Recommendation Monkey
Data and Decisions
Sponser: Bill.com
Team Inspector Royale

Derek Garcia  
Project Lead

Noa Kim  
Scribe  
Database

Carson Coley  
Front-End

Jonas Zhang  
Statistics

Joe Zhuang  
Backend
Challenge

Architect and develop a scalable data science lab to analyze, predict, and recommend the most promising U.S. small and medium size industries and businesses to pursue

Data Set Characteristics

- 10.4m real transactions representing $36B in business transactions
- 5 years of financial transactions (2015 to 2019)
- 100K+ entities and 100’s of Dun and Bradstreet industries
What We Did and How We Did It

**Fall 2021** - “Build the Model T” and drive it around the block

- Design the architecture/code from scratch
- Test lab against 4M record data set

**Winter 2022** - “Build the Ferrari” and race it

- Refactor lab for scalability
- Created a Platform
- Test lab against 10M record data set

“Bill.com is going to immediately use this scalable data science lab platform and continue to extend and scale this lab into the AWS EC2 Cloud”

-stu ogawa
System Architecture

Big Data

- SQLite
- DB Browser

Port Forwarding

Analyze/Interact

- Python
- PyTorch
- Matplotlib
- Jupyter

Recommend

Invest in these Industries
1. Company1
2. Company2
3. Company3

Invest in these companies:
4. Company1
5. Company2
6. Company3
Refactored Jupyter User Experience

- Easier user experience navigation
- Quicker results and recommendations
- Faster computations
- Improves efficiency of statistical algorithms
- Visualization of results more user-friendly
3.5 CPU vs GPU “head to head” compute experiment

- Dell laptop (2015),
  - AMD 4 core CPU
    - 16GB of RAM
  - GTX 6 core GPU
    - 4GB of RAM
    - 500GB SSD
- PyTorch ver. 1.10.1
- GPU has 10x reduction in compute time = superior hardware for scalability
- K-means computation for one industry takes ~300ms via optimization

CPU: 10M records processed in 10 minutes
GPU: 10M records processed in 1 minute
3.3 & 3.4: Derivative Analysis on Middle Cluster of K-means

- 1st graph = change in trans amt over time
- 2nd graph = change in trans velocity over time
- We can see which vendors are increasing at an increasing rate = potential for exponential growth

X-axis = Vendor ID
Y-axis = Net change in Trans Velocity/Acceleration
Results 3 - Predict which vendors/customers increasing at an increasing rate

- Time Series moving average for long term trends
  - Four-year data forecast two-year transactions

Business recommendation: use these top 20 business vendors as a “seed” / training data to perform subsequent machine learning and deep learning predictions
SARIMA: “moving-average” machine learning model for predicting transaction amounts of an industry

1st Step:
Train model w/ 3 years of data
- **Blue** -> training data
- **Green** -> model’s prediction
- **Orange** -> actual values
- **Grey** -> upper/lower bounds
4. Prioritize and Recommend Vendor/Customer pairs with top subscription revenue

- top ten vendors with subscription-based revenue
  - used LAG FUNCTION
  - Between 25-35 days, +or- 10% difference in transaction amount

Business recommendation: use the top 20 business vendors as training data to perform subsequent machine learning.

X-axis = Sender ID -> Receiver ID
Y-axis = Total transaction value of recurring transactions
2. Regression Analysis for Best Fitting Function & Future Prediction

Use Polyfit to find best fitting function with designated order/derivative

Predict future performance of a certain industry
Final Recommendations

Top 5 Growth Potential
- K-means & derivatives
- Regression analysis
- Arts, Entertainment and Rec.
- Professional, Science and Tech
- Education

Top 5 Consistency
- Time-series analysis
- Subscription analysis
- Health Care
- Manufacturing
- Transportation

Information Administrative
Special thanks:

Kevin Pierce  
Tech Advisor

Stu Ogawa  
Mentor

Chris Nishimura  
Tech Advisor