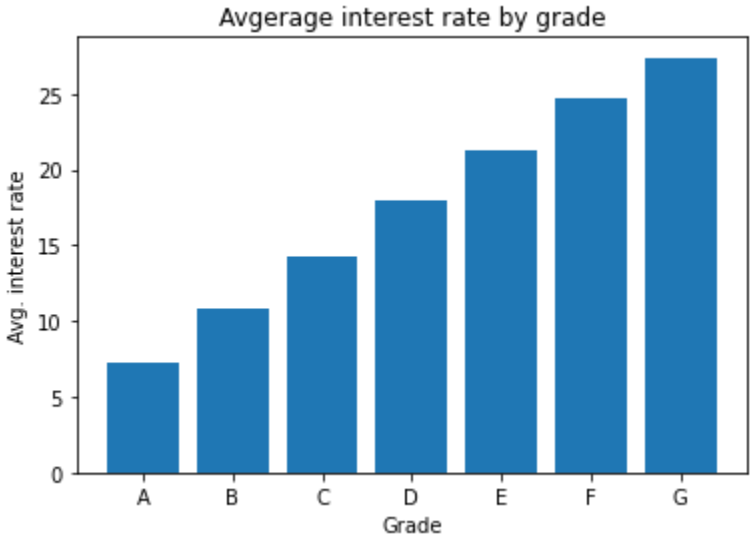
	perc_default
A	6.476934
B	13.461626
C	21.839563
D	29.096783
E	36.587276
F	41.742393
G	45.154899



The default rate for each grade is displayed above. The distribution shows an increasing tendency to default as the grade value increases alphabetically. A grade "A" loan has a chance of defaulting of 6%, while a grade G loan has a chance of defaulting of 50%. These stark differences are something that we must consider when choosing loans. As seen in part i, very few loans are in the categories E, F, and G and we now see they have high-interest rates. With only this information, they would seem to be unattractive loans, however, these grades often have the highest interest rates. Depending on the buyer's level of risk, these loans may be attractive.

(iii) What is the average interest rate in each grade? How do you interpret those numbers?

	avg_int_rate
A	7.219852
B	10.885972
C	14.205114
D	17.988371
E	21.227476
F	24.767634
G	27.401644



In the real world, the interest rate is a representation of the risk presented by the borrower, market conditions, collateral, and the underlying product. In our data, we clearly see that for a downgrade in credit rating, the interest rate increases. In the traded products domain, AAA bonds typically yield lower than CCC bonds. This is represented in our data by looking at the increase in interest rates from A-G.

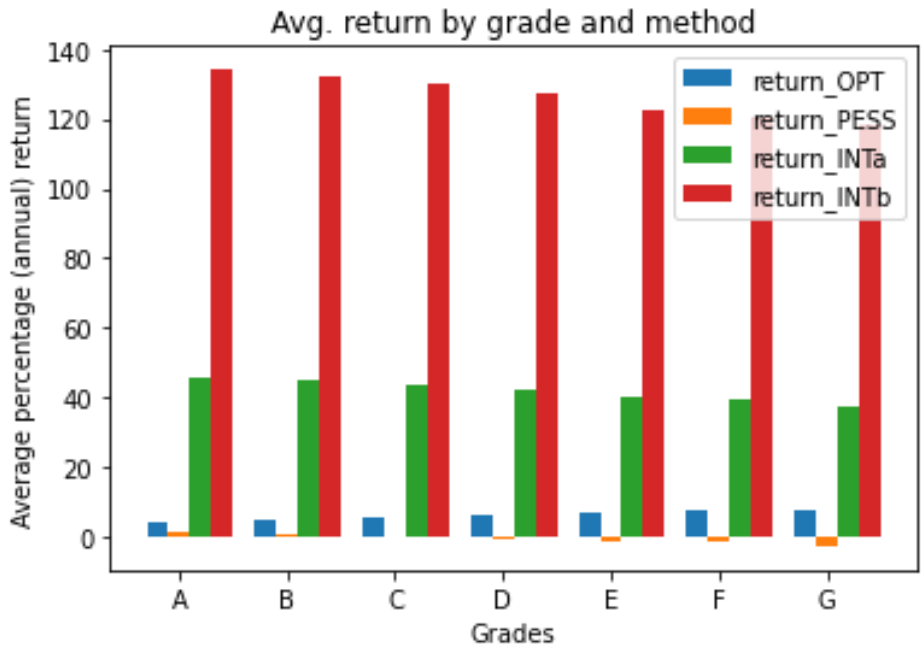
Pulling data from the bond benchmarks in the Wall Street Journal, we can actually see this trend in current market conditions for current yields (see below). According to Investopedia, the yield on new issue debt security reflects interest rates when they were issued, which is applicable to our use case. AA corporate bonds are yielding 3.28% while Baa corporates are yielding 4.010%.

	CLOSE	% CHG	YTD TOTAL RETURN	52-WK % CHG	YIELD (%), 52-WEEK RANGE			SPREAD, 52 WEEK RANGE		
					LATEST	LOW	HIGH	LATEST	LOW	HIGH
Broad Market Bloomberg Fixed Income Indices										
U.S. Government/Credit	2444.06	-0.88	-7.34	-5.15	**3.070**	1.240	3.070	n.a.	33.00	56.00
U.S. Aggregate	2091.64	-0.87	-6.94	-5.37	**3.070**	1.340	3.070	n.a.	29.00	50.00
U.S. Corporate Indexes Bloomberg Fixed Income Indices										
U.S. Corporate	3121.36	-1.00	-8.58	-5.51	**3.750**	1.910	3.760	n.a.	80.00	145.00
Intermediate	2892.71	-0.59	-5.95	-4.90	**3.530**	1.260	3.530	n.a.	58.00	122.00
Long-term	4473.87	-1.67	-12.61	-6.45	**4.140**	2.900	4.290	n.a.	117.00	185.00
Double-A-rated (AA)	621.44	-1.13	-8.87	-5.58	**3.280**	1.650	3.280	n.a.	47.00	93.00
Triple-B-rated (Baa)	831.06	-0.99	-8.79	-5.43	**4.010**	2.110	4.030	n.a.	100.00	172.00
High Yield Bonds ICE Data Services										
High Yield Constrained°	497.06	-0.24	-4.75	-1.11	**6.123**	3.796	6.447	328.00	303.00	422.00
Triple-C-rated (CCC)	483.77	-0.09	-3.96	0.13	**9.929**	6.304	10.490	715.00	579.00	830.00
High Yield 100	3315.30	-0.33	-4.57	-1.19	**5.652**	3.162	6.055	272.00	247.00	370.00
Europe High Yield Constrained	332.69	-0.02	-4.68	-3.37	**4.319**	2.304	4.733	392.00	101.00	494.00
Global High Yield Constrained	434.60	-0.15	-5.56	-4.01	**6.252**	3.968	6.731	396.00	343.00	499.00

(iv) What is the average percentage (annual) return per grade (as calculated using the three methods in part 6.)? (Assume two different yearly rates for M3: (i = 0.023) and (i = 0.04))

The raw values were multiplied by 100.

	return_OPT	return_PESS	return_INTa	return_INTb
A	3.986987	1.201130	45.577675	134.107600
B	5.121760	1.152820	45.014553	132.149286
C	5.657940	0.264614	43.763157	130.057631
D	6.367915	-0.354900	42.404218	127.204594
E	6.800662	-1.001098	40.568384	122.744257
F	7.624870	-1.253729	39.460259	120.123653
G	7.994172	-2.586403	37.707052	118.320612



The average percentage (annual) returns per grade are shown above. Looking at the optimistic return rate, we see that these values are greater than the other three return rate methods. The return_OPT shows that the value increases as the grade lowers from A to G. This is expected as the optimistic outlook would project the lower, riskier grades would do better than a safe grade, like “A”, due to its optimistic calculation. This is in contrast to the other return methods that show a decreasing average return rate as the grade lowers.

When analyzing the 3rd method of calculated returns, we see a similar story unfold to that of the pessimistic rating method. The returns decrease as we decrease in grade. Based on the formula provided, we can interpret this in the context of the high default rate and low residual capital leftover to compound in future periods. We can also see the sensitivity of the 3rd method’s return calculation with respect to interest rates. As we increase the reinvestment interest rate (part B), the overall return increases, as expected. This being said, picking an accurate reflection of future interest rates is critical if we were to use this method as our evaluation criteria. In the market today, we are expecting further hikes in the Federal Funds Rate through the remainder of 2022, which further complicates this calculation and presents an increased interest rate risk for fixed income securities. Please see the last paragraph of the next question for additional commentary on the returns for method 3.

(v) Do these numbers surprise you? If you had to invest in one grade only, which loans would you invest in?

If these bonds were actually tradeable in the secondary market, we see that B has a decent return and a low default rate. Another factor this analysis did not consider is liquidity. It is expected that the risk teams at LendingClub carefully optimize their portfolios and match market demand with market supply. So, if we did need to liquidate our position or perform a portfolio reallocation, there would most likely be a substantial amount of buyers looking at this class. Due to this, we would not incur high execution costs in the form of blown-out bid-ask spreads by the market maker (LendingClub).

Overall, the return values do surprise us. The optimistic values still show relatively low levels of return, while the method 3 return rates show extremely high return values. In-depth debugging was conducted to ensure the methods were calculated properly, but the outcome has persisted. As we continue with the project, we will continue to validate method calculations and ensure all features and outcome variables are accurate before modeling.