Finding Startup Unicorns with Web Traffic Data

Intern Name
Company Name

Data has become overwhelming for investors

- Across the industry, more alternative data is being used to find investment opportunities and keeping up with all the data has become overwhelming. This is true with our analysts having to take on web traffic data.

- With so many new companies starting every year, it becomes difficult to find which startups have high growth potential when analysts have to manually assess each company

This internship streamlined this data navigation issue

- Created monthly automated reports for web traffic data, our newest data source, detailing the condition of a company's website using the website SimilarWeb

- Developed a model to assess an e-commerce and/or online company's website data to predict their likelihood of raising more capital in the future as a way of identifying potential investment opportunities without as much manual research

Web traffic data can now be easily gathered with an API

- As a way to prevent investment analysts from having to manually gather data from SimilarWeb, an API wrapper was created as a simple solution to **corral all the web data with just one line of code**

- This was made by **leveraging our existing API wrappers** from other alternative data sources such as People Data Labs

- Also used the provided documentation within the SimilarWeb platform

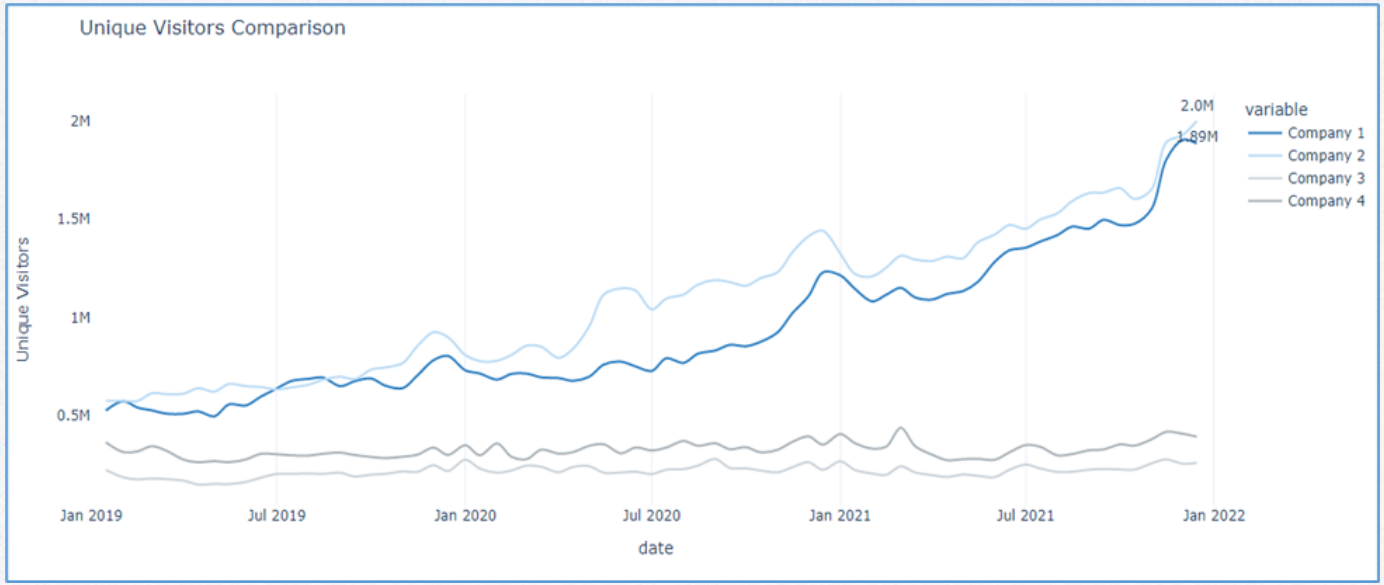
All gathered data is stored within our existing databases

- The **API wrapper was connected to a relational database** to store all gathered information as a way to control against gathering duplicated data and wasting API credits

- This was done using a snowflake schema where the company name is the primary key as a way to **remain consistent with our larger database system**

- **All data gathered now will be inserted into our AWS data**, allowing web traffic data to be used in any future tools or research

The wrapper and database created the automate report



Report Information

- Monthly Report
- Available for all Companies
- Automated via AWS
- Cost $6 per report

Building the model was split into three steps

Model Definition

Define the desired input and output of the model

Explore Relationships

Initial exploratory analysis to find most meaningful variables

Test Different Models

Test different models and parameters to find most accurate

Both web traffic and financial data for 71 companies framed the model

Data gathered for 71 private e-commerce and/or online companies

Dependent

Web Traffic Data

Transformed to periodic growth or average

- Site Visits
- Unique Visitors
- Pages per Visit
- Bounce Rate
- Average Visit Duration
- Traffic Sources

Financial Data (Various Sources)

- Estimated Revenue
- Estimated Units

Independent

Did a company raise money in:

Proxy for finding companies with high growth potential

Next month
Next quarter
Next 6 months
Next year

Predicting funding in six-months was most practical and meaningful

Dependent

Web Traffic Data

- **Site Visits**
- Unique Visitors
- Pages per Visit
- **Bounce Rate**
- Average Visit Duration
- Traffic Sources

Financial Data (Various Sources)

- Estimated Revenue
- Estimated Units

Other Information (Manually Entered)

- **Type of Product Offering**
- **Industry**
- **Seasonality**

Independent

Did a company raise money in:

Proxy for finding companies with high growth potential

Next month
Next quarter
Next 6 months
Next year

The final model has a 50% batting average and has been put in production

	Predicted No Raise	Predicted Yes Raised
Actual No Raise	**7**	3
Actual Yes Raised	1	**3**

Overall Accuracy: **71%**

Recall: **75%**

Batting Average (Precision): **50%**

Typical VC Batting Average: **15%–20%**

This model has been included in the monthly report to help improve investment decisions

Next Steps:
Expand Modeling and Automation Bias

- Expand this modeling to include companies who aren’t web based to make it more impactful and less niche

- Educate investment analysts on automation bias to avoid bad investment opportunities or overlooking good ones

QUESTIONS