INST 737 - Intro To Data Science
Milestone 1

Bhavesh Bellara
Gaurav Hasija
Tanishka Jain
Danish Mir

a) Research Questions

H1B visa is a great way for US-based firms to recruit foreign talent. The US Department of Labour holds the right to certify or deny any application. Applying for the H1B visa is a time consuming and an expensive process in itself. For the employer it requires him to sponsor the visa to its employee. On employees part, the fact that the company reserves the right to withdraw the sponsorship at any time adds to his vulnerability.

We have exhaustive data regarding H1B applications that were certified or denied by the US Department of Labour. Using this data, we are going to perform predictive analysis using historical data that would be able to give some insight into whether or not the application will be certified. Apart from this, we are also thinking of answering other related socio-economic questions.

The research questions that our analysis will answer are as follows:

1. Is the applicant's success with H1B dependent on the selection of agents?
2. Do chances of getting an H1B visa differ based on the type of job?
3. Does wage difference beyond the cap limit affect the chances of getting an H1B?
    a. Eg: If a certain database administrator is earning $85k and other DBA is earning $90k annually then are the chances of getting an H1B for the DBA who earns $90k higher than one who earns $85k?
4. Based on the job domain, which worksite/geographic region should an individual target?
5. Are H1B employees paid less as compared to native employees for the same job?

b) State of Art

i) “Are H1B Visa Workers Paid Less than Similarly Employed Natives?” - Sperry, Will

The research is based on US Department of Labor’s H1B Disclosure Dataset for fiscal year 2015 and the Occupation Employment Survey from 2015 which compares salaries of employees working in banking, computer science, sciences, architecture and engineering sectors. The author discovered that in computer science, sciences, and architecture sectors, H1B employees earn less ($6000, $16000) than native employees whereas in the banking sector, they earn more ($2000). The difference could also be attributed to difference in skill level, however, there is not enough evidence (limited data) for the same.

We will extend the efforts and compare the data from our dataset to data available publicly on the web (glassdoor,ziprecruiter). We will compare the wage data in our dataset with the data present online and answer the research question.

ii) An allotment of H1B work visa in USA using machine learning - Pooja Thakur, Mandeep Singh, Harpreet Singh, Prashant Singh Rana

The analysis is based on OFLC's (Office of Foreign Labour Certification) yearly data. Following is the methodology that they followed:

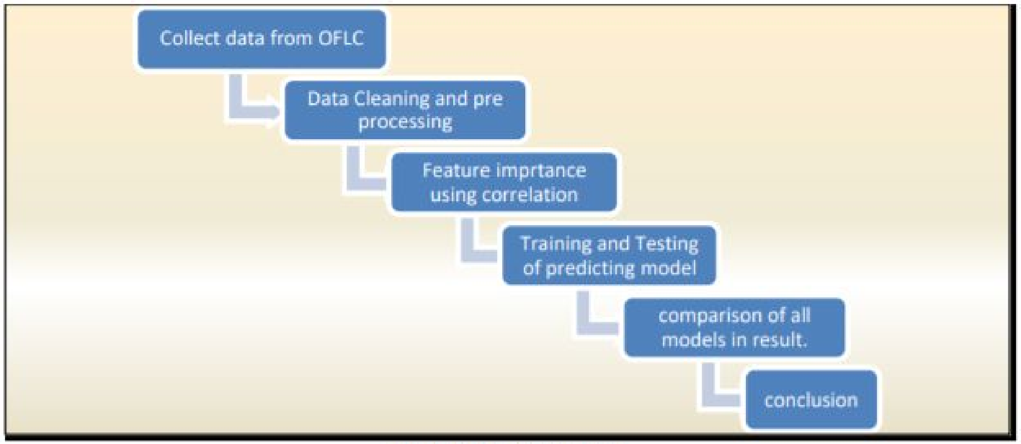

Collect data from OFLC
Data Cleaning and pre processing
Feature imprtance using correlation
Training and Testing of predicting model
comparison of all models in result.
conclusion

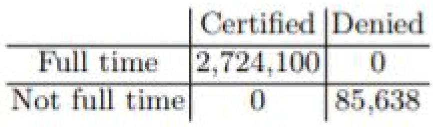
Fig. 8: Methodology Used.

We will be following a similar methodology. We have cleaned the data and will be predicting variables mostly using regression (We might change the procedure as our work proceeds).

iii) An Exploration of H-1B Visa Applications in the United States - Habeeb Hooshmand, Joseph Martinsen, Jonathan Arauco, Alishah Dholasaniya, Bhavik Bhatt

Using the data from the Department of Labour, the authors have done an exploratory data analysis and came up with interesting insights. One of the interesting insights being: a full-time job position is essential to get your visa application certified.

Table 1. Certification and Full time jobs



	Certified	Denied
Full time	2,724,100	0
Not full time	0	85,638

The paper uses 3 features to build the classifier viz. Full time position, Wage and Work site. Our team will also build a classifier model and we will choose our parameters based on our research questions.

iv) Predicting the Outcome of H-1B Visa Applications - Beliz Gunel, Onur Cezmi Mutlu

The analysis is based on the “H-1B Visa Petitions 2011-2016 dataset” on Kaggle. The work gave us good insight into data cleaning and general interpretation of data. For example, the authors pointed out that columns Job title and Soc_name essentially give out the same information and so analysis can just be based on one column. They have used Naive Bayes, Logistic Regression and soft-margin linear SVM for predictive analysis.

Our work will add to this research in terms of adding temporal components till 2019, as well as adding our specific research question related parameters to analysis.

c) Dataset

To get the necessary data for this project we downloaded the data from ”Office of Foreign Labor”. Here is the link to the datasets https://www.foreignlaborcert.doleta.gov/performancedata.cfm#dis

The datasets we plan on using for this course contain information about a candidate's H1B filing. Information such as the filing status, filling date, employer location, details of agents that assisted in filling and employment details such as job title, annual wage, etc. are the variables that are critical for our analysis.

While the data was available from 2008 onwards for the scope of this course we decided to use the data from 2015 to 2019. The data irregularities include descriptions like mismatching columns between data set of 2015 and 2016, and so on. As we cleaned our data, we also finalized the variables, the process was iterative taking in cognisense our research questions. Our final data set has data within these variables from the years 2015-19, combined under a single .csv file.

Based on the data in hand, we divided the database in 6 major categories as follows:

1. Case: This section comprised of the case number, case status, case filling and decision date. Case Status would act as the dependent variable in our project where we would be predicting whether certain parameters influence the case decision.

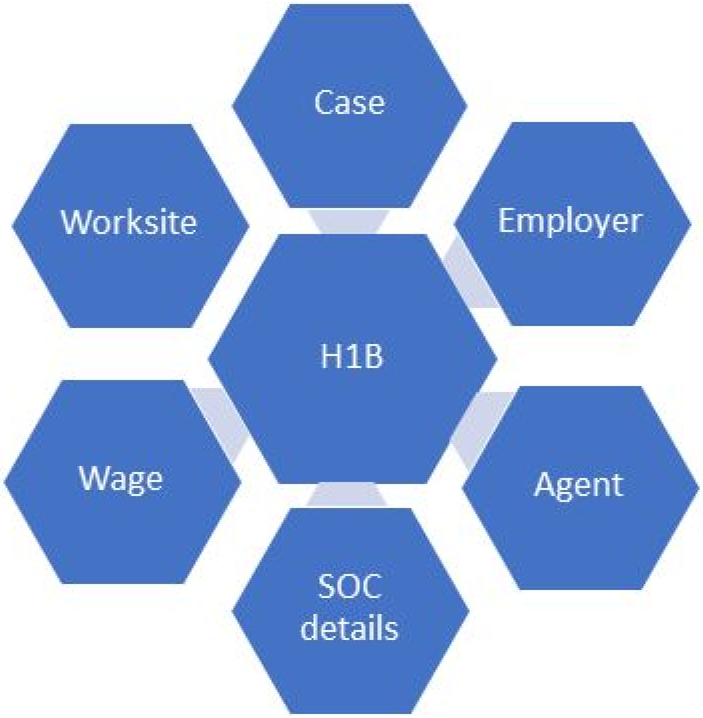
2. Employer: This section comprises employer related information such as employer name, location and phone number.

3. Agent: From our dataset we could identify that the majority of the H1B filings have been done with the help of the agent. Hence, we decided to make a separate category for agents to identify whether agents have an influence on the H1B case result. The category comprises agent details such as name, city and state they operate in.

4. SOC details: SOC is the classification of job types under various categories by the US government. This category contains the majority of data related to candidate’s employment such as job title, SOC title, SOC code, etc.

5. Wage: One of the main criteria when filing for H1B is the minimum wage that the employee needs to earn per year. Because of this we made a separate category for wages as it would be easier to study if there exists any effect on the case status. This category comprises variables such as employee's wage, the average pay in that job category for that area, level of wage etc.

6. Worksite: This category comprises the location related data of the employment location of the candidate whose H1B has been filled. We would like to study whether location of employment has an effect on getting the H1B visa successfully.



1. Attributes of H1B

d) Data cleaning efforts

While looking up the datasets for years 2015 through 2019, we were able to identify the following number of columns per year.

Year	Columns
2015	40
2016	40

2017	52
2018	52
2019	260

To combine all the data into a single file, we cleaned the datasets to standardize them and remove any conflicting variables. We used Python for the cleaning process. We used Pandas library specifically for this task.

After the cleaning process, dataset for each year had 34 columns which we categorized into 6 aforementioned categories i.e. Case, Employer, Agent, SOC details, Wage & Worksite.

e) Other Software Engineering Efforts

To get an extensive understanding of the scenario and gain proper insight into the data, we combined data from 2015 to 2019. As mentioned in the data cleaning section, there were different number of columns in every dataset ranging from 40 to 260. So we performed cleaning and aggregation to build a uniform data set.

Adding to our secondary research, we talked to a few people on H1B visa and double checked our understanding about intricacies like H1B minimum wage policies. We were able to confirm that the minimum wage for the approval of the H1B visa depends on the job type, worksite location among various other factors. That gave us a solid insight for developing research questions.