H1B Case Status Prediction

Project Proposal

Rushikesh Naidu[1], Mihin Sumaria[2], Jinal Jain[3], Janvi Kothari[4], Mihir Sawant[5]

1(Data Science, Worcester Polytechnic Institute, Worcester, ranaidu@wpi.edu)

2(Computer Science, Worcester Polytechnic Institute, Worcester, mssumaria@wpi.edu)

3(Data Science, Worcester Polytechnic Institute, Worcester, jjjain@wpi.edu)

4(Data Science, Worcester Polytechnic Institute, Worcester, jkkothari@wpi.edu)

5(Data Science, Worcester Polytechnic Institute, Worcester, msawant@wpi.edu)

Dataset Description: The H-1B is a visa in the United States that allows U.S. employers to employ foreign workers in specialty occupations. If a foreign worker in H-1B status quits or is dismissed from the sponsoring employer, the worker must either apply for and be granted a change of status, find another employer (subject to application for adjustment of status and/or change of visa), or leave the United States. In carrying out its responsibility for the processing of labor certification and labor attestation applications, the Office of Foreign Labor Certification (OFLC) generates program data that is essential both for internal assessment of program effectiveness and for providing the Department's external stakeholders with useful information about the immigration programs administered by OFLC. This data is made public to access the latest quarterly and annual disclosure data in easily accessible formats for the purpose of performing in-depth longitudinal research and analysis. OFLC case disclosure data is available for download by the federal fiscal year cycle covering the October 1 through September 30 period.
Gathered from https://www.foreignlaborcert.doleta.gov/performancedata.cfm, the project would be using disclosure data from the fiscal years 2013 to 2017.
The data has approximately 520,000 observations for every fiscal year from 2013 – 2017 with 35 variables defining the dataset.

Problem Description:

1. Which are the most significant variables in determining the case status of a new H1B application?
2. How has the impact of these variables fluctuated for the years 2013-2017?
3. If there is a significant change in any particular variable affecting the status of a H1B, then what was the reason of this major fluctuation?
4. **Is the H1B allotment truly random or not?**

We aim at understanding the major causes which affect the eligibility of an individual to get certified for an H1B visa.
USCIS publishes a memo when enough cap-subject applications have been received, indicating the closure of cap-subject application season. The associated random selection process is often referred to as the H-1B lottery. Those who have the U.S. master's exemption have two chances to be selected in the lottery: first, a lottery is held to award the 20,000 visas available to master's degree holders, and those not selected are then entered in the regular lottery for the other 65,000 visas. If the H1B visas depend on 35 variables, we plan to check if the allotment is a lottery indeed**?**

Classification or Regression?

The project is a classification problem since we have to predict whether a given case would be “Certified”, “Withdrawn”, “Certified-Withdrawn” or “Rejected”.

Classification Methods:

We would be trying and testing the following methods: Logistic Regression, Quadratic Discriminant Analysis, KNN, Decision Tree Models – Random Forest & ID3 and Support Vector Machines. Looking at the performance for each method, we would select the best 2 methods.

Dimension Reduction Methods:

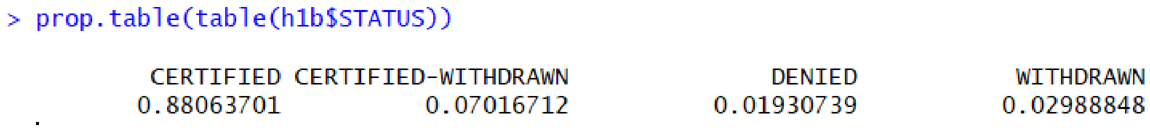
To reduce the number of predictors and to obtain the best model we will be using Ridge Regression, Lasso, and Subset Selection. PCA cannot be performed on the data since the data contains a lot of categorical variables.

Error Metrics:

Predictive Modeling works on constructive feedback principle. You build a model. Get feedback from metrics, make improvements and continue until you achieve a desirable accuracy. To develop the best model to answer our question we would be using the Confusion Matrix, Cross Validation Error and since it is a classification problem that can be solving using decision trees, we would also use the Gini Coefficient.

Comments/Concerns:

Relatively flexible models like logistic regression may suffer when applied to our data because there is a severe class imbalance. The response variable, status, for 88% of the cases is 'Certified'. Accuracy will not be an appropriate error metric, because if our model predicts the case status 'Certified' for each of the test cases, then we will get an accuracy of 0.88. Therefore, we will use error metrics like AUC-ROC, Sensitivity, Specificity, etc. To inflate minority cases for our model, we will use bagging, boosting, random oversampling, and clustering based oversampling to solve the class imbalance problem.



```
> prop.table(table(h1b$STATUS))

         CERTIFIED CERTIFIED-WITHDRAWN              DENIED           WITHDRAWN
        0.88063701          0.07016712          0.01930739          0.02988848
```

Fig. Case Status variable distribution for the year 2014

Description of each variable

FIELD NAME	DESCRIPTION
LCA_CASE_NUMBER	Unique identifier assigned to each application submitted for processing
STATUS	Status associated with the last significant event or decision. Valid values include “Certified”, “Certified-Withdrawn,” Denied,” and “Withdrawn”
LCA_CASE_SUBMIT	Date and time the application was submitted

DECISION_DATE	Date on which the last significant event or decision was recorded by the ETA National Processing Center
VISA_CLASS	Indicates the type of temporary application submitted for processing. R = H-1B; A = E-3 Australian; C = H-1B1 Chile; S = H-1B1 Singapore. Also referred to as “Program” in prior years.
LCA_CASE_EMPLOYMENT_START_DATE	Beginning date of employment
LCA_CASE_EMPLOYMENT_END_DATE	Ending date of employment
LCA_CASE_EMPLOYER_NAME	Employer's name
LCA_CASE_EMPLOYER_ADDRESS	Employer's address
LCA_CASE_EMPLOYER_CITY	Employer's city
LCA_CASE_EMPLOYER_STATE	Employer's state
LCA_CASE_EMPLOYER_POSTAL_CODE	Employer's postal code
LCA_CASE_SOC_CODE	The Standard Occupational Classification (SOC) code which classifies workers by occupational groups
LCA_CASE_SOC_NAME	Title of the SOC occupational group
LCA_CASE_JOB_TITLE	Job title
LCA_CASE_WAGE_RATE_FROM	Employer's proposed wage rate
LCA_CASE_WAGE_RATE_TO	Maximum proposed wage rate
LCA_CASE_WAGE_RATE_UNIT	Unit of pay for proposed wage rate
FULL_TIME_POS	Y = Full time; N = Part time position
TOTAL_WORKERS	Total number of foreign workers being requested for temporary labor certification
LCA_CASE_WORKLOC1_CITY	Address information of the intended are in which the foreign worker is expected to be employed (location of the job opening)
LCA_CASE_WORKLOC1_STATE	Prevailing wage rate
PW_1	
PW_UNIT_1	Unit of pay
PW_SOURCE_1	Collective bargaining; SESA; Other
OTHER_WAGE_SOURCE_1	Description of the Other wage source (online wage library, OES, employer provided survey, etc.)
YR_SOURCE_PUB_1	Collective bargaining; SESA; Other
LCA_CASE_WORKLOC2_CITY	Address information of the second location in which the foreign worker is expected to be employed (location of the job opening)
LCA_CASE_WORKLOC2_STATE	Prevailing wage rate - second location
PW_2	
PW_UNIT_2	Unit of pay - second location
PW_SOURCE_2	Collective bargaining; SESA; Other - second location

OTHER_WAGE_SOURCE_2	Description of the Other wage source (online wage library, OES, employer provided survey, etc.) – second location
YR_SOURCE_PUB_2	Year that the prevailing wage data was published – second location
LCA_CASE_NAICS_CODE	Industry code associated with the employer requesting permanent labor certification, as classified by the North American Industrial Classification System (NAICS)