Week 2 Report

Kristina Schiffhauer, Sagar Dubey, Hewitt Trinh, Jinwoo Oh, Stefanie Montgomery

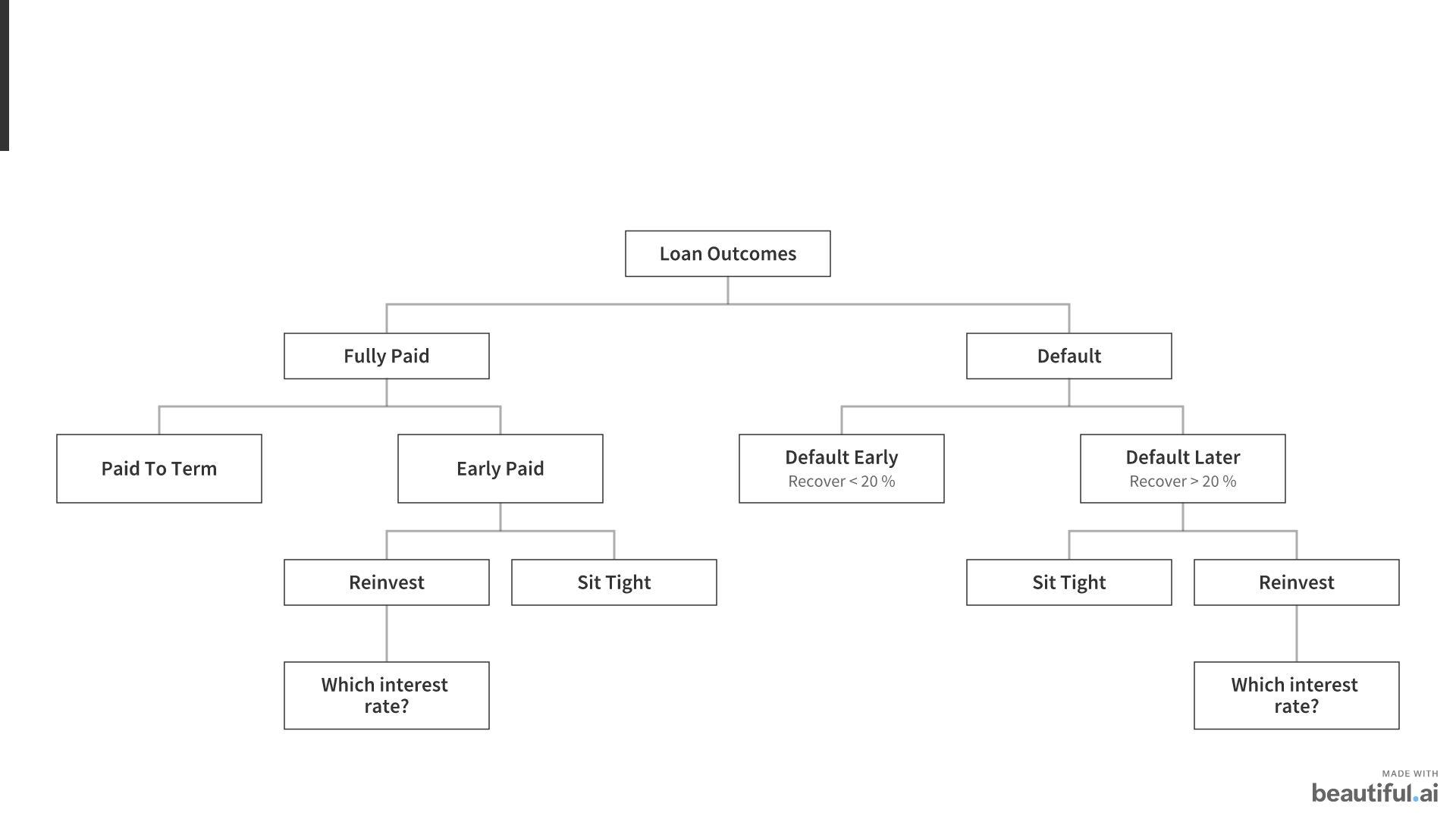
## 1. Multiple variable summary and visualization

This week, we explored multiple variable summaries and visualizations. We saw that loan interest rate rises as the loan grade increases from A to G, clearly because the risk of default rises across the loan grades. We also explored how total payment is related to other features, as it is a strong measure of the loan quality. Unsurprisingly, it has a strong correlation with funded amount, and also installment. Higher total payments also tend to increase chances of paying off the loan fully (fully paid status), whereas lower payments tend to favor charging off a loan. Total payment amount tends to be higher for 5-year loans, rather than 3-year loans. Though total payment amount is highly correlated with loan status, it is not a predictive feature as it is not known at loan investment time.

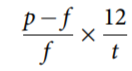
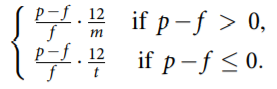
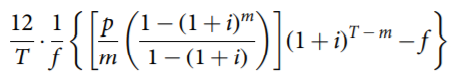
Curiously, we were unable to find any relation between FICO scores and credit span (time since the first credit line). As expected, a high FICO score is associated with lower interest rate and revolving balance utilization. Also, funded amount and loan amount have a very strong correlation, so we were careful to use only one in downstream analyses.

## 2. Calculate return on loans

As mentioned in our report last week, there are several possible outcomes for loans once funded by investors. Ultimately the return on each loan depends on assumptions about how an investor would respond to an investment outcome. The tree graph belows capture the complex nature of loan outcomes and investors’ response:

****

From the graph above, calculating the return is complicated by two edge cases: (1) how to account for defaulted loans, which are usually partially paid off, and (2) how to account for loans that have been paid early. There are three obvious methods to simplify and convert this dynamic problem into a static one:

* Method 1: 
* Method 2: 
* Method 3: 
* *p* is the total amount repaid and recovered from the loan
* *f* is the total amount invested in the loan
* *t* is the nominal length of the loan in months
* *m* is the actual length of the loan in months

Following is our analysis of the different methods:

|  |  |  |  |
| --- | --- | --- | --- |
| **Methods** | **Description/Assumptions** | **Advantages** | **Disadvantages** |
| **M1 (Pessimistic)** | Assume once the loan is paid back, investor is forced to sit with the money until the term of loan ends | -Handle defaults gracefully without inflating the negative returns  -Treat 5-yr and 3-yr loans in the same ways if loans go to term | -Hardly realistic as investors often reinvest in other types of assets  -For loans that are repaid early, this favors short-term loans because the gain is spread over shorter period  -For loans that default early, this method favors long-term loans as the loss is spread out over greater period |
| **M2 (Optimistic)** | Assume once the loan is paid back, investor can immediately invest in another loan with exactly the same return | -Simple and takes into account that the funds would (or could) be reinvested  - Treat 5-yr and 3-yr loans equally and regardless being of repaid early or going to term | -Not realistic to assume cash repaid early to be reinvested at the same rate  -If a loan defaults early, it assume investor will invest in a loan that lose at the same rate |
| **M3 (Fixed Time Horizon)** | Consider a fixed time horizon (T months= 60*)* and calculate the return on investing in a particular loan. Assume any revenues paid out from loan are immediately reinvested at a yearly rate of i%, compounded monthly | -Quite close to what would realistically happens  -Equalize all differences between loans of different lengths and correctly accounts for defaulted loans | -Undervalue the potential return of other alternative investments, especially on loans that are repaid early |

Therefore, to match a typical investor’s response as realistically as possible, we model our return calculation based on the loan outcomes and normal psychological responses:

|  |  |  |
| --- | --- | --- |
| **Loan Outcomes** | **Calculation Methods** | **Rationale** |
| **Paid to Term** | **M3 (T=60, i=5%)** | If a loan is performing well, a reasonable investor would invest in a relatively stable, medium returns assets like Bond ETFs |
| **Early Repaid** | **M2** | A loan is repaid early, investor would likely to experience euphoria and seek another loan of the same group |
| **Early Default** | **M3 (T=60, i=2.5%)** | A loan defaults, investor would likely to experience shock and become risk-averse and invest in the safest asset class at prime rate |
| **Default close to Term** | **M3 (T=60, i=5%)** | Throughout the horizon, investor would not know beforehand whether a loan would default close to term or paid to term, thus he/she is likely to invest on the same asset as Paid to Term category |

The default rate has a high positive correlation with M3, while having a highly negative correlation with the M2 and M3 return approaches. So, an investor may want to adopt the M1 or M3 approach, if they are predominantly investing in loan grades A, B, or C (which are less risky). On the other hand, if the investor wants to invest more in riskier loan grades such as D, E, F, or G, they may want to adopt the pessimistic or hybrid return approach. The hybrid return approach favors users investing in loan grades B and C. All the four approaches are plotted against the default rate below. Loan grades A - G are plotted left-to-right in the x-axis.

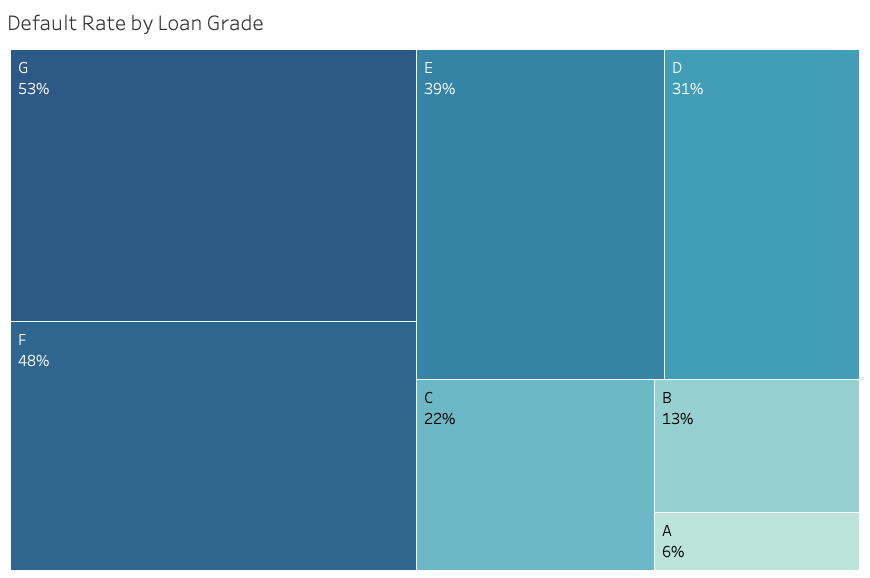
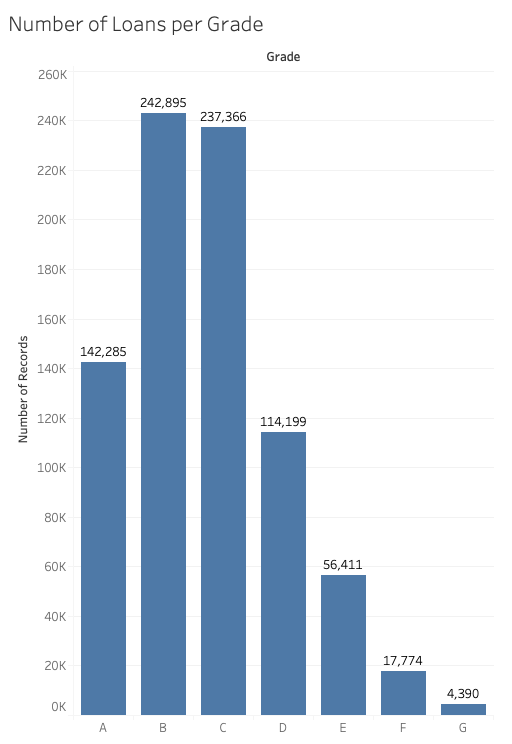
|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |

## 3. Diagnostic Analysis:

In order to understand the different loan grades, we explored how each behaves through a variety of metrics. The chart below shows, respectively, the percentage of total loans that are in each grade, the average default rate of each grade, the average interest rate of each grade, and the average percentage annual return calculated using the varying methods explored above.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Grade** | **Percent** | **Default** | **Interest** | **Return M1** | **Return M2** | **Return M3 (2.5%)** | **Return M3 (5%)** | **Return (Hybrid)** |
| A | 17% | 5.7% | 6.9% | 2.2% | 3.6% | 3.9% | 6.9% | 5.9% |
| B | 30% | 12.9% | 10.2% | 2.3% | 4.7% | 4.1% | 7.1% | 6.3% |
| C | 29% | 21.7% | 21.7% | 2.0% | 5.4% | 4.0% | 7.0% | 6.6% |
| D | 14% | 30.8% | 30.8% | 1.6% | 6.0% | 3.7% | 6.7% | 6.7% |
| E | 7% | 39.3% | 39.3% | 0.9% | 6.3% | 3.3% | 6.1% | 6.7% |
| F | 2% | 48.5% | 48.5% | -0.2% | 6.3% | 2.2% | 5.0% | 6.3% |
| G | 0.5% | 53.2% | 53.2% | -0.8% | 6.5% | 1.6% | 4.4% | 6.3% |

The charts below show the number of loans in each grade and the average default rate of each loan grade.

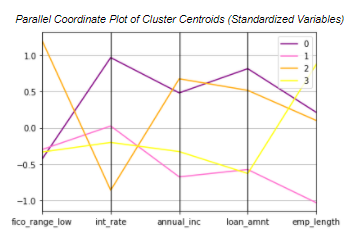


Overall, these numbers make sense. We see default rates increase fairly steadily as grade decreases, and higher interest rates for lower grade loans. There are some surprises, however. Using methods one and three, we see a slightly lower average return for grade A than grade B, and then the return rate gradually drops with the grade. Ideally, we would see the return be one-directional as the grade decreases, as we see mainly through method two. However, the average return doesn’t change much between grades E and F using method 2. This pattern deviation may be due to the smaller sample size of A, E, and F, and will be something to keep in mind as we move forward. At this time, it is difficult to conclude that one grade is the best to invest in. Since our objective is to maximize the return, and multiple methods of calculating the return rate show Grade B with the highest returns for this sample, higher interest rates than A, and lower default rates than the grades below, this may lead us to believe that Grade B would be the best to choose. However, there are other considerations that must be taken into account including whether our return calculations are the best to achieve our objective. We also may choose to follow the hybrid approach for Jasmin’s portfolio, and we will be exploring these avenues further in the weeks to come.

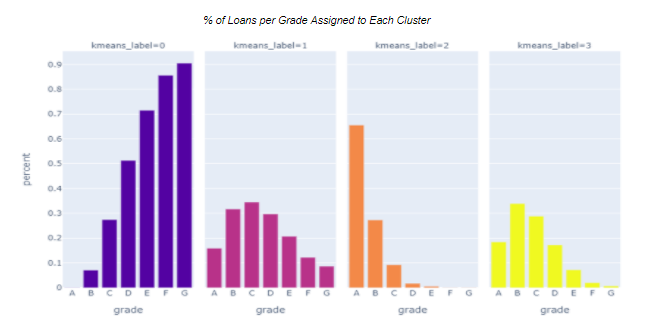
## 4. Descriptive Analysis

### K-Means Clustering

We applied K-Means clustering with the goal of reproducing the loan grade structure. We found that using the following (transformed and standardized) variables and four clusters produced the best separation by grade: "fico\_range\_low", "int\_rate", "annual\_inc", "loan\_amnt", and "emp\_length".



The two charts here summarize the K-means clustering results. Cluster 0 (purple) represents the low-grade loans. It contains a majority of the D, E, F, and G grade loans, and is distinctive due to its high interest rate and low fico score. Cluster 2 (orange), on the other hand, represents the majority of the A grade loans and is characterized by very high fico scores and low interest rates. Interestingly, both the very high and very low grade loans tend to have higher average income. This tells us that high income alone does not make a good borrower - in fact, high income combined with low fico score is a signal for a risky loan. Clusters 1 (pink) and 3 (yellow) are the most similar in terms of grade, and together contain most of the B and C grade loans. Cluster 3 skews slightly more toward lower grades (C and D), and the coordinate plot tells us that longer employment length and slightly higher annual income are differentiators. Cluster 1 skews slightly more toward C grade loans and is characterized by lower annual income and shorter employment length compared to cluster 3. This tells us that, all else equal, a longer employment length and higher income can result in a slightly boosted loan grade evaluation.



### PCA

After analyzing clusters with K means, we applied PCA to see how it identifies differently in terms of the loan grade structure. The same variables that were selected for K-Means were used. A line graph shows that 58% of variance is explained by the first two components, but more than 90% of variance is explained by 4 components. The table summarizes the result of the component loadings. Interest rate and annual income have the highest weights in the first component, whereas fico score and interest rate have the most impact on the second component. Overall, fico score has a strong impact on each component. Besides fico score, annual income, loan amount, and interest rate are the important factors to determine loan grade structures. Lastly, a biplot shows key variables in the first two components. It shows that fico score and interest rate are in the opposite direction. It makes sense since a high fico score usually results in a low interest rate, and vice versa.

