Analyzing VC Funding for Women-Owned Businesses

Team #43

1. Introduction

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

Our goal for this project is to help female entrepreneurs from fundraising perspectives by understanding the VC fundraising capabilities and investigating gender disparity. This leads us to our overarching question: what does it take for a women-owned business to receive VC funding? We explored this question in 2 directions:

1. Prediction: We produced 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 studied the relationship between the number of female partners that a VC has and the likelihood that the VC will invest in women-owned businesses.

After exploring this question, we will provide recommendations to both women-owned startups as well as VCs to contribute to closing the gender gap in VC funding.

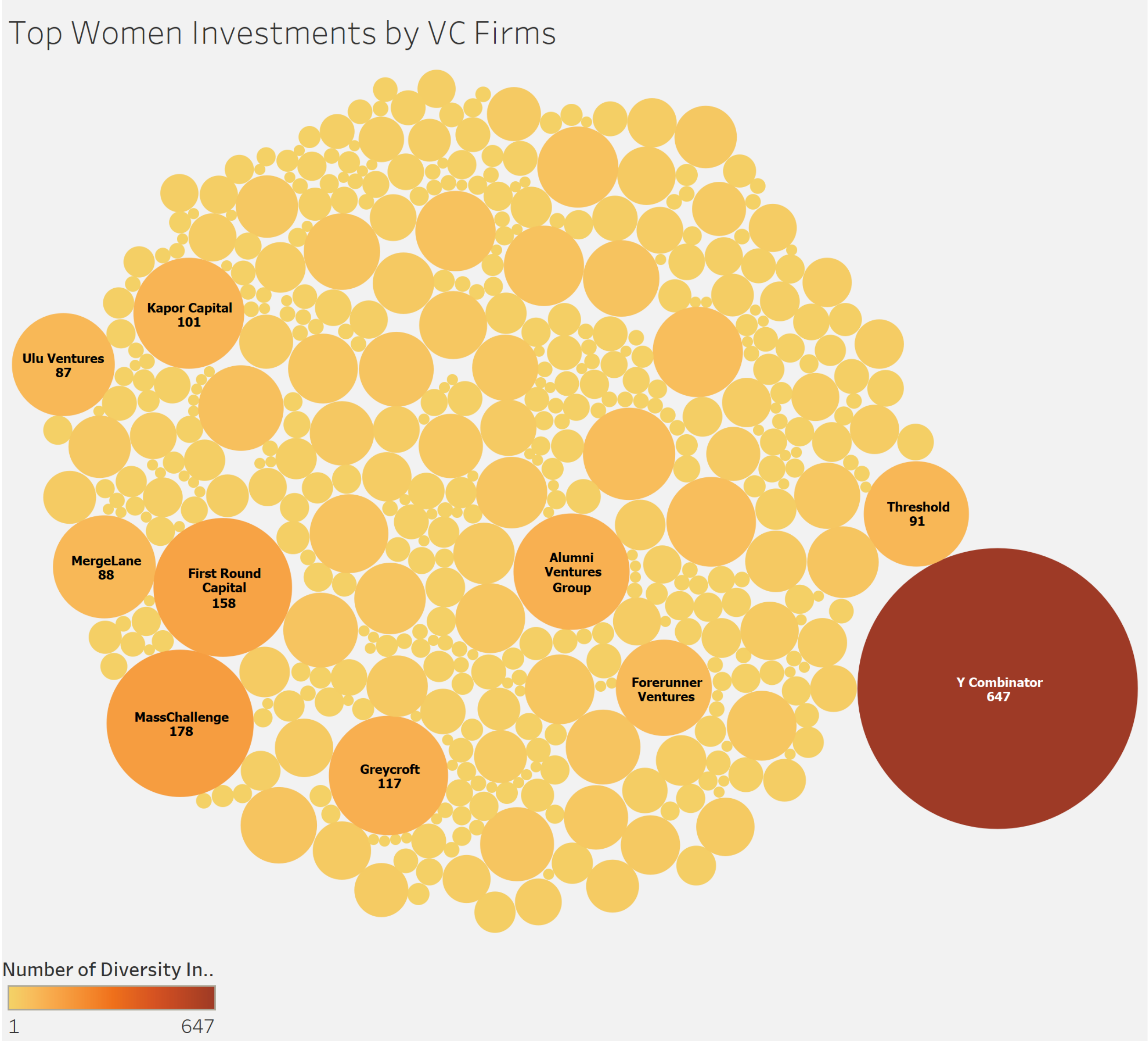
2. Data Analysis & Computation

A. Datasets + Data Wrangling & Cleaning

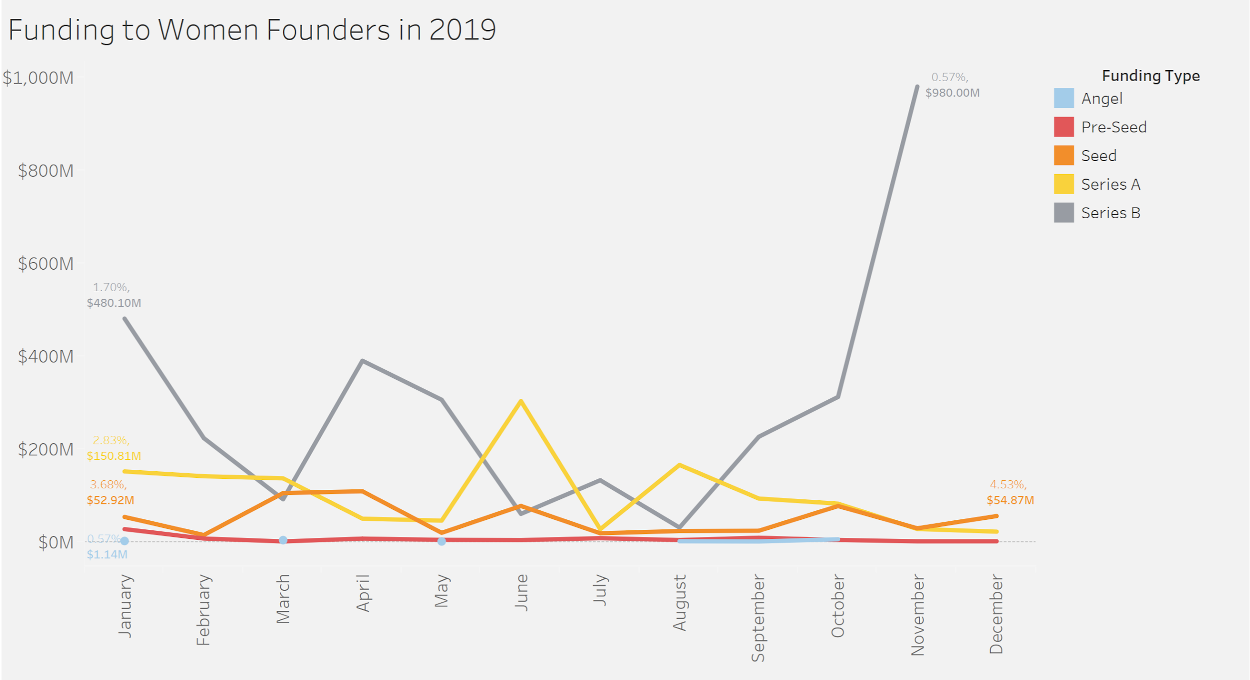
We leveraged an array of quantitative and qualitative data to create this tool. The primary source of our data was from Crunchbase. We selected Crunchbase because it contains highly available data and insights from early stage startups and Fortune 1000 companies. We scoped our data to only companies headquartered in the United States and investments only in 2019. One of the disadvantages of this dataset is that there is a lack of the qualitative data needed to understand why investors committed to or did not commit to funding organizations.

The dataset includes:

Investors - All investments made by investors, including active investors, number of diversity investments (51k records)



Funding Rounds - Details for each funding round/history for companies in the dataset and which investors participated in the most funding round (15k records)



Companies - Organization profiles available on the Crunchbase platform (including Diversity Spotlight data) (100k records)

We started data cleaning each of these tables. For the Investors table, we removed columns that were unrelated to our research so that we could reduce the runtime for our models and focus on the following columns:

- Organization/Person Name (column A)
- Organization/Person Name URL (column B)
- Number of Investments (column C)
- Number of Exits (column D)
- Location (column E)
- Gender (column I)
- Diversity Spotlight (column AD)
- Investor Type (column AS)
- Industry Groups (column BC)

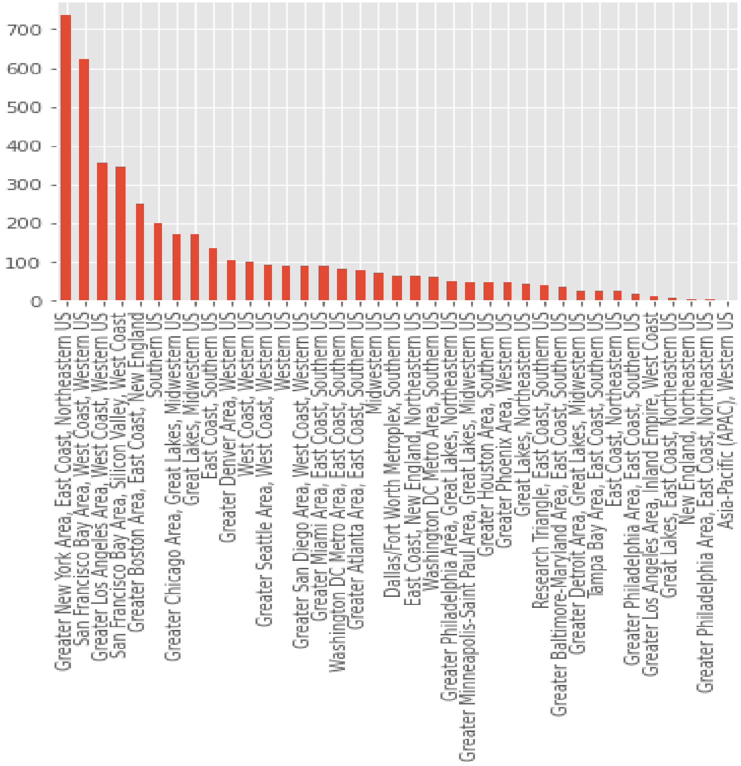
For the Companies table, we also removed columns that were unrelated to our research so that we could focus on the following columns:

- Organization Name (column A)
- Organization Name URL (column B)
- Headquarter Location (column D)
- Industry Groups (column AU)

- Total Funding Amount Currency (in USD)(column I)
- Founded Date (column J)
- Diversity Spotlight (US Only) (column M)
- Estimated Revenue Range (column N)
- Operating Status (column O)
- Number of Articles (column R)
- Company Type (column Y)
- Investor Type (AD)
- Number of Alumni (column AT)
- Number of Founders (column AV)
- Number of Employees (column AX)
- Number of Founding Rounds (column AY)
- Number of Investors (column BO)
- Number of Acquisitions (column BQ)
- Price Currency (in USD) (column BZ)
- SEMrush – Monthly Visits (column CW)

B. Exploratory Data Analysis

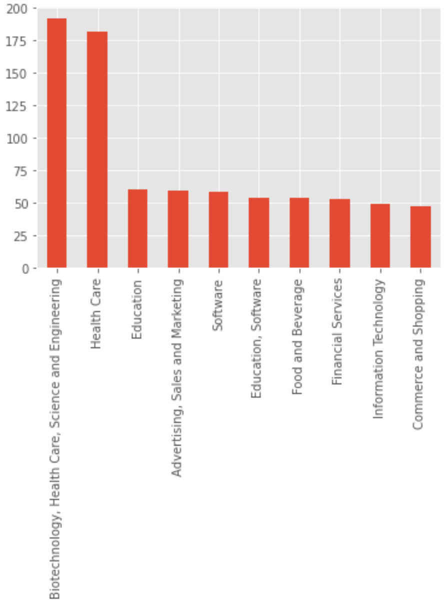
We first explored the popular headquarters for women-founded companies:



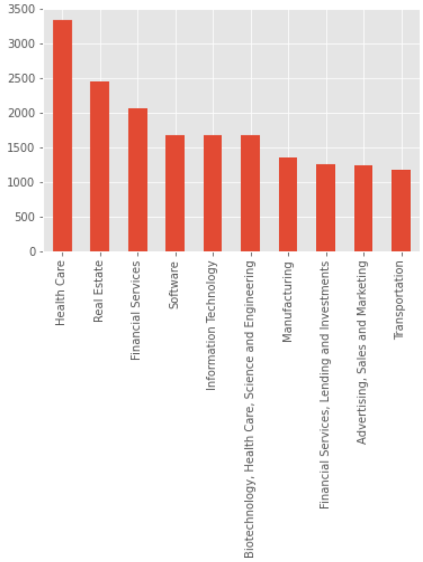
From these graphs, we can see that cities like New York, San Francisco, and Los Angeles appear to be the top choices of headquarters. The top headquarter choices may be attributed to more networking opportunities for fundraising.

Next, we looked at the top 10 industries in which women founded/led companies are in:

Women-led



Men-led

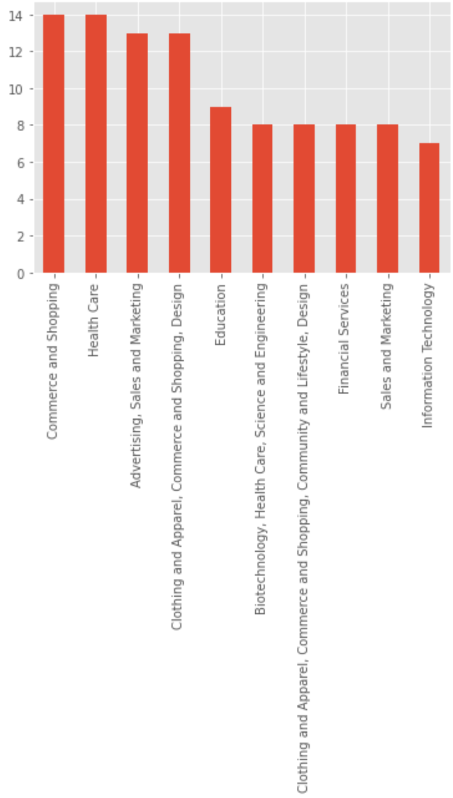


The top industry for both women-led and men-led companies is biotech/health care. Top industries where a large number of women-led companies are receiving funding are healthcare, education and advertisement, whereas those of men-led are real estate, software, and financial services. This may reflect the difference between the women's and men's focuses and biases prevalent in the industries.

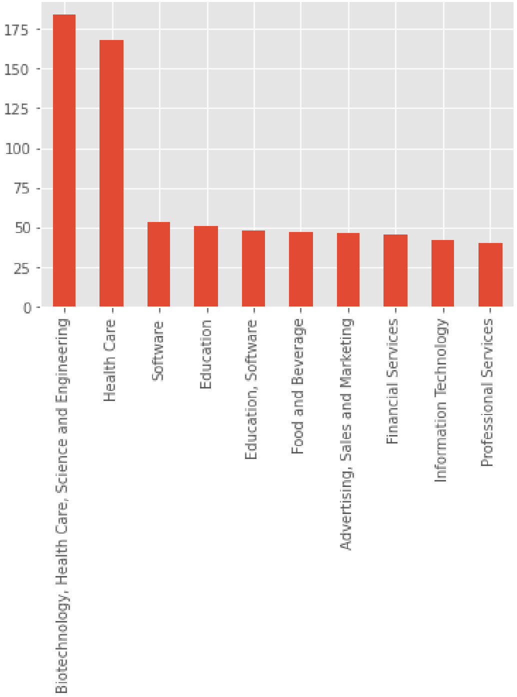
Then, we combined our analysis of both to focus on a city case study:

Top 10 Industries for NY-based Companies

Women-led

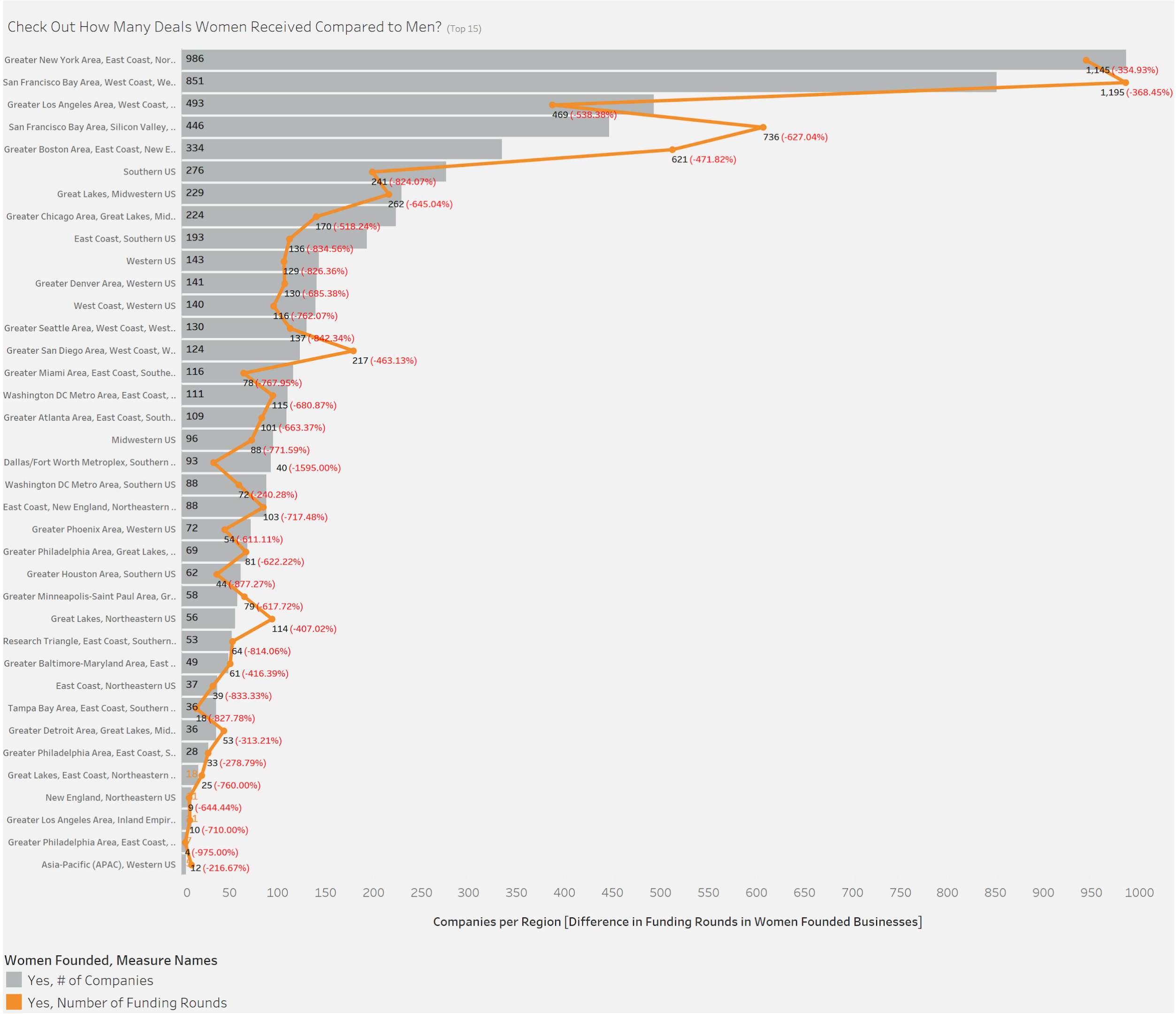


Men-led



The popular industries where women-led companies are getting funded in New York City are Shopping, Health Care, Advertising, Education, etc., while the popular industries for men are Biotechnology, Health Care, Software, etc. Financial Services appears to be at an equal position for both (7th most popular).

Lastly, we investigated the popularity of the different industry groups in general. We looked at the number of companies involv



Check Out How Many Deals Women Received Compared to Men? (Top 15)

ed in each industry group as well as the total funding for each industry. We found that the industry with the most funding was Science and Engineering, and the industry with the most companies was Software.

C. Statistical Analysis & Machine Learning

C.1. Selection of Factors

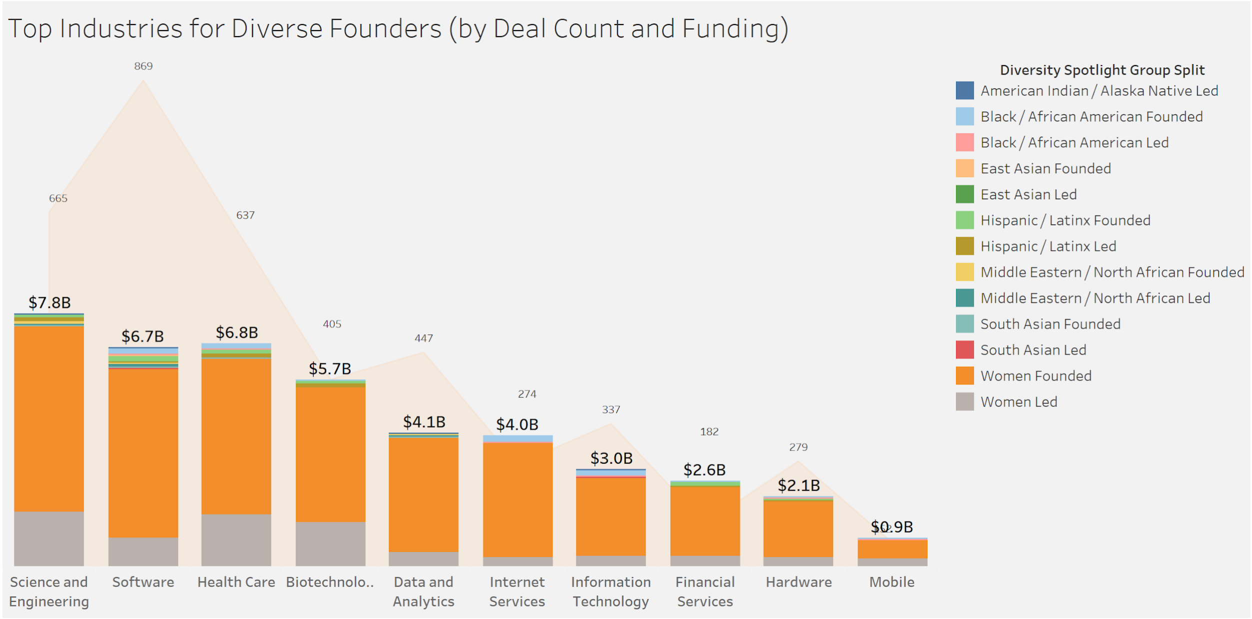
By conducting interviews with VC (Venture Capital) investors, the following four variables are identified to be the most important success factors for the women-founded/led startups:

- Estimated Revenue Range
- Total Funding Amount
- Number of Investments

- Number of Exits (IPO)

According to the VC interview results, women-founded/led businesses get more fundraising probabilities if their estimated revenue ranges are over 1 million USD. However, a commonly observed situation is that women founders and leaders tend to underestimate their potential revenue and thus have lower estimated revenue than they can actually achieve. Such underestimate reduces their funding opportunities.

In addition to the estimated revenue range, whether the women-founded/led startups have received 200,000 USD from the Angel Funding is a key differentiator of whether they attract VC investors' attention. As claimed by our interviewees (i.e., VC investors), many women entrepreneurs could not raise investors' interests not due to the lack of strong abilities but because of their current funding of less than 200,000 USD.

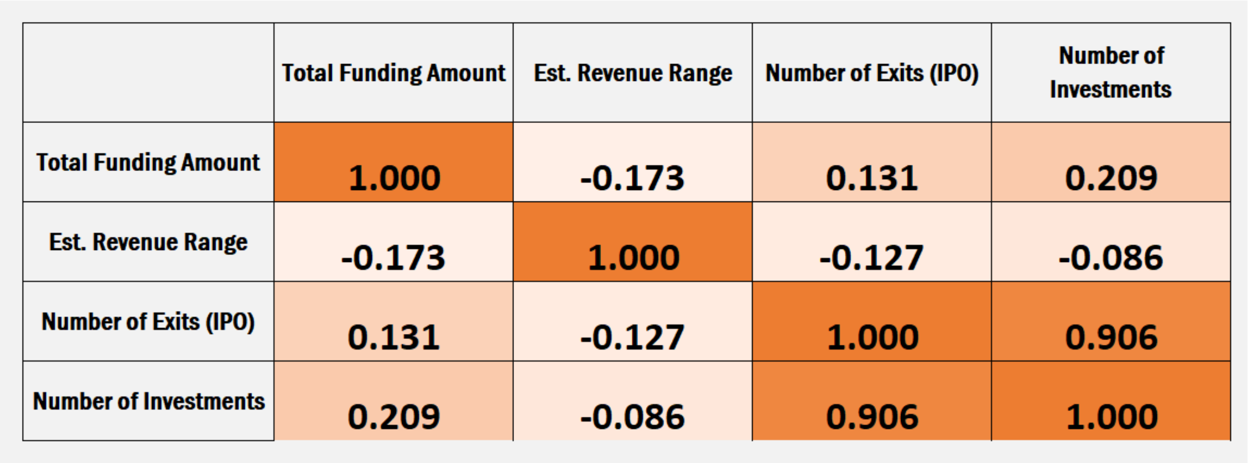


The number of investments that women-founded/led startups have already achieved is identified as a significant factor by VC investors to investigate the founders' abilities and the startups' prospects. The higher the number of investments, the higher the fundraising probability. Similarly, the number of exits (IPO) is used as funding criteria by VC investors to assess the companies' profitability.

C.2. Verification of Factors

To investigate and prove the importance of the factors Estimated Revenue Range and Total Funding Amount for women entrepreneurs, we first extracted the 6,019 data entries of women-founded/led businesses out of 109,043 Crunchbase data points in total. By utilizing the power of pandas dataframes, we get that 4,059 (67.43%) of the women-founded/led startups either received no funds or received some funds but less than 200,000 USD. Additionally, it is derived from the data that among 3839 women-founded/led companies which reported their estimated revenue range, 1758 (45.79%) of them had the estimated revenue of less than 1 million USD. Considering that the ratios of 67.43% and 45.79% are relatively high, the Estimated Revenue Range and the Total Funding Amount are vital variables to be included in the model in order to give useful insights to women entrepreneurs.

Moreover, the following correlation matrix is computed to investigate the collinearity between the variables:



	Total Funding Amount	Est. Revenue Range	Number of Exits (IPO)	Number of Investments
Total Funding Amount	1.000	-0.173	0.131	0.209
Est. Revenue Range	-0.173	1.000	-0.127	-0.086
Number of Exits (IPO)	0.131	-0.127	1.000	0.906
Number of Investments	0.209	-0.086	0.906	1.000

In the above correlation matrix, the diagonal elements from the top left to the bottom right corner are the variances of a variable with itself, which equals 1. The other elements in the matrix are the correlation between the variables in the x and y axises. It can be concluded that Total Funding Amount and Estimated Revenue Range are slightly and negatively correlated (-0.173). A similar situation is shared by the correlation between Estimated Revenue Range and Number of Exits (IPO) (-0.127). On the other hand, both Total Funding Amount and Number of Exits (IPO), and Total Funding Amount and Number of Investments have a sligh and positive correlation (0.131 and 0.209 separately). It is notable that the positive correlation between Number of Investments and Number of Exits (IPO) are relatively high (0.906). Such a high correlation is understandable in the real world situations.

C.3. Data Pre-processing for the Selected Model

Combining the insights gained from VC interviews with the time constraints and available data, we decided to classify the fundraising probability of women-founded/led businesses into three categories: low, moderate, and high. Thus, we chose the Random Forest classifier, which constructs a multitude of decision trees at training time and outputs the mode of the classifications.

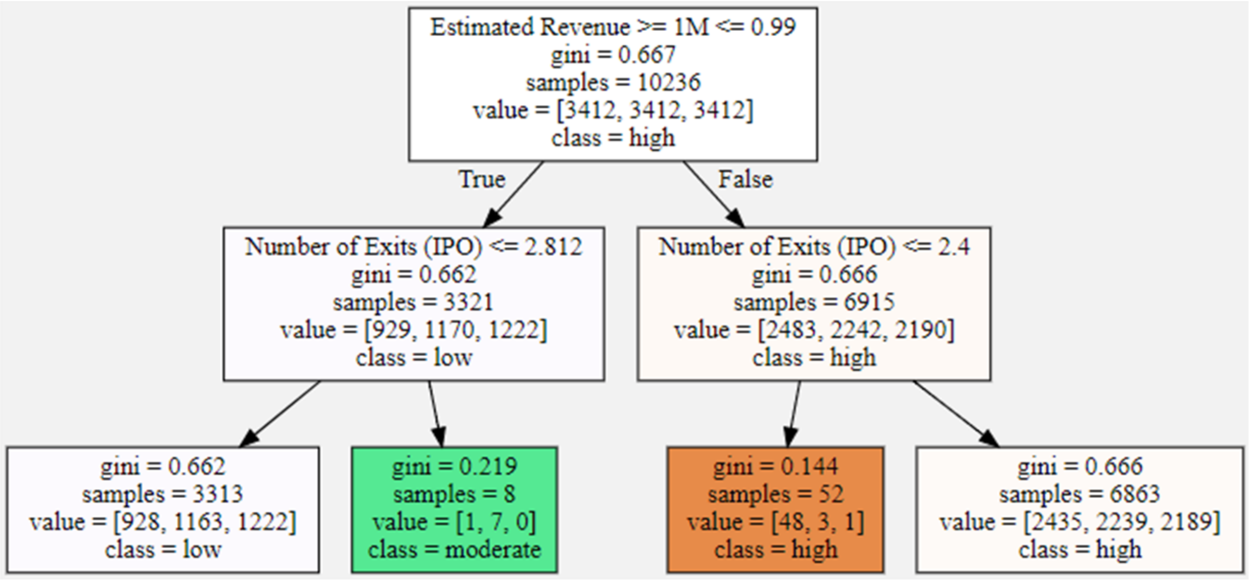
Since the original data is a mixture of data types such as floats and strings, we first pre-processed the 6019 data points of women-founded/led businesses by encoding the four variables discussed above (i.e., Total Funding Amount, Estimated Revenue Range, Number of Exits (IPO), and Number of Investments). As supported by the VC interview results, two of the thresholds of whether the companies could raise VC investors' interests are if the companies' estimated revenue is over 1 million USD and if their total funding amount (in USD) is over 200,000. Thus, we encoded the estimated revenue range of a company as 1 if it is over 1 million USD, and 0 otherwise. We used the factor Total Funding Amount to categorize the data into 3 classes: low, moderate, and high. 0.25, 0.5, and 0.75 quartiles of the total funding amount are adopted to differentiate the low, medium, and high classes separately. Additionally, whether the startup has received any investments or not is another major assessment criteria used by VC investors, which is represented by the variables Number of Investments and Number of Exits (IPO). In consonance with the criteria, we encoded the number of investments of a company as 1 if it has at least one investment, and 0 otherwise. Similarly, we encoded the number of exits (IPO) of a company as 1 if it has at least 1 IPO exit, and 0 otherwise.

Furthermore, KNN (k-nearest neighbours) imputation is adopted to predict the null values. The initiative of performing KNN imputation is that a drawback of the Crunchbase dataset is lots of null values in each column. The null values occur when companies choose not to report their status or not to participate in the funding rounds. Thus, it is inappropriate to fill the null values as 0 if companies have raised some funds or have estimated revenue but do not choose to report. It is also inappropriate to simply remove all null values as almost 95% of the Crunchbase data is composed of null values in at least one column. To solve the null value obstruction, we chose KNN imputation to predict the missing value by finding the mode of its k-nearest neighbours. The KNN algorithm has been proven to be generally effective to predict the missing values in a range of different models.

C.4. Modeling Process and Results

After preprocessing the data with encoding and KNN imputation, we encountered an imbalance class problem. There were many "low" and "moderate" classes in the dataset, which made the “high” class as a minority class. To address this problem, We used Synthetic Minority Oversampling Technique (SMOTE) to upsample the minority class (i.e., the “high” class).

As mentioned above, the factor Total Funding Amount is used to categorize the data into 3 classes (i.e., low, moderate, and high) according to the 0.25, 0.5, and 0.75 quartiles. Thus, the encoded results of the remaining variables Estimated Revenue Range, Number of Investments, and Number of Exits (IPO) act as three predictors in the random forest classifier. By setting the maximum depth of decision trees as 3 to avoid overfitting, we get a validation accuracy of 78.6%, which indicates that our random forest model classifies and predicts the fundraising probability of women-founded/led businesses with a generally high accuracy. This validation accuracy also proves that the selection of the variables can effectively predict the fundraising probability.

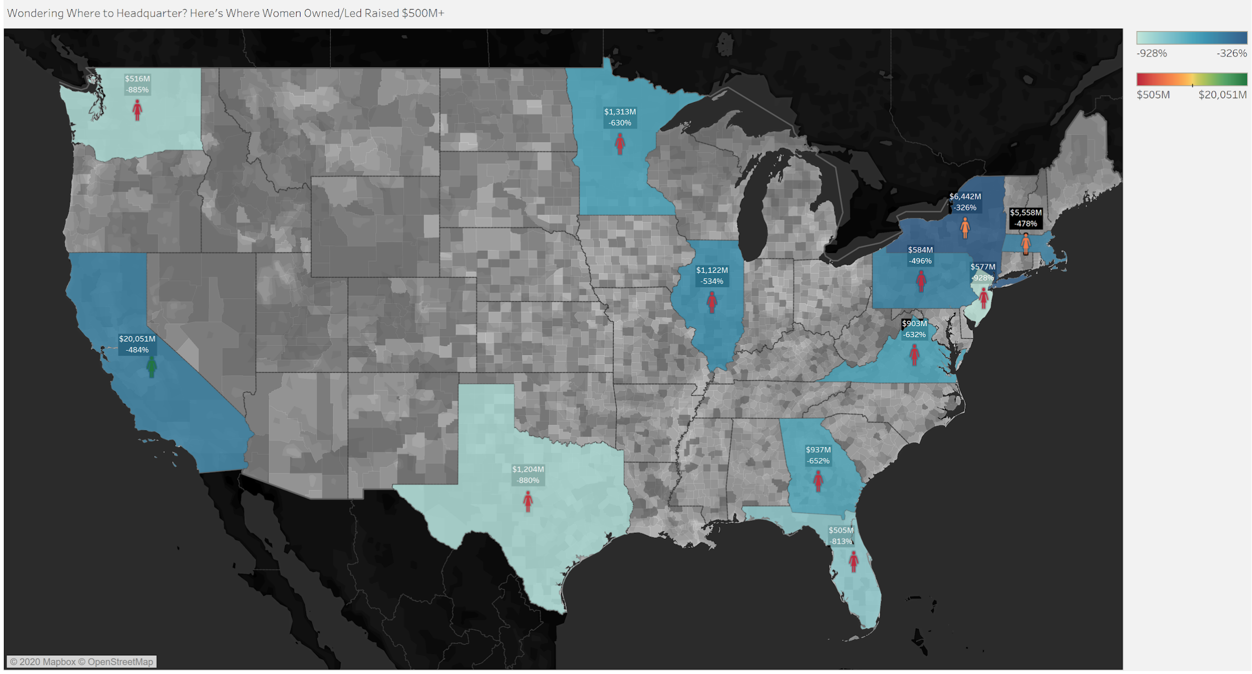


The above figure illustrates the decision process and the results produced by the random forest classifier. The model arranges the three predictors in the order which can distinguish and classify the data with a shortest path and in the most accurate way. In each node of the decision tree, the decision criteria is listed at the top and a gini index is followed to indicate the quality of the node. The gini index represents the probability of an object being wrongly classified. Thus, the lower the gini index, the better the quality of the node and the order of decision criterias. The above figure also gives the node information such as how many samples are contained in the node, how many objects are in each class in the node (which is shown by the “value” list corresponding to the number of objects in the “high”, “moderate”, and “low” class separately), and which class (i.e., high, moderate, or low) dominates the current node according to the number of objects in each class.

According to the order of decision criterias produced by the model, it is vitally important for women entrepreneurs to confidently and positively estimate their revenues instead of underestimating the expected revenue, especially when they pitch their startups before VC investors. Additionally, the number of exits (IPO) also plays a significant role when women founders/leaders want to maximize their fundraising probabilities. Since the number of investments is highly and positively correlated with the number of exits (IPO), it is reasonable to use only the number of exits (IPO) in the decision criteria instead of adopting both. The usage of the number of exits (IPO) and the discard of the number of investments also indicate that the decision tree and random forest classifiers perform well with variables that are highly correlated with each other.

3. Conclusions and Future Work

From our analysis, we were able to conclude that the gender disparity in funding does exist, but there are ways for women-owned businesses to increase their VC funding. For example, from our graphs, since we noticed that there were states where there was a disproportionately low number of deals received by women in comparison to men, we would suggest women-owned businesses to have their company headquarters in other states where the ratio is higher, such as California, New Jersey, and New York. Some states like New Jersey also have startup incubator or accelerator programs, which would increase the probability that the startup receives a higher amount of funding.



Wondering Where to Headquarter? Here's Where Women Owned/Led Raised $500M+

In order to improve our work, we would use more data and more qualitative data because there is a huge deficit in the availability of data. There was a significant amount of null values in numerous fields. Ensuring more data is recorded and is available for analysis can help us in understanding the start-up and funding culture better which in turn will help in providing better recommendations to women founded and minority founded businesses. Furthermore, our model currently only categorizes the funding that a female founder would receive as low, moderate, or high. For future work, we would improve the model so that we can quantify the estimated funding amount that women-owned businesses will receive. That way, we can provide more targeted recommendations for the businesses.