Understanding General Trends in Permanent Visa Applications and Predicting Visa Decisions using SAS Enterprise Miner.

ARUN TEJA BAIREDDLAPALLI KRISHNA REDDY

OKLAMOHA STATE UNIVERSITY

Contents

ABSTRACT

Lately, employing foreign workers in US companies has become a cumbersome task due to uncertainty in permanent work visa approvals by Department of Homeland Security's U.S. Citizenship and Immigration Services (USCIS). Before submitting an immigration petition to USCIS, the employer must submit a labor certification application to the Department of Labor (DOL) indicating there are not enough US workers qualified and willing to accept the job opportunity and employment of foreign workers will not pose any threat or affect the working conditions and wages of similar US workers. This research paper is intended to help US employers understand the general trend and factors that affect visa decisions and predict visa application decisions based on employer, company, wage, position, employee education background, and past visa history. The data covers decisions of permanent visa applications from 2012-2017 with more than 300,000 records and 152 variables. Base SAS, SAS Enterprise Guide and SAS Enterprise Miner have been utilized for data cleaning, exploration and data modeling.

INTRODUCTION

The number of permanent visa applications submitted in US has been increasing at unprecedented rate and has become a major concern for employers due to constantly changing rules and regulations regarding work permit. This study helps uncover different factors that affect permanent visa decisions and predict visa application decisions by Department of Labor.

The dataset was obtained from Kaggle. The data was originally collected and distributed by the US Department of Labor. Dataset covers permanent visa applications form 2012-2017 and contains 152 explanatory variables with more than 300,000 visa applications and includes information on employer, offered wage, employee position, type of visa application and job location.

Questions that are answered in this study

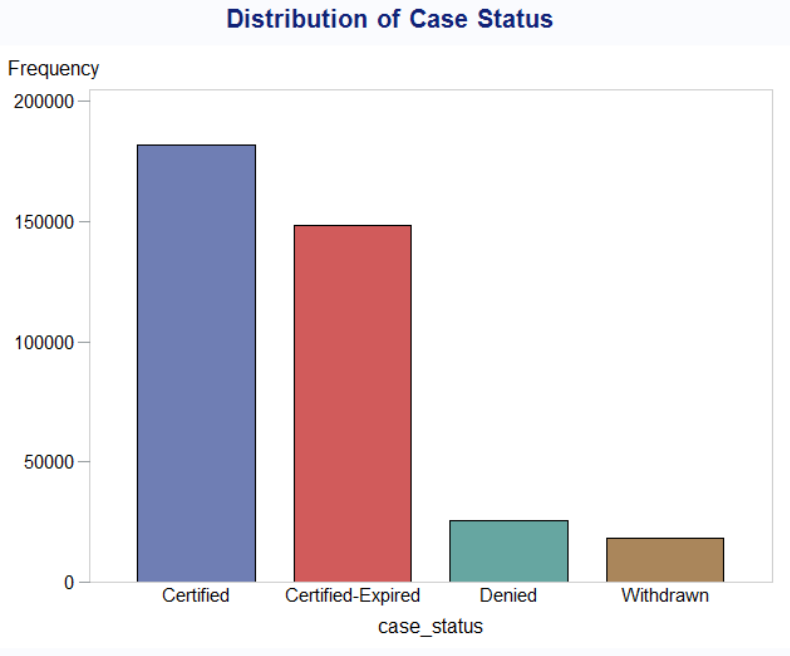
1. What are the key factors that affect permanent visa decisions?

2. Can we build a model that predicts permanent visa decisions based on key factors?

Data Cleaning

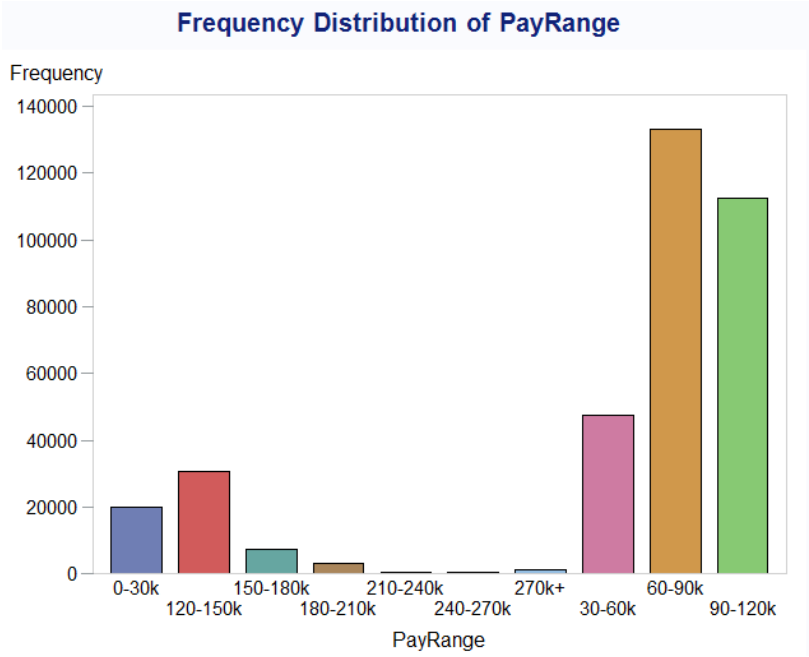
1. Exploratory data analysis was initially done and variables that had less than 15% missing values were used, and remaining variables were dropped form the dataset.

2. Target variable Case_Status had 4 levels namely Certified, Certified-Expired, Denied and Withdrawn as shown in the fig below. Employers will have 6 months to file immigration petition before status expires and becomes Certified-Expired. So Certified-Expired and Certified has been combined together and Case_Status Withdrawn has been removed from the dataset.



3. Variable PW_Unit_of_Pay had more than different levels of pay like Bi-Weekly, Monthly, Weekly, Hourly. All the levels were converted to Year.

4. New variable called PayRange has been created to look at the range of wages as shown in below figure

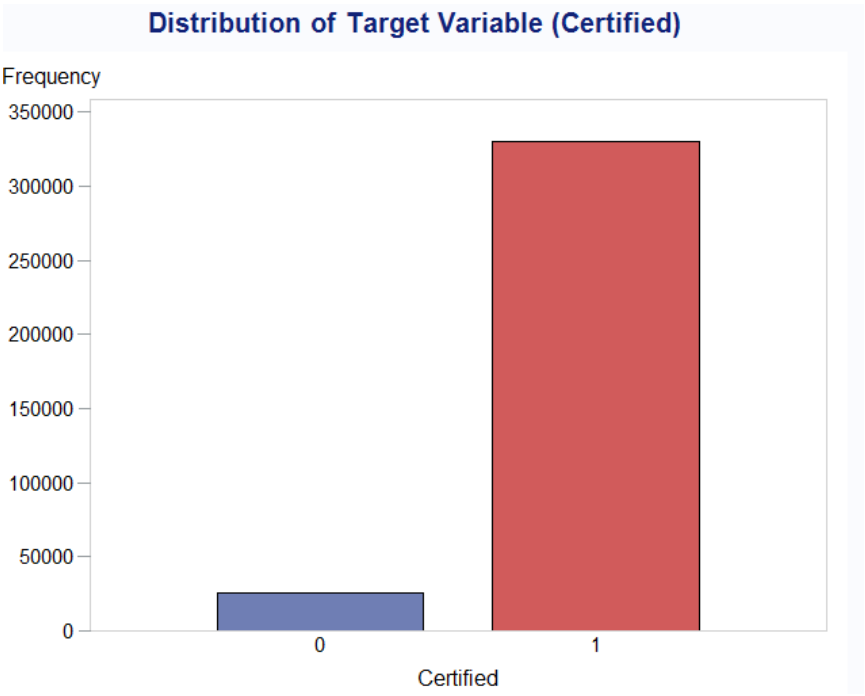


5. State names were in different formats like shown in figure below. They been converted into state code format

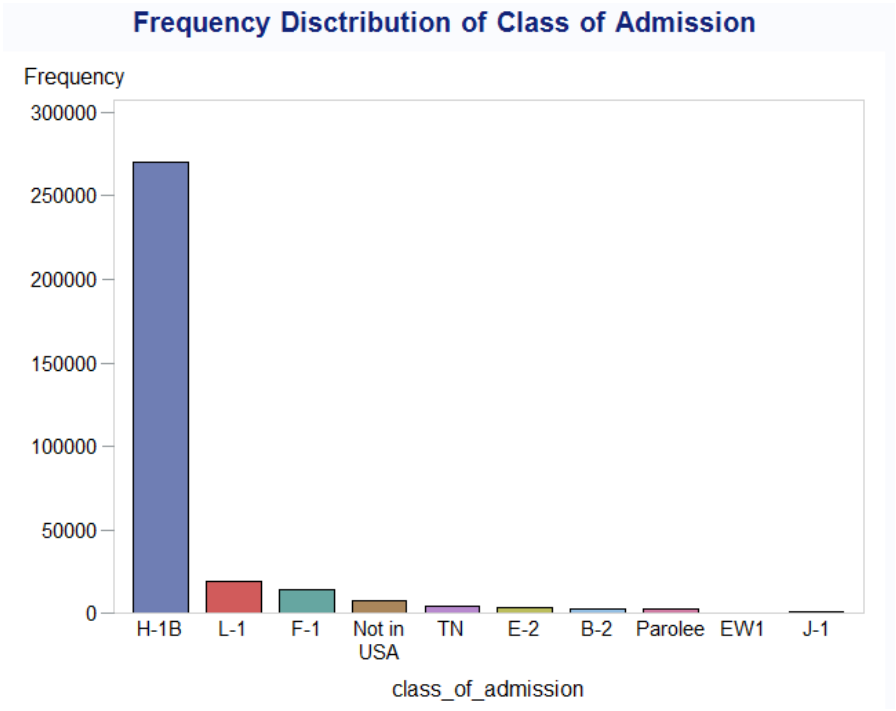
Obs	employer_state	_TYPE_	_FREQ_
1		0	356131
2	CALIFORNIA	1	48312
3	CA	1	40868
4	TEXAS	1	23865
5	TX	1	19369
6	NEW JERSEY	1	15318
7	NEW YORK	1	14676
8	NJ	1	13320

Exploratory Data Analysis

The figure below shows the distribution of the target variable Certified. Out of all visa applications 7.2% of the applications were denied and 92.8% of the applications were accepted.

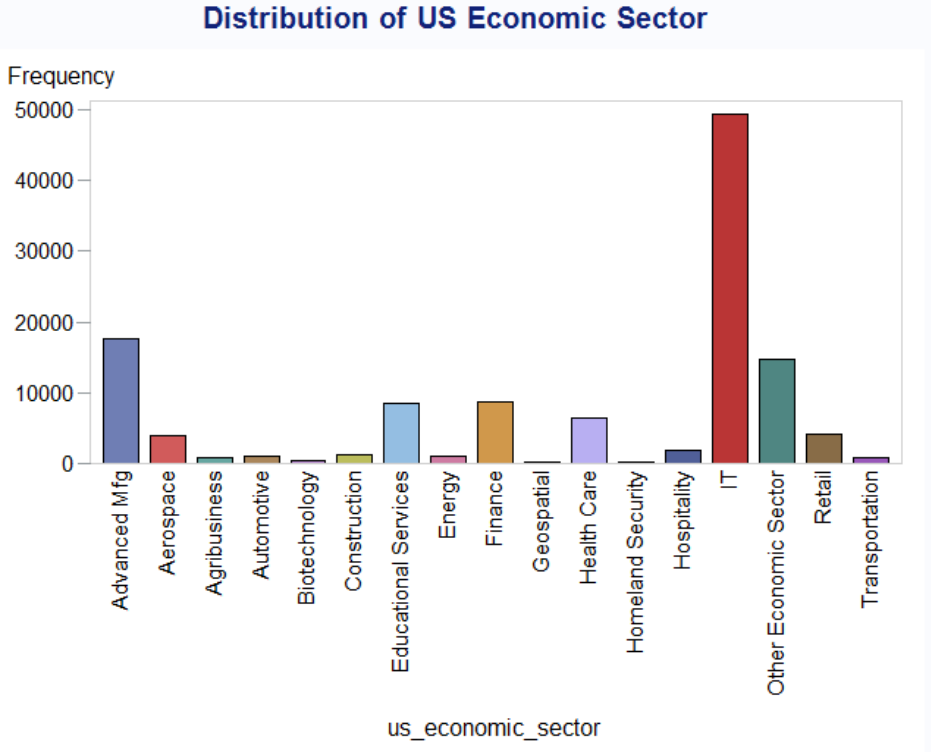


As expected, working visa, H1B tops the list in class of admission as shown in the figure below. Some of the other type of visas were L-1 and F-1 (Under OPT).

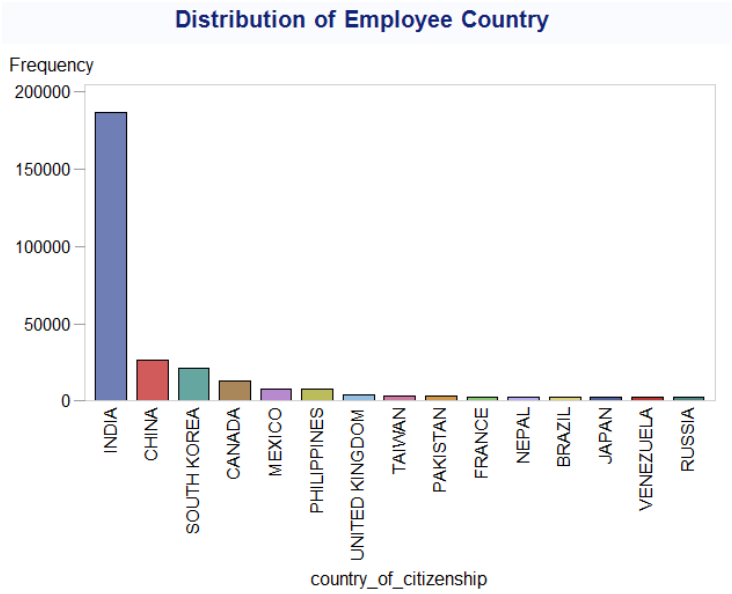


The figure below shows the distribution of the number of employees from different economic sectors. Usually the IT sector dominates other sectors in most of the developed countries and developing countries. As seen in the figure below number of applications from IT sector is almost

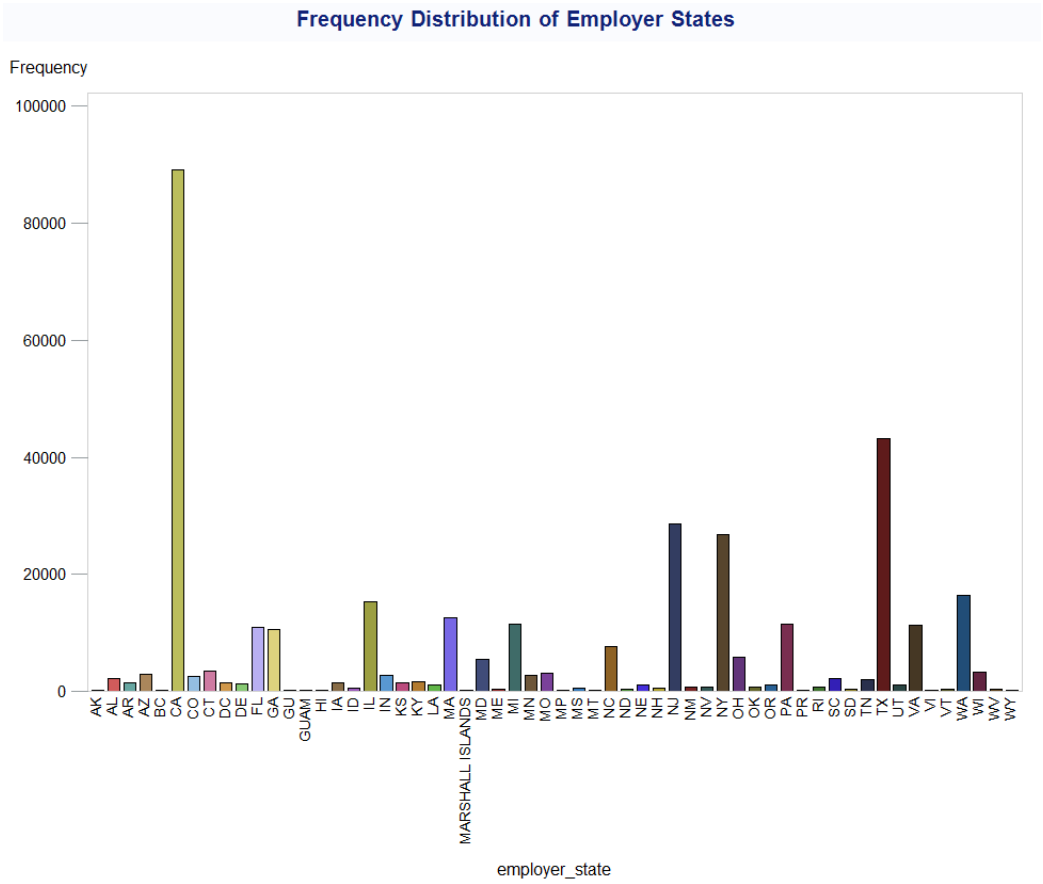
twice than the second highest which is Advanced Manufacturing. Some of the other sectors were Finance, Retail and Educational Services.

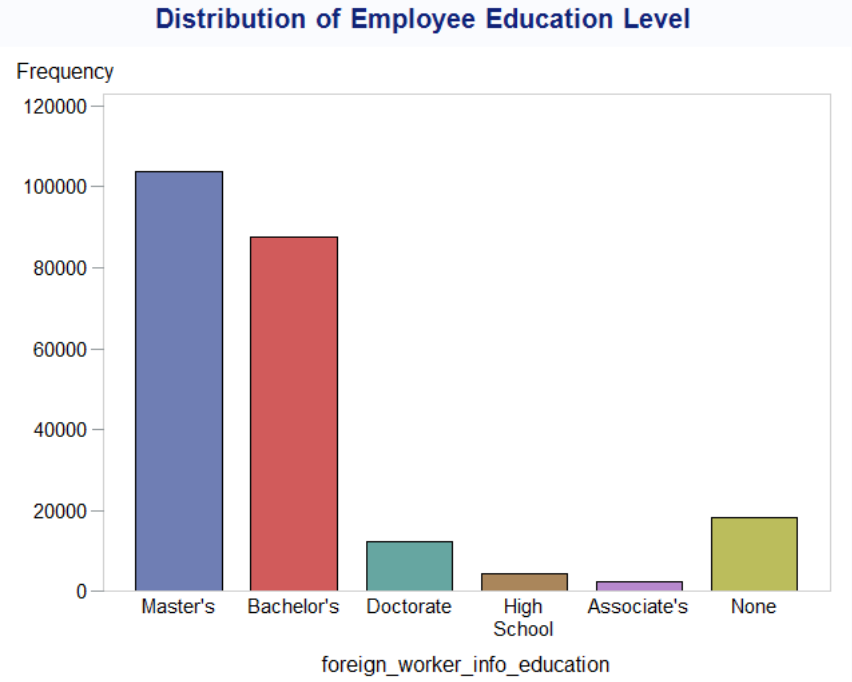


The below output shows the foreign workers from top 15 countries. People from populated countries like India and China go to developed countries like US, UK in search of jobs. As we supposed, we can see India, China and South Korea are in top 3. Surprisingly applications for Indian Employees are almost thrice that of China.



The number of applications were highest from California one of the biggest states of US shown in figure below. It is expected as Silicon Valley is in California. Some of the other big states are TX, NY and NJ.

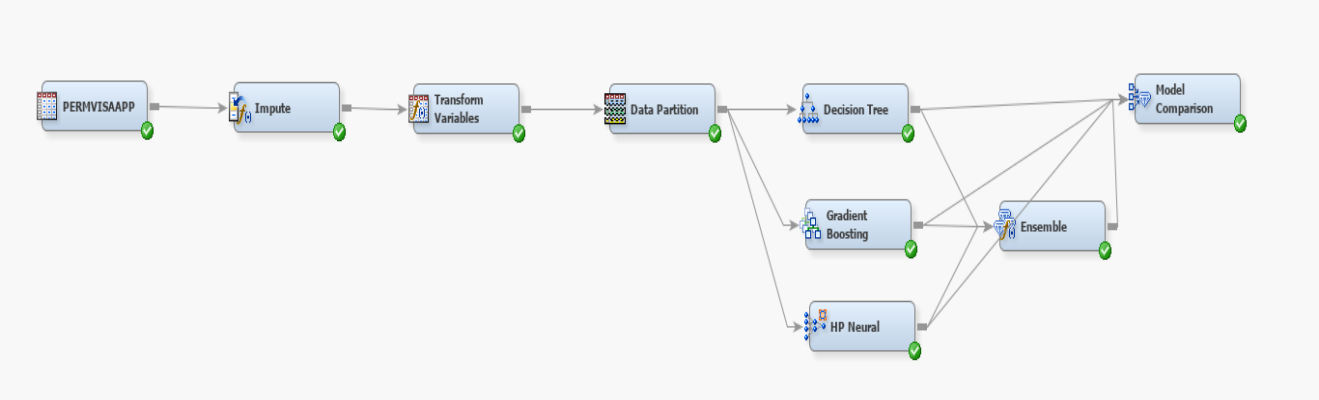




The above figure shows the education level of the foreign workers. As we can see master's and bachelor's students outnumber other education levels as many people from countries like India and China come to US to pursue master's or bachelors.

Methodology

Below figure shows the predictive modelling methodology used in this study.



Input Data

Cleaned Dataset PERMVISAPP has been imported into SAS Enterprise Miner using Data node. The below figure shows the roles and levels of the final selected variables.

Name	Role	Level
Certified	Target	Binary
PayRange	Input	Nominal
case_status	Rejected	Nominal
class_of_admission	Input	Nominal
country_of_citizenship	Input	Nominal
decision_date	Rejected	Interval
employer_address_1	Rejected	Nominal
employer_city	Input	Nominal
employer_name	Rejected	Nominal
employer_state	Input	Nominal
job_info_work_city	Rejected	Nominal
job_info_work_state	Rejected	Nominal
pw_amount_9089	Input	Interval
pw_soc_code	Input	Nominal
pw_soc_title	Rejected	Nominal
pw_source_name_9089	Rejected	Nominal
pw_unit_of_pay_9089	Input	Nominal
year	Input	Nominal

Impute

Except target variable Certified all the variables had missing values initially. Missing values reduces the number of usable observations. To overcome this issue, impute node has been used and missing values for both continuous and categorical variables were imputed using Tree based imputation method.

Transform Variables

The interval variable Pw_amount_9089 which is employee age had high skewness and kurtosis values. To make the variable normally distributed and minimize skewness and kurtosis max normal transformation was initially employed. Using max normal Log 10 transformation has been selected as best transformation and used to transform the variable.

Data Partition

The dataset has been divided into 70% training and 30% validation using Data Partition node.

Model Comparison

After employing Decision Tree, Gradient Boosting, Neural Networks and Ensemble nodes Model comparison node has been used to compare model nodes.

Results

Important Variables

Using decision tree important variables that affect the target variable have been identified and the same can be seen in the figure below. PW_SOC_CODE is the most important variable which is the code for Occupation of employees. It is followed by PW_AMOUNT_9089 which is employee wage per year followed by Employer_city. The results shows that higher the job level of employee better the chances and more the number of work experience of employee more the chances. H1B and L1 visa applications have better chances of getting approved compared to

other visa type applications. Employees working in big cities like Atlanta, New Jersey, Dallas, California, New York have more chance of getting approved compared to other cities.

Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
IMP_pw_soc_code	Imputed pw_soc_code	3	1.0000	1.0000	1.0000
LG10_pw_amount_9089	Transformed pw_amount_9089	11	0.5938	0.6102	1.0275
IMP_employer_city	Imputed employer_city	5	0.5334	0.4661	0.8738
year		4	0.3990	0.4412	1.1056
IMP_class_of_admission	Imputed class_of_admission	1	0.1604	0.1683	1.0488
IMP_employer_state	Imputed employer_state	1	0.0317	0.0861	2.7127

Model Comparison Results

We see that based on Misclassification rate Neural Network has been selected as the best model as per Model Comparison results. Multiple models have been employed like Decision Tree, Gradient Boosting, Neural Network, Ensemble models. The misclassification rate of Neural Network for train dataset was 6.17% and for validation dataset was 6.36%. Here we don't see overfitting. Though Neural Network has been selected as the best model it is best to go with Decision Tree because of the interpretability.

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Train: Misclassification Rate	Train: Average Squared Error	Train: Divisor for ASE	Train: Maximum Absolute Error	Train: Sum of Frequencies	Train: Root Average Squared Error
Y	HPNNA	HPNNA	HP Neural	Certified		0.061701	0.050775	498632	0.999257	249316	0.225334
	Tree	Tree	Decision ...	Certified		0.062748	0.054627	498632	0.974403	249316	0.233724
	Ensmbl	Ensmbl	Ensemble	Certified		0.064717	0.053523	498632	0.966899	249316	0.23135
	Boost	Boost	Gradient ...	Certified		0.072009	0.066824	498632	0.927991	249316	0.258503

Limitations

Due to a greater number of levels in many variables Logistic Regression and Neural Network were not used.

Conclusion and Future Scope

- We have seen the important all the important variables like Employee Wage, Occupation Code, Employee State that affect Visa decisions. These variables make intuitive sense as in general perspective these are the variables that affect visa decisions.
- This model can be used in predicting future visa decisions with an accuracy of 93.8%.
- In future I would like to analyze job description and title of the job by leveraging text analytics to predict visa decisions.

References

1.US Department of Labor

2. Practical Business Analytics Using SAS: A Hands-on Guide

Contact Information

Feedback, comments and questions are highly encouraged. Contact the author at:

Arun Teja Baireddlapalli Krishna Reddy

MS in Business Analytics

Oklahoma State University, Stillwater

abaired@okstate.edu