US Permanent Visa Applications Analysis

Venkata Sai Santosh Perumalla
Student ID: 801029180
+1 980-202-8148
vperumal@uncc.edu

Priyanka Taneja
Student ID: 801046756
+1 980-267-5328
ptaneja@uncc.edu

Rashi Jain
Student ID: 801046743
+1 980-267-5004
rjain12@uncc.edu

Sonika Kannegalla
Student ID: 801053058
+1 704-421-5436
skannega@uncc.edu

Akhil Kumar Reddy Are
Student ID: 801045078
+1 704-501-6965
aare@uncc.edu

Saketh Kumar Kappala
Student ID: 801046556
+1 704-501-7737
skappal2@uncc.edu

ABSTRACT

The Department of Labor (DOL) issues a permanent labor certification which allows an employer to hire a foreign worker for working in the United States permanently. The DOL must certify to the U.S. Citizenship and Immigration Services (USCIS) that there are not sufficient U.S. workers able, willing, qualified and available to accept the job opportunity in the area of intended employment and that employment of the foreign worker will not adversely affect the wages and working conditions of similarly employed U.S. workers. This permanent visa applications rejection and acceptance can be dependent on various factors as location, education qualification, prevailing wages, employer, job title etc.

With this paper, few hypothesis are created and illuminated utilizing this dataset. The plan of this investigation is to examine the dataset, explore the visualizations and apply distinct modelling techniques using XLMiner in quest of finding the solutions for the generated hypothesis.

1. INTRODUCTION

United States lawful permanent residency, informally known as having a green card, is the immigration status of a person authorized to live and work in the United States of America permanently. Green cards are valid for 10 years for permanent residents, and 2 years for conditional permanent residents. After this period, the card must be renewed or replaced. The application process may take several years. An immigrant usually has to go through a three-step process to get permanent residency that includes petition and processing.

This data set contains the permanent visa applications to the United States over a period of 9 years (2006-2015) from around the world and from all the class of admissions counting to 3,57,187 records. The term paper is about visualizing and analyzing the dataset. It covers the analysis of three hypothesis made from the dataset. Initially the data is preprocessed to 66,000 records and further reduced according to individual hypothesis needs. The second section explains the pre-processing of data followed by visualizations of the data. The hypothesis are briefed and business use case is explained in the third. Section four briefly explains the various modelling techniques applied on the data. In the fifth section, classification, association and prediction models are applied on the data based on the hypothesis. A set of three models are applied on each hypothesis and then the reports are validated to select the best model based on various measures such as error percentage, area under curve, lift charts, and others. Examining all the models, the final section covers the issues faced and recommendations are listed accordingly.

2. DATA SET

The US Permanent Visa Applications data is collected from Kaggle [1]. It was originally collected and distributed by the US Department of Labor. This dataset provides information about the people applying for US permanent visa, their location, education level, prevailing wage, job title etc. It consists of more than 3 lakh records which are pre-processed and used to create classification models and thus predicting the solutions of the hypothesis questions.

A. Pre-Processing of Data

The dataset contains some of the redundant columns and some columns with all the values as null, these columns are ignored and following columns are considered for further analysis. The proportion of the dependent variable values were maintained properly when the data set is being reduced. The data where few columns have null values have also been removed.

- **application_type**: Mode of application.
- **case_received_date**: Date the applications was received by the ETA National Processing Center

- **case_status**: Status associated with the last significant event or decision. Valid values include "Certified," "Certified-Expired," "Denied," and "Withdrawn"
- **class_of_admission**: Indicates the class of immigration visa the foreign worker held at the time the permanent labor certification application was submitted for processing (if applicable)
- **country_of_citzenship**: Country of citizenship of the foreign worker being sponsored by the employer for permanent employment in the United States.
- **decision_date**: Date on which the last significant event or decision was recorded by the ETA National Processing Center
- **employer_country,employer_decl_info_title,employer_name,employer_state** : Contact information of the employer requesting permanent labor certification
- **employer_num_employees**: Total Number of employees employed by employer.
- **foreign_worker_info_alt_edu_experience**: Indicates whether the foreign worker possesses the alternate combination of education and experience
- **foreign_worker_info_birth_country**: Foreign Worker's country of birth
- **foreign_worker_info_education:** Highest Education achieved by the foreign worker
- **foreign_worker_info_major**: Major field(s) of study in reference to the highest level achieved by the foreign worker
- **foreign_worker_info_req_experience**: Indicates whether the foreign worker has the experience as required for the requested job opportunity
- **foreign_worker_info_state**: State of the foreign worker
- **job_info_alt_combo_ed**: The alternate level of education that is acceptable, in combination with experience (if applicable)
- **job_info_alt_combo_ed_exp**: Indicates if an alternate combination of education and experience will be acceptable in lieu of minimum level of education requirement
- **job_info_experience**: Identifies whether experience in the job offered is a requirement
- **job_info_experience_num_months**: The number of months of training that is required (if applicable)
- **job_info_foreign_ed**: Indicates if a foreign educational equivalent is acceptable
- **job_info_foreign_lang_req**: Indicates if knowledge of a foreign language is required to perform the job duties
- **job_info_job_title**: Common name or payroll title of the job being offered.
- **job_info_work_state**: State information of the foreign worker's intended area of employment
- **Duration**: Number of days for getting decision

The number of samples considered are 66582. Initially, all the rows were removed which contain the value *null*. This can be achieved by missing data field in the XLMiner.

B. Data Visualization

Initial analysis of the dataset is done and following visualizations are done based on the different predictors selected:

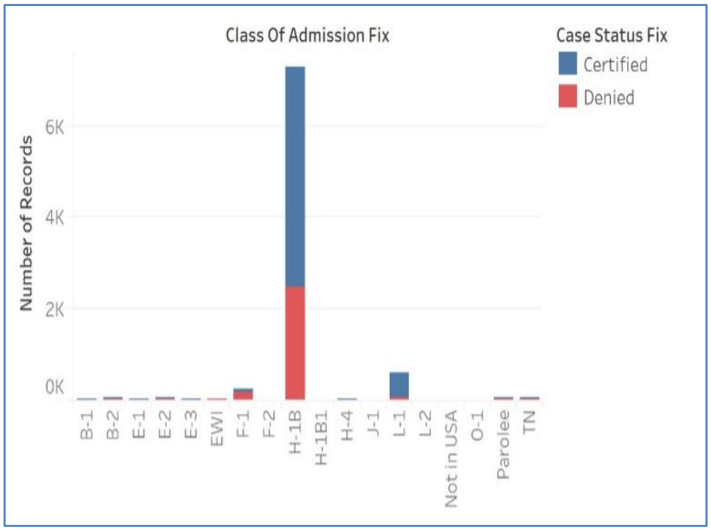


Fig 1: The histogram shows the case status i.e. Certified or Denied based on the class of admission of the applicants, it can be inferred that most of the applications to permanent visa are made by the H-1B class people and that the certified to denied ratio is high for the L1 category.

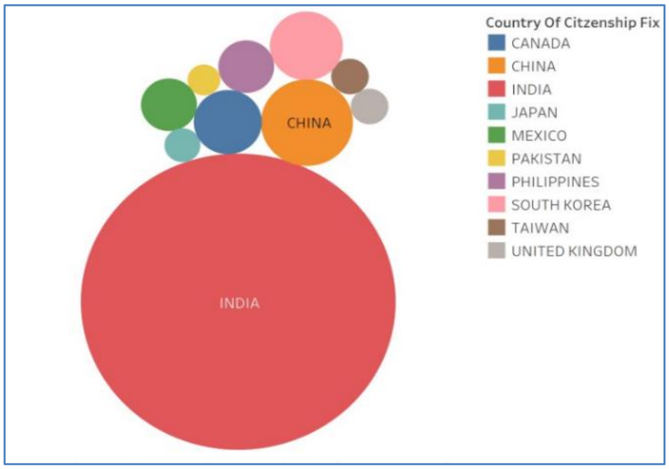


Fig 2: The bubble graph represents the permanent visa applications country wise, it can be inferred that even among the top 10 countries applications, India leads with a huge margin contributing more than fifty percent of the data.

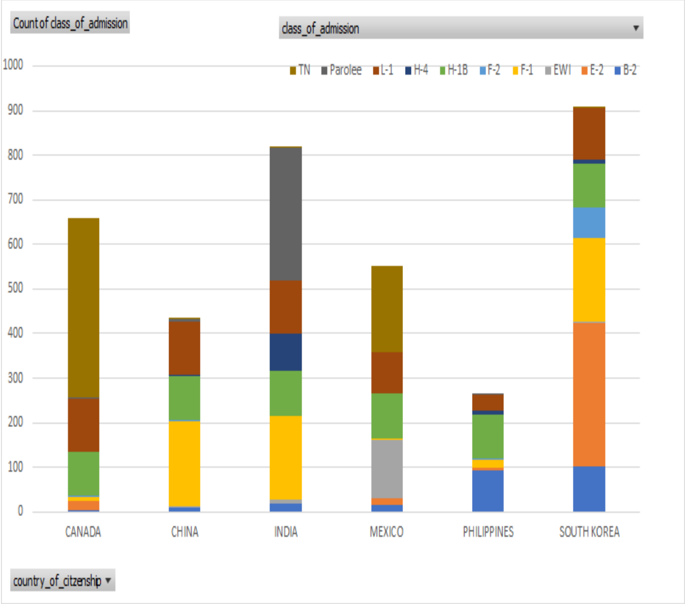


Fig 3: The above Bar Graph shows the top six countries with individual's Class of admission applying for the USA permanent visa.

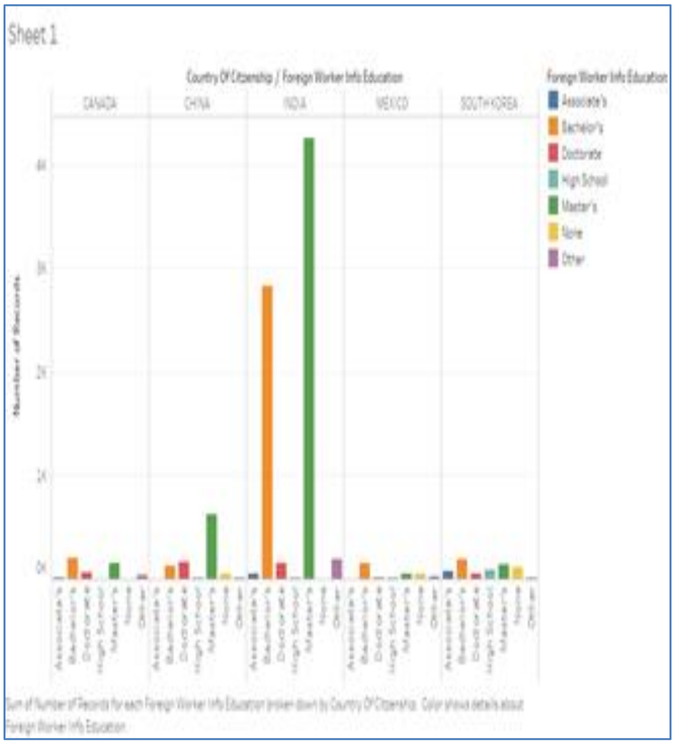


Fig 4: The histogram shows the education level of people from top 5 countries that applied for US permanent visa. It could help government in comparing the education level differences between various countries and their own. This would further help in bringing out new policies and better decision making.

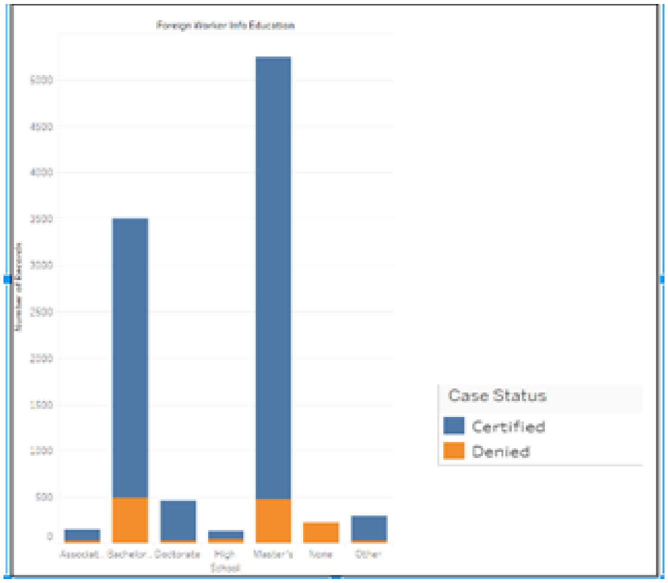


Fig 5: The above histogram shows the case status i.e. certified or Denied based on the education level of the people.

3. HYPOTHESIS QUESTIONS

i. *Analyzing the chances of getting permanent visa from a specific region for various visa categories and based on experience.*

   This analysis can assist the government to frame and implement certain policies on getting the permanent visa over a region in the future.

ii. *Predicting the duration for getting the USA Permanent Visa decision based on Class of Admission and country of citizenship for the top six countries and top ten class of admission.*

   This prediction estimates the average time required t to process ones application and helps an individual to plan accordingly.

iii. *To find the chances of getting permanent residency based on the education level of the people filing the petition for topmost 5 countries.*

   This analysis can assist the government to frame and implement certain policies on getting the permanent visa based upon the education level of the people applying.

Business Use Case:

The objective for the project is to prepare a dataset to help the US government in analyzing the data and the patterns of it so that it can aid them in making policies of issuing a permanent visa by examining the current trends and predictors from the dataset. This data was collected and distributed by the US Department of Labor.

Our data mining project will classify the output variable based on the predictors used. One of main instances of data investigation will be finding a pattern designated to specific region and specific class of admission where it can found if more/less permanent visas are being issued. In case of any imbalance, the Government can refrain its current policy and update a new one.

The main reason to issue a permanent visa is when there is a shortage of skilled U.S. workers. With this analysis, the Government can modify the education system to fill the gap.

4. MODELS

There are diverse demonstrating procedures that can be connected on the pre-prepared dataset for anticipating and breaking down the information. All the three hypothesis questions are addressed utilizing these demonstrating procedures by picking the best pertinent model for a specific theory. Following are the general modelling techniques:

Clustering:

In data mining, Clustering is commonly used for unsupervised learning technique. It is a way of locating similar data objects into clusters based on some similarity.

Modelling techniques under clustering:

1. K – means clustering
2. Hierarchical clustering

Classification:

Classification models predict categorical class labels. It is used to classify each item in a set of data into one of predefined set of classes or groups. Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of this model is to accurately predict the target class [2].

Prediction*:*

Predictive models are used to examine existing data and trends to better understand customers and products while also identifying potential future opportunities and risks Model continuous-valued functions, i.e., predicts unknown or missing values[3].

Association*:*

Association is a popular data mining method. It is used in finding relationships between variables. It is also known as market basket analysis. The relationships between co-occurring items are expressed as association rules. Association rules can be achieved using Apriori, FP growth, Ecat Algorithm [4].

5. MODEL EVALUATION

Various models can be applied on each hypothesis. The most relevant model with the optimized solution is considered. Different variables/attributes are considered for each hypothesis which play an important role in decision making.

Below are the different modelling techniques that are considered for each hypothesis.

Hypothesis 1:

Analyzing the chances of getting permanent visa from a specific region for various visa categories and based on experience.

This analysis can assist the government to frame and implement certain policies on getting the permanent visa over a region in the future.

The final number of records for creation of model were 8625 and it is divided into training data set (5175) and the validation dataset (3450)

Independent variables:

class_of_admission_fix,
country_of_citzenship_fix,employer_num_employees_fix,
employer_state_fix,job_info_experience_fix,
job_info_experience_num_months_fix

Dependent variables:

Case status (Variable: case_status_fix)

The dependent (output) variable can have 2 expected values i.e. Denied and Certified.

In the hypothesis data set preprocessing stage, the columns and its data not related to the hypothesis are removed. The dependent variable has the values Certified and Denied. The proportion of values in the final hypothesis subset in maintained as that of the original subset. The null values in the dataset are handling by Missing data handling method. Dummy variables have been created for the categorical data.

Rationale:

As hypothesis states, the analysis is done based on class of admission, country of citizenship and experience. Hence our main predictors are these tree variables (class_of_admission_fix, country_of_citzenship_fix, job_info_experience_fix)

Limitations of the modelling techniques:

Regression modelling techniques can be applied only on numerical and continuous data.

Association rules is not best as it does not always result in case status as output variable.

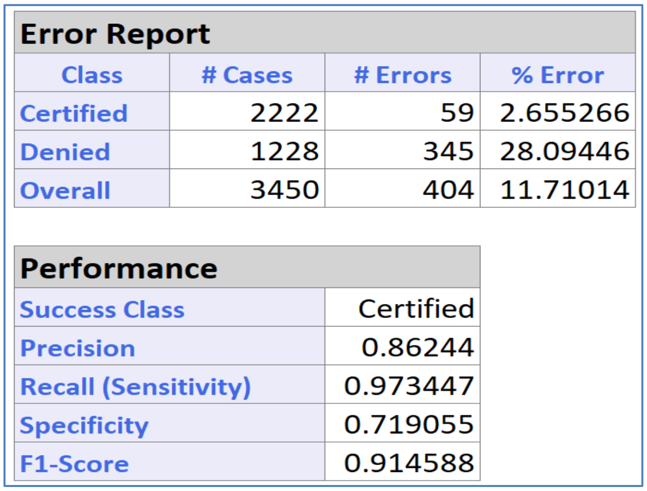
In Clustering, the order of data and initial seeds have strong impact on the final results.

Best Model:

Three modelling techniques were performed on the hypothesis. They are k-nearest neighbor algorithm (k-NN), Boosting algorithm and Single tree under classifications models. The best model among the three is Single tree classification tree.

The error percentage is the least for single tree. On comparing the F1 score, precision, makes the model using Single tree algorithm as the best for this hypothesis.

Statistics based on Single tree classification model:



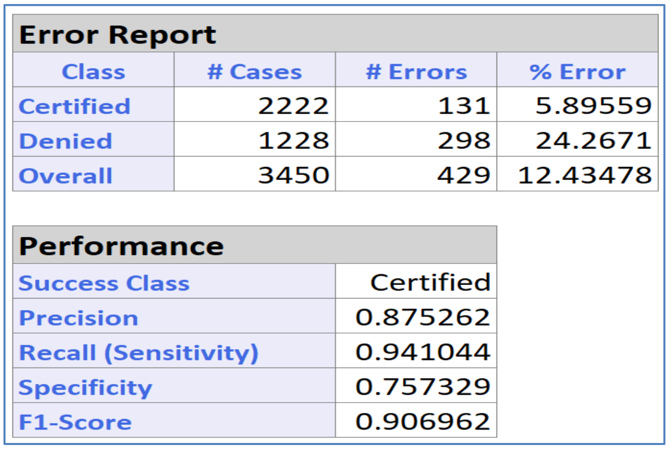
Error Report

Class	# Cases	# Errors	% Error
Certified	2222	59	2.655266
Denied	1228	345	28.09446
Overall	3450	404	11.71014

Performance

Success Class	Certified
Precision	0.86244
Recall (Sensitivity)	0.973447
Specificity	0.719055
F1-Score	0.914588

Statistics based on Boosting Classification model:



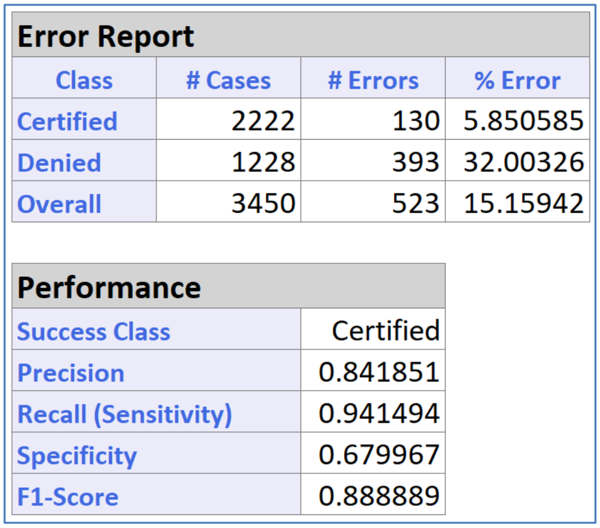
Error Report

Class	# Cases	# Errors	% Error
Certified	2222	131	5.89559
Denied	1228	298	24.2671
Overall	3450	429	12.43478

Performance

Success Class	Certified
Precision	0.875262
Recall (Sensitivity)	0.941044
Specificity	0.757329
F1-Score	0.906962

Statistics *based on k-NN model:*



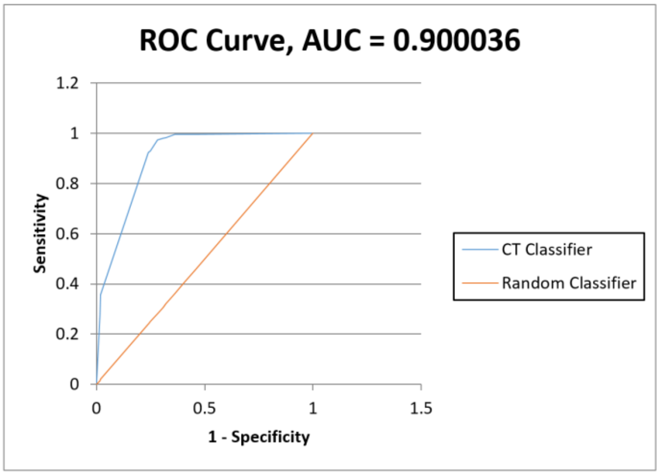
Error Report

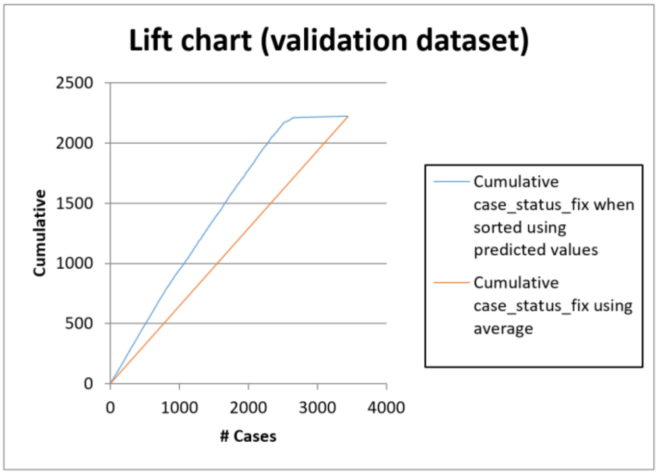
Class	# Cases	# Errors	% Error
Certified	2222	130	5.850585
Denied	1228	393	32.00326
Overall	3450	523	15.15942

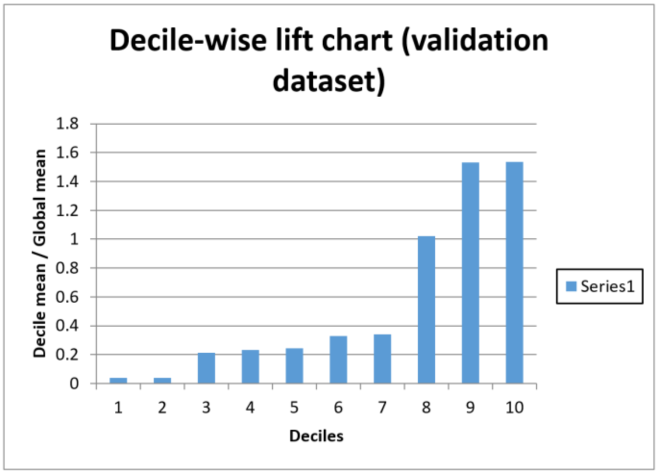
Performance

Success Class	Certified
Precision	0.841851
Recall (Sensitivity)	0.941494
Specificity	0.679967
F1-Score	0.888889

Charts for Validation data using Single tree (Best Model):







Hypothesis: 2

Predicting the duration for getting the USA Permanent Visa decision based on Class of Admission and country of citizenship for the top six countries and top ten class of admission

The final number of records for creation of model were 3638 and it is divided into training data set (2183) and the validation dataset (1455)

Independent variables:

case_received_date, decision_date, Country_of_citizenship, class_of_admission, case status

Dependent variables:

Duration

The Output variable i.e. Dependent variable has different numerical values based on class of admission and country of citizenship.

Dummy variables are created for all the categorical data and the correlation between the attributes is taken into consideration. There are no variables which are highly correlated to output variable.

Rationale:

For predicting the duration for decision, case_recieved date, decision date is considered. Duration variable gives a clear idea for an individual till what time he/she need to wait to get the final decision.

Limitations of the modelling techniques:

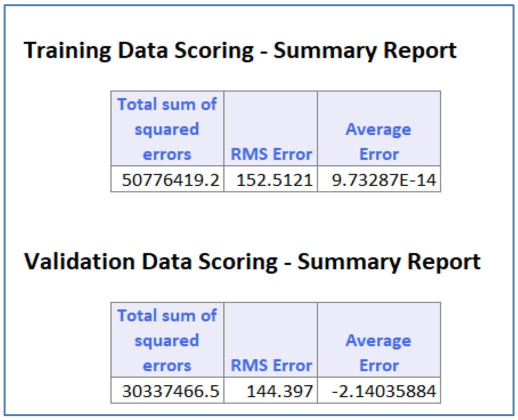
Association rules is not best for numerical predictor because output variable has more than two categories.

Classification works better when output variable has only two categories whereas duration variable has many categories which when implemented may give erroneous results.

Best model:

Regression is the best technique that can be applied to the given hypothesis because it examines the relation between dependent variable and independent variable, and after analysis it predicts the values for the new data based on the training data.

Summary report using Multiple linear regression:



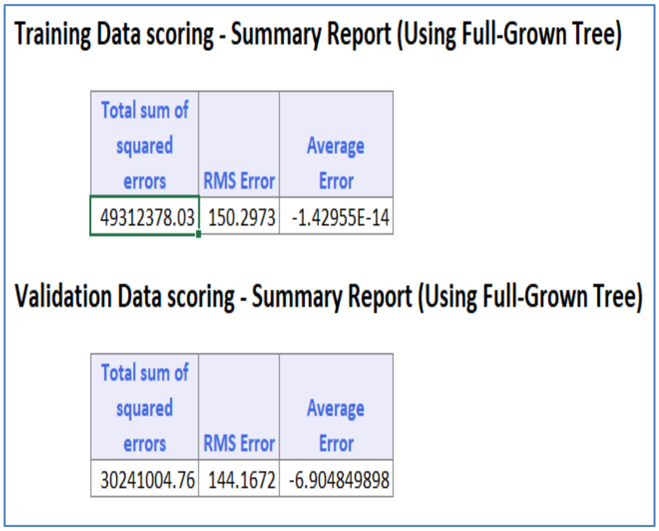
Training Data Scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
50776419.2	152.5121	9.73287E-14

Validation Data Scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
30337466.5	144.397	-2.14035884

Summary report using Regression tree:



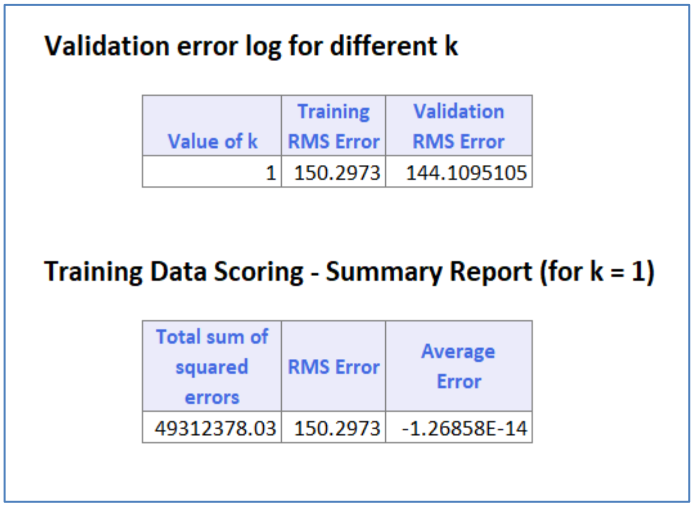
Training Data scoring - Summary Report (Using Full-Grown Tree)

Total sum of squared errors	RMS Error	Average Error
49312378.03	150.2973	-1.42955E-14

Validation Data scoring - Summary Report (Using Full-Grown Tree)

Total sum of squared errors	RMS Error	Average Error
30241004.76	144.1672	-6.904849898

Summary report using K- nearest neighbors:



Validation error log for different k

Value of k	Training RMS Error	Validation RMS Error
1	150.2973	144.1095105

Training Data Scoring - Summary Report (for k = 1)

Total sum of squared errors	RMS Error	Average Error
49312378.03	150.2973	-1.26858E-14

Hypothesis 3:

To find the chances of getting permanent residency based on the education level of the people filing the petition for topmost 5 countries.

The final number of records for creation of model were 10,000 and it is divided into training data set (6000) and the validation dataset (4000)

Independent variables:
foreign_worker_info_alt_edu_experience
country_of_citzenship, foreign_worker_info_education,
foreign_worker_info_req_experience,
job_info_alt_combo_ed_exp, job_info_experience,
job_info_experience_num_months

Dependent variables: Case status

The dependent (output) variable can have 2 expected values i.e. Denied and Certified

Dummy variables are created for all the categorical data and the correlation between the attributes is taken into consideration. The one variable out of 2 highly correlated variables is removed. Blank values for number of months of experience is replaced with 0 to work upon the models.

Rationale:

Since case status for the individual applying needs to be checked on basis of education, the main predictor variable is foreign_worker_info and the other independent variables might have an impact on the dependent variable as per the hypothesis.

Limitations of the modelling techniques:

Regression cannot be performed as the hypothesis does not have continuous variables associated with it.

In Clustering, the order of data and initial seeds have strong impact on the final results.

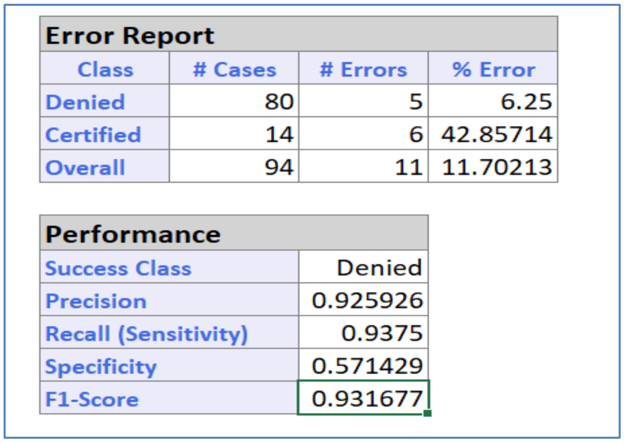
Association rules will not give "case status" as specific output. Therefore it cannot be considered for our hypothesis.

Best model:

The k-nearest neighbor algorithm (k-NN) under classification is the best modelling technique that can be applied to this hypothesis.

The error percentage is the least for k-NN model. On comparing the F1 score, precision, makes the model using k-NN algorithm as the best for this hypothesis [5].

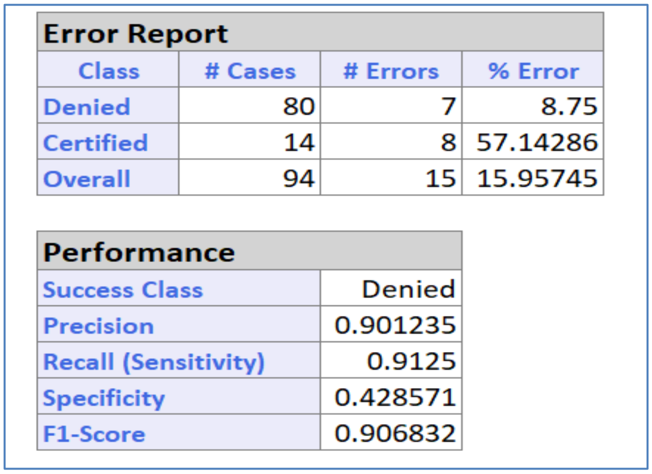
Model based on k-NN model:



Error Report			
Class	# Cases	# Errors	% Error
Denied	80	5	6.25
Certified	14	6	42.85714
Overall	94	11	11.70213

Performance	
Success Class	Denied
Precision	0.925926
Recall (Sensitivity)	0.9375
Specificity	0.571429
F1-Score	0.931677

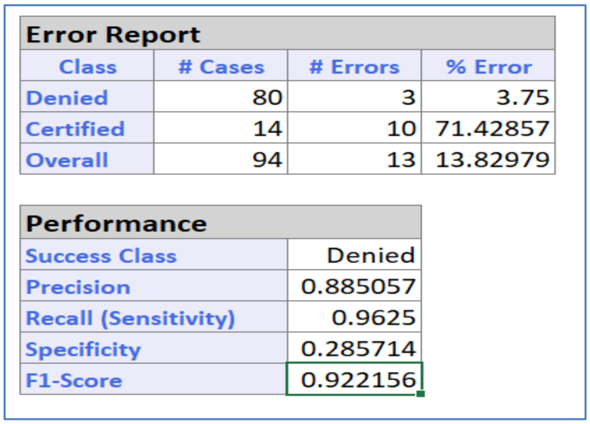
Model based on Naive Bayes model:



Error Report			
Class	# Cases	# Errors	% Error
Denied	80	7	8.75
Certified	14	8	57.14286
Overall	94	15	15.95745

Performance	
Success Class	Denied
Precision	0.901235
Recall (Sensitivity)	0.9125
Specificity	0.428571
F1-Score	0.906832

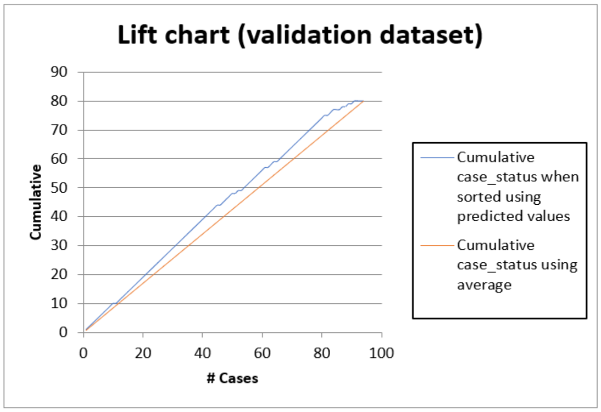
Model based on Classification Tree:

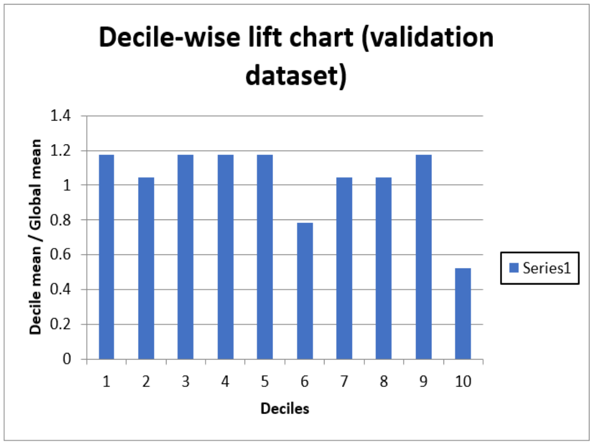


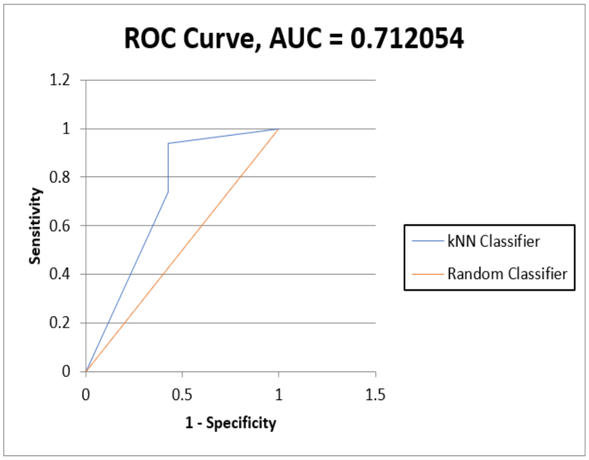
Error Report			
Class	# Cases	# Errors	% Error
Denied	80	3	3.75
Certified	14	10	71.42857
Overall	94	13	13.82979

Performance	
Success Class	Denied
Precision	0.885057
Recall (Sensitivity)	0.9625
Specificity	0.285714
F1-Score	0.922156

Charts for Validation data using k-NN (Best Model):







6. STRATEGIC RECOMMENDATIONS

One of the recommendation that can be made is on the variables type and the data collected. In the dataset, majority of the variables are categorical variables. There are very few variables which are numerical and hence Prediction and clustering models cannot be applied directly. For instance, Regression models cannot accept categorical data, only Classification models can be applied. If prediction or clustering models are to be applied, we need to pre-process the data like creating the dummies or binning the data in case of data that has range and that can be segregated. The dataset contains two date variables i.e. *case_start_date*, *case_end date* and in this case it is hard to know the number of days an individual needs to wait for the decision. So, duration variable is derived from start and end date which introduces the time constraint in data analysis. So, a dataset should have equal distribution of numerical as well as categorical variable this can help in predictive analysis as well as classification.

Another recommendation can be made on the dependent variable.

The initial dataset has the following dependent variable values: Certified, Certified and Expired, Withdraw and Denied. The proportion of the values were very unequal which leads the dataset to be biased. The subset of the data which was considered has proportionate values which makes the model appropriate to the data taken. This can be verified by the ROC curves of the models used.

[1]

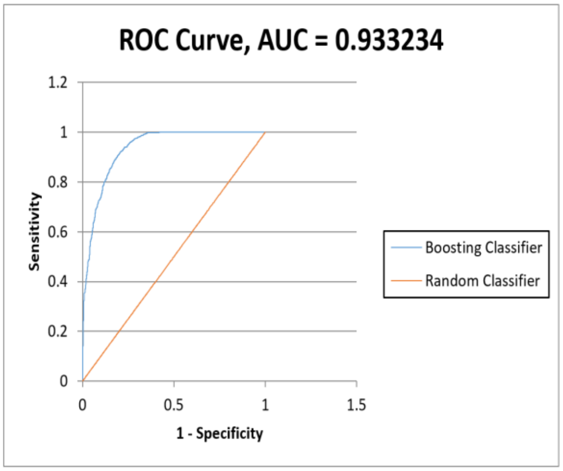


Fig 6: ROC Curve for Validation data (Boosting classification tree)

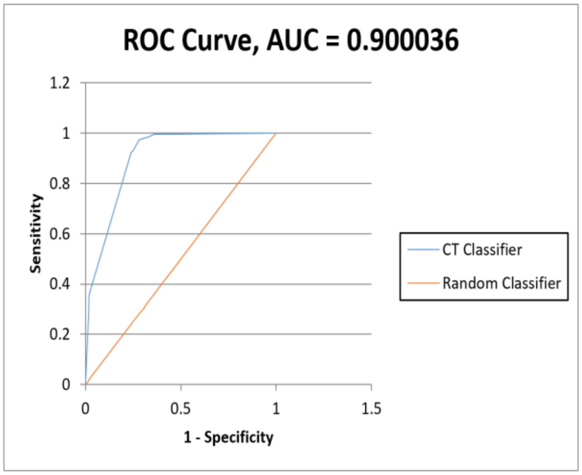


Fig 7: ROC Curve for Validation data (Single classification tree)

[1]

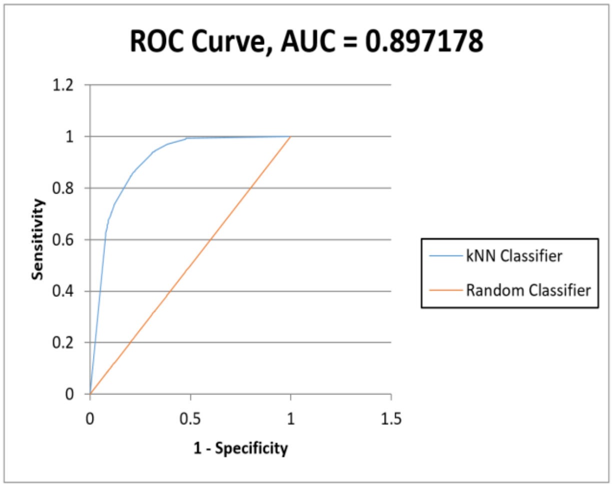


Fig 8: ROC Curve for Validation data (k-NN classification model)

Another recommendation can be made on the input data collected. For example, India is dominant in Permanent Visa applications when compared to other countries. Data isn’t in equal distributions. So, model can give an output which may be biased and over fitted for India country data. In order to have a proper model, the data should be equally distributed among all the countries available.

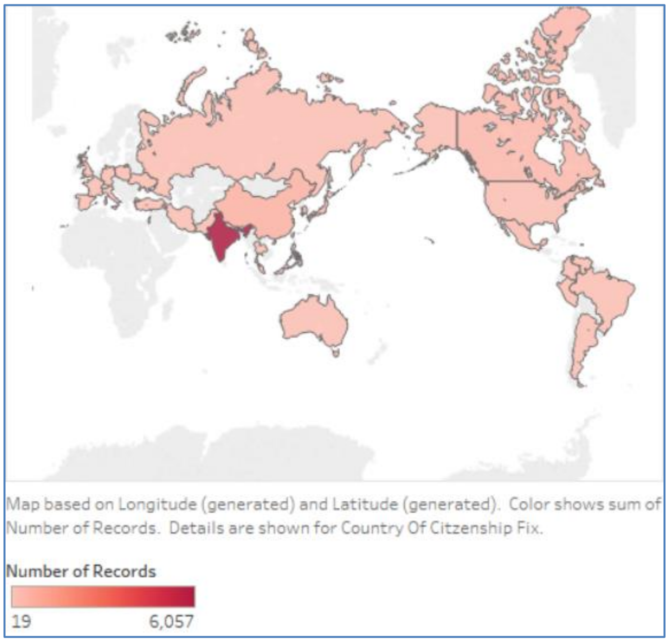

Map based on Longitude (generated) and Latitude (generated). Color shows sum of Number of Records. Details are shown for Country Of Citzenship Fix.

Fig 9: The above map depicts the distribution of visa applications from all the countries in the world

7. REFERENCES

[1] https://www.kaggle.com/jboysen/us-perm-visas

[2] G. Kesavaraj and S. Sukumaran, "A study on classification techniques in data mining," 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), Tiruchengode, 2013, pp. 1-7.doi: 10.1109/ICCCNT.2013.6726842

[3] H. Gulati, "Predictive analytics using data mining technique," 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, 2015, pp. 713-716.

[4] C. Song, "Research of association rule algorithm based on data mining," *2016 IEEE International Conference on Big Data Analysis (ICBDA)*, Hangzhou, 2016, pp. 1-4.doi: 10.1109/ICBDA.2016.7509789

[5] Galit Shmueli, Peter C. Bruce, Nitin R. Patel Data Mining for Business Analytics: Concepts, Techniques, and Applications in XLMiner®, Third Edition