User Accounts and Derogratory Marks for Credit Sesame

Carolyn Chen, Man-Lin Hsiao, Jose San Martin, Michael Tan *

Duke University, Department of Statistical Science

Abstract. We were given information on user profiles, first session activity, and 30-day user engagement activity from Credit Sesame. We found that users most behind in their credit payments originated from 3 states in the midwest and Vermont. Users with any derogratory accounts, regardless of how many, were still more similar to each other than users with no derogratory accounts. Age should not be a major predictor for deliquency. Once one deliquency noted, intervention should occur to stem further late payments.

Keywords: credit monitoring, user engagement, derogratory accounts

1 Introduction

Credit Sesame is a credit score company that provides users with free information and monitoring of their credit score. The company makes money by advertising third-party sites that have loan offers a user can apply for. Our team was given three datasets by this company: User Profile, User First Session, and User Engagement. As a team of data scientists, we aimed to support the business arm of the company by analyzing how Credit Sesame could reduce its costs, as well as better understanding the demographics of its users in order to increase its revenue. With this goal in mind, we narrowed the scope of our analysis to a few pertinent areas, which we identified as: 1) demographic analysis of users and 2) understanding clusters of users 3) finding significant factors that predict delinquent credit behavior . Overall, we found that users most behind in their credit payments originated from Nevada, Utah, Wyoming and Vermont. Users with any derogratory accounts, regardless of how many, were still more similar to each other than users with no derogratory accounts. Age, although correlated with the type of loans a user was taking out, was not strongly correlated with deliquent accounts. Therefore CS should

* A big thank you to Professor Steorts and Duke Undergraduate Machine Learning Club!

not necessarily be wary of younger users. Once one deliquency noted, intervention should occur to stem further ones.

In section 2, we describe the dataset from Datathon. In section 2.1, we provide an exploratory data analysis (EDA) of user profile data and user engagement before considering any methods or algorithms. Based upon our EDA, we considered a Hiearchical model and Poisson Regression model in section 3. In section 4, we consider the methods from 3 and evaluate them with K-Means Clustering and looked at p-values respectively . In section 5, we summarize our findings and provide suggestions for future analysis utilizing Principal Component Analysis.

2 Data set

The three datasets we were given were User Profile, First Session and 30-Day User Engagement. User Profile gave a snapshot of a users's demographic, including their credit score bucket, at the time of sign-up. This dataset contained 285,619 rows and 38 columns. First session gave detailed actions that the user took on their first session on the site, with a session being defined as time the user logged on to their logout or automatic timeout. This dataset contained 8,755,480 rows and 16 columns, with each column being an action a user took. The 30-Day Engagement dataset summaried the actions of a user on their first 30 days of the site. This dataset contained 1,179,988 rows and 40 columns. Among the three datasets, we decided to focus on User Profile and 30-Day User Engagement since we believed that the 30-Day User Engagement dataset, since it was taken over a longer period of time, was more descriptive of the actions of a given user and general activity on the site. We performed EDA on these two datasets. We first cleaned and merged the data in 2 ways: 1) We combined user profile and user engagement data through linkage in the user id field to match the user actions with their profile. 2) We transformed variables (by taking averages of the age and credit score buckets to make credit score bucket and age bucket continuous.

2.1 Exploratory Data Analysis

In this section, we provide an EDA of user profile to better understand the demographics of Credit Sesame users. Understanding the demographics of the users is essential to CS as a business in order to better target marketing of certain loans.

We see above that of the loan types, student loans seem to have the greatest range in number of loans taken out by a given user. Males take

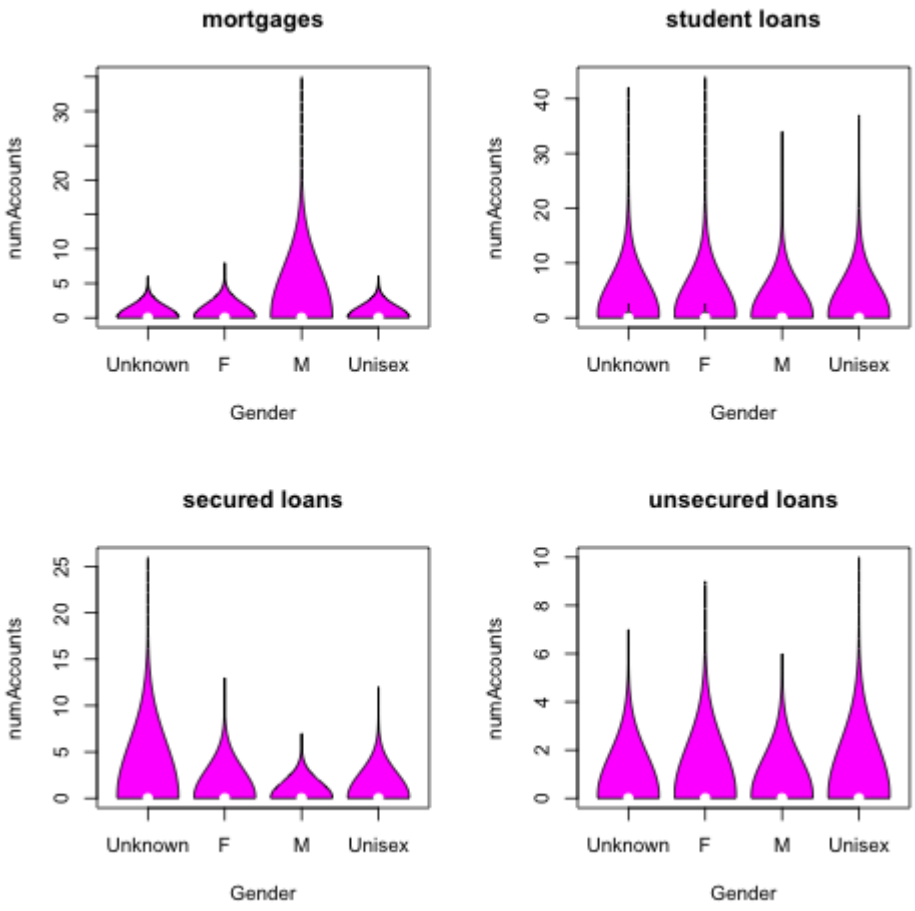

Fig. 1. Loan types by gender

out more mortgages than the other gender demographics.Those who did not to include their gender take out the greatest range in secured loans. In general, mortgages, student loans and secured loans seem to be more popular than unsecured loans. The prevelance of the ”unknown” and ”unisex” gender demographic was an attribute in the data that we would have liked to address, had we more time (see Appendix).

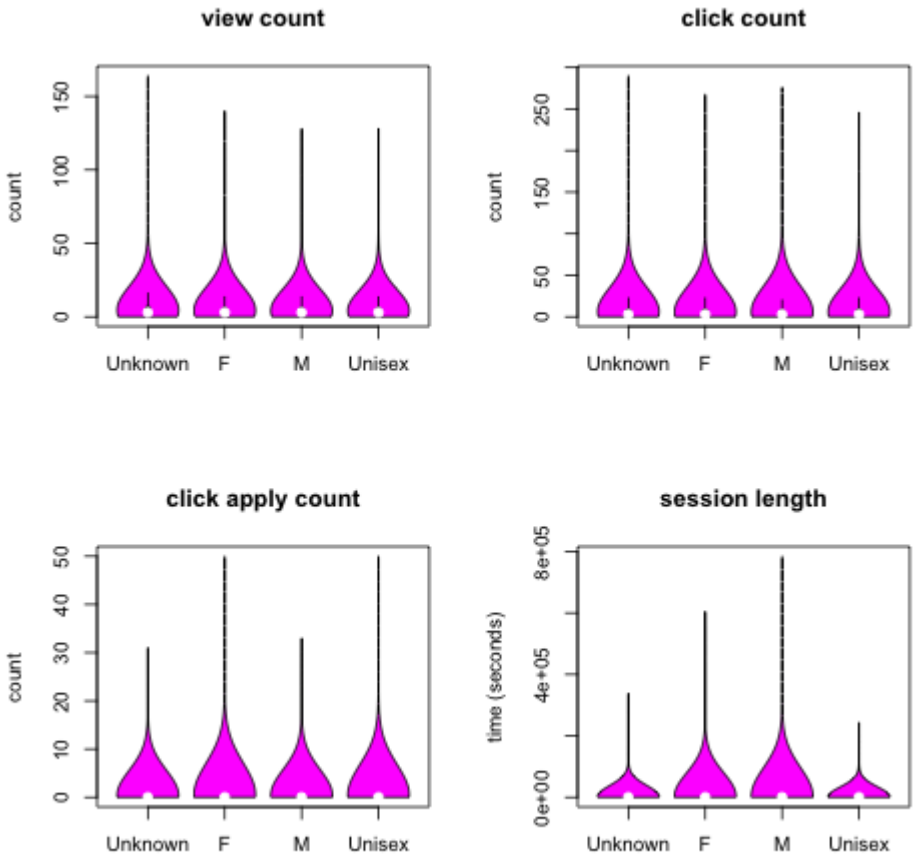

Fig. 2. User engagement by gender

The graphs above show us that females tend have clicked ’apply’ more than males, the upper range of session length was longer for men than for women.

Choropleth Maps: We then generated choropleth maps of the United States from the given data. First, we calculated the average of the Average Mortgage Loan Balance, Average Student Loan Balance, Average Auto Loan Balance, and the Average Credit Card Amount Past Due for each state from the data. This geographical debt data is valuable, as high debt levels are correlated with many other informative characteristics about a stateâĂŹs users as a whole. Low debt levels may also indicate less opportunity for banking companies to offer services. Thus, the map also allows for one to find possible 'Goldilocks states', where borrowing behavior is high enough for there to be a good consumer appetite for credit companies, yet not so high as to enter into dangerously high debt level territory. In all maps, states with low levels of whatever debt is being measured are colored more white, while deeper color indicate higher levels of debt.

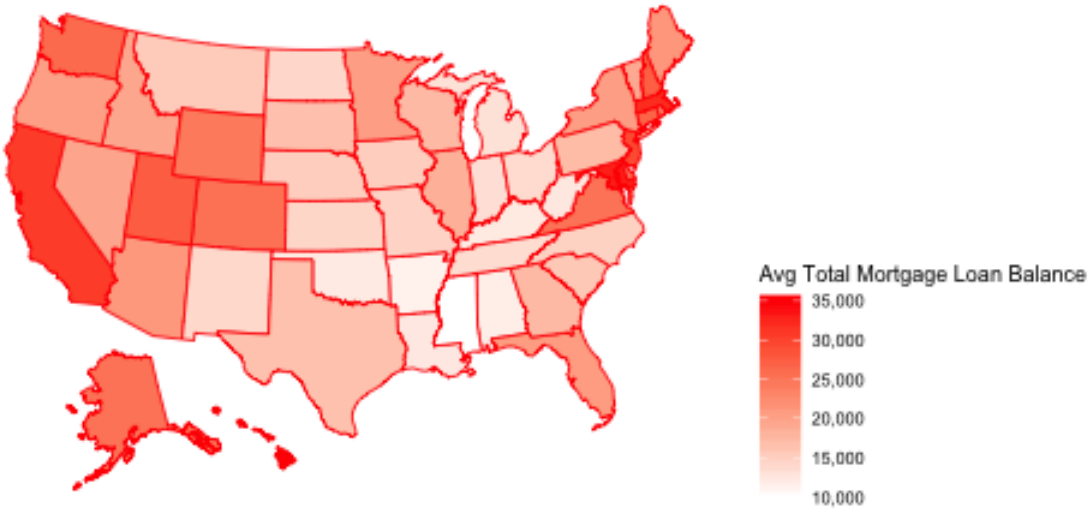

Fig. 3. Average Mortgage Loan Balance

From the choropleth of the US with Average Total Mortgage Loan Balance by state, we can see that states like California and Massachusetts exhibit the highest averages, while states like Alabama and Arkansas are among the lowest. This seems to be consistent with the distribution of median home prices by state – higher median prices mean that homebuy-

ers need to take out higher mortgage balances (California is expensive!). By combining this choropleth with existing data on real estate markets in different states, one can gain some color on which states would be the best for the mortgage business. We can see very plainly that there is high demand for mortgages in places like California and Massachusetts, but not in Alabama. Now let us briefly go through the other loan types.

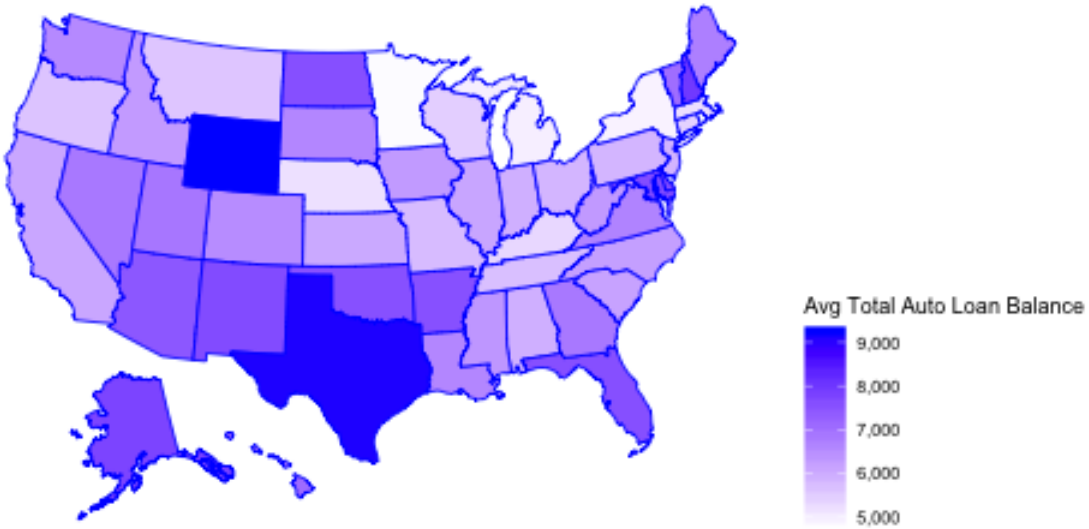

Fig. 4. Average Auto Loan Balance

Texas and Wyoming appear to have particularly high concentration of auto loans. New Hampshire is also high.

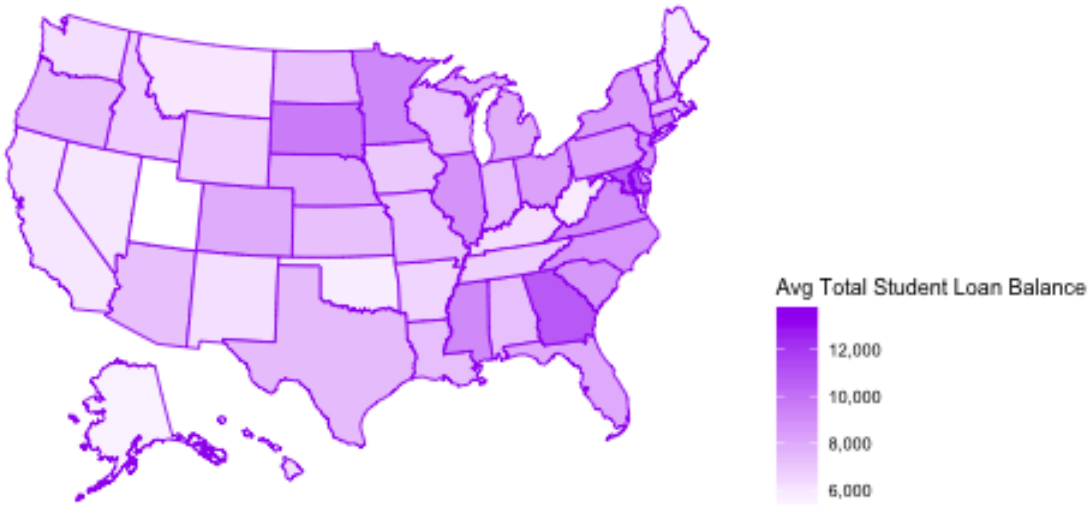

Fig. 5. Average Student Loan Balance

Student loan debt appears to be concentrated more in the midwest and eastern seaboard. This makes sense given the large concentration of schools on the East Coast. California also has very affordable public universities for in-state residents.

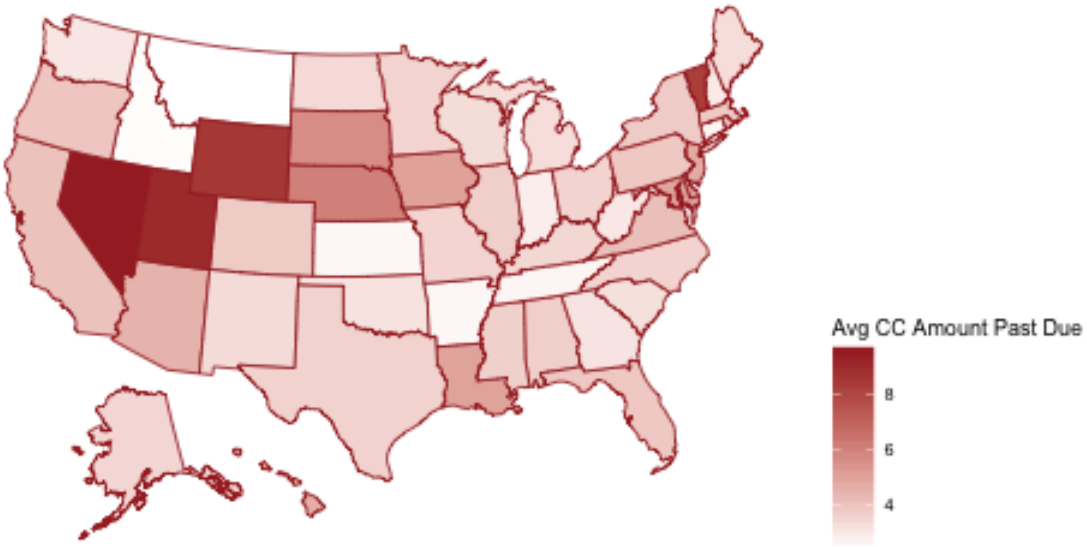

Fig. 6. Average Credit Card Payment Past Due

Nevada, Utah, and Wyoming have the highest average amounts of past due credit card payments. Vermont is also surprisingly high. In general it appears that several states in the midwest tend to have higher average amounts of past due credit card payments.

3 Methods

In this section, we consider applying the methods of (1) Hierarchical clustering and (2) Poisson regression to a merged dataset of user profile and user engagement. We evaluate the Hierarchical clustering model with a K-means model to visualize the grouping of the data from the Hierarchical model. We evaluate our Possion regression model by the significance of the p-values of the predictors. Before applying each method to the merged dataset of user profile and user engagement, we review each method below.

First we considered K-means clustering. We wanted to see if we could obtain a model that accurately predicts what type of category a user falls into in terms of number of derogatory accounts. Not only that, but we wanted to see if we should treat customers who only obtain a couple derogatory accounts the same as those who obtain many. We thought this method would be the best way to visualize our user clusters. The thresholds we set were as follows: 1) 'None' = No derogatory accounts 2) 'Some' = Between 1 and 2 derogatory accountsâĂİ 3) 'Many' = More than 2 derogatory accounts

We only used a small amount of variables in our two models, or else our models ended up not being able to differentiate groups. The variables we used were whether the user is a homeowner, their credit score, their age, and the number of times they viewed their credit card account in the first 30 days.

Next, we created a Poisson Regression for the purpose of predicting the number of derogatory accounts that a user is expected to open. Since the response variable is a count value we employed Poisson regression.

4 Application to User Profile and Engagement

Hierarchical Modeling and K-Means Clustering:

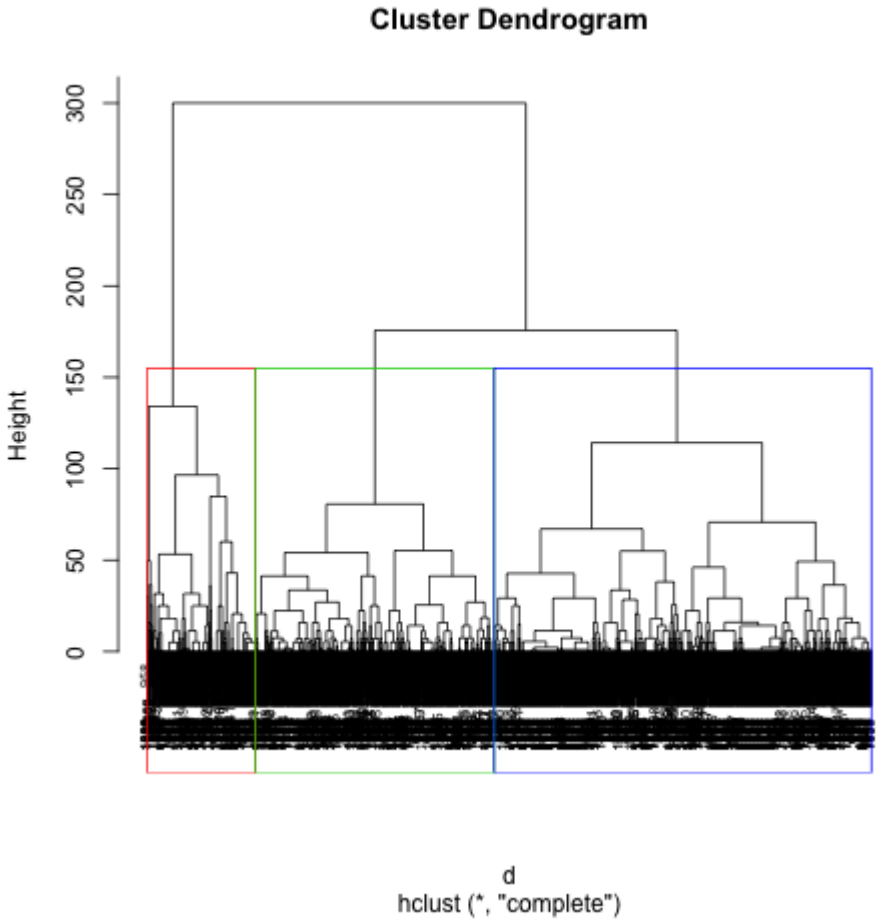

Fig. 7. K Means Clustering Model

Based on the clustering above, we found that the group of customers in the Some category of derogratory accounts and the ones in the Many category are similar to each other in profile, and noticably different from users in the None category . This tells us that the customers that obtained only a few derogatory accounts are not much different than those that took out many, and as a result they could also end up taking many. Credit Sesame should be careful even with users that have only few deliquencies, as this could be the beginning of a pattern for more in the future.

Poisson Model:

```
Call:
glm(formula = count_tradelines_condition_derogatory ~ Age1 +
```

```
    gender + creditScore1 + avg_cc_utilization_ratio + total_auto_loans_balance +
    total_student_loans_balance + total_mortgage_loans_balance +
    total_auto_loans_balance:total_student_loans_balance + total_auto_loans_balance:total_mortgage_loans_balance,
    family = "poisson", data = pois_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.8847  -1.7164  -0.9259   0.5725  29.8838

Coefficients:
                                                          Estimate Std. Error
(Intercept)                                             7.178e+00  1.827e-02
Age1                                                    9.863e-03  1.134e-04
genderMale                                             -1.372e-01  3.769e-03
genderUnisex                                           -7.007e-02  4.360e-03
creditScore1                                           -1.005e-02  2.868e-05
avg_cc_utilization_ratio                               -4.917e-01  4.447e-03
total_auto_loans_balance                                1.033e-06  1.277e-07
total_student_loans_balance                             1.704e-06  5.295e-08
total_mortgage_loans_balance                           -1.384e-06  3.168e-08
total_auto_loans_balance:total_student_loans_balance   -5.854e-12  2.749e-12
total_auto_loans_balance:total_mortgage_loans_balance  -6.705e-12  1.246e-12
                                                        z value Pr(>|z|)
(Intercept)                                             392.913  < 2e-16 ***
Age1                                                     86.978  < 2e-16 ***
genderMale                                              -36.392  < 2e-16 ***
genderUnisex                                            -16.071  < 2e-16 ***
creditScore1                                           -350.334  < 2e-16 ***
avg_cc_utilization_ratio                               -110.569  < 2e-16 ***
total_auto_loans_balance                                  8.087 6.13e-16 ***
total_student_loans_balance                              32.178  < 2e-16 ***
total_mortgage_loans_balance                            -43.693  < 2e-16 ***
total_auto_loans_balance:total_student_loans_balance     -2.129   0.0332 *
total_auto_loans_balance:total_mortgage_loans_balance    -5.379 7.48e-08 ***
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 598802  on 116760  degrees of freedom
Residual deviance: 443679  on 116750  degrees of freedom
  (133958 observations deleted due to missingness)
AIC: 682568

Number of Fisher Scoring iterations: 6
```

Overall, the poisson model ended up having many significant predictors; all of the predictors were significant. However that could also be due to the fact that there was a lot of data. The coefficient of the Age 1 variable, 9.863e-03, tells us that for every 10 years of age we add to a user, we expect the number of derogatory accounts to change by a multiplicative factor of e(10*9.863e-03) = 1.1. This does not seem like a big change, and it is good because it tells us that we probably should not assume a user who is older will take out more accounts that become derogatory. The output is below.

Let us interpret the coefficient of the creditScore1 predictor, which is -1.005e-02. This tells us that when we increase the credit score of a user by 100 points, the expected number of derogatory accounts changes by a multiplicative factor of 0.36. This is a pretty significant change, and so credit score should have more weight when deciding whether or not to lend.

5 Discussion

With the datasets provided, we set out to understand factors that predicted derogatory credit behavior, user types, demographic analysis and user engagement. We found that demographically, the highest mortgage balances are found in California and Massachusetts while the states with most past due credit are Nevada, Utah, and Wyoming. These demographic insights could be used to tailor marketing of certain types of loans to users in different states. It is also an opportunity to increase frequency of advertising certain types of loans to states that have less penetration. Regarding derogatory accounts, we found that users with non-zero derogatory accounts tended to have similar characteristics regardless of how many of those accounts they had. Users with one derogratory account are at risk for more, and therefore intervention should come early. Age was also not be a significant predictor of user derogatory behavior, which could be counterintuitive given the different borrowing behaviors of each age group. Credit score is a strong predictor of derogatory behaviour even though it appears those in the lower credit score ranges also apply for slightly more loans and are slightly more engaged users, so Credit Sesame should balance the potential increase in profit they may recieve from a more engaged user with the fact that this user is at higher risk of ultimately costing the compnany money.

6 Looking Forward

One thing we were not able to do during datathon, that we wanted to do, was a Principal Component Analysis. In essence, we wanted to perform one so that we would have gotten a better sense of the data and what it looked like. Doing a PCA would also give us some insight into how we should split the data up. For example, we can employ this on user profile data. We will use every numerical variable in that dataset, so this excludes zip code and recent bankruptcy date. We took a sample of 5000 points from the dataset, obtained their principal components, and plotted the PC1 vs PC2.

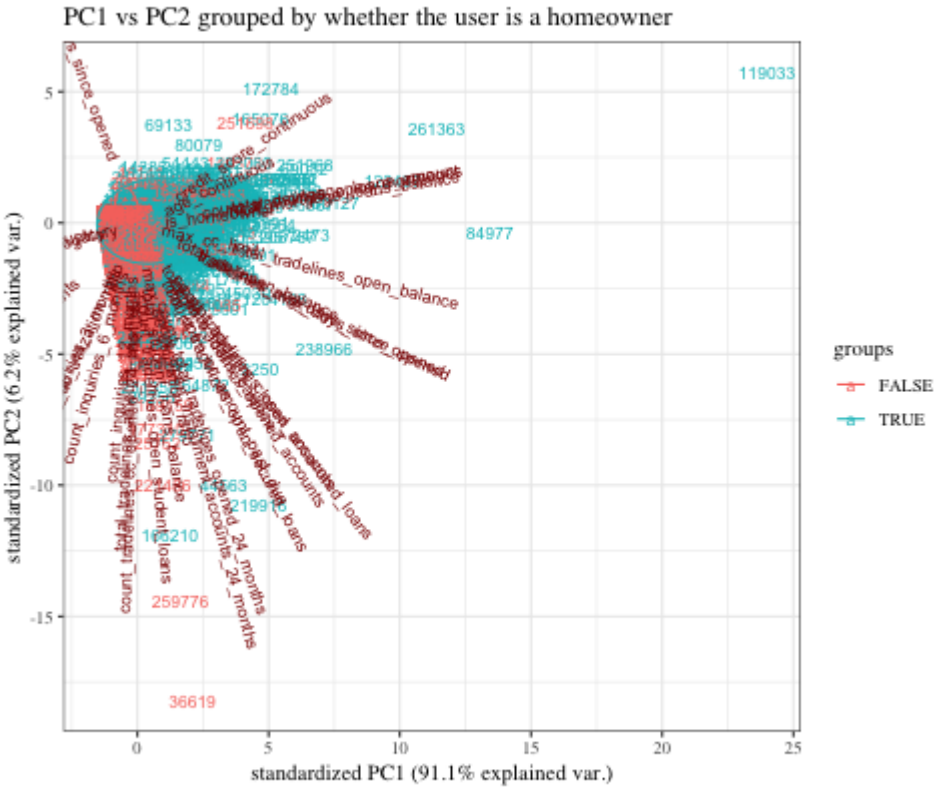

Fig. 8. PCA

Looking at the two plots, we can see that when we plot PC1 by PC2 and filter them by whether the user is a homeowner, we can see a clear separation. The first PC has 88.7 percent of all of the variance while rhe second PC has 6.3 percent. The users who are not homeowners seem to follow a certain, linear pattern upwards. The users that are homeowners, in the blue, go in every other direction. Therefore there is something specific about users that are not homeowners that make them very homogenous. Credit Sesame could use this information to group users that are not homeowners and provide them with very similar content since it appears they may be interested in similar offers, wheras the content

provided for homeowners should be more tailored based on a user's click history.

Appendix

Some limitations in the data cleaning were that there were ”unisex” and ”unknown” genders were significant in the user profile data. The unisex gender cateogry was unclear and we are uncertain if gender was unknown or unisex at random.This could impede interpretability of models with gender as a predictor.

A Algorithms

```
>
> #user_profile = fread("user_profile.csv")
> #first_session = fread("first_session.csv")
> #user_engagement = fread("user_engagement.csv")
> #
> ##Function that takes in bracketed range values, such as age and credit score, and returns
> ##the average in that bracket
> #sub<-function(x){
> # x <-x%>%
> #   gsub("[(]", "",.)%>%
> #   gsub("[]]", "",.)%>%
> #   gsub(" ", "", .)%>%
> #   strsplit(split = ",")
> # for(i in 1:length(x)){
> #   x[[i]] = as.numeric(x[[i]][1]) + as.numeric(x[[i]][2])
> #   x[[i]] = x[[i]]/2
> #   }
> #   return(unlist(x))
> #}
> #user_profile$credit_score_continuous = sub(user_profile$credit_score_bucket)
> #user_profile$age_continuous = sub(user_profile$age_bucket)
> #
> #
> ##Barplot of Credit Score Buckets
> #barplot(table(user_profile$credit_score_bucket), main = "Credit Score Frequency")
> #
> ### LIST OF VARIABLES CONTAINING NUMBER OF OPEN TRADELINES FOR DIFFERENT LOAN ACCOUNT TYPES:
> #total_open_loan_accounts = list(user_profile$count_tradelines_open_mortgages,
> #user_profile$count_tradelines_open_student_loans, user_profile$count_tradelines_open_secured_loans,
> #loan_list = c("mortgages", "student loans", "secured loans", "unsecured loans")
> ###Vector indicating order of loan list
> #count = 1
> #par(mfrow=c(2,2))
> #for(numAccounts in total_open_loan_accounts){
> #  vioplot(numAccounts[user_profile$gender==""], numAccounts[user_profile$gender=="Female"],
> #numAccounts[user_profile$gender=="Male"], numAccounts[user_profile$gender=="Unisex"],
> #names = c("Unknown", "F", "M", "Unisex"))
> #  title(main=loan_list[count], ylab = "numAccounts", xlab = "Gender")
> #  count=count+1
> #}
```

```
> #
> ##User Engagement visualization by gender
> ##merge user profile and user engagement datasets
> #c = merge(user_profile, user_engagement, on="user_id", how="outer")
> #
> ##Make list of user engagement stats
> #engagementStats = list(c$view_count, c$click_count, c$click_apply_count, c$session_length)
> #engagementNames = c("view count", "click count", "click apply count", "session length")
> #count = 1
> #par(mfrow=c(2,2))
> #y = c("count", "count", "count", "time (seconds)")
> ##Loop through each column name and seperate by gender
> #for(ct in engagementStats){
> #  vioplot(ct[c$gender==""], ct[c$gender=="Female"], ct[c$gender=="Male"],
> #ct[c$gender=="Unisex"], names = c("Unknown", "F", "M", "Unisex"))
> #  title(main=engagementNames[count], ylab = y[count])
> #  count=count+1
> #}
> #user_profile$credit_score_continuous = sub(user_profile$credit_score_bucket)
> #user_profile$age_continuous = sub(user_profile$age_bucket)
> #
> ##summarize mortgage loans by state
> #AvgMortgageLoanByState = user_profile %>% group_by(state) %>% summarize(mean(total_mortgage_loans_b
> #map_with_data(AvgMortgageLoanByState, values = "mean(total_mortgage_loans_balance)", include = c(),
> #plot_usmap(data = AvgMortgageLoanByState, values = "mean(total_mortgage_loans_balance)", lines = "r
> #  scale_fill_continuous(
> #    low = "white", high = "red", name = "Avg Total Mortgage Loan Balance", label = scales::comma
> #  ) + theme(legend.position = "right")
> #
> ##summarize auto loans by state
> #AvgAutoLoanByState = user_profile %>% group_by(state) %>% summarize(mean(total_auto_loans_balance))
> #
> ##plot US map with average auto loan balance as shading
> #plot_usmap(data = AvgAutoLoanByState, values = "mean(total_auto_loans_balance)", lines = "blue") +
> #  scale_fill_continuous(
> #    low = "white", high = "blue", name = "Avg Total Auto Loan Balance", label = scales::comma
> #  ) + theme(legend.position = "right")
> #
> ##summarize student loans by state
> #AvgStudentLoanByState = user_profile %>% group_by(state) %>% summarize(mean(total_student_loans_bal
> #
> ##plot average student loan balance as shading by state
> #plot_usmap(data = AvgStudentLoanByState, values = "mean(total_student_loans_balance)", lines = "pur
> #  scale_fill_continuous(
> #    low = "white", high = "purple", name = "Avg Total Student Loan Balance", label = scales::comma
> #  ) + theme(legend.position = "right")
> #
> ##summarize past due credit card balannce by state
> #AvgTotalCCAmountPastDue = user_profile %>% group_by(state) %>% summarize(mean(total_open_cc_amount_
```

```
> #
> ##plot past due credit card amount as shading by state
> #plot_usmap(data = AvgTotalCCAmountPastDue , values = "mean(total_open_cc_amount_past_due)", lines =
> #  scale_fill_continuous(
> #    low = "white", high = "brown", name = "Avg CC Amount Past Due", label = scales::comma
> #  ) + theme(legend.position = "right")
> #
> ##create categories for number of derogratory tradelines
> #cluster_data <- c %>%
> #  mutate(derogatory = ifelse(count_tradelines_condition_derogatory==0, "None",
> #                             ifelse(count_tradelines_condition_derogatory==1 |count_tradelines_cond
> #  select(derogatory,is_homeowner,credit_score_continuous, age_continuous, view_cc_details_count)%>%
> #  na.omit()
> ##Take a sample of 2000 observations to create our models
> #cluster_data <- cluster_data%>%
> #  #scale(.)%>%
> #  slice(1:2000)%>%
> #  na.omit()
> #
> ##Hierarchial Modeling
> ##Create the distance matrix using euclidean length
> #d <- dist(cluster_data[,2:5], method = "euclidean")
> ## Hierarchical clustering using Ward's method
> #res.hc <- hclust(d)
> #grp <- cutree(res.hc, k = 3)
> ##table(grp, cluster_data$derogatory)
> ## Visualize Our hierachical model with 3 groups
> #plot(res.hc, cex = 0.6) # plot tree
> #rect.hclust(res.hc, k = 3, border = 2:5) # add rectangle
> #
> ##poisson regression model to find predictors of derogratory tradeline count
> #pois_data <- user_profile%>%
> # filter(gender != "")
> #creditScore1 = sub(pois_data$credit_score_bucket)
> #Age1 = sub(pois_data$age_bucket)
> #m1 <- glm(count_tradelines_condition_derogatory ~ Age1 + gender + creditScore1 + avg_cc_utilizatio
> #summary(m1)
> #
> ##Data cleaning
> #user_pcaData = user_profile%>%
> #  mutate(gender = ifelse(gender == "Male", 1 , 0))%>%
> #  mutate(zipcode = as.numeric(zipcode))%>%
> #  mutate(derogatory = ifelse(count_tradelines_condition_derogatory==0, 0, 1))%>%
> #  select(-c(recent_bankruptcy_date, count_tradelines_condition_derogatory, credit_score_bucket, age
> #  na.omit()
> #user_pcaData<- user_pcaData[sample(1:nrow(user_pcaData), 5000,replace=FALSE),]
> ##Model
> #users.pca <- prcomp(user_pcaData[,c(6,8:38)],center = TRUE)
> #users.latent.sem = users.pca$rotation
```

```
> #summary(users.pca)
> ##We can see that the first PC has 88.7% of all the variance, while the second PC has 6.3% of all va
> ##Now plot them and split the data up based on whether it is derogatory or not, and choose 5000 data
> #ggbiplot(users.pca, ellipse=TRUE, labels=rownames(user_pcaData),groups=user_pcaData$is_homeowner) +
```