SAS® GLOBAL FORUM 2018

USERS PROGRAM

Who Is Likely to Succeed: Predictive Modeling of the Journey from H-1B to Permanent US Work Visa

Shibbir Dripto Khan

Clark University , Graduate School of Management , Worcester , MA, USA

April 8 – 11 | Denver, CO

#SASGF

Predictive Modeling of the Journey from H-1B to Permanent US Work Visa

Shibbir Khan

Clark University , Graduate School of Management , Worcester , MA, USA

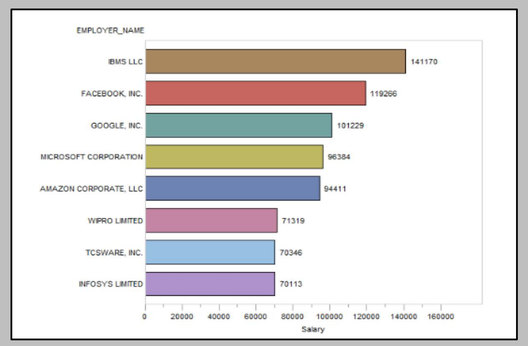
ABSTRACT

The purpose of this Study is to help US employers and legislators predict which employees are most likely to succeed in US Job market and therefore making sure that the time and money spent is for the most eligible ones. A permanent labor certification issued by the Department of Labor (DOL) allows an employer to hire a foreign worker to work permanently in the United States. Here we build several predictive models to signal among the applicants who has been awarded with the opportunity of H-1B visa which is a temporary work permit, are likely to sustain and succeed eventually. Then we shall do a probability measure of each H1B visa holders which we would later use as a valuation factor. This gives an idea as to which areas the U.S. government should emphasize to encourage domestic students to develop more local workforce in those job domains for the future.

The model will allow legislators and employers in US job market to pass legislation and conduct hiring process based on and targeted towards those workers who are possibly going to make the most out of the process.

DATA & METHODS

The H1B data set and the Permanent work authorization dataset is collected from United States Department of Labor - Office of Foreign Labor Certification (OFLC). The Office of Foreign Labor Certification (OFLC) generates disclosure data that is useful information about the immigration programs including the H1-B visa. we removed all the data points which had empty values for case status, In our models, we only included the cases 'CERTIFIED' and 'DENIED'. We decided to ignore 'CERTIFIED- WITHDRAWN' and 'WITHDRAWN' since those were decisions taken by the applicant and/or employer. The data set includes 40 columns and covers a total of 3 million records spanning from 2011-2016.



Basic Statistical Measures

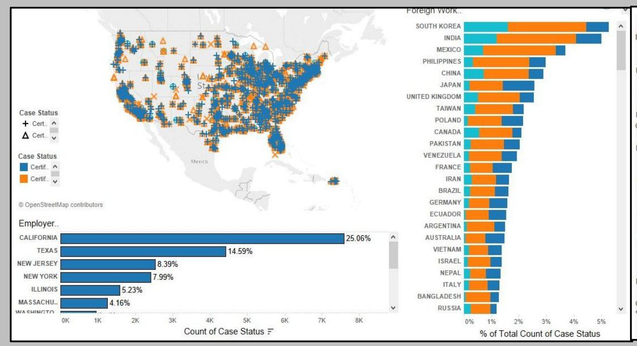
Location		Variability	
Mean	70611.53	Std Deviation	21471
Median	66165.00	Variance	460986889
Mode	60000.00	Range	150000
		Interquartile Range	25417

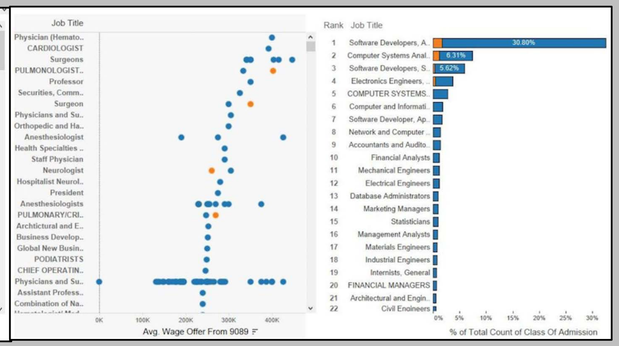
Basic Confidence Limits Assuming Normality

Parameter	Estimate	95% Confidence Limits	
Mean	70612	70570	70653
Std Deviation	21471	21441	.
Variance	460986889	459732894	.

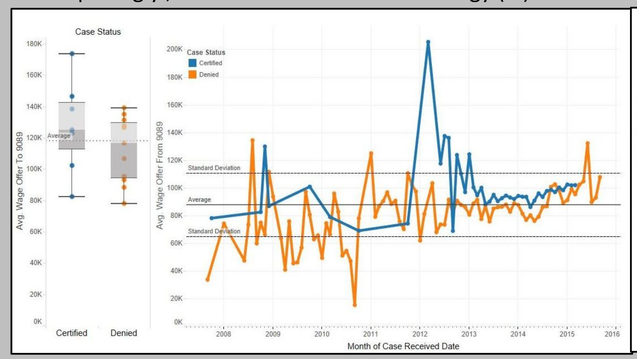
EXPLORATORY ANALYSIS

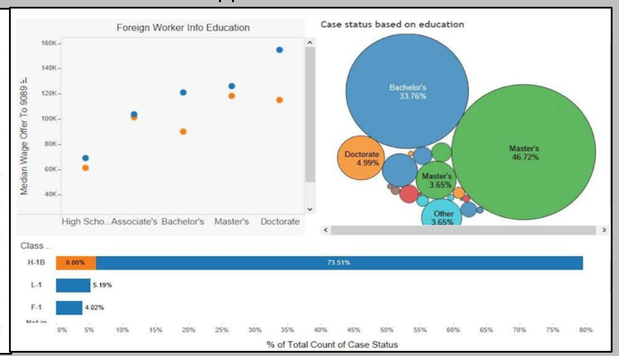
Exploratory Analysis : We analyzed the distribution of various features, their relation with the number of applications, and with the acceptance rate. These charts show significant trend in terms of these features.





Further exploratory analysis were done in terms of case status based on level of graduate and undergraduate education , year of approval , pay scale and jobs that hire he most These charts show significant trend in terms of these features. Not so surprisingly , Various Information Technology (IT) roles dominate the H-1B visa applications.





We have also used PROC summary in SAS Enterprise guide to get some basic statistical measures. Such as Mean , Variance and confidence intervals. We found the descending order of employers based on the number of applications submitted. We also found with PROC summary most frequent occupation code in the application pattern. We also used some data preparation tools such as string to number, Row filters and column filters. Missing value node was used to decide measure and imputation. Since logistic regression will not work with missing values, we derived Non Null distribution through R.

Predictive Modeling of the Journey from H-1B to Permanent US Work Visa

Shibbir Khan

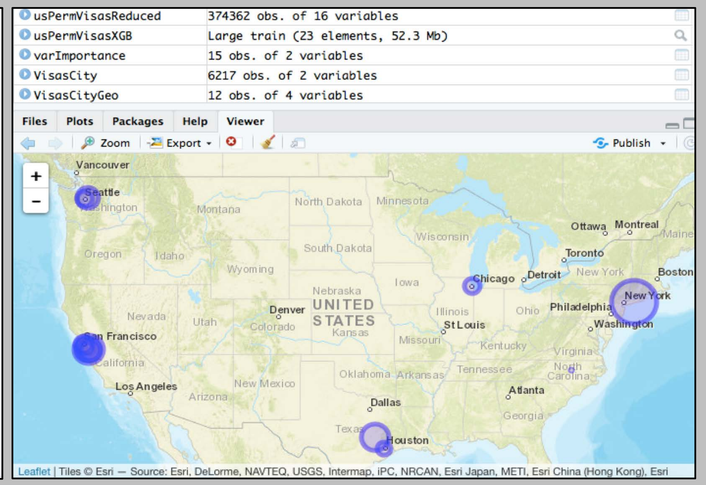
Clark University , Graduate School of Management , Worcester , MA, USA

EXPLORATORY ANALYSIS

The following map shows top cities with Visa Applications for Case Status as Certified. Large number of applications is indicated by a larger circle , we included Cary North Carolina. We used R's free third-party basemaps which were added using the addProviderTiles() function, this is implemented using leaflet-providers minimap plugin. minimap shows map tiles; markers, polygons, and other layers on the main map .We used Esri USGS WorldStreet minimap here

```
proc summary data=tmp00001.h1_b_data nway;
    class employer_name / descending order=freq;
    output out=count;
    run;
proc print data=count(obs=30);
    run;
```

Obs	EMPLOYER_NAME	_TYPE_	_FREQ_
1	INFOSYS LIMITED	1	48680
2	TATA CONSULTANCY SERVICES LIMITED	1	23125
3	IBM INDIA PRIVATE LIMITED	1	19026
4	WIPRO LIMITED	1	18644
5	CAPGEMINI AMERICA INC	1	16725
6	ACCENTURE LLP	1	16018
7	DELOITTE CONSULTING LLP	1	12021
8	TECH MAHINDRA (AMERICAS),INC.	1	9254
9	HCL AMERICA, INC.	1	8853
10	MICROSOFT CORPORATION	1	8116
11	ERNST & YOUNG U.S. LLP	1	7399
12	COGNIZANT TECHNOLOGY SOLUTIONS U.S. CORPORATION	1	7399
13	GOOGLE INC.	1	7305
14	IGATE TECHNOLOGIES INC.	1	5331
15	LARSEN & TOUBRO INFOTECH LIMITED	1	4569
16	IBM CORPORATION	1	3946
17	AMAZON CORPORATE LLC	1	3762
18	UST GLOBAL INC	1	3695
19	INTEL CORPORATION	1	2938
20	JPMORGAN CHASE & CO.	1	2508
21	L&T TECHNOLOGY SERVICES LIMITED	1	2479
22	DELOITTE & TOUCHE LLP	1	2456



RESULTS CONTINUED

For the predictive modeling we initially had a couple of different procedures in mind including Clustering and nearest-neighbor methods , but usually those are ideally suited for use with numeric data. However, our dataset mostly consists of categorical values, so using a probabilistic method, such as the Naive Bayes Classifier (NBC) is likely to provide more accurate results.

We have used four different classifiers for the pre- diction task and evaluated the results of each.:

A. Gaussian Naive Bayes
B. Logistic Regression
C. Gradient Boosted Classifier
D. Tree Ensemble Learner Classifier

RESULTS CONTINUED

Partitioning : before diving into the predictive modeling we have partitioned the datasets. We have performed sampling of 70% and 30% for training and testing and we have chosen stratified sampling based on the strata in case status column

A. Gaussian Naive Bayes

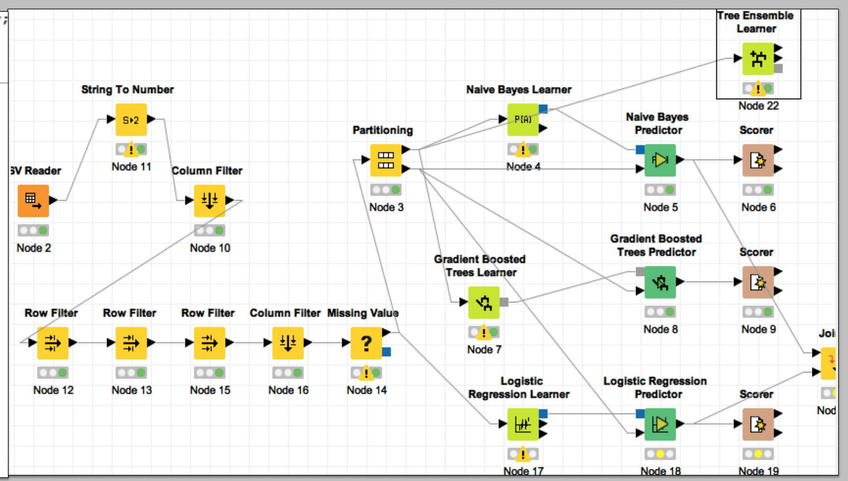
Naive Bayes assumes all features are conditionally independent given labels. It calculates p(x|y = 1), p(x|y = 0) and p(y) by taking their maximum likelihood estimates in the joint likelihood of the data. While making a prediction, it calculates p(y = 1|x) both for p(y=1) and p(y=0) using the Bayes rule and compares the two. For the Naïve Bayes learner we have created 4.2 PMML compatible model and we have selected case status as the classification column as it is our target column

B. Logistic Regression

For this we have again picked target column as 'case status' and reference category as 'certified'. We have selected solver as 'stochastic average gradient' In logistic regression, probability of the response taking a particular value is modeled based on combination of values taken by the predictors. Estimation in logistic regression chooses parameters that maximize the likelihood of observing the sample values. The advantage of Logistic Regression model is that it gives the confidence of prediction as a probability. The disadvantage is that it assumes that the classes are linearly separable in feature space.

```
proc summary data=tmp00001.h1_b_data nway;
    class job_title / descending order=freq;
    output out=count;
    run;
proc print data=count(obs=20);
    run;
```

Obs	JOB_TITLE	_TYPE_	_FREQ_
1	PROGRAMMER ANALYST	1	91745
2	SOFTWARE ENGINEER	1	49275
3	SOFTWARE DEVELOPER	1	21391
4	COMPUTER PROGRAMMER	1	20965
5	SYSTEMS ANALYST	1	20693
6	BUSINESS ANALYST	1	15520
7	COMPUTER SYSTEMS ANALYST	1	12480
8	TECHNOLOGY LEAD - US	1	10524
9	TECHNOLOGY ANALYST - US	1	10204
10	SENIOR SOFTWARE ENGINEER	1	9919
11	DEVELOPER	1	8750
12	CONSULTANT	1	8470
13	PROJECT MANAGER	1	7605
14	ASSISTANT PROFESSOR	1	6990
15	SENIOR CONSULTANT	1	6925
16	DATABASE ADMINISTRATOR	1	6331
17	ANALYST	1	4767
18	LEAD ENGINEER	1	4727
19	SYSTEMS ENGINEER - US	1	4704
20	PHYSICAL THERAPIST	1	4685



C. Gradient Boosted Classifier

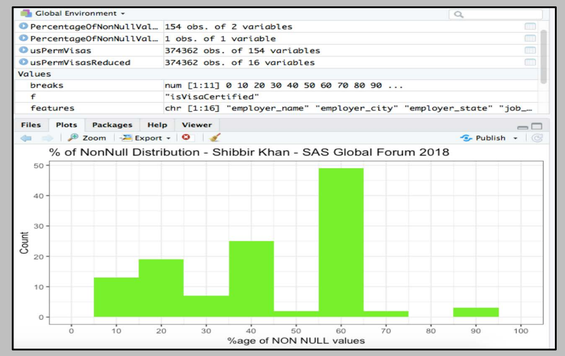
For this we picked as missing value handling method : 'XGBoost' , which we have later reproduced for H1B dataset through R Studio. we used a static random seed for this classifier. we changed prediction column name as *Prediction (case status) GB*

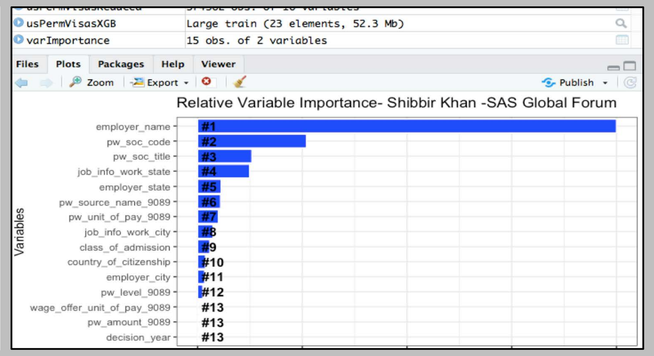
Predictive Modeling of the Journey from H-1B to Permanent US Work Visa

Shibbir Khan

Clark University , Graduate School of Management , Worcester , MA, USA

RESULTS CONTINUED

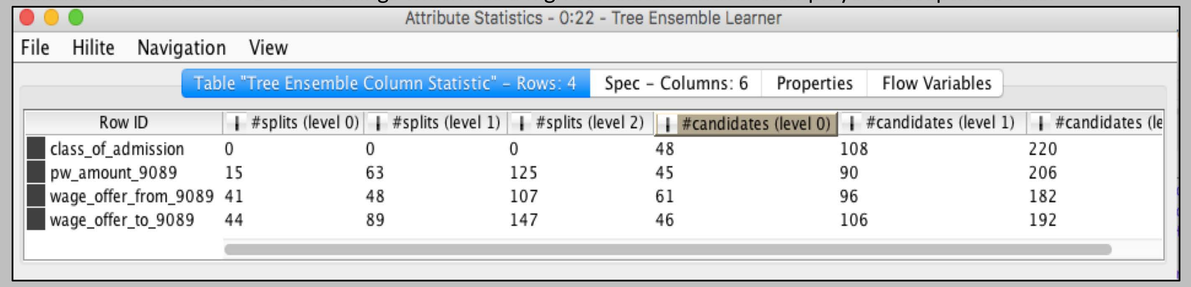




XGBoost is a scalable and accurate implementation of gradient boosting machines It is a library for developing fast and high performance gradient boosting tree models.

Feature Importance Graph.

shows the feature importance of the variables in the model. We observed that the most important feature to consider for is the acceptance ratio for the employer and the number of petitions filed by the employer. This clearly indicates the trends of H1- B visa filings which has a high correlation with the employer's acceptance rate.

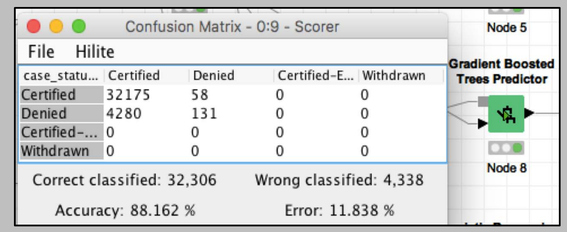


Row ID	#splits (level 0)	#splits (level 1)	#splits (level 2)	#candidates (level 0)	#candidates (level 1)	#candidates (le
class_of_admission	0	0	0	48	108	220
pw_amount_9089	15	63	125	45	90	206
wage_offer_from_9089	41	48	107	61	96	182
wage_offer_to_9089	44	89	147	46	106	192

RESULTS CONTINUED

D. Tree Ensemble Learner Classifier:

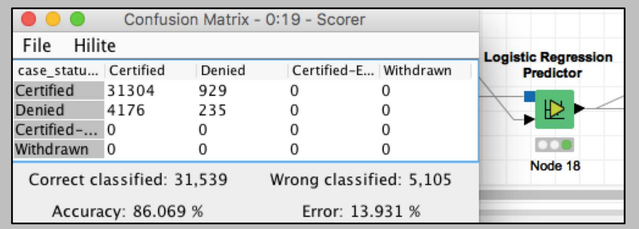
we picked Information Gain ratio as split criterion , we could also have picked Gini Index. We used binary splits for nominal columns. We asked to ignore columns without domain information. This classifier gives us an ensemble of decision trees. The output model is applied in the corresponding predictor node using the selected aggregation mode to aggregate the votes of the individual decision trees.



case_statu...	Certified	Denied	Certified-E...	Withdrawn
Certified	32175	58	0	0
Denied	4280	131	0	0
Certified-...	0	0	0	0
Withdrawn	0	0	0	0

Correct classified: 32,306 Wrong classified: 4,338

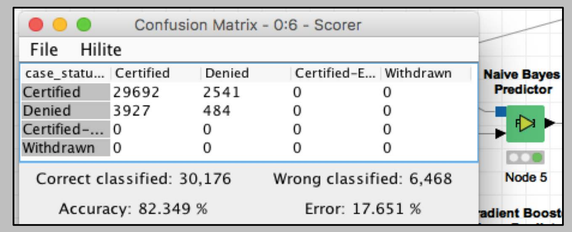
Accuracy: 88.162 % Error: 11.838 %



case_statu...	Certified	Denied	Certified-E...	Withdrawn
Certified	31304	929	0	0
Denied	4176	235	0	0
Certified-...	0	0	0	0
Withdrawn	0	0	0	0

Correct classified: 31,539 Wrong classified: 5,105

Accuracy: 86.069 % Error: 13.931 %



case_statu...	Certified	Denied	Certified-E...	Withdrawn
Certified	29692	2541	0	0
Denied	3927	484	0	0
Certified-...	0	0	0	0
Withdrawn	0	0	0	0

Correct classified: 30,176 Wrong classified: 6,468

Accuracy: 82.349 % Error: 17.651 %

CONCLUSIONS

In this work, Gaussian Naive Bayes, Logistic Regression, Tree Ensemble Learner Classifier and Gradient Boosted Classifier were considered for determining the status of PERM visa applications from H1B visa applications . Gradient Boosted Classifier performed the best in terms of accuracy as 88.16% were correctly classified. We achieved a 86.069% classification accuracy on validation data for Logistic Regression. Although we had high expectation for Naïve Bayes Classifier but we ended up having an accuracy of 82.34% , this can be due to the fact that it assumes that all our features are independent and we might have at least some form of correlation between the columns . We also inferred from our relative variable importance plot that : Particular well known Employer , prevailing wage , prevailing SOC occupation codes (type of occupation) play an important role in determining the case status of the visa application.

REFERENCES

[1] https://www.kaggle.com/nsharan/h-1b-visa/version/2
[2] https://www.foreignlaborcert.doleta.gov/performancedata.cfm
[3] https://www.kaggle.com/ambarish/eda-us-permanent-visas-with-feature-analysis
[4] https://blog.bigml.com/2013/10/01/using-text-analysis-to- predict-h1-b-wages/
[4] http://cs229.stanford.edu/proj2017/final-reports/5208701.pdf
[5] https://www.ischool.berkeley.edu/projects/2016/project-alien- worker
[6] https://cseweb.ucsd.edu/classes/wi17/cse258-a/reports/a054.pdf

SAS® GLOBAL FORUM 2018

April 8 – 11 | Denver, CO
Colorado Convention Center

Acknowledgement :
Professor Pankush Kalgotra phD , GSOM , Clark University

#SASGF