BUSINESS ANALYTICS BATCH 1
TEAM 13

H-1B
Processing Likelihood Estimation

By Abhilash (as5637), Jessica (jj29270), Rohan(rs3874), Suriya(sa3628) & Uday(um2147)

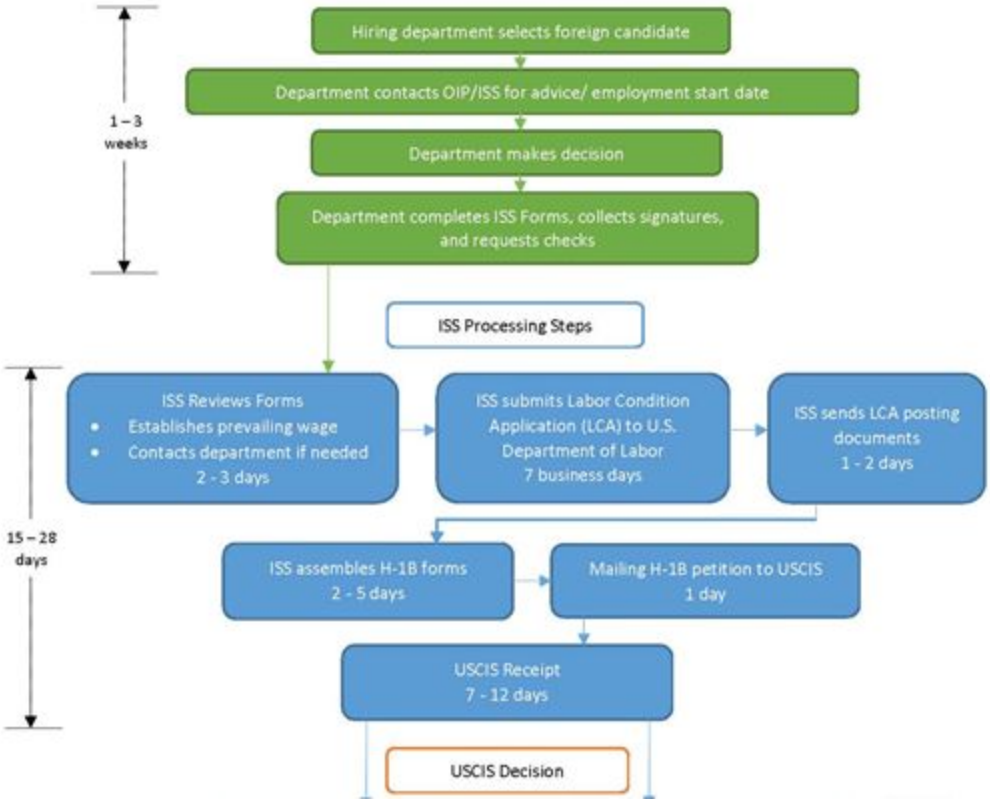
INTRODUCTION

The H1B process is hardly one that's unfamiliar to a majority of us. To those new to the term, the H1B can be considered as a golden ticket to Willy Wonka's Chocolate Factory. With the exception that:

- The ticket expires beyond a period of 6 years
- We're not future heirs to the factory. We're more like the Oompa Loompa in them.

Childish analogies aside, the H1B Visa process and the inflow of applications is a testament to the popularity of the United States for professionals from across the globe. Although the number of H-1B Visas have been capped at 65,000 (20,000 additional for foreigners with degrees

from the US), the number of petitions filed in 2018 alone amounted to 190,098. A general overview of the initial stages of the H-1B process is as follows:



With regards to the scope of the analysis, the focus will be on the Labor Condition Application (LCA) phase of the application process.

What is the LCA?

A Labor Condition Application (LCA) form has important information about the offered job position for the foreign worker as listed below.

- The job title of position offered (including SOC code, title)
- Duration of the job position offered (up to 3 years)
- If the position offered is either full time or not.
- Total number of jobs positions the LCA is applied for (can be any number)
- Rate of Pay / Salary offered for the position
- Location of the job position
- Prevailing Wage for the same position in that area.

- Employers & Attorney contact information.

DATA CLEANSING

“H-1B_Disclosure_Data” is available for the years 2011-2018 in the USCIS website, documenting all the applications submitted for the LCA procedure in the given period.

Combining these datasets, a cumulative database consisting of 3 million rows with 34 variables was created. 13 variables were selected from these that would aide in the analysis and model creation. Cleaning the data to remove rows with missing values (N/A) resulted in a total of 1.6 million rows.

The 13 variables specifically used for this analysis are,

'CASE_STATUS' - Status associated with the last significant event or decision. Valid values include “Certified”, “Certified-Withdrawn”, “Denied”, and, “Withdrawn”

'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.

'EMPLOYER_NAME' - Employer’s name

'JOB_TITLE' - Job Title

'SOC_CODE' - The Standard Occupational Classification (SOC) code which classifies workers by occupational groups

'SOC_NAME' - Title of the SOC occupational group

'FULL_TIME_POSITION'- Y = Full time; N = Part time position

'PREVAILING_WAGE' - The present wage rate on offer

'PW_UNIT_OF_PAY' - The unit of pay for the wage. Yearly or Hourly.

'H1B_DEPENDENT' - Y’ if dependents exist, ‘N’ otherwise.

'WILLFUL_VIOLATOR' - Whether employers who have committed either a willful failure or a misrepresentation of a material fact when hiring foreign workers

'WORKSITE_STATE' - Location (state) associated with the present position.

'YEAR' - Year at which the LCA was filed

Feature Engineering:

Based on the variables available in the data, feature sets were created to quantify the observations and perform the necessary analyses.

rate.x:

Success rate per employer. A percentage quantity obtained on computing the number of applications with CASE_STATUS=’Certified’ for every unique value corresponding to “EMPLOYER_NAME”

Application_number:

Number of applications filed per employer. A count of the number of rows corresponding to each unique value in the “EMPLOYER_NAME” column.

rate.y:

Success rate per SOC_CODE. A percentage quantity obtained on computing the number of applications with CASE_STATUS=’Certified’ corresponding to every unique “SOC_CODE” value.

Application_number_SOC:

Number of applications filed per SOC_CODE. A count of the number of rows corresponding to each unique value in the “SOC_CODE” column.

```
q <- group_by(t1, Var1) %>% mutate(rate = Freq/sum(Freq), Application_number = sum(Freq))

certified <- q[q$Var2 == "CERTIFIED",]
data_certified <- merge(data, certified[c('Var1','rate','Application_number')],
                        by.x="EMPLOYER_NAME", by.y="Var1", sort=FALSE)

p <- table(data$SOC_CODE, data$CASE_STATUS)
g <- as.data.frame(p)
g1 <- g[order(g$Var1),]
k <- group_by(g1, Var1) %>% mutate(rate = Freq/sum(Freq), Application_number_SOC = sum(Freq))

certified <- k[k$Var2 == "CERTIFIED",]
data_SOC_employer_certified <- merge(data_certified,
                                     certified[c('Var1','rate','Application_number_SOC')],
                                     by.x="SOC_CODE", by.y="Var1", sort=FALSE)
lca_certified <- data[(data$CASE_STATUS=="CERTIFIED"),]
lca_denied <- data[(data$CASE_STATUS=="DENIED"),]

certified_sample <- sample(1:nrow(lca_certified),2*nrow(lca_denied))
lca_certified_sample <- lca_certified[certified_sample,]

lca_combined <- rbind(lca_certified_sample,lca_denied)

#rate.x = acceptance per employer
#rate.y = acceptance per SOC

names(lca_combined) <- tolower(names(lca_combined))
lca_combined$factor_cs <-as.factor(ifelse(lca_combined$case_status %in% c("CERTIFIED"),1,0))
lca_combined$factor_ftp <-as.factor(ifelse(lca_combined$full_time_position == 'Y',1,0))
lca_combined$h1b_dependent <-as.factor(ifelse(lca_combined$h1b_dependent == 'Y',1,0))
lca_combined$willful_violator <- as.factor(ifelse(lca_combined$willful_violator == 'N',1,0))
lca_combined$worksite_state = as.factor(lca_combined$worksite_state)
```

This is the snippet of the code used in the creation of these feature sets.

MODELLING

Logistic Regression:

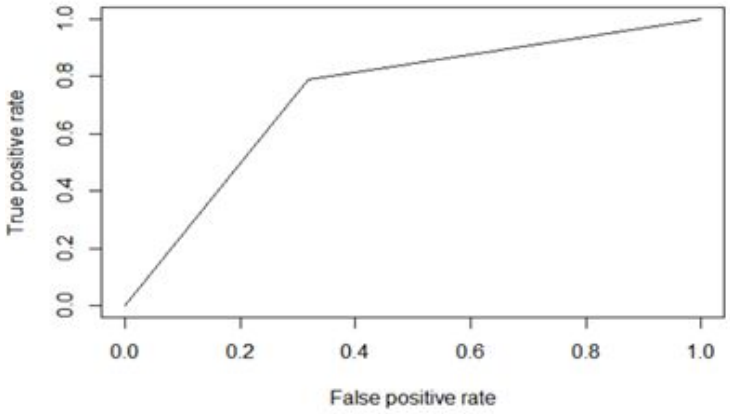
The original dataset consisted of 1.6 Million values on which we performed sampling to get 50,008 "Certified" Applications and 25,004 "Denied" Applications. All 1/0, Yes/No values were converted to factors (factor_cs, h1b_dependent1,factor_ftp1,willful_violator1) Splitting the data into Train – Test sets (75:25), a Logistic Regression was applied to the 4 new factors – Success Rate Per Employer, Number of Applications per Employer, Success Rate per SOC Code, Number of Applications filed per SOC Code. and other relevant columns of the original data set (Prevailing wage, H-1B Dependent, Willful Violator).

Summary as shown:

```
Call:
glm(formula = factor_cs ~ prevailing_wage + h1b_dependent + willful_violator +
    rate.x + application_number + rate.y + application_number_soc +
    factor_cs + factor_ftp, family = binomial, data = train_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6405  -0.6261   0.5206   0.6843   2.7570

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)          -5.163e+00  3.941e-01 -13.100  < 2e-16 ***
prevailing_wage      -2.412e-06  3.247e-07  -7.430 1.09e-13 ***
h1b_dependent1        4.527e-01  2.673e-02  16.939  < 2e-16 ***
willful_violator1     4.908e-01  3.305e-01   1.485    0.137
rate.x                6.238e+00  7.156e-02  87.171  < 2e-16 ***
application_number    1.976e-05  1.428e-06  13.839  < 2e-16 ***
rate.y               -1.156e-01  2.504e-01  -0.462    0.644
application_number_soc 6.007e-07  7.334e-08   8.191 2.60e-16 ***
factor_ftp1           2.815e-01  6.114e-02   4.604 4.15e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 71611  on 56258  degrees of freedom
Residual deviance: 52805  on 56250  degrees of freedom
AIC: 52823

Number of Fisher Scoring iterations: 12
```

Looking at the p-values of the summary, shows that 'prevailing_wage', 'h1b_dependent1', 'rate.x', 'application_number', 'application_number_soc' and 'factor_ftp1' were the significant variables. However, it would be crucial to note that 'rate.y' a feature assumed to be of key importance proved to have no effect on the chances of an individual being certified in this phase of the application (factor_cs = 1).

Classifier experimentation with various ranges



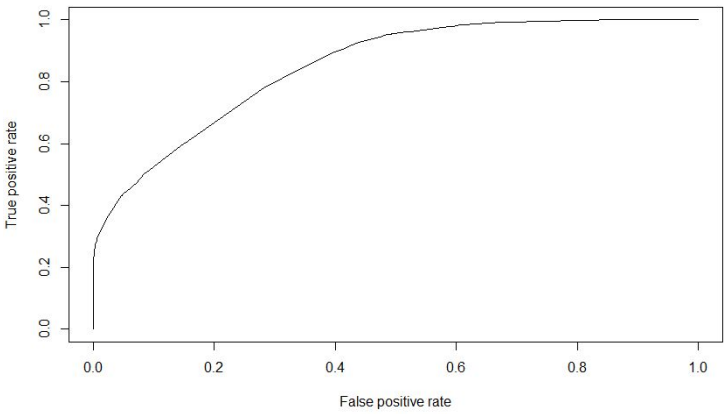
Predict/True	0	1
0	4272	1985
1	2643	9853

For p > 0.7, Probability is 1; For p < 0.7, Probability is 0 : AUC found to be 0.73

This methodology works well, delivering a high AUC value (Area Under the Curve; a measure of model accuracy) and a higher likelihood of correct predictions.

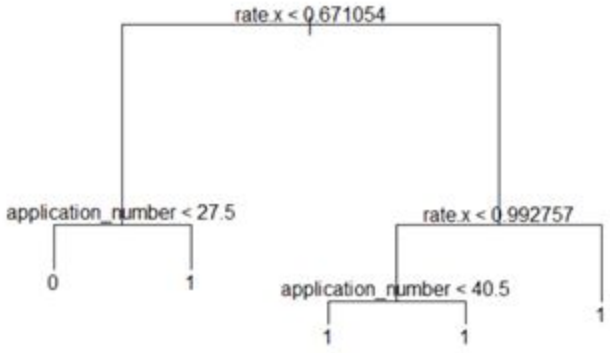
Decision Tree:

Using a Decision Tree on the 4 variables described in the logistic regression process to identify probabilities of 1 or 0 occurring. Assigning these values based on classifier, the value of the AUC curve generated is then compared to the AUC values generated through Logistic to identify the classifier that provides the highest AUC value.



Predict/True	0	1
0	4836	1499
1	3775	8643

```
Classification tree:
tree(formula = factor_cs ~ ., data = training1)
Variables actually used in tree construction:
[1] "rate.x"             "application_number"
Number of terminal nodes:  5
Residual mean deviance:  0.8997 = 50610 / 56250
Misclassification error rate: 0.211 = 11871 / 56259
```



Although there is a reduction in the number of predictions of a certified application, the number of predictions of denied applications has improved. Overall, the AUC score remains unchanged (0.73).

Alternate Methodologies

We could improve this model with the use of a Random Forest, a methodology that stands to improve the quality of our predictions. Neural Network and SVM/Naive Bayesian Classifiers could greatly improve this model.

What use is this to you?

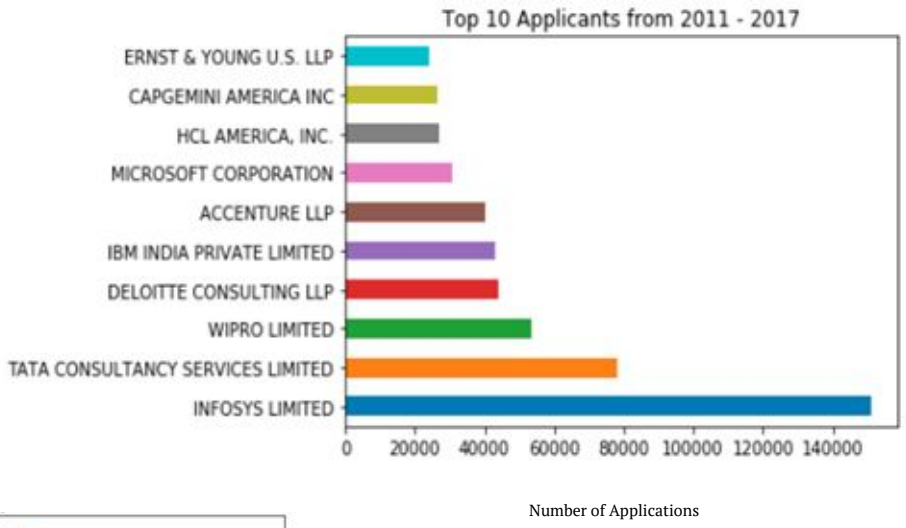
Although the H-1B process is a lottery, the LCA procedure that preludes it is not. This model assesses the chance that an individual's application would cross this crucial stage that plays a key role in it being accepted if picked in the lottery. This data would come in handy for corporations that spend millions of dollars in applications for each of their candidates, allowing them to make better decisions. For the future applicant, it stands to highlight the roles, wage and industries that have a higher likelihood of success in this stage. Here's hoping that after the long wait period, you can finally stand at that interview booth and hear the golden words:

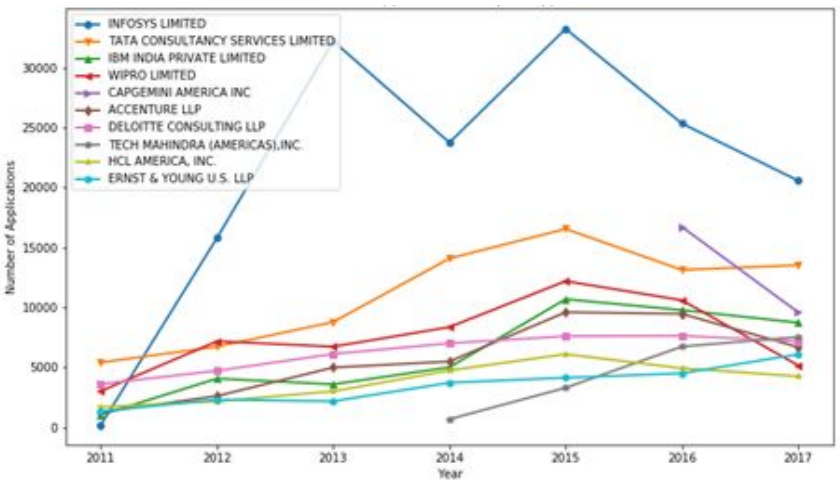
"Welcome my friends, welcome to my chocolate factory."

EDA

The following trends are observed from the dataset:

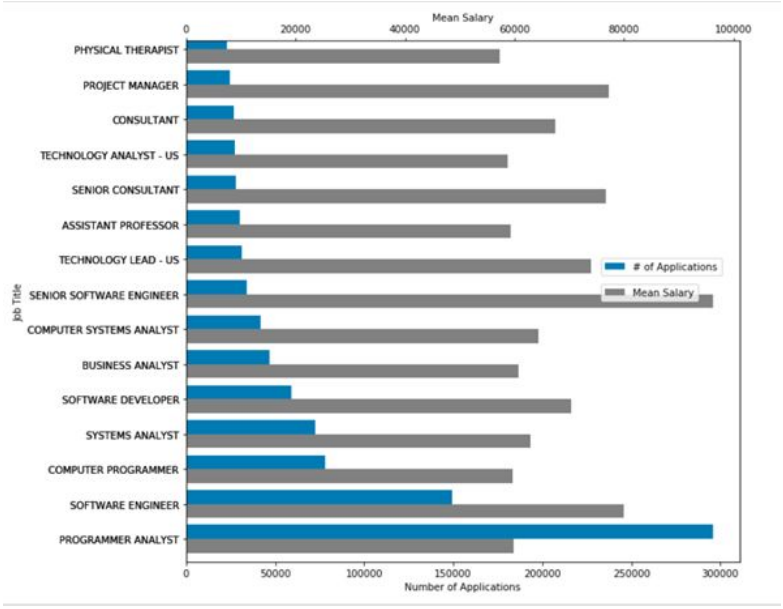
Infosys Ltd and Tata Consultancy Services Ltd are the top employers by the number of applicants.





Infosys consistently has the highest number of applications over the period of 2011 -2017

Top jobs from top 10 applications are mostly IT jobs; both in development and analysis.



APPENDIX

Database:
https://www.foreignlaborcert.doleta.gov/performancedata.cfm

Sources:
"TIMELINES FOR H-1B PETITION APPROVAL"
http://international.utsa.edu/utsa-policy/timelines-for-h1-b-petition-approval/

"What is H1B LCA ? Why file it ? Salary, Processing times – DOL"
https://redbus2us.com/what-is-h1b-lca-why-file-it-salary-processing-times-dol/

"Charlie and the Chocolate Factory in West End debut"
https://www.bbc.com/news/av/entertainment-arts-23056473/charlie-and-the-chocolate-factory-in-west-end-debut

"Welcome my friends, welcome to my chocolate factory."
https://www.rottentomatoes.com/m/willy_wonka_and_the_chocolate_factory/quotes/