Credit Sesame Consumer Report

Audreya Metz, Calvin Ma, Hari Rajan, Paul Giroud

1 Introduction

Determining the appropriate consumer to give loans to is a difficult task that can depend on many factors. We analyzed data from Credit Sesame, a company that calculates credit scores to determine options for credit cards, mortgage rates and loans, to determine the best candidates for loans. Their dataset included three parts: User Demographics, First Session Information and 30-day User Engagement Data. The user demographics data included mostly credit profile information including loan histories, tradeline details, and also some personal information.The First Session data provided action logs of each user's first interaction with the Credit Sesame website. Finally, the 30-day user engagement gave insight into the actions of users in each one's first 30 days with similar features as the first session data. We believe that identifying accurate credit scores with both financial and non-financial data can be of use to Credit Sesame. If Credit Sesame could approximate credit scores from certain characteristics, they could choose better candidates to give loans to. Our ultimate goal was to see what variables were strong indicators of credit score and delinquencies. We also found it relevant to explore the engagement habits of certain demographics groups with the Credit Sesame website. To accomplish these tasks, we used 4 tools: PCA, Linear Regression, Lasso Regression and K-means clustering. PCA, linear regression and lasso regression were used with the user demographics data to explore which features had a significant impact on the credit score, and by how much. The k-mean clustering was used with both the user demographics and 30-day engagement to cluster similar users based on their engagement and then compare the difference between users of differing clusters. The credit score of an individual is heavily linked to their homeowner status, length of opened accounts, credit card limits, and number of accounts turned over to collection agencies. While Credit Sesame can use any/all of these factors to estimate individual's credit score, the results of our regression are not extremely convincing. The strength of our regression leads us to believe the results would be best used to refine an individual's calculated credit score or group individuals in larger credit score buckets for targeted advertising. Credit Sesame should also consider more targeted advertising on certain parts of their website after we've seen different customer traits based on what they use the website for.

2 Data set

2.1 Exploratory Data Analysis

HARI, PAUL, CALVIN ADD IN EDA. (We should mention that some variables are very correlated - the correlation matrix is easy to make) We started our exploratory data analysis by first looking at the distribution of each of the variables in the dataset, as shown below. As shown by the histograms of the column variables above, we noticed that almost all the non-categorical data was heavily skewed to the right as many of the variables had a high frequency for lower values and significantly lower frequency for higher values. The only non-cateogrical variables that were not heavily skewed were `age` and `credit_score`. This was important to keep in mind as we proceeded in our analysis as the distribution of each variable can impact the results of our analysis methods. We plotted the variables we were considering for our regression against one another to see if there was any collinearity between variables. From the plots we saw one concerningly strong correlations between open collection accounts and tradelines condition derogatory with an R value of 0.82. We ultimately eliminated tradelines condition derogatory from out model. The next strongest R value between our variables was -0.21 for tradelines derogatory and max cc limit both of which we ultimately ended up keeping within the model.

3 Methods

3.1 PCA: We used Principle Component Analysis because it is a powerful tool that can be used for both exploratory data analysis and also making conclusions about the dataset.We wanted some method to visualize the relationship between data points, but the dataset's dimensionality was too large to plot. Using PCA, we could plot Principle Components in order to have some sort of visualization of the data, and can offer some intuition of the relationships between data. We looked at the biplot and the loadings and saw how correlated various features are within each principle component, we used the correlated variables to determine which features to keep in or leave out of the regression.

3.2.2 Linear Regression: Based on our EDA and PCA results, we modeled a regression on credit score using the given demographic variables. Using our PCA results, we were able to to load the initial model with variables that ppeared to be correlated with credit score. This initial model had a decent R-squared of about .36 then we began to remove the least significant predictors to create a leaner model and ensure no colinearity. Every time a variable was removed we would check that the R-squared $> .30$. Finally, we reviewed our residual plot and found that we had some clustering indicative of colinearity in the model.

3.2.3 k-Means: K-means clustering allows us to find groups of similar datapoints. By dictating a number of clusters we expect to see, the algorithm will find the closest subset of datapoints for each of clusters. This model is useful because it gives us a definitive way to find which datapoints are related. For our dataset, we can take datapoints for different clusters and view them as a collective, and find the demographics of the clusters based on its datapoints.

4 Applications

4.1 Applications for PCA: After doing PCA on the user_demographics, the first thing that we looked at was the plot of PC1 vs. PC2. The plot is clearly just a large clump of datapoints (Figure SOMETHING). To find some meaning from it, we also plotted the biplot of the PCA. The biplot revealed much more information about the loadings and relationships between values. We can see that the entire first loading is negative. The second loading is much more interesting - there is a split between the PC2 loadings. It is difficult to interpret the meaning of the loadings as a whole, but we can see that the positive correlations easily differentiable from the negative loadings. Next, we looked at the loadings themselves and scree plot of the PCs. From the scree plot, we see that PC4 is the last significant PC that provides a large decrease in variance. Finally, we interpreted the Principle Components loadings by looking at PC1 through PC4. The values of the loadings correspond to the correlation coefficient of the variable with that principle component. What we are looking for in the loadings are values that are related to credit_score. From the first PC, we see that all of the correlation values are negative, some more than others. Credit_score has a loading of -0.23, which means it is moderately related to PC1, but we cannot make any conclusion with respect to the other variables. PC2 gives us much more information. The correlation for credit_score in PC2 is 0.29, and we can see other variables that are relatively large positive numbers and small negative numbers. Looking at the large positive correlations, we can see that tradelines_avg_days_since_opened, max_cc_limit, and total_mortgage_loans_balance are strongly positively correlated with credit_score. This means that as these values increase, we would expect to see credit_score increase as well. And we see that the negative loadings would be negatively correlated with credit score. We can continue down the line with PC3 and PC4, but the interpretability goes down as you analyze more principle components and they become less significant. From our PCA, we determined which features are correlated with each other, which provided us with useful information that was used for future analysis. Although the interpretability of PCA is lower than other models, we can still determine relationships between variables and datapoints.

4.2 Applications for Regression: From applications, we can see how our group removed the features that eventually led to our current model. Initially when we fit all of the variables, each one was determined to be significant (p value < 0.05). However, we did not want to have so many variables in our models, especially after seeing collinearity in our plots. This is why we decided to remove variables based on R-squared. Our current regression model indicates that being a homeowner, your number of open collection accounts,

your maximum credit card limit, and the average days since your account has been open, are the most significant indicators of credit score and future delinquent payments. We expected open collections to have a strong predictive relationship with delinquencies and credit score since open collections are caused by missed payments. Our results indicate that each account turned over to a third party for collections leads to a 5-point decrease in credit score. Homeowners tend to have a credit score that is 20 points above non-homeowners. This makes sense because homeowners tend to be more financially stable. Every dollar increase on a credit card limit tends to indicate a 0.0045-point increase in credit score. This is probably explained by the fact that people with better credit scores who are more financially stable are given larger credit card limits. We also found that with every day an individual has an account open, his credit score increases by 0.007135 points. This last finding is also logical because accounts that stay open are accounts that have gone longer making payments. Unfortunately, we only reached around 0.35 R-squared, which could be improved. The further limits of this regression lie in the skewed overall distribution of this data, which resulted in residual plots for the model that indicate that we must be wary using this model.

4.3 Applications for k-means Clustering: From our preprocessed data of clicks on different parts of the website, we start to run the k-means clustering algorithm. We found the within-sum-squares of k = 1 through 20 (Figure SOMETHING).Between 5 and 8 number of clusters seems to be where the amount where the within-cluster sum of squares difference starts to drop off. Therefore, we will use 8 clusters. Looking at FIGURE SOMETHING, we see that cluster 7 and 4 have relatively high values for click_count_credit_card and click_count_personal_loan, respectively. This means that the clustered groups tend to click on the these sites more often. The other clusters are all more centered to the origin. We then compare the user demographic for users of different groups. FIGURE SOMETHING shows all of the features between the two groups. We can see that in several categories there is quite a large difference between the mean variables of the two clustered groups. The users who click on the credit pages more often have a lower credit score, have inflated values that are associated with good credit, and the click_loan group tend to have higher loan amounts. From the k-means clusters, we were able to determine that the demographics of people who use Credit Sesame's website differently and access different pages are in fact different.

5 Discussion

Our PCA and Linear Regression Analysis both support the conclusion that longer opened accounts and higher credit card limits are two of the strongest indicators of credit scores. They also support homeowner status and number of accounts turned over to collection agencies as strong predictors of credit score. Our linear regression interpretation in the applications section defines our estimate for the strength of these indicators based off of how they are measured. However, the concerning linear cluster in out residual plot lead us to take the coefficients lightly. While we are confident that these factors are strongly linked to an individual's credit score, using them to estimate an individual's exact credit score could be problematic. We believe that it would be best to use our results to estimate more general credit score buckets for targeted advertisements since the precise effect of our variables is questionable.

Our k-means clustering concluded that Credit Sesame's website clientele that use the credit card links are different from the people who used the website for loans. Credit Sesame can use this information to more specifically target advertise credit cards to one group and loans to another. The clustering also concluded that people who look at their credit tend to have worse credit scores, and more specifically have more unpaid balances that negatively impacted their credit. Naturally the clustering also indicated that people who use the loan feature tend to have higher amounts of loans. All of this information can help credit sesame advertise to specific groups of customers in different areas of their website.

Future Work: Moving fowards, we can work to improve the current models that we used, namely the linear regression and clustering. In our linear regression, we starting modeling knowing that variables were highly correlated. Using other regression methods such as Lasso or ElasticNet would mean we wouldn't have to do feature selection by hand (See Appendix for Lasso result). For the k-means clustering, other clustering methods could be used such as hierarchical clustering. It is hard to say if they would produce better results, but they would give us another perspective on the data. We could also run a logistic regression using

the results of our linear regression to decide whether or not to give loans to individuals based on their demographics.