H1B Petition Prediction Midterm Review

Hyung C. Park (hcp084)
5/15/2019

Highlights

○ More research of the problem statement showed that H1-B approval data is not easily accessible and had to be scrapped so the petition data was used. I learned that certified petition is different from H1-B approval because petition just allows people to file for H1-B.
○ EDA/Feature Engineering (more details in the demo and analysis slides)
- The number of certified petitions grew significantly for the past few years, while the number of rejections and withdraw stayed the same
- There were many outliers to the pay level, so a new feature was created that maps pay to bin based on quartiles to keep the number of datapoints per bucket the same
- Computer related jobs seem to dominate the H1B petitions

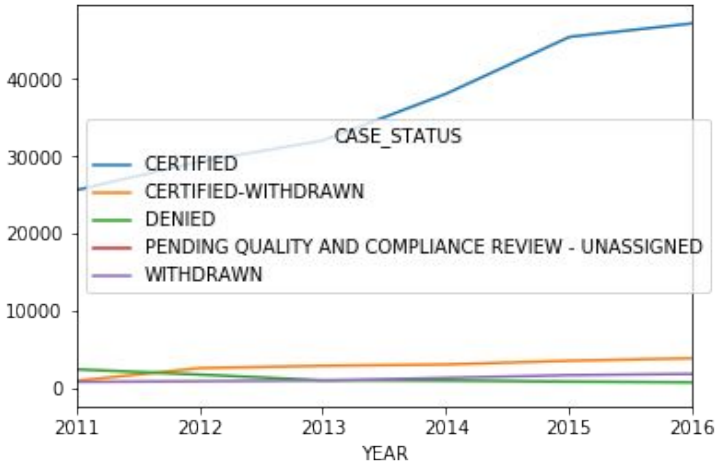
○ Data has been secured on Kaggle and uploaded to AWS S3
○ A table has been created in RDS based on the features generated during and after EDA

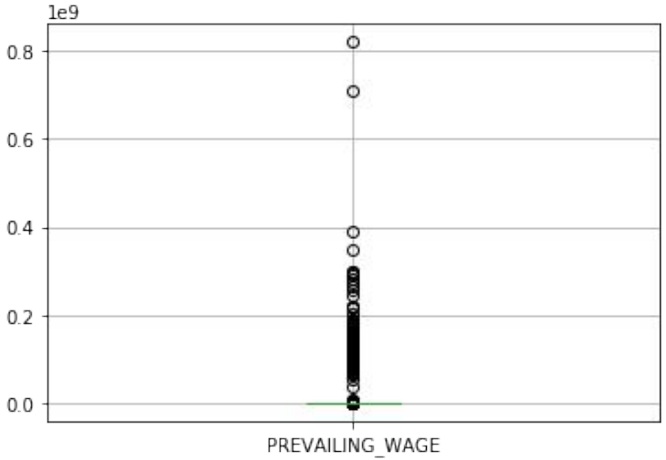
Review Process

- Epic 1: Understanding the Problem
  - Story 1: Understanding the H1B Process
- Epic 2: Data Collection and Exploration
  - Story 1: Finding the dataset and downloading an available data set
  - Story 2: EDA
- Epic 3: Feature Engineering
  - Story 1: Created new features for clustering (categorizing numerical variables to bins)
- Epic 4: Data in AWS
  - Story 1: Raw data in S3
  - Story 2: Table in RDS
  - Story 3: Connected platform between S3, EC2, RDS, and MSiA 423 server

Demo/Analysis

1067	SOFTWARE DEVELOPERS, SYSTEMS SOFTWARE	3195
771	Management Analysts	3229
1131	Software Developers, Systems Software	3803
292	Computer Occupations, All Other	5806
197	COMPUTER OCCUPATIONS, ALL OTHER	8977
204	COMPUTER PROGRAMMERS	14321
1128	Software Developers, Applications	16296
222	COMPUTER SYSTEMS ANALYSTS	17815
1058	SOFTWARE DEVELOPERS, APPLICATIONS	18516
297	Computer Programmers	18902
311	Computer Systems Analysts	24237





Lessons Learned

- I was unfamiliar with setting up S3 and RDS and how to leverage these tools alongside EC2, and this project has shown why they are so useful
- Setting up credentials and learning what should be on the server what should not and setting up connections between multiple nodes has been very helpful
- I learned how useful logging can be because it helped me troubleshoot what parts of the code is not functioning the way I want it to
- It was very interesting to learn how to manage environments and allow someone else to replicate your results, and how important it is to convey the process step by step
- I learned a lot about H1-B visa process and how complicated it is (Also some fun facts like how applicants working in computer related fields dominate the number of visa petitions)

Recommendations

Story 5: Model Building (5/20)
Story 6: Uploading the Model on AWS to get inputs and output predictions (5/22)
Story 7: Build a Flask App for interacting with users (5/24)
Story 8: Host Flask App on AWS and experiment that it can take inputs from users (5/26)
Story 9: Deploy Flask App (5/27)
Story 10: Compile Demo (5/29)
Story 11: Finalize Demo and Presentation (6/1)
Story 12: Present (6/3)