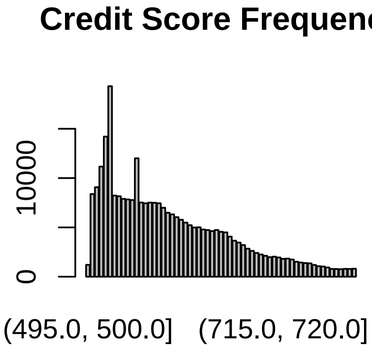
Understanding Credit Card Delinquency

Carolyn Chen, José San Martin, Michael Tan, Man-Lin Hsiao

11/7/2018

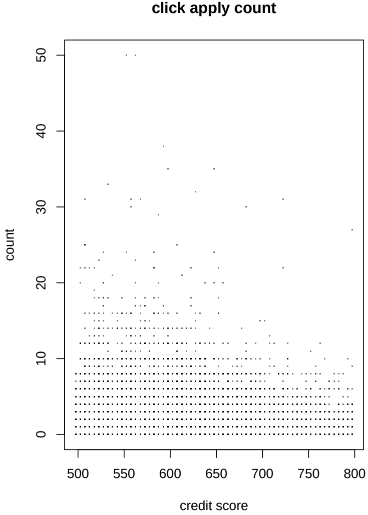
Data Set Visualizations

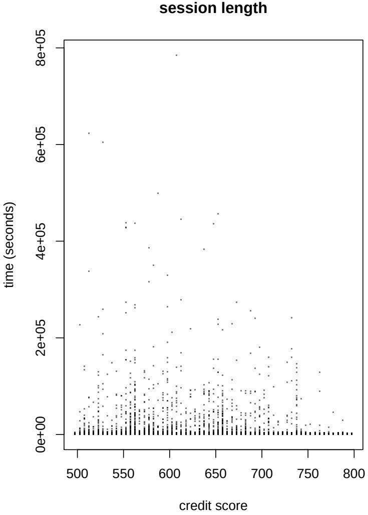
- Credit Sesame is a credit and loan-management platform
- Datasets: User Profile, First Session, 30-Day User Engagement
- First, we wanted to understand the demographics of Credit Sesame users.
- Data cleaning for ease of visualization
- Histograms, Dot Plots, Violin Plots, Choropleth Maps



Exploratory Data Analysis (cont.)

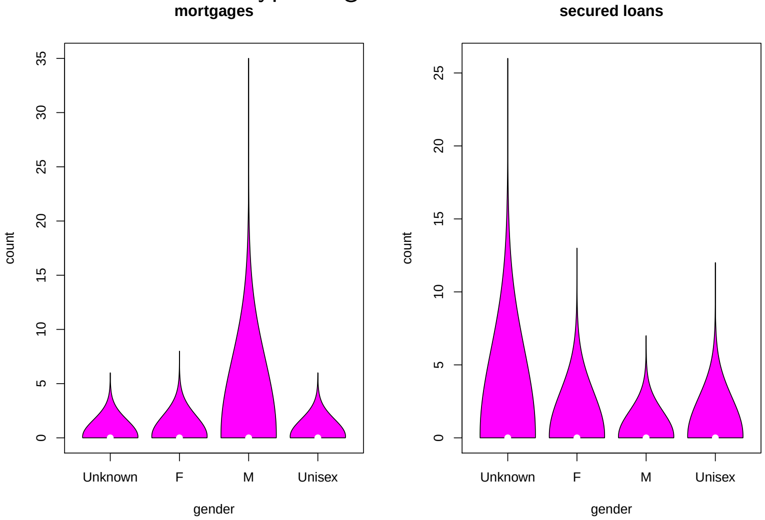
- Dot plots of engagement stats versus credit score





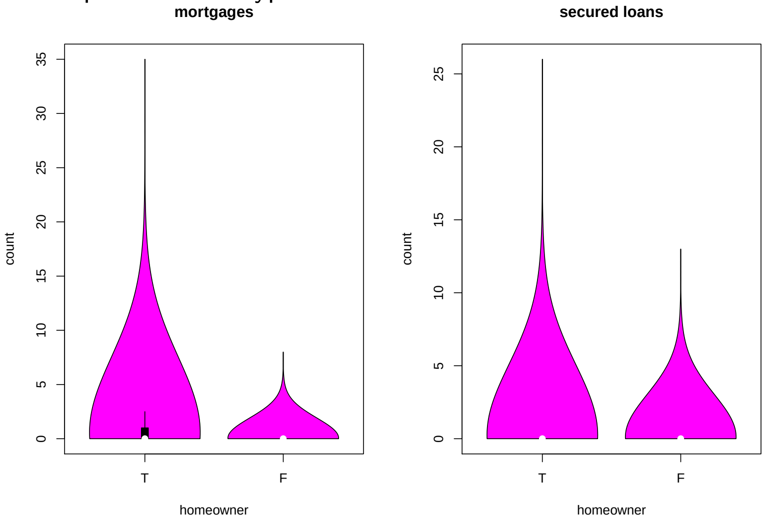
EDA Visualizations (cont.)

- Violin Plots of loan type vs gender



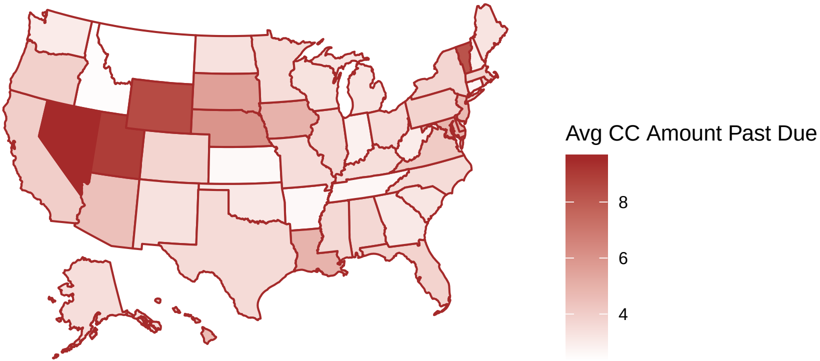
EDA Visualizations (cont.)

- Violin plots of loan type vs homeowners



Visualizations (cont.)

- Choropleth Map shows us geographical distribution of credit card debt trends
- Delinquency: user has missed 2 consecutive payments
- What are profiles of delinquent vs. non-delinquent users and within levels of delinquency?

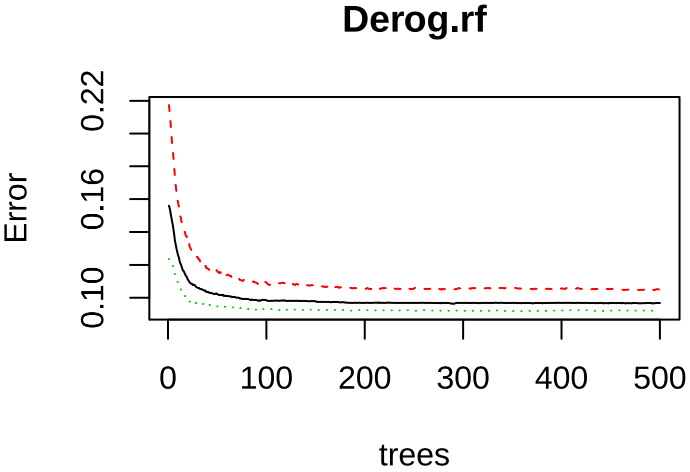


Next Step: Random Forest Model

- From our EDA we could already tell that there was a difference in profile between people with and without derogatory accounts, and the next logical step would be to create a predictive model
- Appropriate model given we have response variable for derogratory variable

Random Forest Model

- Accuracy of about 90%, obtained a pretty small out-of-bag error rate



Random Forest Model

- Now we look at other values from our fitted Random Forest model
- Using a classifier to determine if a person has obtained a derogatory account, or not at all. We concluded that having even just one derogatory account is cause of concern for the bank.
- Based on the importance output, the 5 variables of the highest importance are homeownership, tradelines average days since opened, tradelines maximum days since opened, tradelines minimum days since opened, and number of closed tradeline accounts

Poisson Model

$$
\begin{aligned}
\text{DerogratoryTradelines} = \alpha &+ \beta_1 \text{Age} \\
&+ \beta_2 \text{Gender} \\
&+ \beta_3 \text{CreditScore} \\
&+ \beta_4 \text{CreditCardUtilizationRatio} \\
&+ \beta_5 \text{AutoLoansBalance} \\
&+ \beta_6 \text{StudentLoansBalance} \\
&+ \beta_7 \text{MortgageBalance} \\
&+ \beta_8 \text{MortgageLoan} * \text{AutoLoan} \\
&+ \beta_9 \text{AutoLoan} * \text{StudentLoan}
\end{aligned}
$$

- Age: For every 10 years 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 shows age is not a significant factor.

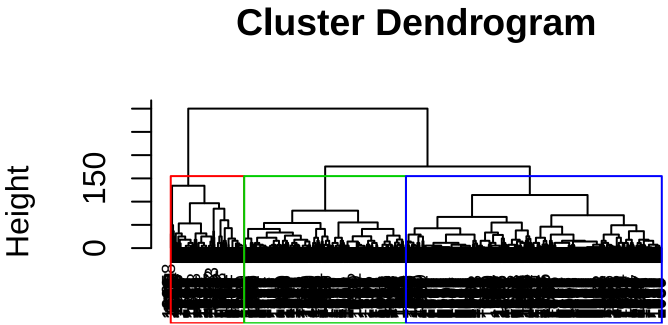
When we increase the credit score of a user by 100 points, the expected number of derogatory accounts changes by a multiplicative factor of 0.36.

Conclusions

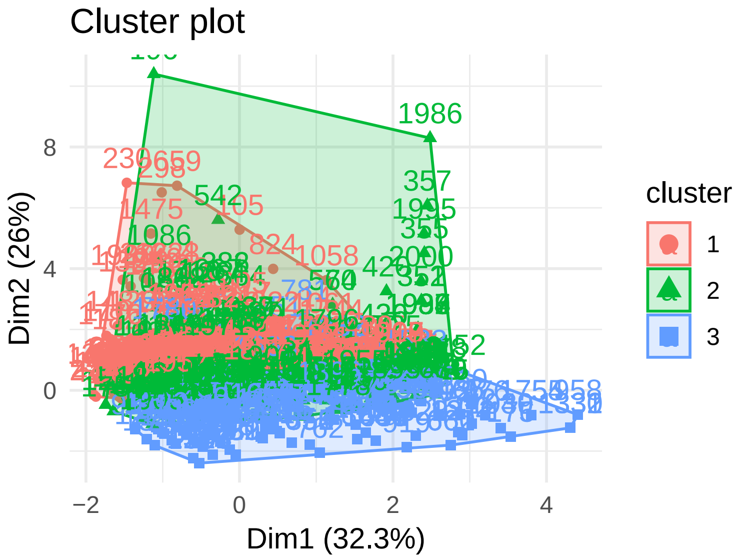
- Users with non-zero derogatory accounts tended to have similar characteristics regardless of how many of those accounts they had. Once one deliquency noted, intervention should occur to stem further ones.
- Age should not be a significant predictor of user derogatory behavior.
- Credit score is a strong predictor of derogatory behaviour, but user base of CS is also right-skewed.
- States with most past due credit card accounts are Nevada, Utah, Wyoming and Vermont. Could target reminders to people from those states.

Original Model: Clustering

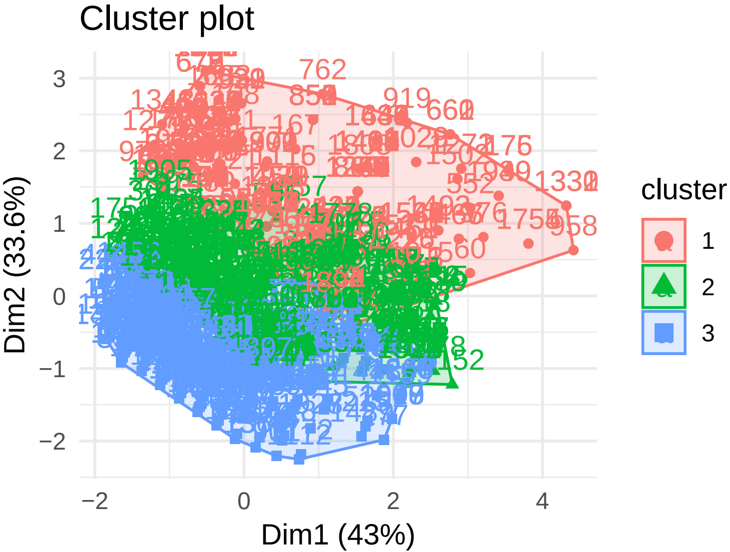
- K means clustering to identify similarity of deliquent vs non-deliquent users
- Looked at users with 'none' (0), 'some' (1-2) and 'many' (>2) deliquencies
- Users with 'some' deliquencies still quite similar to those with 'many'



K-means Visualization



K-means Visualization



Problems with Clustering Model

- The variables we chose were arbitrary and the hierarchical model was too naive.
- It wasn't a good method to use with the rest of our analysis because it didn't tell us anything substantial.
- We did not split into training/testing sets and use Cross Validation to check the model.
- We arbitrarily cut our trees at an unmotivated point.