Credit Sesame Consumer Report

Audreya Metz, Calvin Ma, Hari Rajan, Paul Giroud

Abstract

The analysis of the Credit Sesame data was done using various methods including principle component analysis (PCA), linear regression, and k-means clustering. It is important to not that the given data is shown to be heavily skewed to the right which affected much of the results of our analysis methods. The main focus of the study was to identify factors that impact an individual's credit score and to better understand how different users interact with the Credit Sesame website differently. From the principle component analysis and the linear regression, we were able to identify variables that were positively and negatively correlated with credit score which gives us insight as to what factors can help to predict an individuals credit score. In particular, it was found that homeowner status, length of opened accounts, credit card limits, and number of accounts turned over to collection agencies were heavily correlated to credit score. From the k-means clustering, we were able to distinguish between users that use the credit card links and those whoused the website for loans and understand how the demographics ofusers from these two groups differ (ie. loan amount, credit score, etc.).

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

We started our exploratory data analysis by first looking at the distribution of each of the variables in the dataset. We noticed that almost all the non-categorical data were heavily skewed to the right as many of the variables had a high frequency for lower values and significantly lower frequency for higher values. From the histograms (Figures 6, 7), we can see how most features have many zeroes with other values typically outside of the 75th percentile. 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. Although the sparcity of the data could negatively impact our data, we decided to keep them in because it is an inherent characteristic of the data. Knowing that we wanted to fit a linear regression of all the features with credit score, we plotted each feature vs. credit score for all users after preprocessing (Figures 4,5). From the plots, we saw that a few of the features had a clear linear relationship with credit_score. We used this information for feature selection later during the linear regression, making sure to keep features that had a clear relationship. However, we also had to be wary of the non-normal distribution of the credit_score - we can see the long tails of the qqplot of the credit_score (FIGURE 17). We also used the correlation matrix to help us determine what features were highly correlated with others (Figure 18). Since linear regression struggles with correlated variables, we kept the values that were highly correlated (correlation coefficient > 0.7) in mind, and we planned to remove these first during the linear regression.

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 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.

3.3 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 the biplot of the PCA (Figure 9). The biplot revealed much 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 (Figure 10), 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 (Figure 11). 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 (Figure 12) 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 (Figure 13) 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 14).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 15, 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 16 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.

6 Appendix

Preprossesing

We did different forms of preprocessing for each of the different methods (from working separately on different parts).

- To clean the user profile data and make it more manageable, we removed any rows with any NA's and removed users with gender marked as "unisex". We removed the unisex because there was an extremely large number of unisex users, which led us to believe that unisex represented the users who didn't want to choose which sex they were. We also removed the user_signup_timesteamp, state and zipcode features since we decided they were unnecessary. For other features, there seemed to be overlapping information that were redundant and colinear- Cases like avg_days vs. max_days were not both necessary since avg_days accounts for max_days, so we removed rows that represented the same information. Finally, in dealing with bucketed age and credit scores, we parsed the min and max of the buckets and had numerical values as the average of the min & max of the bucket. This means that we don't have to deal with numberous dummy variables.
- The regression required minimal preprocessing as any variabes that we felt were not useful or colinear could simply be excluded from the regression model. The bulk of the preprocessing done for the regression was in creating dummy variables for the categorical that were to be included in the model and converting categorical bins for age and credit score to numerical variables.
- The k-means clustering was meant as a way to cluster users based on their engagement. We decided that the best way to gauge engagement is based on the number of clicks per page. So, we took the 30 day engagement dataset and only took the features that represented clicks per various pages like credit cards, loans, etc. Although the dataset was very sparse, we didn't need to preprocess it any further.

Lasso:

Regular linear regression cannot work properly when features are highly correlated. However, Lasso is a form of linear regression that allows us to input a dataset with correlated variables. Lasso will account for these variables and remove them based on the lambda regularization. We used this method because our process of removing features by hand using the R-squared was slow and had a lot of room of error.

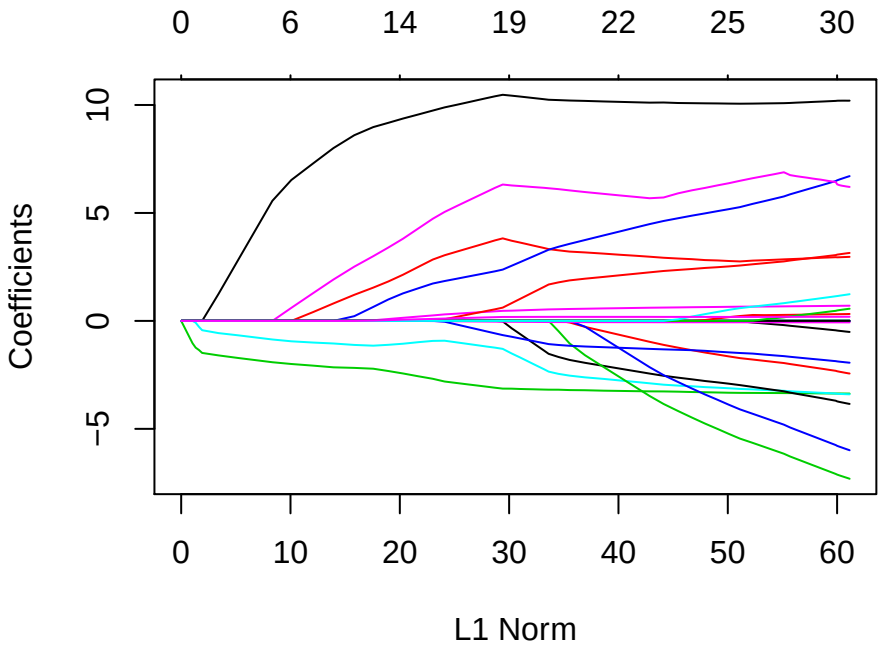

Figure 1: Lasso - Coefficient Values vs. Lambda Regularization

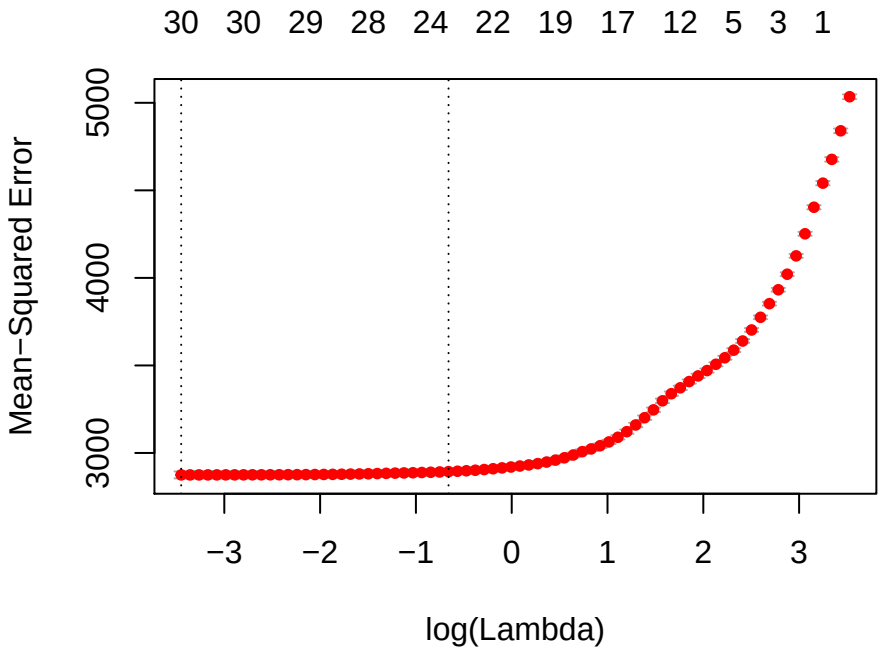

Figure 2: Lasso - MSE vs. Log(Lambda)

```
## 31 x 1 sparse Matrix of class "dgCMatrix"
```

```
##                                                    1
## (Intercept)                                5.781374e+02
## is_homeowner                               9.744147e+00
## gender                                     .
## tradelines_avg_days_since_opened           3.529964e-03
## tradelines_max_days_since_opened           1.316892e-03
## tradelines_min_days_since_opened           .
## count_tradelines_closed_accounts           2.536919e-01
## count_total_tradelines_opened_24_months    .
## count_tradelines_cc_opened_24_months       2.844563e+00
## count_tradelines_condition_derogatory     -2.683211e+00
## count_open_installment_accounts_24_months  1.733562e+00
## count_tradelines_open_collection_accounts -9.263168e-01
## count_tradelines_open_mortgages            4.735609e+00
## count_tradelines_open_student_loans        .
## count_tradelines_opened_accounts           1.072332e-03
## count_tradelines_open_secured_loans        .
## count_tradelines_open_unsecured_loans      .
## total_tradelines_amount_past_due          -2.540258e-04
## total_open_cc_amount_past_due             -8.392652e-03
## total_cc_open_balance                     -5.229511e-04
## total_tradelines_open_balance              .
## max_cc_limit                               3.793950e-03
## total_mortgage_loans_amount                .
## total_mortgage_loans_balance               .
## total_auto_loans_balance                   4.215533e-05
## total_student_loans_balance                .
## count_inquiries_3_months                   .
## count_inquiries_6_months                   .
## count_inquiries_12_months                 -5.175509e-04
## count_bankruptcy                           .
## age                                        9.223010e-02
```

Figure 3: Feature Coefficients for Lasso with Lambda = 4

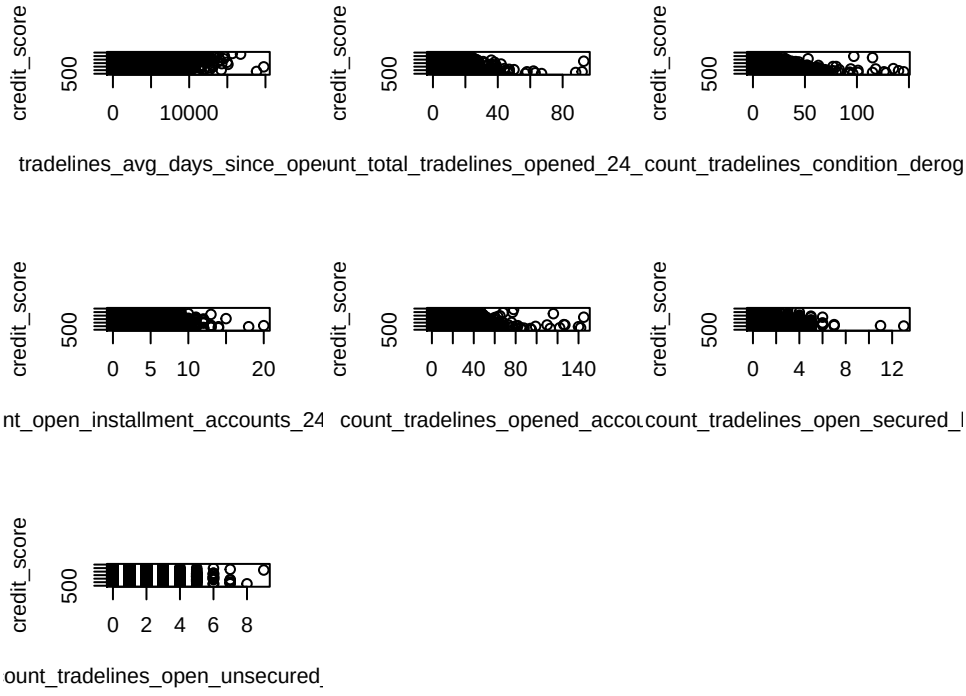

Figure 4: Scatter plot of Features vs. Credit Score Part 1

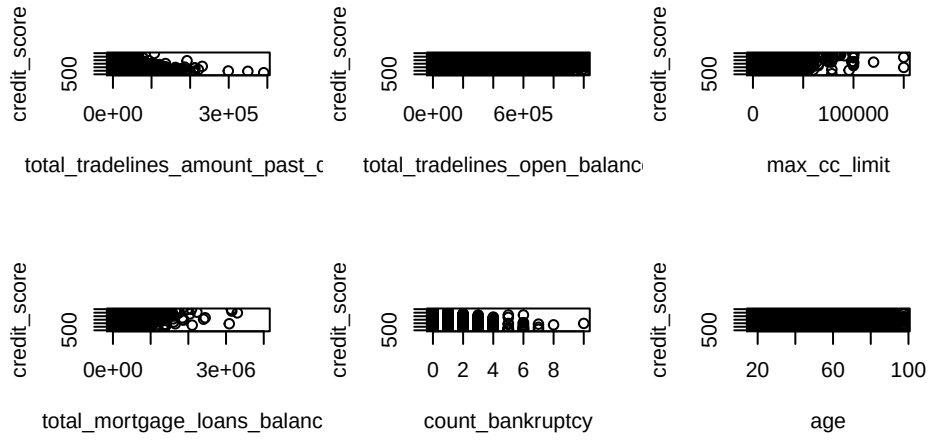

Figure 5: Scatter plot of Features vs. Credit Score Part 2

```
## Warning: attributes are not identical across measure variables;
```

```
## they will be dropped
```

```
##       key                                      value
## 1 user_id 50991631a5e7fafd8b5856fc15e3d1a3af5dcf98
## 2 user_id cabee62f0c4f26bb088f4a48d9ca5efa3a4f96e3
## 3 user_id 6da929725c76c01aa151d97060df2e6bd051e31e
## 4 user_id e8a6717452a88ec8d699c0a4181637c67d247e84
## 5 user_id 03c209fbb349633c40826a83874f92e302382b13
## 6 user_id ae0ebe7492c5af1fec00c8ecd59f83cc5a659fb2
```

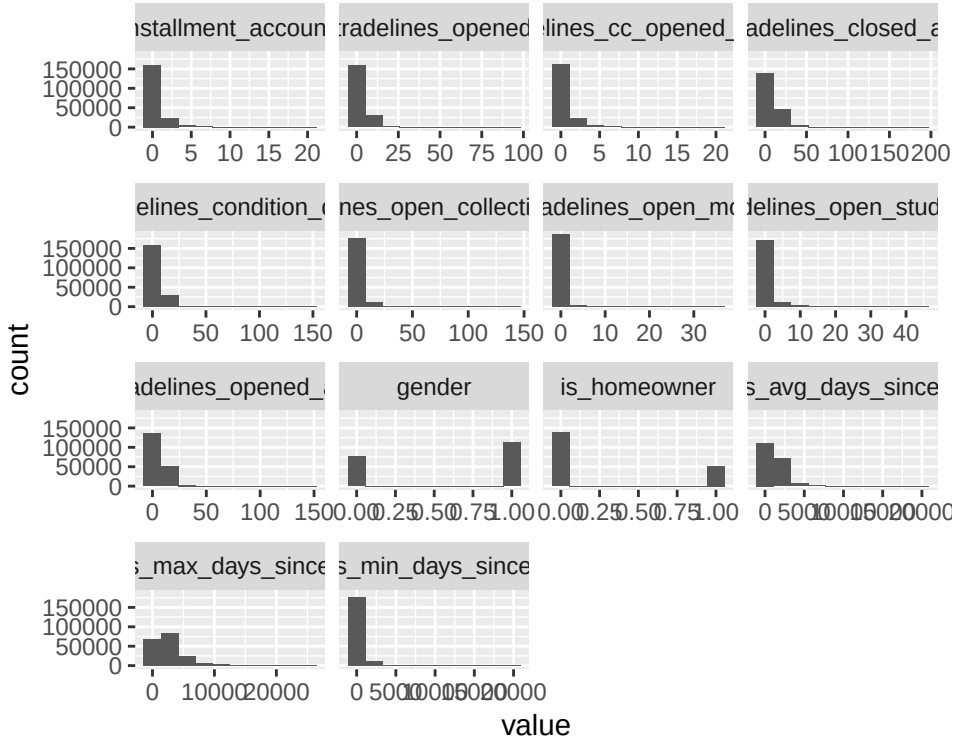

Figure 6: Histogram of Features Part 1

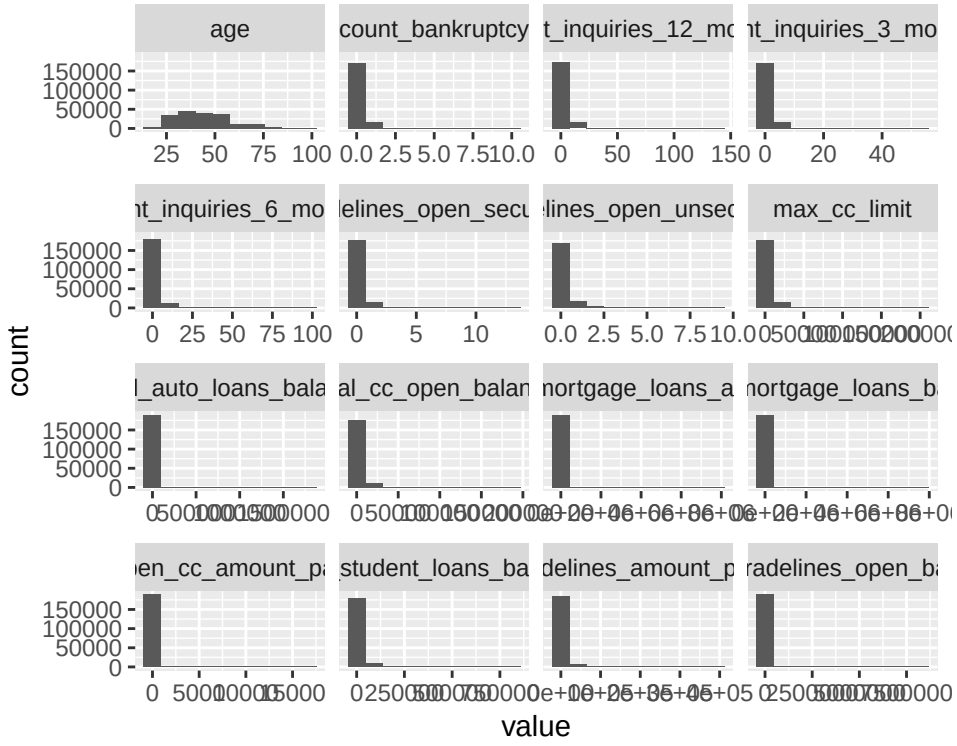

Figure 7: Histogram of Features Part 2

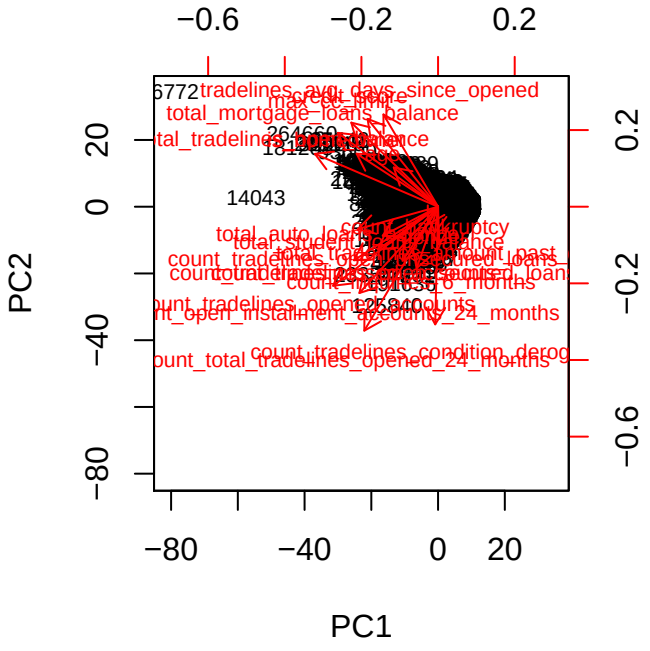

Figure 8: Biplot of PC1 and PC2

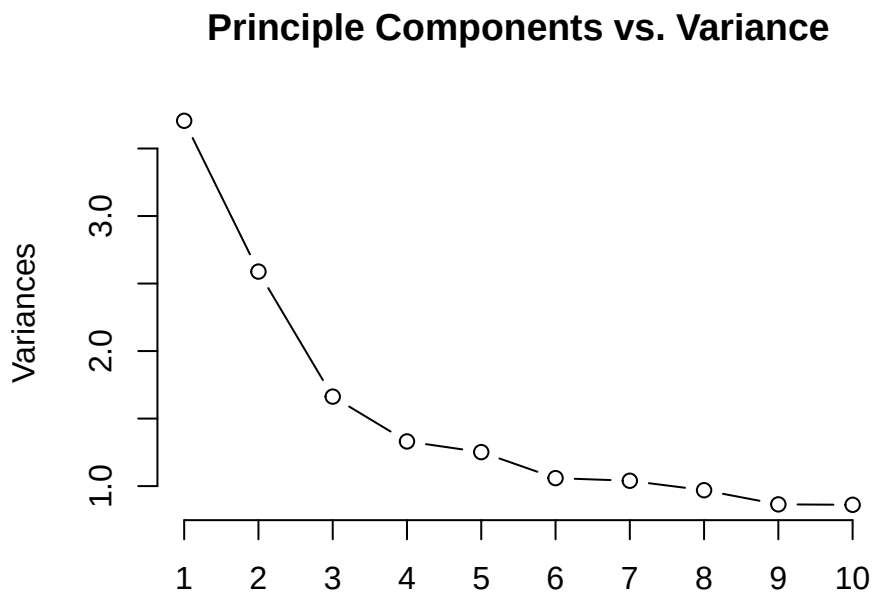

Figure 9: Screeplot of the PCAs

```
##                                                PC1         PC2
## is_homeowner                           -0.265749640  0.17305440
## gender                                 -0.010274389 -0.08327323
## tradelines_avg_days_since_opened       -0.178215179  0.30098061
## count_tradelines_closed_accounts       -0.274483231 -0.17461720
## count_total_tradelines_opened_24_months -0.240645508 -0.40437433
## count_tradelines_condition_derogatory  -0.009301021 -0.38383500
## count_open_installment_accounts_24_months -0.256623494 -0.28157493
## count_tradelines_opened_accounts       -0.342340376 -0.25795940
## count_tradelines_open_secured_loans    -0.115586155 -0.17626423
## count_tradelines_open_unsecured_loans  -0.196337473 -0.13971221
## total_tradelines_amount_past_due        0.004046851 -0.12392095
## total_tradelines_open_balance          -0.401728005  0.17239936
## max_cc_limit                           -0.283081287  0.27432410
## total_mortgage_loans_balance           -0.327026202  0.24094675
## total_auto_loans_balance               -0.261479615 -0.07301565
## total_student_loans_balance            -0.180515221 -0.09485020
## count_inquiries_6_months               -0.069988959 -0.20048977
## count_bankruptcy                       -0.033354933 -0.05813271
## age                                    -0.144567316  0.12806396
## credit_score                           -0.229915247  0.28650043
##                                                PC3         PC4
## is_homeowner                           -0.11400276  0.18499807
## gender                                 -0.09514876 -0.02343193
## tradelines_avg_days_since_opened       -0.24973575  0.11615267
## count_tradelines_closed_accounts       -0.29015770  0.21957401
## count_total_tradelines_opened_24_months  0.06936418 -0.04979841
## count_tradelines_condition_derogatory  -0.46932016 -0.05413354
```

```
## count_open_installment_accounts_24_months  0.37651626  0.07721219
## count_tradelines_opened_accounts          -0.09670775 -0.09930531
## count_tradelines_open_secured_loans        0.14120492  0.30923177
## count_tradelines_open_unsecured_loans      0.22036394  0.23597644
## total_tradelines_amount_past_due          -0.41837417 -0.07128415
## total_tradelines_open_balance             -0.05549340 -0.37604110
## max_cc_limit                               0.01678692  0.03489727
## total_mortgage_loans_balance              -0.07544896 -0.33184781
## total_auto_loans_balance                   0.22543807 -0.07189954
## total_student_loans_balance               -0.06203506 -0.28662612
## count_inquiries_6_months                   0.14411127  0.01653901
## count_bankruptcy                          -0.15005839  0.32376684
## age                                       -0.25775070  0.47478425
## credit_score                               0.19549777  0.23517635
```

Figure 10: PCA Loadings for PC1 through PC4

```
##
## Call:
## lm(formula = credit_cat ~ home_cat + count_tradelines_open_collection_accounts +
##     max_cc_limit + tradelines_avg_days_since_opened, data = user_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -173.589   -9.141   -1.967    7.520  254.818
##
## Coefficients:
##                                           Estimate Std. Error t value
## (Intercept)                               1.577e+01  3.965e-02  397.79
## home_cat                                  3.962e+00  5.325e-02   74.41
## count_tradelines_open_collection_accounts -9.482e-01  6.159e-03 -153.95
## max_cc_limit                              9.073e-04  4.441e-06  204.31
## tradelines_avg_days_since_opened          1.427e-03  2.486e-05   57.42
##                                          Pr(>|t|)
## (Intercept)                                <2e-16 ***
## home_cat                                   <2e-16 ***
## count_tradelines_open_collection_accounts  <2e-16 ***
## max_cc_limit                               <2e-16 ***
## tradelines_avg_days_since_opened           <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.57 on 272813 degrees of freedom
##   (12673 observations deleted due to missingness)
## Multiple R-squared:  0.3227, Adjusted R-squared:  0.3227
## F-statistic: 3.25e+04 on 4 and 272813 DF,  p-value: < 2.2e-16
```

Figure 11: summary of the Final Linear Regression

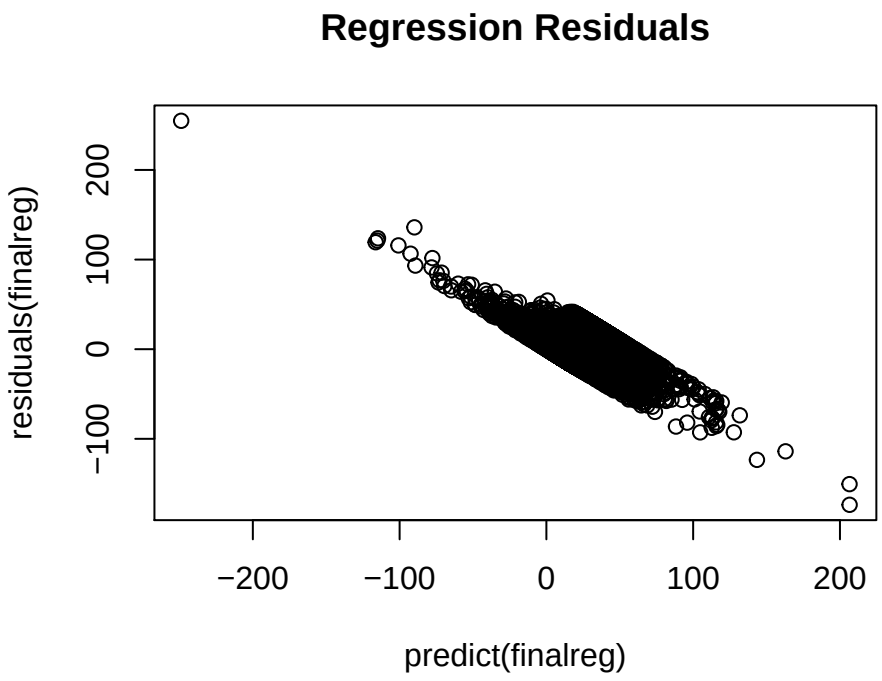

Figure 12: Residuals for the Linear Regression

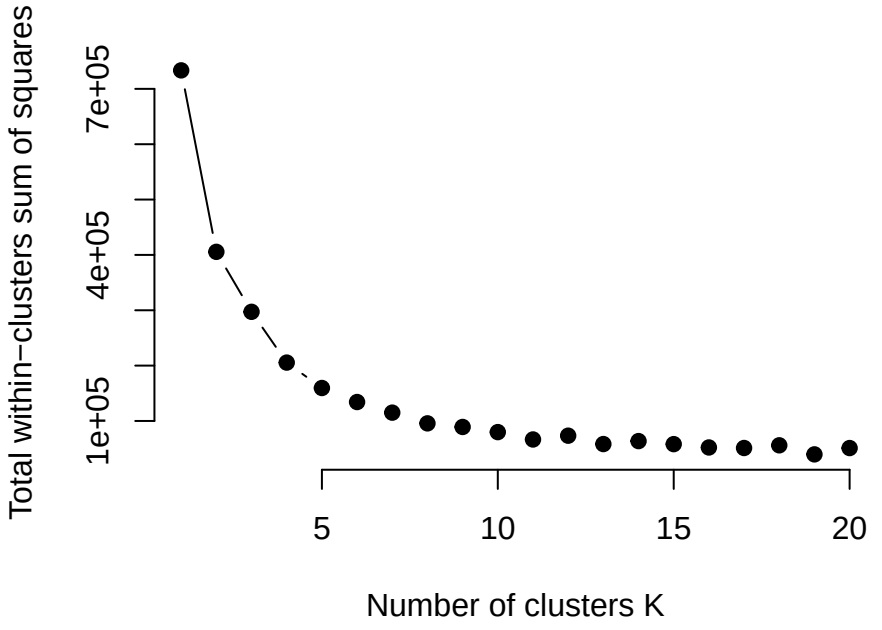

Figure 13: WSS for various K number of Clusters

```
##   click_count_credit_card click_count_personal_loan click_count_mortgage
```

```
## 1              2.42372881               0.24250326          0.043024772
## 2              0.00000000               0.00000000          0.001996741
## 3              0.07725445               1.46600373          0.013497990
## 4              0.52521178               5.06978620          0.034288019
## 5              0.08707440               0.08149270          0.017542461
## 6              4.00561366               0.11665641          0.025146109
## 7             10.54363284               0.21913381          0.045248869
## 8              1.26445073               0.03114592          0.009925624
##   click_count_credit_repair click_count_banking click_count_auto_products
## 1                1.40286832        0.0000000000              0.0234680574
## 2                0.00000000        0.0003855379              0.0002551072
## 3                0.03212952        0.0081849512              0.0077900632
## 4                0.14199274        0.0189592578              0.0254134732
## 5                1.21122717        0.0015947692              0.0051829998
## 6                0.02230083        0.0066133497              0.0029221778
## 7                0.07175178        0.0051712993              0.0058177117
## 8                0.00000000        0.0016718225              0.0015665109
```

Figure 14: Cluster Coordinates

```
##                                          click_credit     click_loan
## is_homeowner                               0.20084567     0.28468900
## gender                                     0.55391121     0.49441786
## tradelines_avg_days_since_opened        1050.47312896  1204.30214514
## tradelines_max_days_since_opened        2180.80126850  2823.69059011
## tradelines_min_days_since_opened         353.07610994   281.17145136
## count_tradelines_closed_accounts           7.63636364     8.06858054
## count_total_tradelines_opened_24_months    3.39746300     3.59011164
## count_tradelines_cc_opened_24_months       0.78646934     1.30382775
## count_tradelines_condition_derogatory      5.26744186     2.29425837
## count_open_installment_accounts_24_months  0.66490486     0.91467305
## count_tradelines_open_collection_accounts  2.66384778     0.89314195
## count_tradelines_open_mortgages            0.09619450     0.20175439
## count_tradelines_open_student_loans        1.04862579     1.09250399
## count_tradelines_opened_accounts           6.81395349     7.96012759
## count_tradelines_open_secured_loans        0.05179704     0.07575758
## count_tradelines_open_unsecured_loans      0.09619450     0.22727273
## total_tradelines_amount_past_due        3731.25158562  1217.49282297
## total_open_cc_amount_past_due              1.66173362     0.41706539
## total_cc_open_balance                   1520.51268499  4543.93779904
## total_tradelines_open_balance          35731.41331924 60459.47368421
## max_cc_limit                            1834.28541226  4855.63556619
## total_mortgage_loans_amount            14611.93234672 34770.76794258
## total_mortgage_loans_balance           12857.08139535 30600.40590112
## total_auto_loans_balance                6974.32346723 10398.56698565
## total_student_loans_balance            10737.63424947 10666.91467305
## count_inquiries_3_months                   1.14799154     0.82775120
## count_inquiries_6_months                   1.99260042     1.41706539
## count_inquiries_12_months                  3.36469345     2.76634769
## count_bankruptcy                           0.10465116     0.09569378
## age                                       37.45771670    37.22488038
## credit_score                             587.52114165   657.94657097
```

Figure 15: Demographic Comparison between Click Credit Users and Click Loan Users

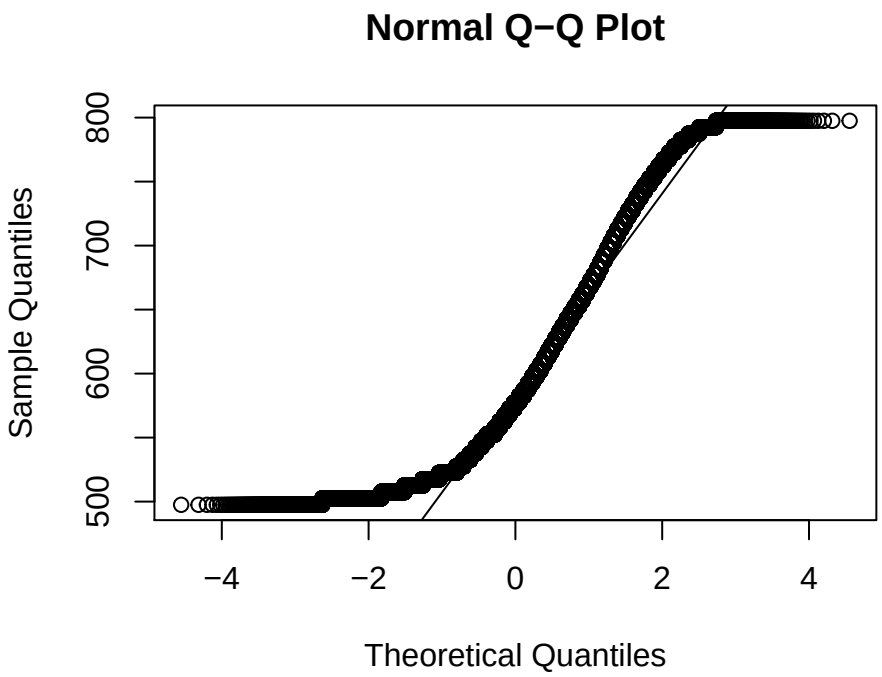

Figure 16: QQ-Plot of Credit Score

```
##                                 is_homeowner       gender
## is_homeowner                     1.00000000 -0.015910117
## gender                          -0.01591012  1.000000000
## tradelines_avg_days_since_opened  0.23933727  0.007070176
## tradelines_max_days_since_opened  0.31693507  0.022786618
## tradelines_min_days_since_opened  0.01039840 -0.032495261
## count_tradelines_closed_accounts  0.21566015  0.067415058
##                                 tradelines_avg_days_since_opened
## is_homeowner                                         0.239337268
## gender                                               0.007070176
## tradelines_avg_days_since_opened                     1.000000000
## tradelines_max_days_since_opened                     0.804840567
## tradelines_min_days_since_opened                     0.556355218
## count_tradelines_closed_accounts                     0.143899705
##                                 tradelines_max_days_since_opened
## is_homeowner                                         0.31693507
## gender                                               0.02278662
## tradelines_avg_days_since_opened                     0.80484057
## tradelines_max_days_since_opened                     1.00000000
## tradelines_min_days_since_opened                     0.14237474
## count_tradelines_closed_accounts                     0.25752938
##                                 tradelines_min_days_since_opened
## is_homeowner                                         0.01039840
## gender                                              -0.03249526
## tradelines_avg_days_since_opened                     0.55635522
## tradelines_max_days_since_opened                     0.14237474
## tradelines_min_days_since_opened                     1.00000000
## count_tradelines_closed_accounts                    -0.09484388
```

```
##                                 count_tradelines_closed_accounts
## is_homeowner                                          0.21566015
## gender                                                0.06741506
## tradelines_avg_days_since_opened                      0.14389971
## tradelines_max_days_since_opened                      0.25752938
## tradelines_min_days_since_opened                     -0.09484388
## count_tradelines_closed_accounts                      1.00000000
##                                 count_total_tradelines_opened_24_months
## is_homeowner                                                 0.06806776
## gender                                                       0.07324640
## tradelines_avg_days_since_opened                            -0.23183652
## tradelines_max_days_since_opened                             0.04166821
## tradelines_min_days_since_opened                            -0.35184871
## count_tradelines_closed_accounts                             0.22523673
##                                 count_tradelines_cc_opened_24_months
## is_homeowner                                             0.108610920
## gender                                                  -0.006006735
## tradelines_avg_days_since_opened                        -0.073942880
## tradelines_max_days_since_opened                         0.103780708
## tradelines_min_days_since_opened                        -0.190858779
## count_tradelines_closed_accounts                         0.108622063
##                                 count_tradelines_condition_derogatory
## is_homeowner                                            -0.07508032
## gender                                                   0.07287800
## tradelines_avg_days_since_opened                        -0.10665868
## tradelines_max_days_since_opened                        -0.04546076
## tradelines_min_days_since_opened                        -0.12488403
## count_tradelines_closed_accounts                         0.32525286
##                                 count_open_installment_accounts_24_months
## is_homeowner                                                   0.06180630
## gender                                                         0.01912673
## tradelines_avg_days_since_opened                              -0.10978584
## tradelines_max_days_since_opened                               0.03838887
## tradelines_min_days_since_opened                              -0.20691611
## count_tradelines_closed_accounts                               0.23347621
##                                 count_tradelines_open_collection_accounts
## is_homeowner                                                  -0.10880496
## gender                                                         0.05962075
## tradelines_avg_days_since_opened                              -0.13023008
## tradelines_max_days_since_opened                              -0.07030984
## tradelines_min_days_since_opened                              -0.13350044
## count_tradelines_closed_accounts                               0.02201817
##                                 count_tradelines_open_mortgages
## is_homeowner                                         0.484636609
## gender                                              -0.035306357
## tradelines_avg_days_since_opened                     0.278138934
## tradelines_max_days_since_opened                     0.367141020
## tradelines_min_days_since_opened                     0.001456113
## count_tradelines_closed_accounts                     0.198691956
##                                 count_tradelines_open_student_loans
## is_homeowner                                             -0.03121494
## gender                                                    0.10133909
## tradelines_avg_days_since_opened                          0.17238630
## tradelines_max_days_since_opened                          0.15178009
```

```
## tradelines_min_days_since_opened                        -0.05590490
## count_tradelines_closed_accounts                         0.11963562
##                                 count_tradelines_opened_accounts
## is_homeowner                                           0.1818955
## gender                                                 0.1110252
## tradelines_avg_days_since_opened                       0.1700417
## tradelines_max_days_since_opened                       0.3736582
## tradelines_min_days_since_opened                      -0.2344851
## count_tradelines_closed_accounts                       0.3266479
##                                 count_tradelines_open_secured_loans
## is_homeowner                                            0.07060915
## gender                                                 -0.01444588
## tradelines_avg_days_since_opened                       -0.05484871
## tradelines_max_days_since_opened                        0.01262783
## tradelines_min_days_since_opened                       -0.08796818
## count_tradelines_closed_accounts                        0.23175194
##                                 count_tradelines_open_unsecured_loans
## is_homeowner                                             0.100665813
## gender                                                  -0.007172969
## tradelines_avg_days_since_opened                        -0.004613578
## tradelines_max_days_since_opened                         0.098511031
## tradelines_min_days_since_opened                        -0.107176648
## count_tradelines_closed_accounts                         0.242726914
##                                 total_tradelines_amount_past_due
## is_homeowner                                          -0.024312552
## gender                                                -0.036447692
## tradelines_avg_days_since_opened                       0.022906806
## tradelines_max_days_since_opened                       0.023188590
## tradelines_min_days_since_opened                       0.002605908
## count_tradelines_closed_accounts                       0.231791352
##                                 total_open_cc_amount_past_due
## is_homeowner                                        0.024506643
## gender                                              0.004048482
## tradelines_avg_days_since_opened                    0.028304452
## tradelines_max_days_since_opened                    0.039107896
## tradelines_min_days_since_opened                   -0.001406001
## count_tradelines_closed_accounts                    0.032266533
##                                 total_cc_open_balance
## is_homeowner                                0.237627772
## gender                                     -0.006285972
## tradelines_avg_days_since_opened            0.311235554
## tradelines_max_days_since_opened            0.409575872
## tradelines_min_days_since_opened           -0.022271401
## count_tradelines_closed_accounts            0.151503914
##                                 total_tradelines_open_balance
## is_homeowner                                      0.37175792
## gender                                           -0.03207733
## tradelines_avg_days_since_opened                  0.26034088
## tradelines_max_days_since_opened                  0.36292505
## tradelines_min_days_since_opened                 -0.05122748
## count_tradelines_closed_accounts                  0.25319929
##                                 max_cc_limit total_mortgage_loans_amount
## is_homeowner                     0.26229817                 0.380749837
## gender                          -0.03199562                -0.050678063
```

```
## tradelines_avg_days_since_opened   0.39332221                0.238519362
## tradelines_max_days_since_opened   0.49339410                0.314156686
## tradelines_min_days_since_opened   0.02089918               -0.003539071
## count_tradelines_closed_accounts   0.15299273                0.156758184
##                                 total_mortgage_loans_balance
## is_homeowner                                     0.37776626
## gender                                          -0.05209186
## tradelines_avg_days_since_opened                 0.21876269
## tradelines_max_days_since_opened                 0.29396138
## tradelines_min_days_since_opened                -0.01113963
## count_tradelines_closed_accounts                 0.15424196
##                                 total_auto_loans_balance
## is_homeowner                                 0.154079338
## gender                                      -0.050142925
## tradelines_avg_days_since_opened             0.009406946
## tradelines_max_days_since_opened             0.130380737
## tradelines_min_days_since_opened            -0.141087478
## count_tradelines_closed_accounts             0.194090155
##                                 total_student_loans_balance
## is_homeowner                                     0.01158660
## gender                                           0.08720679
## tradelines_avg_days_since_opened                 0.14125303
## tradelines_max_days_since_opened                 0.15720901
## tradelines_min_days_since_opened                -0.04411014
## count_tradelines_closed_accounts                 0.24147967
##                                 count_inquiries_3_months
## is_homeowner                                -0.003500862
## gender                                      -0.024392127
## tradelines_avg_days_since_opened            -0.091224341
## tradelines_max_days_since_opened            -0.038624612
## tradelines_min_days_since_opened            -0.103253662
## count_tradelines_closed_accounts             0.071101732
##                                 count_inquiries_6_months
## is_homeowner                                -0.001344471
## gender                                      -0.028077941
## tradelines_avg_days_since_opened            -0.115512173
## tradelines_max_days_since_opened            -0.045561625
## tradelines_min_days_since_opened            -0.131907862
## count_tradelines_closed_accounts             0.091082776
##                                 count_inquiries_12_months
## is_homeowner                                  0.01271050
## gender                                       -0.02356541
## tradelines_avg_days_since_opened             -0.13318878
## tradelines_max_days_since_opened             -0.04120558
## tradelines_min_days_since_opened             -0.15791254
## count_tradelines_closed_accounts              0.13234808
##                                 count_bankruptcy          age
## is_homeowner                       0.0613982862  0.291340348
## gender                             0.0108299593 -0.004122074
## tradelines_avg_days_since_opened  -0.0007565558  0.252139445
## tradelines_max_days_since_opened  -0.0063279979  0.307202043
## tradelines_min_days_since_opened   0.0119331624  0.063089242
## count_tradelines_closed_accounts   0.1492748033  0.202873445
##                                 credit_score
```

```
## is_homeowner                        0.26869920
## gender                             -0.03255488
## tradelines_avg_days_since_opened    0.30113516
## tradelines_max_days_since_opened    0.33716245
## tradelines_min_days_since_opened    0.05833415
## count_tradelines_closed_accounts    0.11265252
```

Figure 17: Correlation Matrix of Variables