Team 22 Report

Introduction

Credit bureau companies offer *free* financial reporting for anyone who signs up for their service - so how do they make money? Credit Sesame, a service that offers free credit score and credit report analyses, employs a business model that is sustained by revenue generated from targeted advertisements to consumers. These advertisements focus on products ranging from credit cards to loans to mortgages, and are heavily suggested to users who would like to better their financial situations. Credit Sesame profits when users choose to consult these external offers by clicking 'apply' on a Credit Sesame recommendation page.

Our team sought to analyze the most important factors that drive consumers to click 'apply' and therefore become what we henceforth label a 'valuable customer' in the eyes of Credit Sesame. In doing so, we uncovered some interesting relationships between the overall action of a valuable customer (whether they did or did not apply) and his/her demographics, financial situation, and use of the service in the past 30 days. We present a few visualizations of these relationships as well as evidence for why a few of these factors are most important in analyzing whether or not a given user will become a valuable customer for Credit Sesame.

Data Engineering Process

We manipulated our dataset to extract both the demographics and engagement behavior of each user. This was done by aggregating web behavior on a user basis in the 30-day `user_engagement` dataset, then appending these results to the `user_profile` dataset. This way, the new `user_profile` dataset has one row for every user in the database, with columns representing information about their demographic background, as well as columns representing their aggregated first-30-day engagement behavior on the Credit Sesame website. Non-numeric attributes were not considered for simplicity sake, but would have been included had we more time. The aggregation process is visualized below:
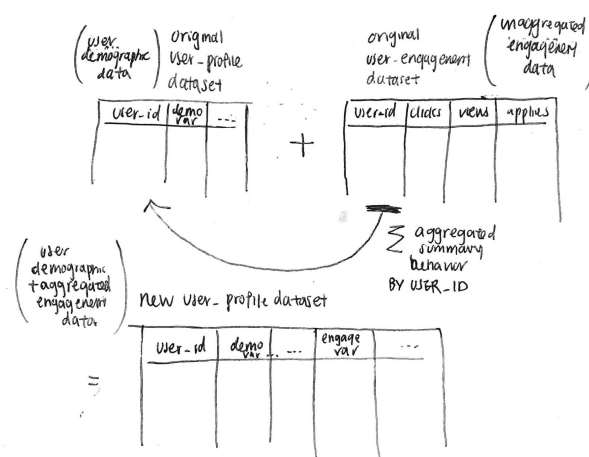
Figure 1: Data Engineering process to obtain demographics and engagement behavior data of users.

Analysis

Our methodology focused on first building a model which could predict whether or not a given user would eventually click 'apply' within their 30 days of usage, and then afterwards determining the most important factors in making that decision. Because of the two-fold nature of our mission, we decided to implement a Random Forest Classifier, because it would present accurate predictions while controlling for overfitting, as well as return a measure of the relative importances of the features that we gave as inputs. The most important of these features became the ones that we focused further analysis on. Another advantage of the Random Forest is its relatively easy interpretability, which lends itself to extracting and analyzing feature importance.

We used the *sklearn* library of Python to build a Random Forest classifier. We first split the data to hold out 20% for testing. With the remaining 80% of non-testing data, we trained our model using 3-fold cross-validation with a 150-tree hyperparameter, then used the model to predict data on our testing set, and compare the predicted test values to the true test values.

In addition, we visualized our high coefficient features in Tableau to better understand the relationship between clicking "apply" and user profile information. These visualizations reinforced our findings and provide a high-level overview for more general audiences on our findings. We also visualized location data with an interactive map which, even though it wasn't a high coefficient feature, is particularly conducive to visual analysis. This allowed us to visualize the distribution of Credit Sesame customers across the US.

Findings

In regards to our classification model, our classifier returned a training accuracy of 71.17%, and a test accuracy of **71.24%**. We also determined an ordering for attribute importance which dictates further interpretation. The five most important features for determining classification are *Total Session Length, Average Session Length, Total Tradelines Open Balance, Total Page Views, and Total Click Counts*. The rest are enumerated below:

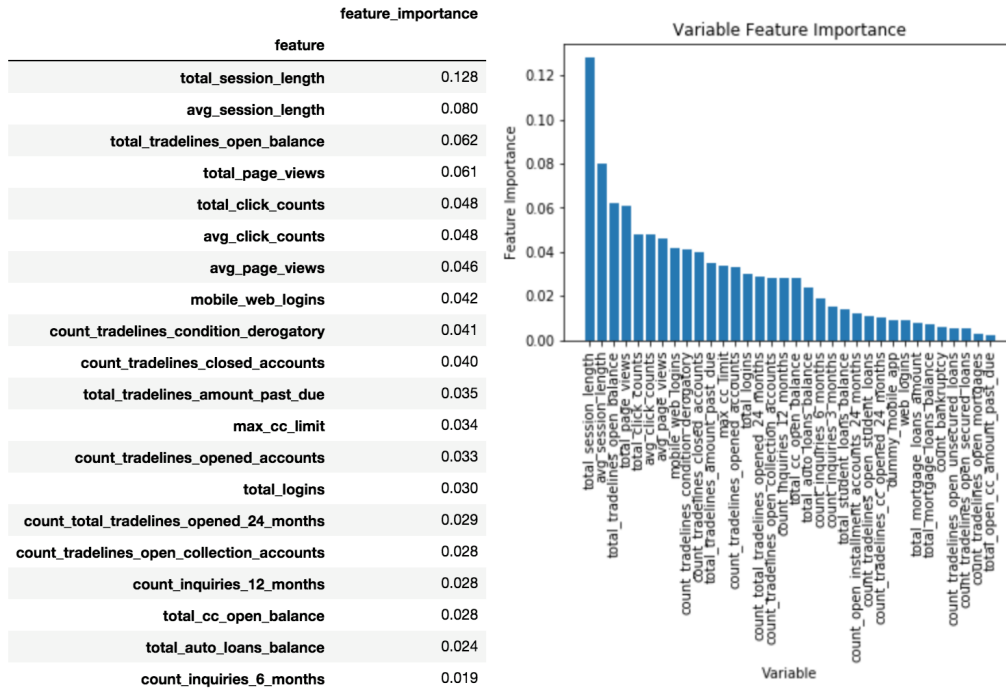| feature	feature_importance
total_session_length	0.128
avg_session_length	0.080
total_tradelines_open_balance	0.062
total_page_views	0.061
total_click_counts	0.048
avg_click_counts	0.048
avg_page_views	0.046
mobile_web_logins	0.042
count_tradelines_condition_derogatory	0.041
count_tradelines_closed_accounts	0.040
total_tradelines_amount_past_due	0.035
max_cc_limit	0.034
count_tradelines_opened_accounts	0.033
total_logins	0.030
count_total_tradelines_opened_24_months	0.029
count_tradelines_open_collection_accounts	0.028
count_inquiries_12_months	0.028
total_cc_open_balance	0.028
total_auto_loans_balance	0.024
count_inquiries_6_months	0.019

Figure 2.a: Table of feature importance (top 20 only), sorted in descending order of importance.
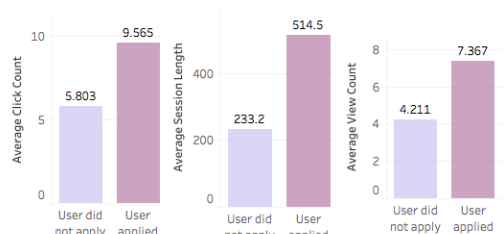*Figure 2.b: Bar Chart of feature importance, sorted in descending order of importance.*

As seen above, our feature importances match intuition; it makes logical sense that behaviors of customers while using Credit Sesame, like session length and page views, are important factors in determining whether one will convert. However, interestingly enough, we see that user profile factors, which users contribute to at the beginning of account creation, are also important; these include tradelines open balance (an estimate of the total debt a person has) and the number of tradelines in derogatory condition (an entry that may be considered negative by lenders because it indicates risk and hurts a person's ability to qualify for credit or other services). These factors are ones that would have been further explored if we had more time, because they are ones that Credit Sesame could potentially consider as soon as someone makes an account, rather than waiting for their 30 day activity.

Our interactive dashboard highlights the key behaviors of users that apply for services on Credit Sesame as well as behaviors that don't seem to be correlated (ex: platform use). The dashboard also offers a comprehensive view of the distribution of Credit Sesame customers across the United States. Screenshots of the dashboard are provided here:

Predicting Prospective Credit Sesame Users

Credit Sesame is a free personal finance tool that allows users to access financial information and assists in services such as credit score monitoring, home loans, and identity theft protection. When users visit the Credit Sesame website or use the mobile application, they are able to apply for certain services.

What characteristics define users that apply for services on Credit Sesame?

Average Click Count
- User did not apply: 5.803
- User applied: 9.565

Average Session Length
- User did not apply: 233.2
- User applied: 514.5

Average View Count
- User did not apply: 4.211
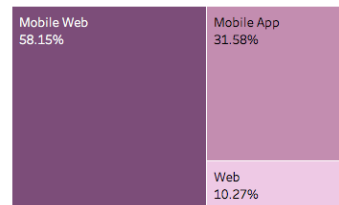- User applied: 7.367

On average, users who apply for a service on Credit Sesame tend to have higher click counts, view more pages, and spend more time on the CreditSesame application than users who do not apply for a service.

Platform Use of Users who Applied for a Service

| Mobile Web 72.74% | Mobile App 16.24% |
| | Web 11.01% |

Platform Choice of Users who Did Not Apply for a Service

| Mobile Web 58.15% | Mobile App 31.58% |
| | Web 10.27% |

All users generally access CreditSesame from the Mobile Web platform. Users who apply for a service on Credit Sesame are even more likely to prefer the Mobile Web application over the Mobile App and Website.

Total Tradelines Open Balance
- User did not apply: 3,294,435
- User applied: 655,147
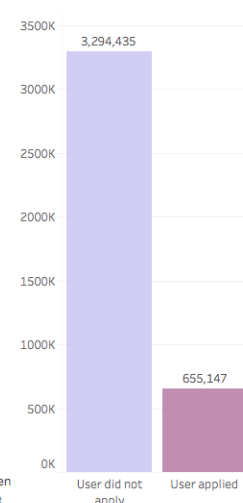
Cumulatively, users that apply for services under Credit Sesame have less total debt than users that choose not to apply.

Predicting Prospective Credit Sesame Users

Though location wasn't a strong indicator of whether or not a user applied for a service on Credit Sesame, it is an interesting variable to visually examine to better understand the general user profile of Credit Sesame users. Use the interactive graphic below to better understand the distribution of Credit Sesame users.
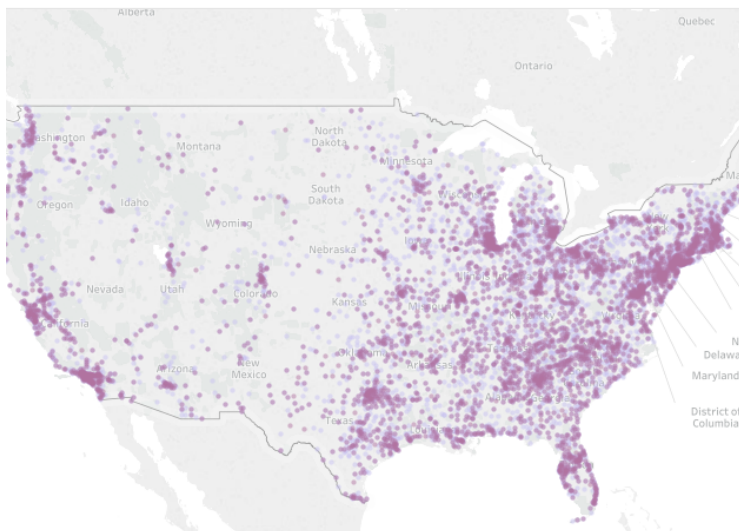
Gender
- ☐ Null
- ☑ Female
- ☑ Male
- ☐ Unisex

Applied for Services:
- ☑ User applied
- ☐ User did not apply

Age Bucket
- ☐ Null
- ☐ (15.0, 20.0]
- ☑ (20.0, 25.0]
- ☐ (25.0, 30.0]
- ☐ (30.0, 35.0]
- ☐ (35.0, 40.0]
- ☐ (40.0, 45.0]
- ☐ (45.0, 50.0]
- ☐ (50.0, 55.0]
- ☑ (55.0, 60.0]
- ☐ (60.0, 65.0]
- ☑ (65.0, 70.0]
- ☑ (70.0, 75.0]
- ☐ (75.0, 80.0]
- ☐ (80.0, 85.0]
- ☐ (85.0, 90.0]
- ☐ (90.0, 95.0]
- ☐ (95.0, 100.0]

Credit Score Bucket
- ☑ (495.0, 500.0]
- ☑ (500.0, 505.0]
- ☑ (505.0, 510.0]
- ☑ (510.0, 515.0]
- ☑ (515.0, 520.0]
- ☑ (520.0, 525.0]
- ☑ (525.0, 530.0]
- ☑ (530.0, 535.0]
- ☑ (535.0, 540.0]
- ☑ (540.0, 545.0]
- ☑ (545.0, 550.0]
- ☑ (550.0, 555.0]
- ☑ (555.0, 560.0]
- ☑ (560.0, 565.0]
- ☑ (565.0, 570.0]
- ☑ (570.0, 575.0]
- ☑ (575.0, 580.0]
- ☑ (580.0, 585.0]
- ☑ (585.0, 590.0]
- ☑ (590.0, 595.0]
- ☑ (595.0, 600.0]
- ☑ (600.0, 605.0]
- ☑ (605.0, 610.0]
- ☑ (610.0, 615.0]
- ☑ (615.0, 620.0]
- ☑ (620.0, 625.0]
- ☑ (625.0, 630.0]
- ☑ (630.0, 635.0]
- ☑ (635.0, 640.0]
- ☑ (640.0, 645.0]
- (645.0, 650.0]

Figure 3.a: Tableau Dashboard showcasing Credit Sesame user information across the United States. If opened in Tableau, this graphic is interactive and can be filtered using the categories on the right.

Conclusion

We present a **random forest model** which can accurately predict user conversion into valuable customer 71% of the time, given only information about their user profile and their behavior using the service in the past 30 days. We also isolate a few factors that are most impactful in user conversion, which include both factors involved at account creation as well as after the 30 day engagement period. Credit Sesame can focus on these factors specifically when interpreting the results from our model.

We also present a handy **Tableau visualization dashboard** which allows Credit Sesame to view these important factors as well as the distribution of customers across the country according to several categorical buckets. This is useful in quickly determining location-based demographic info of current users and could be used in the future to predict patterns in new ones.