# Math 253 Final Project- H-1B Visa Analysis

*Melissa Leong & Muath Ibaid*

*4/18/2017*

## a) Problem Setting:

**The Client:** An H1B visa applicant

**The Situation:** The client walks in and requests that we to predict their chances of being approved for and H1B visa. The client is able to provide information on their: intended work sector, prevailing wage, whether their position is full time or part time amd the year they applied.

**The Data :**

Office of Foreign Labour Certification- Processing Status H-1B Applications filed between 2011 and 2016

**Variables:** - Case Status: Certified, Certified-Withdrawn, Withdrawn, Denied
- Employer Name
- Occupational Sector Code
- Prevailing Wage
- Job Title
- Full Time Position Dummy
- Latitude and Longitude of Worksite
- Year

---

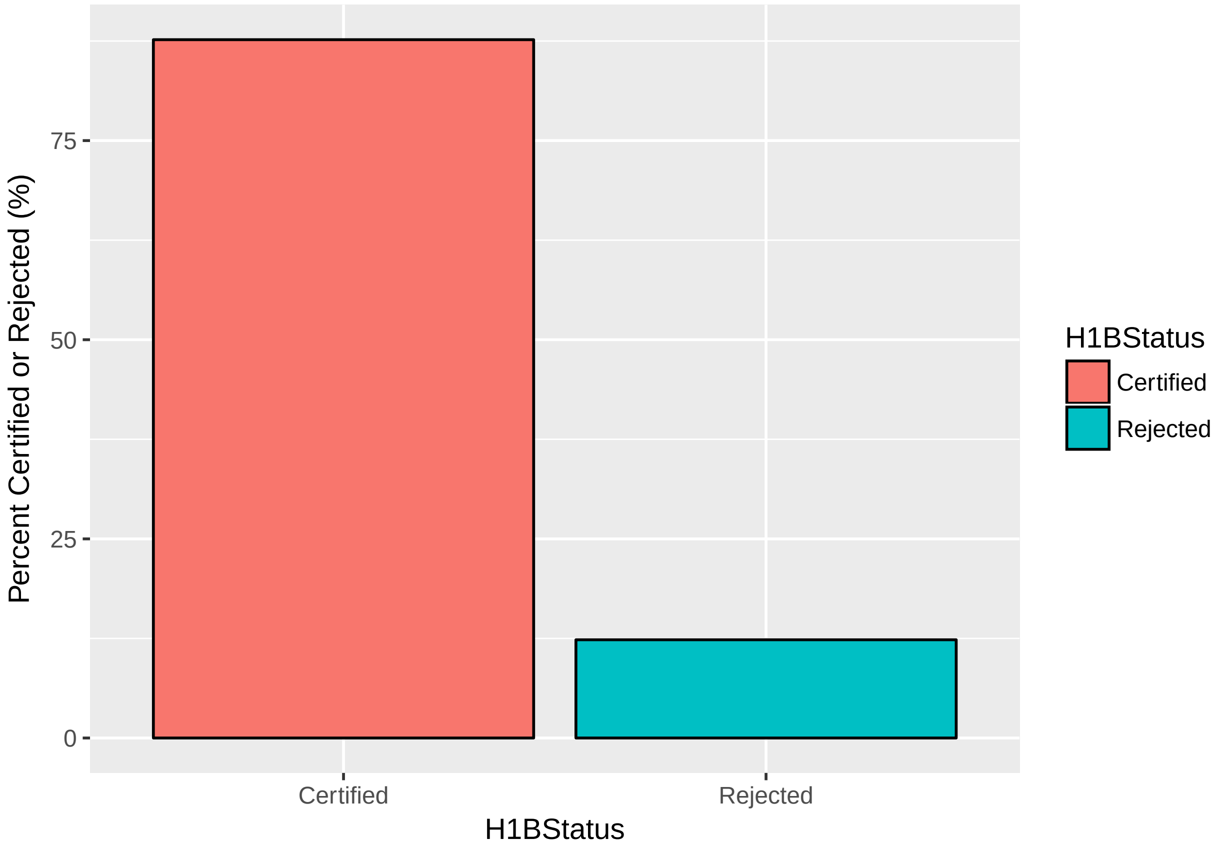
## b) Data Description:

### Declaring Test and training data

### Data Visualizations

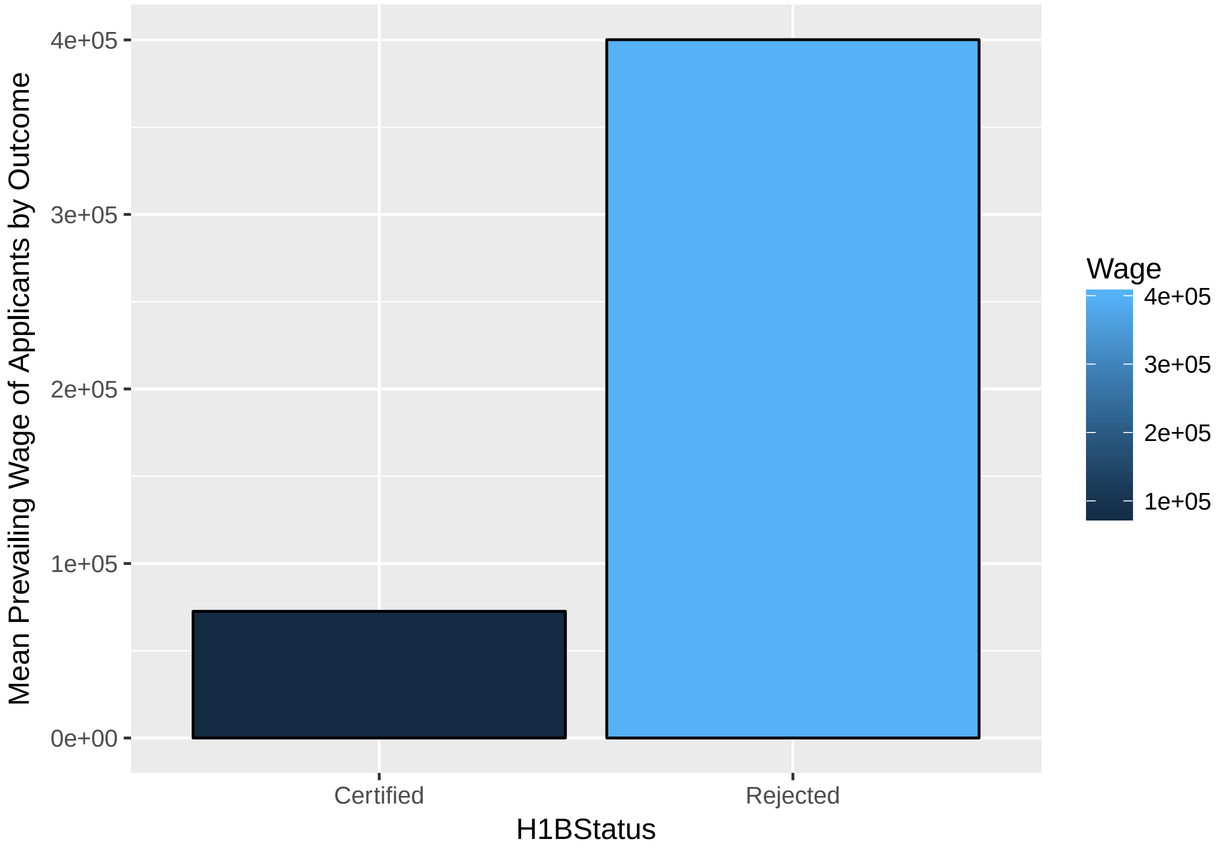
Percentage of "certified" vs. "rejected" visas

```
visual1a<- data.frame(percent=(sum(training$outcome==1)/sum(nrow(training))),H1BStatus="Certified")
visual1b<-data.frame(percent=(sum(training$outcome==0)/sum(nrow(training))),H1BStatus="Rejected")
visual1Total<-rbind(visual1a,visual1b)

ggplot(data=visual1Total,aes(x=H1BStatus,y=(percent*100),fill=H1BStatus))+
  geom_bar(stat="identity",colour="black")+
  ylab("Percent Certified or Rejected (%)")
```



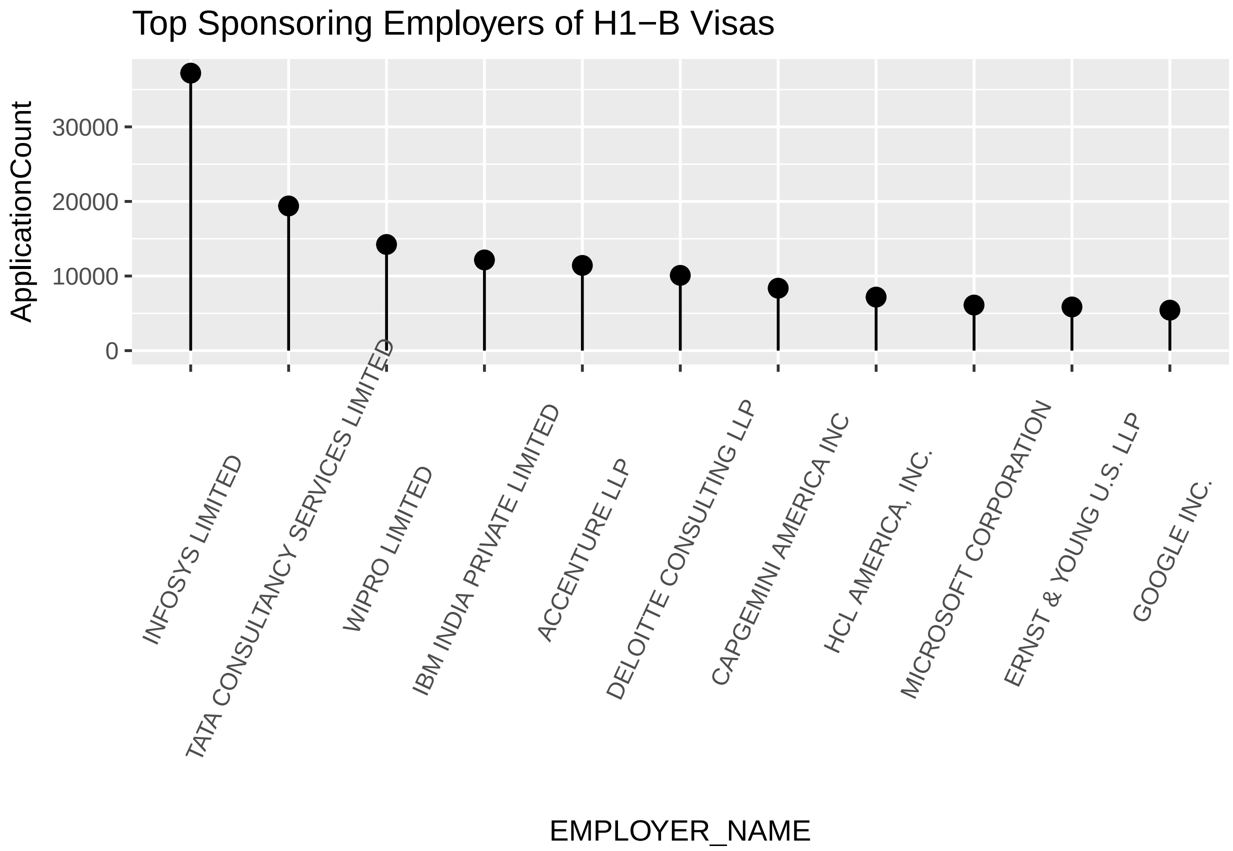
Mean wage of Approved Applicants Vs. Mean wage of Rejected Applicants

```
visual2a<- training[outcome==1,.(Wage=mean(PREVAILING_WAGE),H1BStatus="Certified")]
visual2b<- training[outcome==0 & !is.na(PREVAILING_WAGE),.(Wage=mean(PREVAILING_WAGE),H1BStatus="Rejecte
visual2Total<-rbind(visual2a,visual2b)

ggplot(data=visual2Total,aes(x=H1BStatus,y=Wage,fill=Wage))+
  geom_bar(stat="identity",colour="black")+
  ylab("Mean Prevailing Wage of Applicants by Outcome")
```



Most Popular Employers

```
ggplot(visual3, aes(x=EMPLOYER_NAME, y=ApplicationCount)) +
  geom_point(size=3) +
  geom_segment(aes(x=EMPLOYER_NAME,
                   xend=EMPLOYER_NAME,
                   y=0,
                   yend=ApplicationCount)) +
  labs(title="Top Sponsoring Employers of H1-B Visas")+
  theme(axis.text.x = element_text(angle=65, vjust=0.6))
```



## c) Classification and regression methods:

### c.1 Logistic Regression:

- Looking for a binary response
- Maximum likelihood approach
- p(x)= eˆ B0+B1x / 1+eˆ B0+B1x

### 1) Outcome~ Prevailing Wage

```
summary(modc1Wage)
```

```
##
## Call:
## glm(formula = outcome == 1 ~ PREVAILING_WAGE, family = "binomial",
##     data = training)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0489   0.5122   0.5123   0.5126   4.2708
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)        1.968e+00  3.390e-03 580.592   <2e-16 ***
## PREVAILING_WAGE -6.118e-08  6.184e-09  -9.894   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 614202  on 822222  degrees of freedom
## Residual deviance: 613165  on 822221  degrees of freedom
## AIC: 613169
##
## Number of Fisher Scoring iterations: 7
```

A 1 unit increase in **Prevailing Wage** is associated with an “almost zero” increase in the log odds of H1B approval success.

Prevailing wage is not a meaningful predictor of H1B application outcomes.

### 2) Outcome~ Full-Timeness

```
summary(modc1FullTime)
```

```
##
## Call:
## glm(formula = outcome == 1 ~ FULL_TIME_POSITION, family = "binomial",
##     data = training)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0503   0.5106   0.5106   0.5106   0.5213
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)        1.927186   0.006918 278.569  < 2e-16 ***
## FULL_TIME_POSITIONY 0.044421   0.007910   5.616 1.96e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 614202  on 822222  degrees of freedom
## Residual deviance: 614171  on 822221  degrees of freedom
## AIC: 614175
##
## Number of Fisher Scoring iterations: 4
```

A **Full-Time** position is associated with a 4% higher log odds of H-1B visa approval.

We observe a high Z stat that indicates that is is 5.6 standard deviations away from the mean and a p-value that indicates that full-timeness is a meaningful predictor of H1B application outcomes.

### 3) Outcome~ Big H-1B Employer

```
summary(modc1BigEmployer)
```

```
##
```

```
## Call:
## glm(formula = outcome == 1 ~ bigEmployer, family = "binomial",
##     data = training)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5254   0.2902   0.5436   0.5436   0.5436
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 1.837362   0.003467  529.99   <2e-16 ***
## bigEmployer 1.309413   0.014924   87.74   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 614202  on 822222  degrees of freedom
## Residual deviance: 603199  on 822221  degrees of freedom
## AIC: 603203
##
## Number of Fisher Scoring iterations: 5
```

Being sponsored by a **Big Employer** is associated with a 131% higher log odds of H-1B visa approval.

We observe a high Z stat and a p-value that indicates that full-timeness is a meaningful predictor of H1B application outcomes.

### 4) Outcome~ Wage+ Full-Timeness+ Big H-1B Employer

```r
#4) outcome~Wage*FullTime*Big Employer
summary(modc1Interaction)
```

```
##
## Call:
## glm(formula = outcome == 1 ~ PREVAILING_WAGE + FULL_TIME_POSITION +
##     bigEmployer, family = "binomial", data = training)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5291   0.2930   0.5413   0.5418   4.2371
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.823e+00  6.995e-03 260.659  < 2e-16 ***
## PREVAILING_WAGE     -5.974e-08  6.020e-09  -9.924  < 2e-16 ***
## FULL_TIME_POSITIONY  2.739e-02  7.955e-03   3.443 0.000576 ***
## bigEmployer          1.307e+00  1.494e-02  87.504  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 614202  on 822222  degrees of freedom
## Residual deviance: 602188  on 822219  degrees of freedom
## AIC: 602196
##
## Number of Fisher Scoring iterations: 7
```

Prevailing Wage decreases the log odds of H1B success by almost zero

Full-Time position increases the log odds of H1-B success by 2%

Being sponsored by a big employer increases the log odds of H1-B success by 131%

### c.2 Linear Disciminate Analysis (LDA) / Quadratic Discriminate Analysis (QDA):

```
lda1 = lda(outcome ~ FULL_TIME_POSITION, data = training)
lda1
```

```
## Call:
## lda(outcome ~ FULL_TIME_POSITION, data = training)
##
## Prior probabilities of groups:
##         0         1
## 0.1233266 0.8766734
##
## Group means:
##   FULL_TIME_POSITIONY
## 0           0.7639593
## 1           0.7718754
##
## Coefficients of linear discriminants:
##                          LD1
## FULL_TIME_POSITIONY 2.379555
```

```
lda1Pred = predict(lda1, testing)
ldaPredictionClass = lda1Pred$class

h1bOutcome = testing$outcome
table(ldaPredictionClass, h1bOutcome)
```

```
##                   h1bOutcome
## ldaPredictionClass      0      1
##                  0      0      0
##                  1 101927 720296
```

```
#Testing Error
mean(ldaPredictionClass != h1bOutcome)
```

```
## [1] 0.1239652
```

```
library(grid)
qda1 = qda(outcome ~ FULL_TIME_POSITION, data = training)
qda1
```

```
## Call:
## qda(outcome ~ FULL_TIME_POSITION, data = training)
##
## Prior probabilities of groups:
##         0         1
## 0.1233266 0.8766734
##
## Group means:
##   FULL_TIME_POSITIONY
## 0           0.7639593
## 1           0.7718754
```

```
qda1Pred = predict(qda1, testing)
qdaPredictionClass = qda1Pred$class

h1bOutcome = testing$outcome
table(qdaPredictionClass, h1bOutcome)
```

```
##                     h1bOutcome
## qdaPredictionClass      0      1
##                  0      0      0
##                  1 101927 720296
```

```
#Testing Error
mean(qdaPredictionClass != h1bOutcome)
```

```
## [1] 0.1239652
```

```
#Change posterior threshold!
```

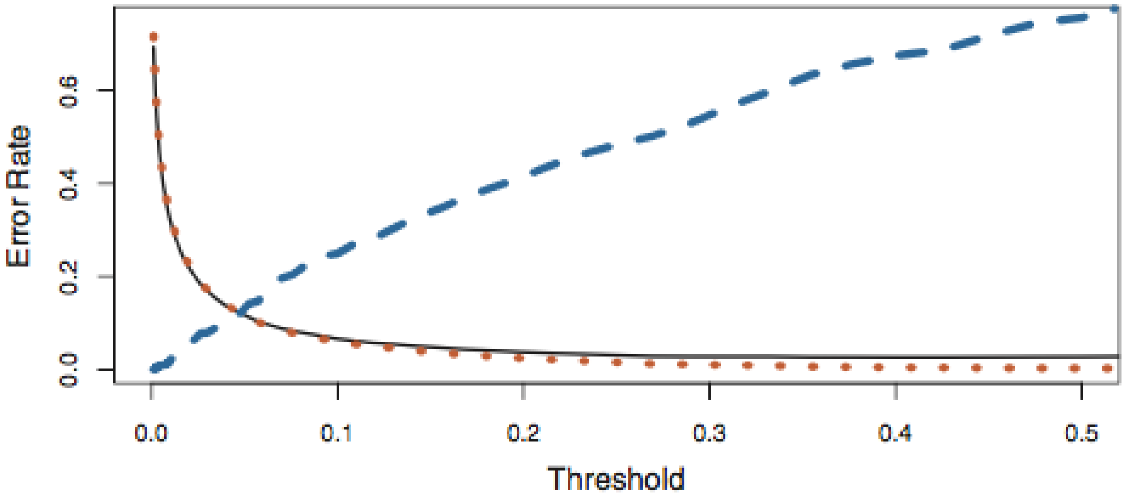


Figure 1: Alt text

# p > 1

```
lda2 = lda(outcome ~ PREVAILING_WAGE + FULL_TIME_POSITION + soc_data, data = training)

lda2
```

```
## Call:
## lda(outcome ~ PREVAILING_WAGE + FULL_TIME_POSITION + soc_data,
##     data = training)
##
## Prior probabilities of groups:
##         0         1
## 0.1233266 0.8766734
##
## Group means:
##   PREVAILING_WAGE FULL_TIME_POSITIONY soc_dataTRUE
## 0       400130.23           0.7639593   0.01301749
## 1        72585.45           0.7718754   0.01391746
##
## Coefficients of linear discriminants:
##                               LD1
## PREVAILING_WAGE     -4.088004e-07
## FULL_TIME_POSITIONY  3.563525e-01
## soc_dataTRUE         4.778003e-01
```

```
lda2Pred = predict(lda2, testing)
lda2Class = lda2Pred$class

table(lda2Class, h1bOutcome)
```

```
##          h1bOutcome
## lda2Class      0      1
##         0    255      1
##         1 101672 720295
```

```
#ldaClass = na.omit(ldaClass)
#mean(ldaClass != testing$outcome)

#Testing Error
table(lda2Class == h1bOutcome) / nrow(testing)
```

```
##
##     FALSE      TRUE
## 0.1236562 0.8763438
```

**Sensitivity & Specificity:**

**Sensitivity: percentage of true approvals that are identified (so 1 & 1) = (720295)/720296 = 99.9%**

**Specificity: percentage of non-approvals that are correcrtly identified = 255/(255+101672) = 0.25%**

**There is a VERY high chance that our model will tell you that you'll be approved but in reality, you'll be rejected! :(**

**this is more dangerous than telling you that you're rejected but you get approved**