H1-B Visa Processing Likelihood

Team 13
Abhilash Somasamudramath (as5637)
Jessica Jacob (jj2927)
Rohan Singh (rs3874)
Suriya Arumugavelan (sa3628)
Uday Marepalli (um2147)

Overview of the H1-B Visa

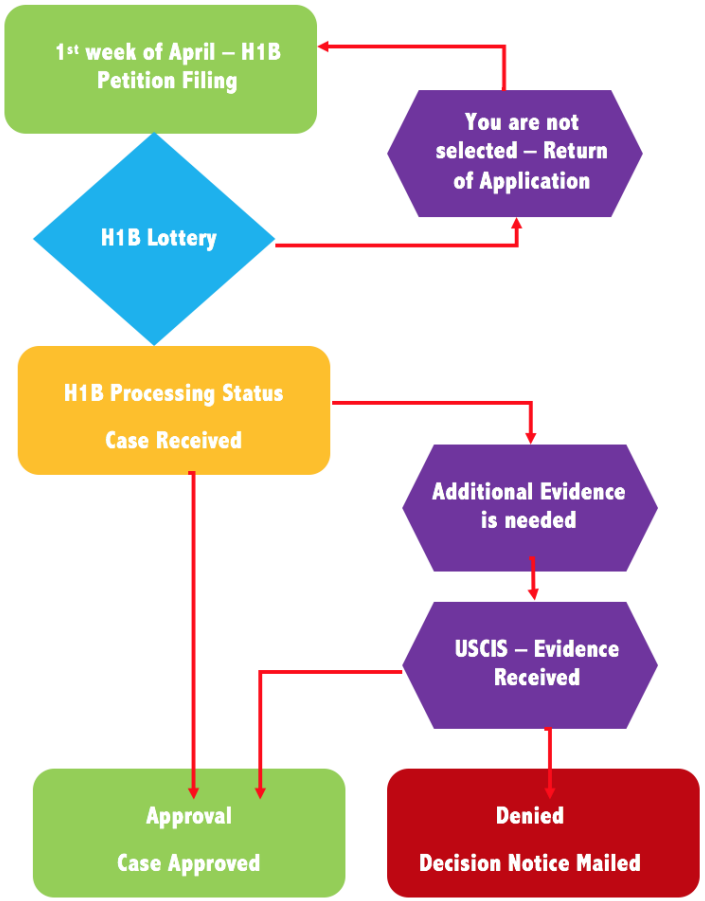
What is the H1-B Visa?

- Temporary **visa** that allows **immigrants to work** in the US in a specialized area of work
- Issued for a period of **3 years**, can be extended for 3 more
- USCIS issues **85,000/year** – **20,000 set aside** for applicants with **Master's degree and above**.
- Heavily **oversubscribed** – 190,098 applications in 2018

Pre-requisites to an H1-B Application

- Before filing H1B application, organizations have to file an **LCA** (Labor Condition Application)
- LCA is required to prove an immigrant worker is required for job – candidates can be rejected at this phase
- No lottery, LCA decisions are completely human centric and determined by immigration officer
- Since the LCA process is manually determined and is not subject to a lottery, it is completely objective

Process Workflow



"The H1B Process Explained, Step by Step." Stilt, Nigel Stevens. Accessed December 10, 2018. https://www.stilt.com/blog/2018/08/h1b-process-explained-step-by-step/

Overview of the H1-B Visa

The LCA Process in Focus

The Trump Effect

- RFE (Request for Further Evidence) – can happen in both the LCA stage as well as H1-B approval process
- Immigration officer asks for clarifications on applications through RFE
- Effective September 11, 2018 (irony?) the Trump administration gave USCIS the authority to deny applications without issuing an RFE.
- What does this imply? – If immigration officer is not personally satisfied with the application, they can deny the application without giving the candidate a chance to defend themselves.

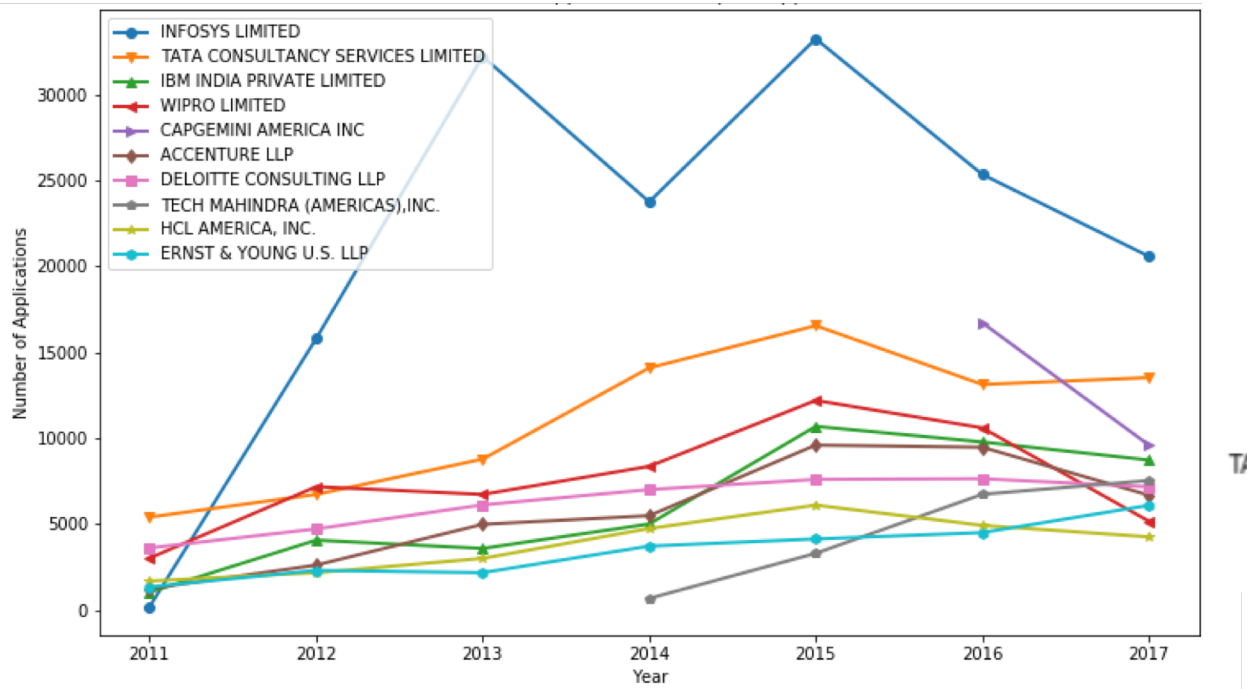
Casual Meme

Project Objective

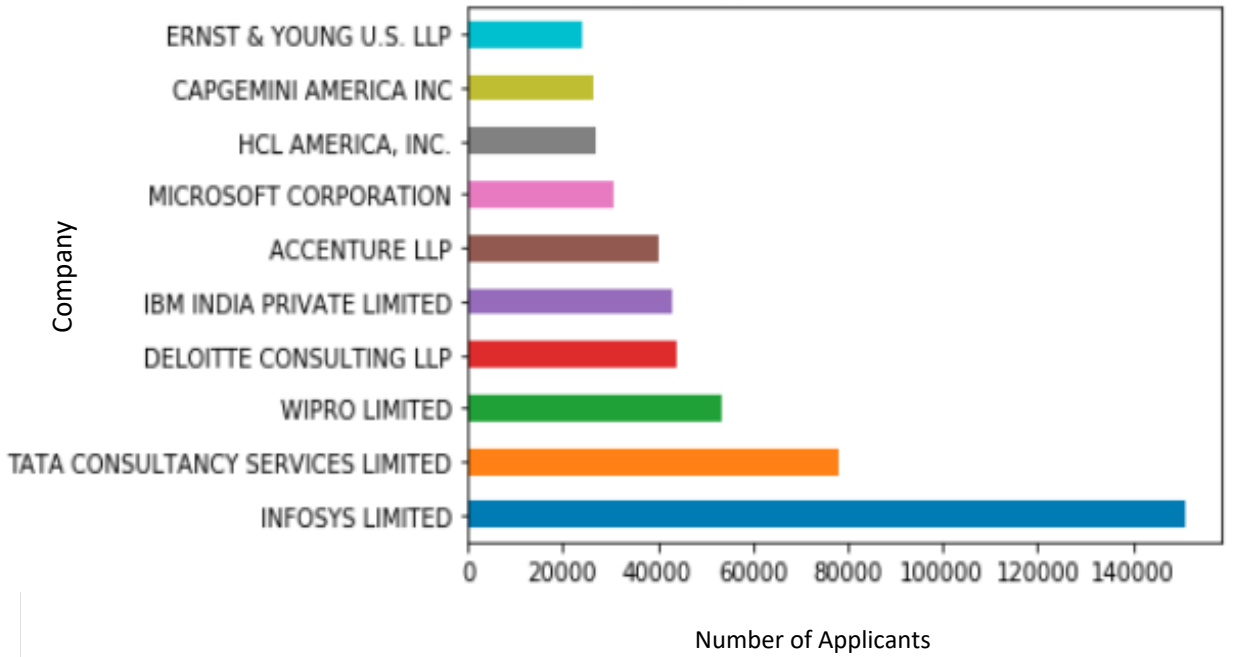
This project works to build a model that can predict the likelihood of a LCA application being accepted based on all the parameters that the USCIS accepts as part of the application.

A Graph Speaks a Million Tuples

Number of Applications from Top 10 Applicants

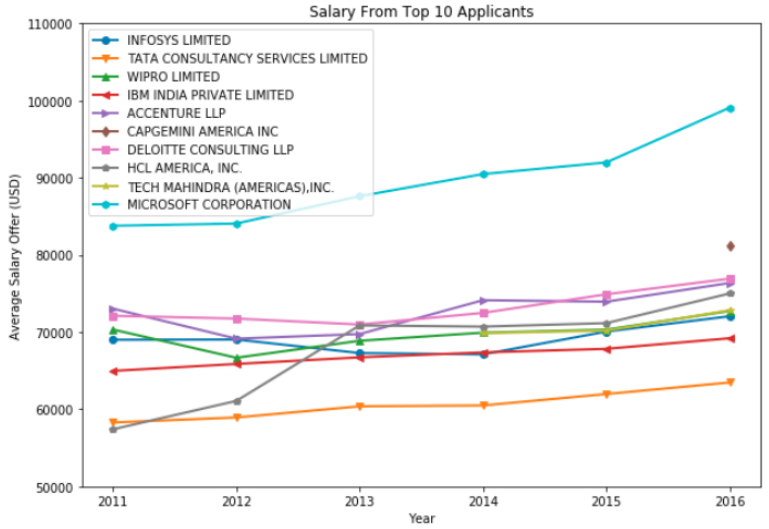


Top Applicants from 2011-2017

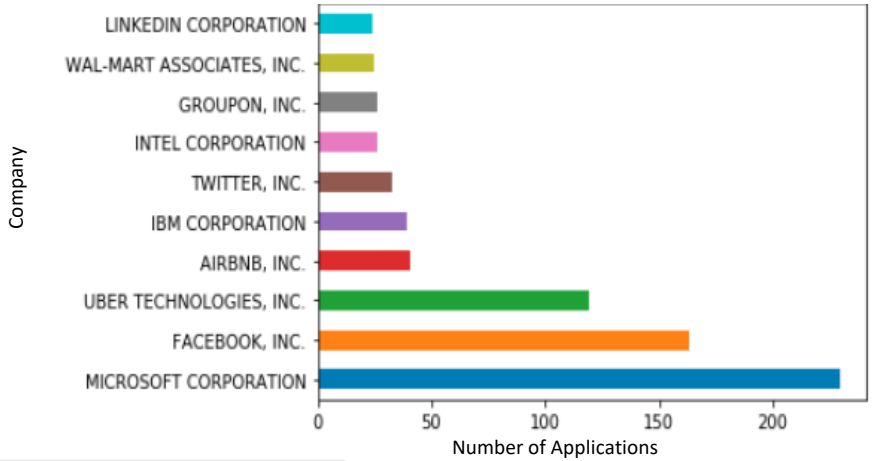


A Graph Speaks a Million Tuples

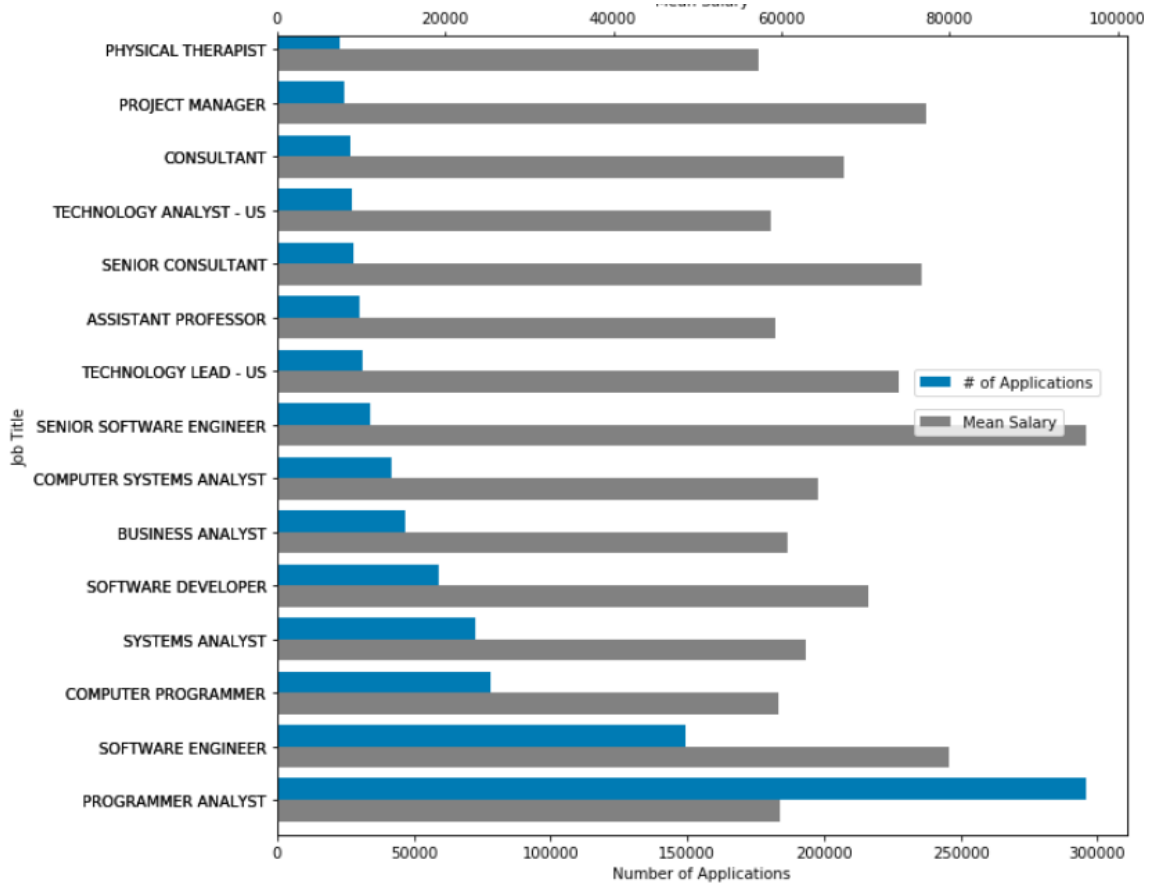
Average Salary Paid by Top 10 Applicants



Top 10 Companies Hiring Data Scientists



Top Job Titles and Mean Salaries

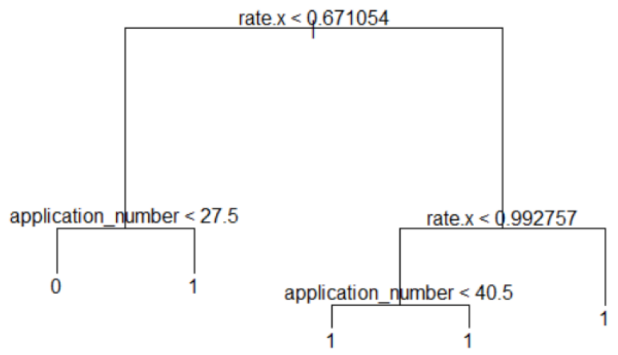


Modeling the Data

Decision Tree

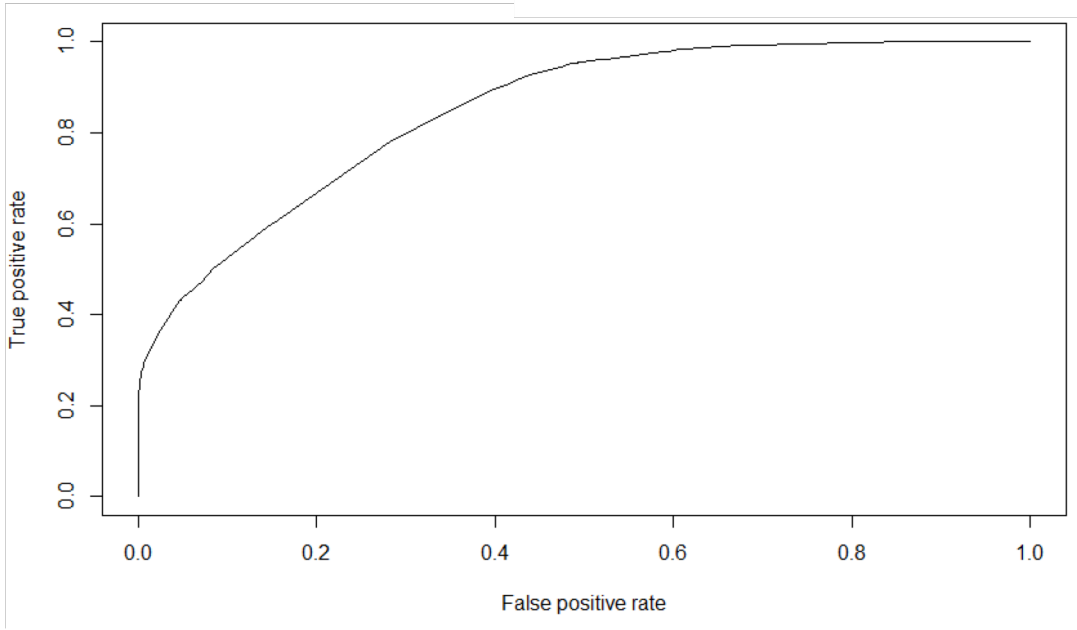
- Original Data with 1.6M values
- Performed sampling to get 500008 Certified Applications and 25004 Denied Applications
- Split into Train – Test : 75:25
- Feature Engineering to create 4 new variables – **Success Rate Per Employer, Number of Applications per Employer, Success Rate per SOC Code, Number of Applications filed per SOC Code**
- Use decision tree to identify probabilities of 1/0
- Assign values based on classifier
- Check values under AUC curve to identify classifier which provides highest AUC value.

```
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
```



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

Predict/True	0	1
0	4836	1499
1	3775	8643



Modeling the Data

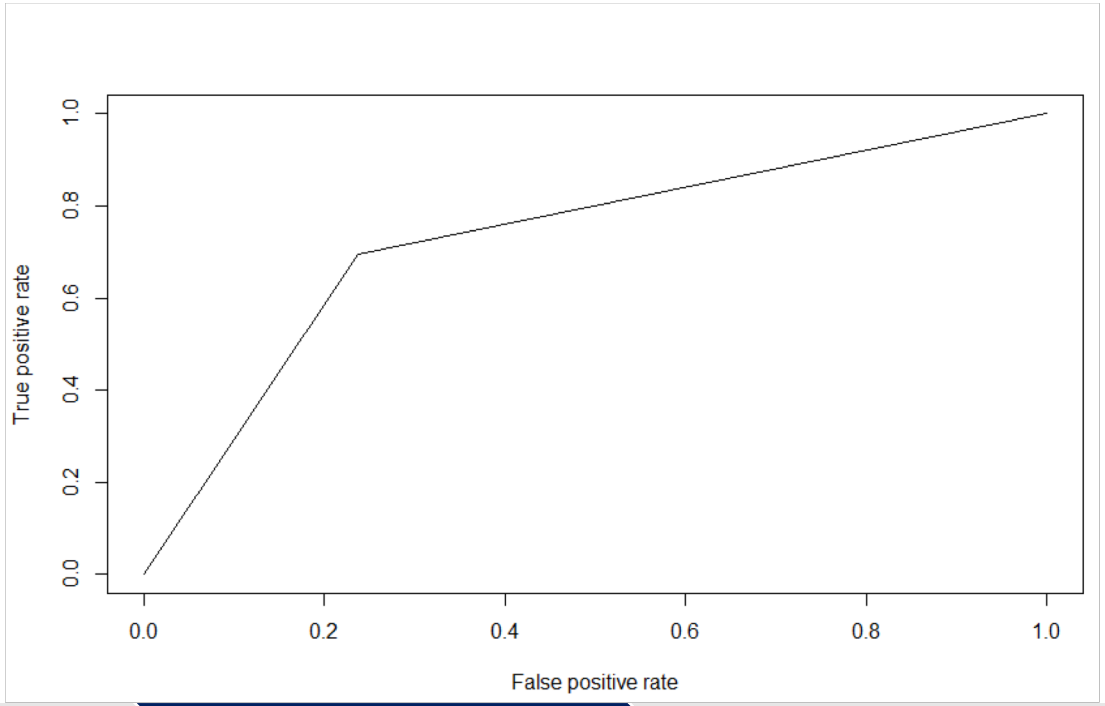
Logistic Regression

- 1/0, Y/N values converted to factors
- Logistic Regression applied to the 4 new variables and other columns of the original data set (Prevailing wage, H1-B Dependent, Willful Violator)
- Summary as shown to the right
- Classifier experimentation with various ranges

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

Predict/True	0	1
0	4272	1985
1	2643	9853

```
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
```



Wrap Up

Results

- Significant indicators critical to predict the likelihood of acceptance are – prevailing wage, H1-B dependent, employer specific success rate, number of applications filed/SOC code
- Company immigration violations in the past do not influence likelihood of success or failure
- Individual SOC types are not significant predictors of success
- Applicants with dependents more likely to be accepted – counter intuitive, can be attributed to experienced adults taking highly specialized roles and being paid higher

How Can We Better Model Performance?

- Random Forest gives 18.75% error rate – How do we

```
Call:
 randomForest(formula = factor_cs ~ ., data = lca_model)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 2

        OOB estimate of  error rate: 18.75%
Confusion matrix:
      0     1 class.error
0 13154 11850  0.47392417
1  2214 47794  0.04427292
```

- Identify other indicators that can enable better prediction
- Use LCA predictors to predict H1-B acceptance rate

Implement Other Models

- Neural Network – Inherently better at explaining complexities
- SVM/ Naïve Bayes Classifiers

Thank you!

Team 13
Abhilash Somasamudramath (as5637)
Jessica Jacob (jj2927)
Rohan Singh (rs3874)
Suriya Arumugavelan (sa3628)
Uday Marepalli (um2147)

Appendix

CTree

Predict/True	0	1
0	3219	3116
1	569	11849