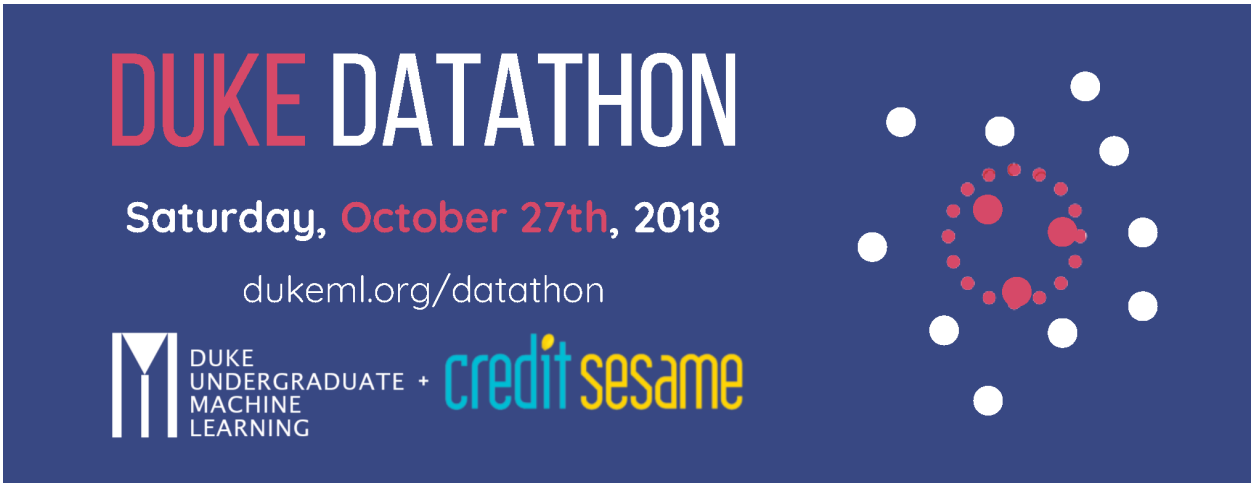


# Duke Datathon 2018

*Sponsored by Credit Sesame*

Credit Sesame (CS) is providing three datasets for Duke Datathon 2018, with credit report information as well as user engagement behavior in the first 30 days after signup. The users in the datasets joined CS during the month of July 2018, and had a credit score upon signup in the range of 500 to 800.

**Your team's goal is to analyze and/or visualize any (or all) of the data in a creative and insightful way.** Pretend you're a team of data scientists at Credit Sesame, and you've been tasked with exploring a few datasets over eight hours to create as much value as possible for the company. Section III details potential questions and prompts for you to explore, but you're free to formulate and pursue any questions or visualizations you think might be interesting! Feel free to ask a mentor for help, and attend a workshop to learn some data science and engineering skills and techniques.

The dataset can be found at dukeml.org/dataset, and the password to the file will be provided at 9:30am on Saturday, October 27, 2018 via email and Slack. Please download the dataset as soon as possible to prevent possible network issues on Duke's wifi!

**Instructions for submission:** submissions will be due at 5:30pm on Saturday, October 27, 2018 at dukeml.org/submit, and projects will be judged on design, creativity, technicality, and presentation. The top five teams will be selected to present, and we'll award prizes to the top eight teams! In Section I, you'll find more details about the competition.

## 1. Datathon Introduction

Below you'll find a basic schedule of the event:

**9am**, Registration & Breakfast
**9:30am**, Datathon Begins (password to dataset released)
**10am-2pm**, Workshops
**12pm**, Networking Lunch
**5:30pm**, Submission Deadline
**5:45pm**, Reception & Dinner
**6:15pm**, Presentations & Judging
**7:30pm**, Winners Announced
**7:30pm**, Event Close

There are also several workshops scheduled at the event, including: Exploratory Data Analysis in R, Exploratory Data Analysis in Python, Storytelling for Action (Data Visualization Best Practices), Introduction to Tensorflow, Applied Time Series Forecasting in R, and Variational Inference. A full schedule, including timings and locations, will be available at the start of the event.

Make sure to take advantage of mentors walking around and through Slack. You are encouraged to discuss your work with other teams, mentors, and students, and can use online and offline resources. However, a maximum of four individuals (the members of your team) should make large, meaningful contributions to your submission in fairness to all teams participating in the competition. **Teams must submit the following materials by the 5:30pm deadline.** It is recommended that teams work continuously from the beginning on their deliverables rather than finish it all within the last hour. You should begin working on the deliverables *at least* three hours before the deadline.

### A. Written Report

Teams must write a report that describes the steps taken to answer their proposed question or prompt. There is no set format for how the report should be written, but example sections of the report can include, but are not limited to, the following:

- Introduction: what question are you answering with the data, and why is it important?
- Data Engineering Process: how did you clean and prepare the data, and what data did you use?
- Analysis: what analytical techniques did you use, and why?
- Findings: what did you discover (include visualizations)?
- Conclusion: what can a layperson at Credit Sesame conclude from your team's work?

At minimum, the report must include the question being answered, findings and visualizations, and a conclusion. There is neither a word nor page limit, but it is recommended that you be as concise as possible. From the report, it should be clear as to how you approached your

analysis. Name the report as **team#_report.pdf**, and do not include any identifying information about the members of your group in the submission. A team number will be assigned to you closer to the submission deadline.

### B. Slide Deck Presentation

Teams will be required to submit a slide deck presentation (up to 10 slides). The goal of the presentation is to guide judges on how you utilized the available data to answer the question you came up with. The presentation should include meaningful visualizations, text, video, and/or other relevant multimedia content. It is up to your discretion as to what kind of material you would like to put in the presentation, but the analytical process, findings, and conclusion should be clear. In general, the content in the presentation should be a condensed version of the written report. Name the presentation as **team#_presentation.pdf**, and do not include any identifying information about the members of your group in the submission. A team number will be assigned to you closer to the submission deadline.

### C. Programs

All programs written during the competition will need to be submitted. The programs can be messy, uncommented, in multiple files, etc., and will not be judged on their quality. Put all programs into a folder, and name the folder as **team#_programs**, and do not include any identifying information about the members of your group in the submission. A team number will be assigned to you closer to the submission deadline.

### D. Submission

In order for judges to properly evaluate a team's performance, it is required that teams submit a .zip file containing the written report, the slide deck presentation, and the programs used for the analysis. If there are difficulties compressing the file into a .zip format, please reach out to a mentor or staff member for guidance. Name the .zip as **team#_submission.zip**, and do not include any identifying information about the members of your group in the submission. A team number will be assigned to you closer to the submission deadline. **Submit your work at dukeml.org/submit by the deadline.**

**Ensure that all materials are submitted by 5:30pm.** *Unfortunately, in fairness to all teams participating in the competition, we cannot offer any extensions to the deadline.* A group of judges will review the submissions and select the top eight finalists. The top five finalists will be selected to present their work at approximately 6:15pm, and a panelist of professors at Duke will rank the finalists. We'll distribute $3,000 in prizes, as well as additional awards provided by sponsors, amongst the top eight teams, and wrap up the event!

## 2. Dataset Summary

Below you'll find high-level summaries of the datasets, including their approximate sizes. Full details on the fields available in each dataset (including name, description, and data type) can be found in the data dictionary file.

### A. User Profile

Description: snapshot of the user's demographic and credit profile information at the time of signup.
Size: 285,619 rows, 38 columns, 1 row per user

### B. First Session

Description: detailed user action logs of each user's first session on the site. A session is defined as an unbroken series of actions on the site (typically referred to as "clickstream"). One-to-many relationship between the user profile and this dataset defined by user ID.
Size: 8,755,480 rows, 16 columns, average 30.75 rows per user

### C. 30-day User Engagment

Description: summaries of the actions taken during each user session that occurred within the user's first 30 days. A session is defined as the time period on the site between login and explicit logout or automated timeout. This should be a fairly sparse dataset for the different types of events. There is a one-to-many relationship between the user profile and this dataset defined by user ID.
Size: 1,179,988 rows, 40 columns, average 4.14 rows per user

**The data dictionary can be found at dukeml.org/dictionary.**

## 3. Example Questions & Prompts

You're free to formulate and pursue any questions or visualizations you think might be interesting, but here are some questions and prompts that you can use as inspiration for your own research on the datasets if you are having trouble coming up with your own analysis.

### A. First Session Questions

1. Create user profiles and segments based on different types of first session interactions. Example attributes to base segments on:
    - Credit profile
    - Demographic
    - Actions taken (login & logout immediately, explore site, apply for a card/loan, etc.)
2. Identify what kinds of users sign up for a card (click apply event) in the first session. What actions lead up to it? Does their credit situation influence this decision?

### B. User Engagement Questions

1. What is the profile of the most engaged users, and what are ways to best predict the level of engagement for new users?
2. What are different patterns of engagement? Are there differences based on device or form factor (desktop vs. mobile web vs. mobile app)? Potential measures of engagement are:
    - Number of logins
    - Number of views/clicks per session
    - Product applications
    - Session length
    - Periodicity (how long between sessions, regularity of sessions)

## 4. Data Glossary

Below you'll find some important terms required to understand the dataset.

### A. General Credit Terms

- **Balance:** the current amount owed to the lender for a given tradeline.
- **Bankruptcy:** a bankruptcy is when consumers or businesses seek legal assistance when bills cannot be paid. There are different types of bankruptcies. In Chapter 7 bankruptcy, debts for an individual or household are discharged. In Chapter 13 bankruptcy, debts of an individual or household are restructured and repaid over three to five years, under bankruptcy court supervision. Chapter 11 bankruptcy allows businesses to restructure.
- **Credit Limit:** the amount of money that can be charged to a credit card.
- **Inquiry:** a credit inquiry is created when a lender pulls someone's credit record. There are hard inquiries which affect your credit score and soft inquiries which do not. Inquiries counted in the datasets will exclusively be hard inquiries.
- **Tradeline:** a tradeline is the most common type of entry found on credit reports. Each tradeline represents a credit account that has been reported to a credit bureau. It contains detailed information about the account, including the type of account, the account number, account owner, and payment status. The tradeline also shows when the account was open (or closed), the credit limit, payment history, balance, and the date of last activity.
- **Utilization Ratio:** the ratio of the user's current balance to their credit limit, and could be phrased as the balance-to-limit ratio or credit-available-to-credit-used ratio. A lower utilization indicates the user is using less credit than they have available and will result in a higher credit score. A higher utilization means that the user is nearing their limit and will result in a lower credit score.

### B. Tradeline Account Types

- **Auto Loan:** a loan taken out for the specific purchase of paying for an automobile.
- **Collection:** if a loan becomes past-due for too long, it will be moved to a collections department or agency and becomes a collection account. Having collection accounts very negatively impacts a person's credit score.
- **Credit Card:** a credit card is a payment card that is accepted by merchants, and which can be read at the point of sale. Credit cards offer revolving lines of credit to cardholders, which means they have the ability to pay balances over time.
- **Derogatory:** this can be any type of account that is currently past-due on its payments.

- **Installment:** an installment loan is a loan in which equal, periodic payments are made for a defined period of time.
- **Mortgage:** a legal agreement by which a bank or other creditor lends money at interest in exchange for taking title of the debtor's property, with the condition that the conveyance of title becomes void upon the payment of the debt.
- **Secured:** a secured debt is one in which a borrower pledges property—most commonly, a home, a car, or cash—as collateral. If the borrower defaults on the loan, the lender may seize the property. In the case of secured credit cards, the collateral is cash.
- **Unsecured:** an unsecured debt is one that is not backed by collateral. Unsecured debt includes credit card debt, medical bills, utility bills, and any other type of credit that was extended without collateral. When a loan is backed by collateral, such as a house or car, it's known as secured debt. Unsecured debt can be wiped out by bankruptcy.

### C. User Engagement Action Types

- **VIEW_PAGE:** the user viewed this page.
- **VIEW_OFFER:** offer was displayed on the page the user viewed (always tied to a VIEW_PAGE event).
- **CLICK:** the user clicked on the page.
- **CLICK_APPLY:** the user clicked on an item being viewed that took them to an affiliate link.