A Chance to Work:
Understanding the composition of foreign workers pursuing specialty occupations on the United States H1-B visa

Team GUNDAM
Alexander Buddenbaum, Tianyu Li, Qinrui Li, Chuanqi Liu, Tianshu Tao

1. What are you trying to do?
   We propose to analyze the past 5 years of US immigrant visa applications (H1-B) to identify trends in applications by several attributes, such as country of origin, job title, sponsoring employer, salary band, and geographic location of the employer.

2. How is it done today; what are the limits of current practice?
   Several online databases such as the website *H1-B Grader* or *1 Point 3 Acres* have aggregated US Department of Labor statistics into textual tables which allow for filtering based on the above-mentioned attributes; however, sufficiently thorough and meaningful interactive visualizations of this information are scarce.

3. What's new in your approach; why will it be successful?
   Our visualization will facilitate immersive exploration of data by the user through various visualization techniques studied in this course. We believe by interacting with the data, the user will gain more meaningful insights to apply in their own career planning.

4. Who cares?
   This study is intended to assist US H1-B visa applicants, especially international students and prospective employees of organizations with a physical presence in the United States, to make better-informed decisions in their career planning.

5. If you're successful, what difference and impact will it make? How do you measure them (e.g., via user studies, experiments, ground truth data, etc.)?
   a. Lead to fewer applications by unqualified applicants, therefore reducing DOL/USCIS processing backlog
   b. Increase the overall quality of applications by enabling applicants to better prepare for work in their target companies
   c. Increase awareness among potential STEM graduate applicants from Sino-US joint university programs

   We will conduct a short survey before and after showing GTSI students the application to evaluate their prior knowledge level and the efficacy of our application.

6. **Tech Stack**
   a. **Data Acquisition**
      We will leverage nearly 2 Million records of public data from the US Department of Labor in .xls and .csv formats.
   b. **Data Cleansing**
      This step includes the formalization and normalization of all fields of data like handling the variations of the same thing, duplicates, encoding issues, dashes, parentheses, delimiters, white spaces etc, which can be done by OpenRefine or Python data libs.
   c. **Database**

An open source database system such as MySQL or SQLite will be used for the persistence of structural data.

d. Visualization

Our technical stack and tool set including but not limited to d3.js and vue.js/react.js open source visualization/front-end libraries and frameworks which support great dynamic and interactive visual effects and provide user-friendliness. Common analytical graphs include linear chart, bar chart and choropleth chart, etc. studied in this course.

e. Frameworks

It could be a typical client-server web application with a front end - back end separation framework. The back-end can be a python flask service which takes charge of the business logic like data accessing, caching and persistent strategy etc and expose the data api to any known clients.

f. Cloud

AWS provides a centralized platform to contain and run the service, including the engineered API of our project and the visual web pages.

7. What are the risks and payoffs?
    a. Risks: Other than the significant time commitment to completing this project, we do not see any real risks to conducting this analysis.
    b. Payoffs: We hope to raise awareness among potential applicants regarding past H1-B application competition to encourage a more well-rounded representation of global talent in the US workforce.
8. How much will it cost?
   H1-B visa application data will be sourced from the US Department of Labor public database. All software and frameworks are free, open source, or free to use under student license.
9. How long will it take?
   We expect approximately eight weeks for our five-member team to complete this project, totaling roughly 300 man hours.
10. What are the midterm and final "exams" to check for success?
    a. By April 1, 2022: Have an initial, working version of choropleth showing applications by employer, their geographic location, and job title.
    b. By April 21, 2022: Release a refined version of choropleth allowing the user to filter based on user-input criteria, such as employer, location, job title, salary band, and country of origin. Include bar graph distribution of applications submitted and accepted based on the above attributes.

Literature Review

Alex

Prediction of H1B Visa Using Machine Learning Algorithms, Swain et al.

Swain et al compared the accuracy of random forest, k-means clustering, and logistic regression algorithms for predicting H1-B application acceptance based on roughly 3 million data points from 2011 to 2016. They also show a variety of static charts, but the charts only compare across one dimension at a time. The data is also nearly ten years old. While this study set us in a good direction for doing some basic analysis, we propose an interactive visualization allowing the user to view results based on multiple attributes such as salary and location.

An allotment of H1B work VISA in USA using machine learning, Thakur et al.

Thakur et al apply seven classification models - Decision tree, C5.0, Random Forest, Naïve Bayes, Neural Network and SVM - to predict the status of each H1-B application. Interestingly, their C5.0 decision tree model proved the most accurate on the same data set used by Swain et al in their research. Their liberal use of bar charts makes some of their

findings hard to follow. We will run a more recent data set to better reflect current trends, and instead of comparing several classifiers for their accuracy, we will focus more on providing meaningful insights to users through the dynamic integration of the charts.

Foreign STEM workers and native wages and employment in US cities, Peri et al.
This paper explains the positive impact on economic growth by immigrant STEM workers in the US. Peri et al show many tables of data much more granular than that which we have for our project, and these tables are complex and perhaps hard to follow. As our target users are international students and young professionals from all academic disciplines, the aim of this reading is mainly to help the team appreciate the positive social and economic impact of the people represented by our data points.

Tianyu
A Deep Learning Based Approach for Predicting the Outcome of H-1B Visa Application, Dombe et al.
In this paper, the authors showed how they built a neural network model to predict the outcome of an applicant's H-1B visa application with an accuracy of 98%, which is really impressive. I think what we can learn here is how the author chose the attributes and designed architecture of the network. In our own work, we could try to add some new attributes, and make our network deeper to see whether we can get a better result.

ANALYSIS OF IMMIGRATION TRENDS IN THE US TO DISCOVER PATTERNS AND MAKE BETTER POLICY DECISIONS, Tandon, A.
In this paper, the author showed a trend of US different-type visa applications in various aspects, including the applicants' education level, employer, job-title, residence place, etc. This paper could give us a hint on what aspects we could choose to analyze our data except that we might only focus on the H1b type visa. Besides, the authors in this paper also gave a comparison of the pros & cons between Tableau and Python when using them to visualize the data, which could also be helpful when we move onto that step.

An analysis of nonimmigrant work visas in the USA using Machine Learning, Sundararaman et al.
In this paper, the authors built a Random Forest Model to predict the H1b visa petition results. This paper provides a good insight on how to preprocess the data before passing to a ML model. The authors showed how they determined the cut-off value, found the priority of companies and labeled the data, in great detail, which we might take into account in our own project. However, we should notice that in this paper, the authors didn't talk too much about the final performance of their prediction model. We should avoid having such a situation in our own project.

Tianshu
An Empirical Analysis of ML Algorithms
This paper dealt with applying Logistic Regression, Decision Tree and Random Forest on h1b visa dataset to show whether a person is eligible to file for an h1b visa based on certain parameters such as the Job title, Prevailing wage, full time position, worksite and year of application. It also compared the accuracy of all models where the decision tree did the best. Further enhancements could be addition of parameters to have better prediction.

An allotment of H1B work visa in USA using machine learning
This paper examines petitions filed from 2015 to 2017 with the goal that a superior prediction model needs to be developed using Decision tree, C5.0, Random Forest, Naïve Bayes, Neural

Network and SVM to foresee the aftereffect of the request which shows whether an appeal is commendable or not. It is found that C5.0 outperform with the best accuracy of 94.62 as a single model but proposed model gives better results of 95.4 accuracies which is built by machine ensemble method and this is validated by 10 fold cross-validation.

Studying to Stay: Understanding Graduate Visa Policy Content and Context in the United States and Australia

Using a summative content analysis approach, this article provides an analysis of the content of the Optional Practical Training and Temporary Graduate Visa programmes situated within the economic, political and social contexts of the United States and Australia respectively. Graduation from a domestic institution is the core requirement of these programmes, directly implicating higher education in immigration issues and placing higher education institutions at the border of the state.

Qinrui

Success of H1-B VISA Using ANN, Priyadarshini Chatterjee et al.

In this paper, Priyadarshini proposed a model based on ANN that works on the data set downloaded from Kaggle.com. They use some special encoding schemes to convert the data set into numerical form, which is good for us to learn. However, their model only achieves an accuracy of 94% in predicting the success of H1-B VISA, which might be kinda poor compared to some well-trained neural networks, I guess it could be their depth of network or function choosing that prevents it from achieving a better performance.

Visualizing Intergenerational Immigrant Assimilation at Work, Are Skeie Hermansen

In this paper, Hermansen used some visualization skills like heat plots to visualize differences in ethnic and socioeconomic characteristics of workplace contexts by immigrant background. More interestingly, the author reveals a striking pattern of intergenerational assimilation in the labor market through the visualization, which gives me some hint of making our project more inspiring for those people who would like to know what would happen if they choose to hold an H1-B visa and stayed in this country for some years, how their culture and mindset is being slightly affected by the truth that some pattern are fixed due to the structure of this society. However, this paper mainly focused on the level of social and cultural phenomena, there are not many skills that can be used in this class.

Data Analysis of H1-B Visa Applications, Roy, Raunak

This paper provides us with some data processing techniques and visualizations of the processed data, and they are using the dataset from Office of Foreign Labor Certification (OFLC) and the Labor Condition Application (LCA), well, the good point is this work focuses on the data science challenge problem of predicting the decision for past immigration visa applications, and they used supervised machine learning for classification, but the structure and the layout of this paper is a little bit rough. All we need is to take what's good in this paper like how the author implements the data cleaning schemes and how to analyze them.

Chuanqi

The H-1B Visa Immigration Program: Analysis and Comments, Edgar W. Butler
In this paper, Edgar analyzed historical data of H-1B Visa applications and approvals from different perspectives to improve this program and give useful suggestions. Since the incomes of US citizens are not affected by the H-1B program and it brings technology, patents, and talents for the US in different areas, the US indeed takes advantage of it. However, the fraudulent and technical violation during the application always occurs and has a high proportion, so it must be an important work for the federal government to distinguish. In this way, we can analyze the denied cases to get information from them.

Big Data analysis on H-1B Visa Application in the United State, JYOTI CHAVDA
This paper shows us the implementation of the MapReduce programming framework with a different design pattern, apache hive, and apache pig are applied on the H-1B application dataset. That helps us in creating and analyzing our own dataset.

A Hybrid Machine Learning Model Approach to H-1B Visa, Akalbir et al.
In this paper, Akalbir et al uses multiple machine learning models to predict the result of the H-1B application and shows good performance of these models. However, what we concern the most is the features he used in the model. These features might show some important information that contribute to the success of the H-1B application and we can use these important features to do our research.

References

Peri, G., Shih, K. Y., & Sparber, C. (2014). Foreign STEM workers and native wages and employment in US cities (No. w20093). National Bureau of Economic Research. https://www.nber.org/system/files/working_papers/w20093/w20093.pdf

Swain, D., Chakraborty, K., Dombe, A., Ashture, A., & Valakunde, N. (2018, December). Prediction of H1B Visa Using Machine Learning Algorithms. In 2018 International Conference on Advanced Computation and Telecommunication (ICACAT) (pp. 1-7). IEEE. https://ieeexplore.ieee.org/abstract/document/8933628

Thakur, P., Singh, M., Singh, H., & Rana, P. S. (2018). An allotment of H1B work VISA in USA using machine learning. International Journal of Engineering & Technology, 7(2.27), 93-103. https://www.researchgate.net/profile/Prashant-Rana-4/publication/328488339_An_allotment_of_H1B_work_visa_in_USA_using_machine_learning/links/5d70f092a6fdcc9961afad48/An-allotment-of-H1B-work-visa-in-USA-using-machine-learning.pdf

Butler, E. W. (2012). The H-1B visa immigration program: Analysis and comments. *International Journal of Business and Social Science*, *3*(9). https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1066.7924&rep=rep1&type=pdf

CHAVDA, J. Big Data analysis on H-1B Visa Application in the United State. https://www.researchgate.net/profile/Jyoti-Chavda/publication/341132236_Big_Data_analysis_on_H-1B_Visa_Application_in_the_United_State/links/5eb0437c45851592d6b884dc/Big-Data-analysis-on-H-1B-Visa-Application-in-the-United-State.pdf

Chadha, A. S., & Shitole, A. (2021, November). A Hybrid Machine Learning Model Approach to H-1B Visa. In *2021 3rd International Conference on Electrical, Control and Instrumentation Engineering (ICECIE)* (pp. 1-8). IEEE. https://www.researchgate.net/publication/357663888_A_Hybrid_Machine_Learning_Model_Approach_to_H-1B_Visa

Dombe, A., Rewale, R., & Swain, D. (2020). A deep learning-based approach for predicting the outcome of H-1B VISA application. In *Machine Learning and Information Processing* (pp. 193-202). Springer, Singapore. https://www.researchgate.net/profile/Drdebabrata-Swain/publication/340109729_A_Deep_Learning-Based_Approach_for_Predicting_the_Outcome_of_H-1B_Visa_Application/links/60f86b17169a1a0103ab192c/A-Deep-Learning-Based-Approach-for-Predicting-the-Outcome-of-H-1B-Visa-Application.pdf

Tandon, A. (2021). ANALYSIS OF IMMIGRATION TRENDS IN THE US TO DISCOVER PATTERNS AND MAKE BETTER POLICY DECISIONS. https://scholarworks.lib.csusb.edu/cgi/viewcontent.cgi?article=2387&context=etd

Sundararaman, D., Pal, N., & Misraa, A. K. (2017). An analysis of nonimmigrant work VISAs in the USA using machine learning. Int. J. Comput. Sci. Secur.(IJCSS), 6. https://dhanasekar-s.github.io/research/3paper.pdf

Jethwani, G., Sachdeva, A., & Goswami, M. (2019, February). An Empirical Analysis of ML Algorithms. In Proceedings of International Conference on Sustainable Computing in Science, Technology and Management (SUSCOM), Amity University Rajasthan, Jaipur-India.

https://deliverypdf.ssrn.com/delivery.php?ID=069113004124118101070098017097014002039056057018054055095126079009111097068067098077102022016029019024016011094008013004076080123002027048085025031003071021117084023003082075000090021025098002095074065112015105117116027005115103070123069071072104087064&EXT=pdf&INDEX=TRUE

Thakur, P., Singh, M., Singh, H., & Rana, P. S. (2018). An allotment of H1B work VISA in USA using machine learning. *International Journal of Engineering & Technology*, *7*(2.27), 93-103. https://www.researchgate.net/profile/Prashant-Rana-4/publication/328488339_An_allotment_of_H1B_work_visa_in_USA_using_machine_learning/links/5d70f092a6fdcc9961afad48/An-allotment-of-H1B-work-visa-in-USA-using-machine-learning.pdf

Grimm, A. (2019). Studying to stay: Understanding graduate visa policy content and context in the United States and Australia. *International Migration*, *57*(5), 235-251. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9664747