Team # 43

DS4AWS: Capstone Project Proposal

Problem Statement: *(please describe the project in 1-2 sentences)*

As of 2019, women-owned businesses are growing 2X faster than all businesses nationwide, and yet, they only receive 2.8% of venture capital (VC) funding.

Our goal is to help female entrepreneurs from fundraising perspectives by producing models to predict fundraising capabilities and investigate gender disparity.

Which question(s) do you want to explore? Why do you think this particular question is interesting?: *(1-2 paragraphs elaborating on the project's relevance)*

An overarching question we are interested in exploring is: what does it take for a women-owned business to receive VC funding? We'll mainly explore this question in 2 directions:

1. Prediction: We would like to produce a model to predict which startups will proceed to further rounds of funding, which startups will get acquired, and which startups will file for an IPO by looking at different characteristics of the startups, such as the industry it is in, and the diversity of the founders/leaders.

2. Bias Identification: We would like to study the relationship between the number of female partners that a VC has and the likelihood that the VC will invest in women-owned businesses.

This project is relevant because the gender disparity in access to financial capital still persists to this day. Venture capital is a critical area to understand because VC is a prime driver for economic growth and employment, which is especially significant now as countries recover from the pandemic's impact on the economy. By analyzing data, we will be able to evaluate the state of diversity in the VC space and quantify why this underrepresentation exists as well as how it impacts female entrepreneurs.

Which datasets do you plan to use? Why? Are there any data sources that you have failed to find? *(List relevant datasets)*

- We are exporting data from Crunchbase (Scope: US and 2019 Only):

  https://drive.google.com/file/d/1Dxjapjck76Pr0WmCPBgm2EhiPgCvtz6X/view?usp=sharing

- Crunchbase data on startup investments:

  https://drive.google.com/file/d/1peEJ3vXdvDdji33DhwGBC9eGWN5XQrI1/view?usp=sharing

- Kaggle's Startup Founders data:

  https://drive.google.com/file/d/1BFHkAFdF20xUq2XjFN9E60VVxT9UCa-V/view?usp=sharing

- NVCA

- U.S. Census Bureau's Survey of Business Owners and Self-Employed Persons (SBO) datasets:

  https://www.census.gov/programs-surveys/sbo.html

- Social + Capital and The Information's data on diversity in venture firms:

  https://docs.google.com/spreadsheets/d/1GT5nUwbW7oPy0-dSAPCtmTF_rg5ug3CJRGpYFsGa-DQ/edit#gid=0

- Kaggle's Startup Investments data:

  https://www.kaggle.com/arindam235/startup-investments-crunchbase

- Kaggle's Failed Startups data (Although it is not a large dataset, we can gain insights from the failed reasons):

  https://drive.google.com/file/d/1BFHkAFdF20xUq2XjFN9E60VVxT9UCa-V/view?usp=sharing

- U.S. Bureau of Economic Analysis on GDP data:

  https://www.bea.gov/

- U.S. Bureau of Labor Statistics on current employment data:

  https://www.bls.gov/ces/

- Although it is not a dataset, another resource that we will look at is PitchBook's VC Female Founders Dashboard.

Please describe the plan or methodology that you will use to answer your question *(1-2 sentence description of statistical analysis techniques)*

We will start with data discovery and data profiling to understand the datasets we have and explore the patterns and trends inside them. Then, we plan to adopt regression analysis, time series analysis, and hypothesis testing:

- Regression analysis for estimating the fundraising capabilities of women-owned businesses
- Time series analysis for tracking and analyzing VC investment trends over time
- Hypothesis testing for investigating if gender disparity affects VC investments in female entrepreneurs

Problem Statement: *(please describe the project in 1-2 sentences)*

Although Lyft is known for acquiring and engaging passengers, it is also crucial for them to acquire and maintain adequate levels of supply—whether in terms of available drivers or scooters or bikes.

For this project, our goal is to calculate a Driver's Lifetime Value (i.e., the value of a driver to Lyft over the entire projected lifetime of a driver).

Which question(s) do you want to explore? Why do you think this particular question is interesting?: *(1-2 paragraphs elaborating on the project's relevance)*

The questions we will be exploring are:

1. What are the main factors that affect a driver's lifetime value?
2. What is the average projected lifetime of a driver? That is, once a driver is onboarded, how long do they typically continue driving with Lyft?
3. Do all drivers act alike? Are there specific segments of drivers that generate more value for Lyft than the average driver?
4. What actionable recommendations are there for the business?

These questions are interesting because Transportation-as-a-Service has been increasingly popular in the recent decades. With Lyft being one of the major businesses in the industry, it is important to understand the relationships that this service is built on, particularly the relationship with the drivers, so that Lyft can continue to thrive.

Which datasets do you plan to use? Why? Are there any data sources that you have failed to find? *(List relevant datasets)*

We will be using the datasets that Lyft provided us in this file:
https://drive.google.com/file/d/1fi97laYlBlilKli01N6ha72v-CuaD1kx/view?usp=sharing

Please describe the plan or methodology that you will use to answer your question *(1-2 sentence description of statistical analysis techniques)*

We will start with data discovery and data profiling to understand the datasets we have and explore the patterns and trends inside them. After that, we plan to analyze the data through:

- Measuring the relationship between the number of weekly rides and the number of weeks since the driver was onboarded

By studying the results from these graphs, we will then be able to provide solid recommendations for the business.