Credit Sesame Data Analysis and Findings

Hari Rajan, Paul Giraud, Audreya Metz, and Calvin Ma

September 27 2018

Dataset

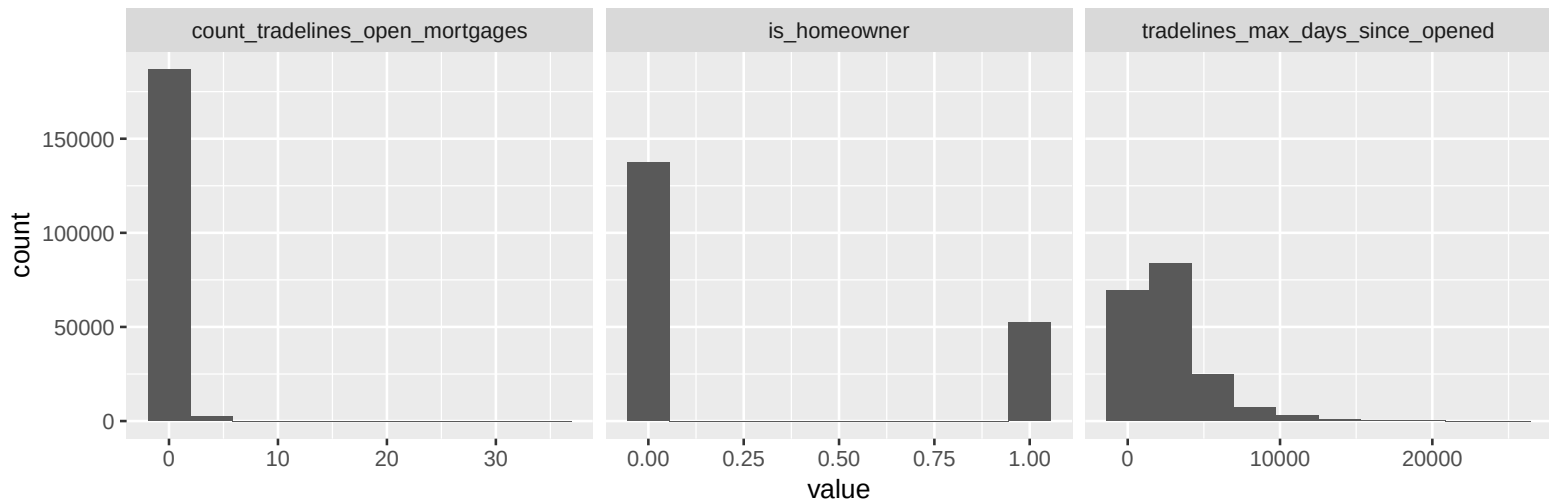
- The data was from Credit Sesame, a company that calculates credit scores to determine options for credit cards, mortgage rates and loans.
- This dataset included:
  - User Demographics
  - First Session Information
  - 30-day User Engagement Data

Our Research Direction

- We wanted to give Credit Sesame an insightful description of their user base
- We believe that they would have an interest in the credit score of their users, since this helps the to predict the best loans, mortgages, etc.

Exploring the Dataset

- We started out EDA by looking at the distribution of the variables in the datset
- We found a significant right skew in the majority of the non-categorical data



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.

Exploring the Dataset

- Plots of Credit Score v. Some Features
- Brief discussion of linear regression

Exploring the Dataset

Add PCA biplots

Challenges of the Dataset

- The skewed distribution of our variables was the biggest challenge for determining how to analyze this data.
- Ultimately we did not let the skew effect our methods but were careful to consider it in our analysis.

Methods

- Linear Regression
  - We determined which features to use in the regression using the plots and PCA
  - Only achieved an R-squared of .35, which is too low to draw meanginful conclusions
  - COuld use other regression methods such as Lasso in the future
- K-means clustering
  - Split users into similar groups
  - Compare within cluster averages to other clusters
  - Find differences

K-means

- How we determined how many K's CH INDEX
- Some visualization for the clusters

Results

- After determining the number of clusters to use and accordingly performing the k-means clustering method, we noticed that in particular, there were two clusters which had relatively high values for `click_count_credit_card` and `click_count_personal_loan`.

Results (continued)

- The users that were grouped into these two clusters had a noticeably large difference in the mean values for several different variables as shown below. Note that these variable values generally correspond to the user's frequent actions on the Credit Sesame website.

Conclusions

- The k-means clustering allowed us to confirm that the demographics of people who use Credit Sesame's website differently and access different pages are in fact different
- There is a distinction between Credit Sesame's website clientele that uses the credit card features and those who use loan features
  - The credit page users tend to have lower credit scores and the loan feature users tend to have higher loan amounts based on grouped findings
  - The largest difference between credit score users and loan feature users are open collection accounts, # of derogatory tradelines, max credit card limit, total credit card balance, and total mortgage loans and balance, which all happen to be strong indicators of credit score

Conclusions

- We can work to improve the current models that we used
  - Regression model can be improved using methods such as Lasso or ElasticNet to facilitate feature selection
  - Hierarchal clustering could have provided more insight into potential variable groups