Team #99 Report

We believe that identifying accurate credit scores with non financial data can be of use to Credit Sesame. If Credit Sesame could approximate credit scores from certain characteristics they could then provide targeted advertisements for specific credit cards or loans. In addition to targeted offerings and advertisements, exploiting more statistically significant data when calculating credit score could lead to a more accurate credit score. 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 demographic groups with the Credit Sesame website. By analyzing how certain groups differ in website navigation we could provide insight to Credit Sesame on their customer base. Credit Sesame could use this when developing internal strategies.

Data Engineering Process

Preprocessing: To clean the user profile data and make it more manageable, we removed any rows with any NA's and removed any user with gender marked unisex. Due to the amount of unisex users, we assumed this was a result of a user not disclosing their gender and disregarded them, which in turn lowered our sample size to a more manageable number. When working with the PCA data, we wanted to cut down on the columns/features that we felt were unnecessary. These include: user_signup_timestamp, state, and zipcode. For other features, there seemed to be overlapping information that were redundant. Cases like avg_days and max_days were not both necessary, so we removed rows that represented the same information. Also, cases where information could be represented by total tradeline data, but also had information split between banking, credit, etc. also presented the same information, so rows were removed in that situation. Also, for PCA specifically, we dealt with the bucketed age and credit scores by parsing the min and max of the buckets, and simply setting the value to be the average of the min and max for the respective datapoint. This way, we didn't have to deal with numerous dummy variables.

PCA: We used PCA because it is a powerful tool that can be used for exploratory data analysis. We wanted some way to visualization the relationships between data points, but the dataset's dimensionality was too large to simply plot. PCA was perfect because we could use then use dimension reduction to plot all of our points on a two dimensional graph. We were hoping that we could see very clear groups, but we ended up have a large clump of data.

Linear Regression: While doing the PCA and working through the results, we wanted a method for determining the magnitude of the relationship between features. From PCA, we already know the correlation, but we were still lacking just how much features were "correlated" or connected with each other. So, we use linear regression to accomplish this. We started by using a full model based on variables that we observed to have a high correlation with credit score in PCA, then used a step-by-step process to remove insignificant variables and any collinear variables. While testing for the potential removable of variables that were significant based on p-value in order to create a leaner model, we ensured that our R-squared never dropped more than .1, and if it did, we would reinsert the variable that caused the drop.

K-Means Clustering: We decided to use k-means clustering on the user_engagement data in order to get a better understanding of how different subsets of users interact with the Credit Sesame website differently. We were able to do this by doing k-means clustering on the columns that corresponded with user actions on the website (click_count_credit_card, click_count_personal, click_count, mortgage, click_count_credit_repair, click_count_banking, click_count_auto_products).

Analysis

PCA: From our PCA, we could see a correlation between values are naturally related to people who seem more responsible and a negative correlation for those who are struggling to keep up with payments. More specifically, there was a strong positive correlation with age, amount of mortgage loans, credit card limit, total open balance, owns a home. These all relate to people who are older and have experience with paying back loans. And despite these people having significantly more mortgage loans, these people have proven to be able to pay them back. The negative correlations include derogatory accounts, increased number of collections, number of opened accounts and past due. These are indicators of people who have been able to pay back their accounts, or seem to be trying various methods to bring up their score.

K-Mean Clustering: From our K-Means, we saw that there is a different in mainly credit score between people who often click on the credit card pages and loan pages. Also, the ages are significantly different as well. This shows that the demographics between people who access the website and engage with it differ based on their needs with their credit, banking, loans, and mortgages. Due to the sparseness of our data, we only had clear clusters around credit card page clicks and loan page clicks.

Conclusion

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.

Using our findings and these characteristics we can fine tune credit scores as well as estimate a score.