Predicting Case status of H1B Visa applications

Arundhathi Patil
Muma College of Business
University of South Florida
Tampa, FL, USA
arundhathi@mail.usf.edu

Palak Tater
Muma College of Business
University of South Florida
Tampa, FL, USA
palaktater@mail.usf.edu

Shruti Sridharan
Muma College of Business
University of South Florida
Tampa, FL, USA
shrutis@mail.usf.edu

Varsha Sharma
Muma College of Business
University of South Florida
Tampa, FL, USA
varshasharma@mail.usf.edu

ABSTRACT

According to information provided by Code.org and the U.S. Bureau of Labor Statistics, there is estimated to be 1 million more computing-based positions than qualified applicants to fill them by 2020. Even if all 170,000 H-1B visas went to the most qualified applicants in the lottery over the next two years, we would still be 870,000 experts short. Every year, the US immigration department receives over 200,000 petitions and selects 85,000 applications through a random process. The application data is available for public access to perform in-depth longitudinal research and analysis. This data provides key insights into the prevailing wages for job titles being sponsored by US employers under H1-B visa category.

Fig 1 – Factors involved in H1B VISA application

INTRODUCTION

The H-1B is a visa in the United States under the Immigration and Nationality Act, section 101(a)(15)(H) that allows U.S. employers to temporarily employ foreign workers in specialty occupations. This is the most common visa status applied for and held by international students once they complete college / higher education (Masters, PhD) and work in a fulltime position. H-1B visa class is very industry relevant and many individuals and companies rely heavily on this yearly allotment. Laws limit the number of H-1B visas that are issued each year. In 2015, there were 348,669 applicants for the H-1B filed, of which 275,317 were approved.

Data subject matter includes personal details of the employer requesting temporary labor certification and the role itself. In this paper we try to classify acceptance and denial of a candidate's H1B application into different categories. This will help public to understand the demographics that goes into an application paper of submission. We have applied various data cleaning methods, classification techniques and algorithms on different timelines of data to predict the chances of a candidate to receive a work visa at the U.S.

1. DATA

1.1 The Enigma Dataset

The dataset used in this project is from the category of "Office of Foreign Labor" from the dataset in https://public.enigma.com. This dataset contains metadata from The United States federal government, Department of Labor. The dataset was fetched using API it is from 2015 to 2017. The dataset size is about 200,000 rows and 35 fields.

1.2 Data Preprocessing

The data fields which we have used for this project are,

1. **Case_Number** - Unique identifier assigned to each application submitted to processing to the National Process Center

2. **Employee_Tenure** - Difference in the employment state date and employment end date in terms of days
3. **Employee_Name** - Name of employer submitting labor condition application
4. **Employer_State** - Employer requesting temporary labor certification-Corporate/Main State
5. **SOC_Name** - Occupational name associated with the SOC Code (the job requested for temporary labor condition, as classified by the Standard Occupational Classification System)
6. **Job_Title** - Title of the job
7. **Total_Workers** - Total number of foreign workers requested by the Employer(s)
8. **Prevailing_Wage** - Prevailing wage for the job being requested for temporary labor condition
9. **H1B_dependent** - Y = Employer is H1-B Dependent; N = Employer is not H1-B dependent.
10. **Willful_Violator** – City information of the foreign worker’s intended area of employment
11. **Year_Of_Case_Filling** - Derived from “Submit Data” field
12. **Days_Of_Case_Filling** - Derived from “Submit Data” field (Can be Monday, Tuesday...Sunday)
13. **Case_Status_ID** - Target Variable – Accepted (1) and Denied (0)

1.3 Data Sampling

Due to various missing data cells and other indefinite attributes, gathering data from the entire population of the dataset was causing imbalance in the dataset. Using MS-Excel, we performed random sampling of about 13,000 rows using the function “RAND” and retrieving the rows by ascending division. In our models, we only included the cases ’CERTIFIED’ and ’DENIED’ and these were labeled ’1’ and ’0’ respectively. We decided to ignore ’CERTIFIEDWITHDRAWN’ and ’WITHDRAWN’ since those were decisions taken by the applicant and/or employer.

1.4 Exploratory Analysis

1.41. StatExplore

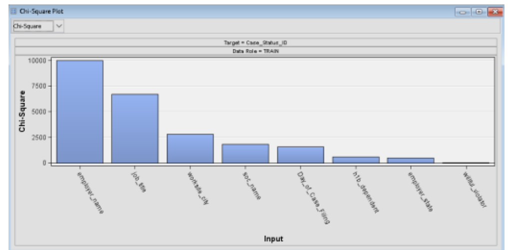


Fig 2: Chi Square Plot

Chi-Square Plot - The Chi-Square statistic shows the strength of the relationship between the target variable and each categorical input variable.
So this plot explains that first few variables explains target variable more accurately.

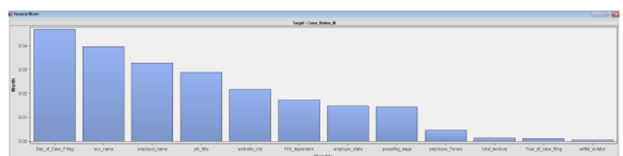


Fig 2: Variable worth plot

Variable worth Plot - This plot displays the worth of each input. The worth is calculated from the p-value corresponding to the calculated Chi-Square test statistic.Both Chi-Square plot and variable worth plot shows that Employer_Name, JOB_TITLE and SOC_Name are the most important variable since it has the highest Chi-Square value and the highest worth.

1.4.2 Variable Selection Node

There are two basic techniques used by the Variable Selection node. They are the R-Square selection method and the Chi-Square selection method. Both these techniques select variables based on the strength of their relationship with the target variable. For interval targets, only the R-Square selection method is available. For binary targets both the R-Square and Chi-Square selection methods are available.

1.5 Tableau Data Exploration

Link to our live tableau data exploration story:
https://public.tableau.com/profile/arundhathi.patil#!/vizhome/AnalysisonH1Bvisaapplications/Story2

We performed data exploration and analysis using Tableau to predict inferences on our dataset.
By exploring data visually, we found few interesting insights about the H1B applications

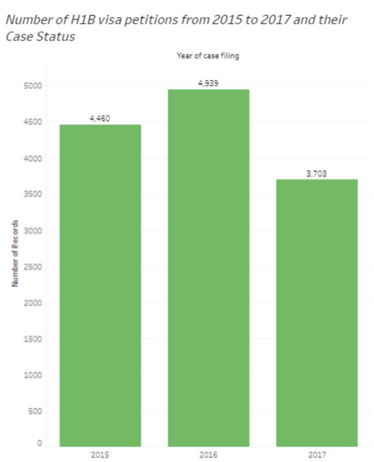


Fig 3: Dataset of 3 years

This graph represents the sample data used from the year 2015 to 2017. The year 2015has 4460 data rows, 2016 has 4939 data rows and 2017 has 3703 data rows.

Fig 4: Statewise H1B visa prediction

This picture talks about the top 10 states that applied for the H1B visa petitions. The leading states are California, Texas and New Jersey. New Jersey has certified cases of about 77% followed by Texas (76%) and California (67%).

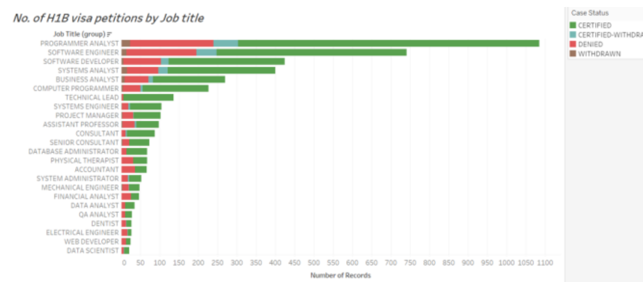


Fig 5: Job Title wise H1B visa prediction

The most popular job titles are Programmer analyst, Software engineer and software developer.
About 781 certified cases have been found out of about 1100 total application in California.

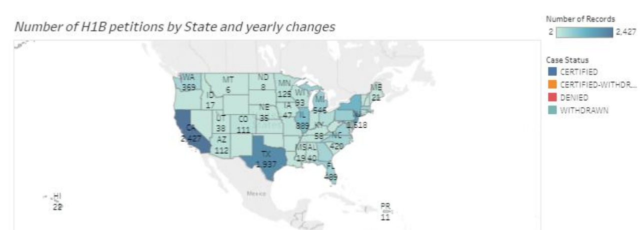


Fig 6: Overall state and yearly prediction

States like California, New York, Texas and Illinois show a decreasing trend in the number of H1B applications. However generally these states have the maximum concentration of employers. Despite the changes in visa regulations proposed by the new government, states like Washington, Montana, Idaho and North Dakota have an increasing trend in the number of H1B applications filed.

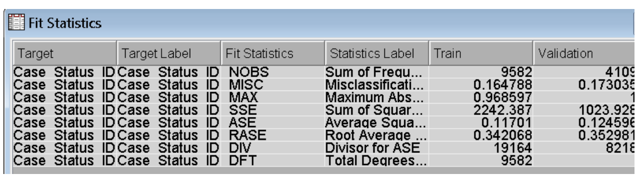
2. METHOD

2.1.1 Decision Tree Model

In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules.

Decision tree run result

Fit Statistics shows that the misclassification rate for training and validation dataset is 0.16 and 0.17, respectively.



Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
Case Status ID	Case Status ID	NOBS	Sum of Frequ...	9582	4109
Case Status ID	Case Status ID	MISC	Misclassificati...	0.164788	0.173035
Case Status ID	Case Status ID	MAX	Maximum Abs...	0.968597	1
Case Status ID	Case Status ID	SSE	Sum of Squar...	2242.387	1023.928
Case Status ID	Case Status ID	ASE	Average Squa...	0.11701	0.124596
Case Status ID	Case Status ID	RASE	Root Average ...	0.342068	0.352981
Case Status ID	Case Status ID	DIV	Divisor for ASE	19164	8218
Case Status ID	Case Status ID	DFT	Total Degrees...	9582	

Fig 7: FIT statistics for decision tree

Classification Chart for CASE_STATUS

Here around 17% of the Denied case status have been classified correctly and around 67% of the Certified case status have been classified correctly.

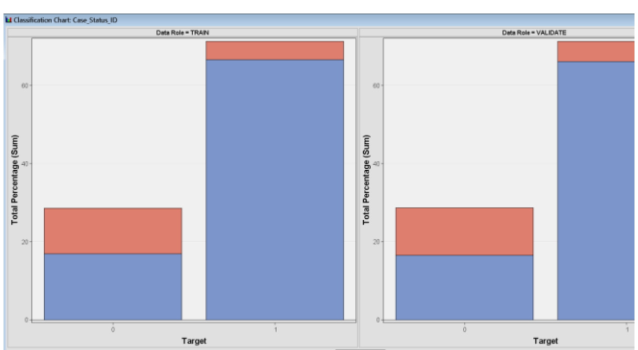


Fig 8: Classification chart using case status in decision tree

Misclassification Rate is inversely proportional to number of leaves. As number of leaves increases, misclassification rate decreases. Subtree assessment plot shows that if number of leaves are between 20 and 40, we will get lowest misclassification rate.

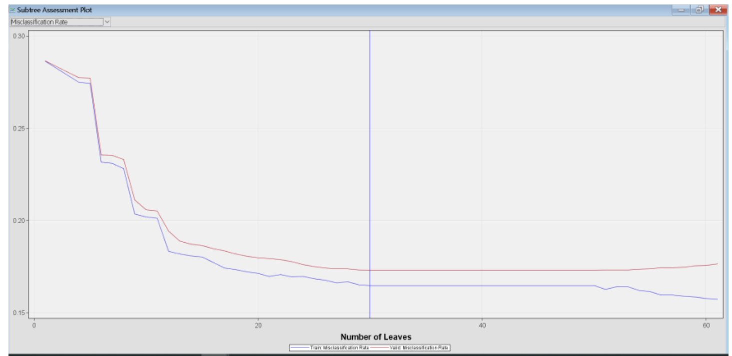


Fig 9: Plot analysis

Confusion Matrix

As number of Denied cases are low compared to number of Registered cases; confusion matrix result will provide better picture of prediction analysis.

Decision Tree Confusion Matrix

Data Role: Train

False Negative	True Negative	False Positive	True Positive
452	1616	1127	6387

Data Role: Validate

False Negative	True Negative	False Positive	True Positive
213	679	498	2719

A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. true positives (TP): These are cases in which we predicted yes (they are certified). true negatives (TN): We predicted no, and they are not certified. false positives (FP): We predicted yes, but they are not certified. ("Type I error.") false negatives (FN): We predicted no, but they are certified.

2.1.2 Neural Network Model

Neural network is a non-linear statistical data modeling tool. It models complex relationships between inputs and outputs and finds patterns in data. Neural network accommodates a wide variety of nonlinear relationship between a set of predictors and target variable.
In this experiment,

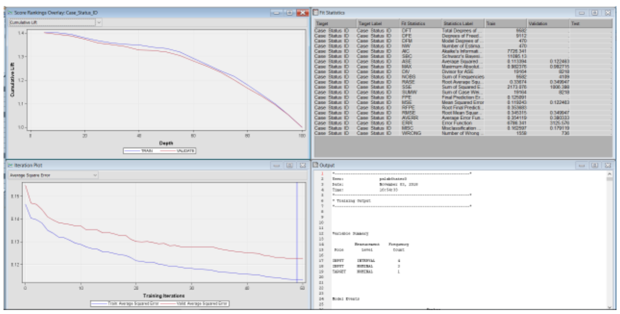


Fig 10: graphical reprentation of output using neural

Here, Increasing the number of hidden nodes to 7 and in the optimization technique, the training technique applied is “back prop” Through this Number of false positives reduce to 1104 from 1143.

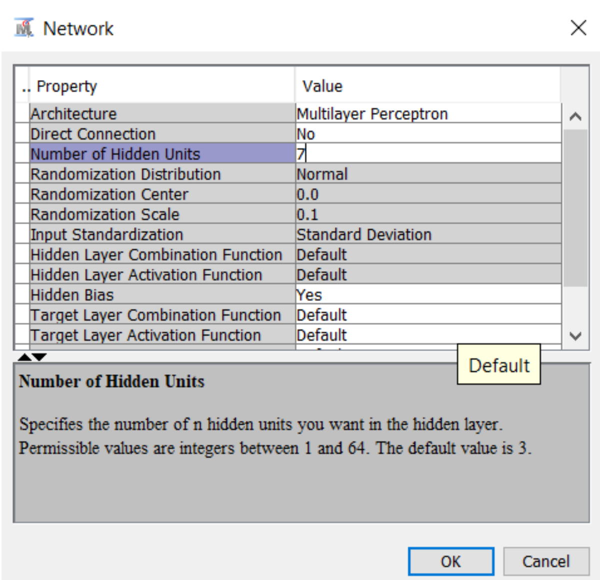


Fig 11: properties of NN

Classification chart using Case Status ID
The misclassification rate for training and validation dataset is 0.162 (Train) and 0.179(Validation), respectively.

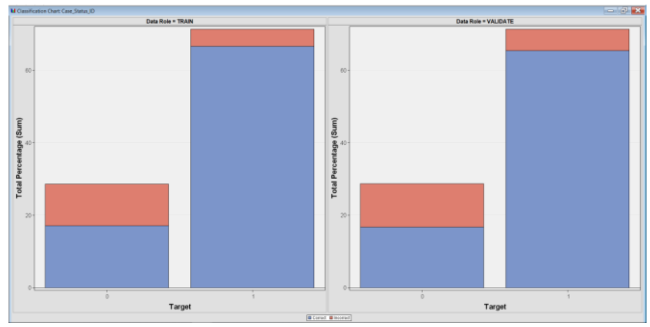


Fig 12: Classification chart using case status in NN

Confusion Matrix

Data Role: Train

False Negative	True Negative	False Positive	True Positive
454	1639	1104	6385

Data Role: Validate

False Negative	True Negative	False Positive	True Positive
245	686	491	2687

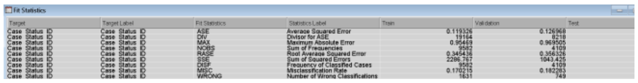
2.1.3 Random Forest Model using HP Random Forest on SAS Miner

Scoring new observations on many trees enables to obtain a consensus for a predicted target value (in our case the prediction of H1B approval) with a more robust and generalizable model.

Set Maximum number of trees to 50 with maximum depth to 50 in order to use get a good share of trees under the forest. Set minimum category size to 5 to make sure the order to use the category in a split search. We have maintained the missing value to “Use in Search” since we shall use missing values as a separate value in split search.

Random Forest run result

Fit Statistics shows that the misclassification rate for training and validation dataset is 0.17 (Train) and 0.18(Validation), respectively.



Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
Case Status ID	Case Status ID	ASE	Average Squared Error	0.119326	0.126968	
Case Status ID	Case Status ID	DIV	Divisor for ASE	19164	8218	
Case Status ID	Case Status ID	MAX	Maximum Absolute Error	0.95469	0.969505	
Case Status ID	Case Status ID	NOBS	Sum of Frequencies	9582	4109	
Case Status ID	Case Status ID	RASE	Root Average Squared Error	0.345436	0.356326	
Case Status ID	Case Status ID	SSE	Sum of Squared Errors	2286.767	1043.425	
Case Status ID	Case Status ID	DISF	Frequency of Classified Cases	9582	4109	
Case Status ID	Case Status ID	MISC	Misclassification Rate	0.170215	0.182283	
Case Status ID	Case Status ID	WRONG	Number of Wrong Classifications	1631	749	

Fig 13: Fit Statistics

Confusion Matrix

Data Role: Train

False Negative	True Negative	False Positive	True Positive
276	1388	1355	6563

Data Role: Validate

False Negative	True Negative	False Positive	True Positive
146	574	603	2786

With 1355 False positive cases, Decision tree and Neural network gave us a better confusion matrix than Random forest experiment.

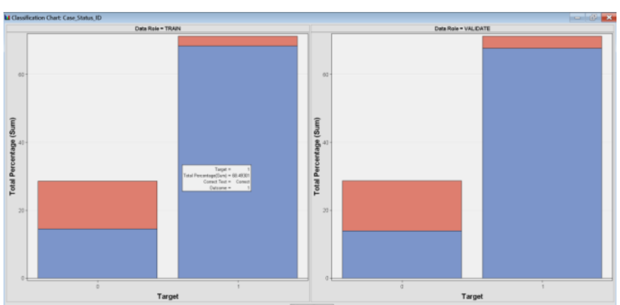
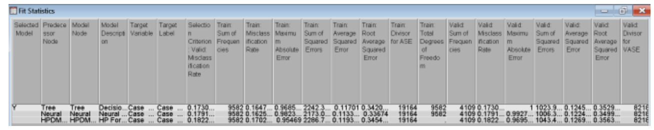


Fig 14: Classification chart in Random Forest

The classification chart using case_status variable has almost same prediction on train as well as validation data.

3. EVALUATION:

We used compare node to draw inferences on Decision Tree Model, Neural Network Model and Random Forest Model. we can conclude that Decision Tree is predicting results more accurately than the others.



Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Train: Sum of Frequencies	Train: Misclassification Rate	Train: Maximum Absolute Error	Train: Sum of Squared Errors	Train: Average Squared Error	Train: Root Average Squared Error	Train: Divisor for ASE	Train: Total Degrees of Freedom	Valid: Sum of Frequencies	Valid: Misclassification Rate	Valid: Maximum Absolute Error	Valid: Sum of Squared Errors	Valid: Average Squared Error	Valid: Root Average Squared Error	Valid: Divisor for VASE
Y	Tree	Tree	Decisio...	Case ...	Case ...	0.1730...	9582	0.1647...	0.9685...	2242.3...	0.11701	0.3420...	19164	9582	4109	0.1730...	1	1023.9...	0.1245...	0.3529...	821[illegible]
	Neural	Neural	Neural ...	Case ...	Case ...	0.1791...	9582	0.1625...	0.9823...	2173.0...	0.1133...	0.33674	19164	9582	4109	0.1791...	0.9927...	1006.3...	0.1224...	0.3499...	821[illegible]
	HPDM...	HPDM...	HP For...	Case ...	Case ...	0.1822...	9582	0.1702...	0.95469	2286.7...	0.1193...	0.3454...	19164		4109	0.1822...	0.9695...	1043.4...	0.1269...	0.3563...	821[illegible]

Fig 15: Fit Statistics of compare model node

Our Fit statistics shows that the training model has done better than the validation model specially when we compare the misclassification rat.

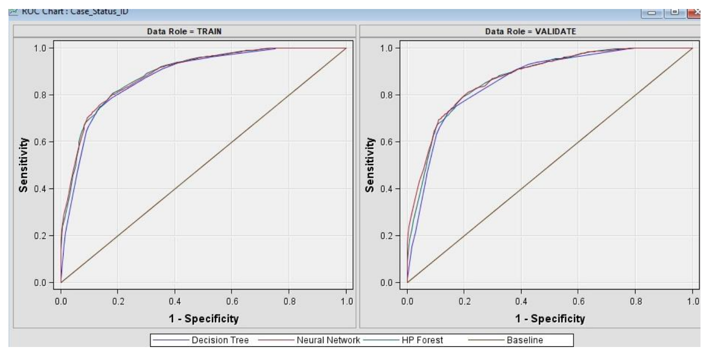


Fig 16: AUC-ROC curve performance

AUC - ROC curve is a performance measurement for classification problem at various thresholds settings. ROC is a probability curve and AUC represent degree or measure of separability. It tells how much model is capable of distinguishing between classes. Higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s. By analogy, Higher the AUC, better the model is at distinguishing between case status to certify and deny. This ROC curve shows that all the three experiments that we have performed has done a good job to differentiate certified and denial cases.

Confusion Matrix of all Models

Model Node	Model Description	Data Role	Target	False Negative	True Negative	False Positive	True Positive
Tree	Decision Tree	Train	Case_Status-ID	452	1616	1127	6387
Tree	Decision Tree	Validate	Case_Status-ID	213	679	498	2719
Neural	Neural Network	Train	Case_Status-ID	454	1639	1104	6385
Neural	Neural Network	Validate	Case_Status-ID	245	686	491	2687
Random Forest	HP Forest	Train	Case_Status-ID	276	1388	1355	6563
Random Forest	HP Forest	Validate	Case_Status-ID	146	146	603	2786

3. CONCLUSION

In this work, Decision Tree, Random Forest and Neural network were considered for determining the status of H1-B visa applications. Model comparison classifier performs a good collation to combine all these models and predict accuracy. We achieved a best of classification accuracy with decision tree model. We inferred that the state of worksite, year of application, prevailing wages, employer name and soc-name play an important role in determining the case status of an H-1B application. We observed that the most important feature to consider for our model 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.

- 50% applications were denied if case is filed on Friday.
- 72% cases were approved if prevailing wage is greater than 42000

Models	FPR	FNR	Misclassification Rate (Validation)	Misclassification Rate (Train)
Decision Tree	1129	213	0.173	0.1647
Neural Network	1104	245	0.1791	0.1625
Random Forest	1355	146	0.1822	0.17

4. REFERENCES

https://public.enigma.com/browse/d582dfbd-4329-4b5e-b0c9-39149f5dd546
https://nycdatascience.com/blog/student-works/h-1b-visa-petitions-exploratory-data-analysis/
https://www.ischool.berkeley.edu/projects
/2016/project-alien-worker