Group Members: Kiana Alikhademi, Mahdi Kouretchian, Diandra Prioleau

Analysis of U.S. Permanent Visa Decisions

Preliminary Data Analysis

For extrapolatory data analysis, two methods - Multidimensional Scaling, MDS, and t-distributed Stochastic Neighbor Embedding (tSNE) - are applied to a subsample of the dataset including 5000 instances. Due to computational limitation, it is not practical to use tSNE on the whole dataset which has around 69K instances. As it is appeared in the following pictures, this dataset has strong non convex and nonlinear elements, and, consequently, linear discriminant methods like LDA and QDA would have poor accuracy relative to other methods and that is the reason they are not used in the data analysis.

Pre-processing Data

Visa application was an imbalanced datasets which needs proper data cleaning and even feature engineering steps before defining and running the models. First step, it was assessed that which variables would be necessary to answer our research questions. This lead to picking 17 variables out of all of them which are listed in following:

1. Case Status
2. Country of Citizenship
3. Decision year
4. Received year
5. Employer state
6. Employer name
7. Employer size
8. Where visa applications is processed
9. Employer year established
10. Applicant's birth country
11. Applicant's education
12. Applicant's major
13. Job major
14. Job_info_alt_combo_ed
15. Job_info_work_state
16. Wage_offer_from_9089
17. Wage_offer_unit

After picking variables, we needed to decide which variables were crucial for the analysis so that we could finalize which instances needed to be remove. The list of the variables in which empty values lead to excluding instances are summarized below:

- case status
- decision date

- foreign_worker_info_education
- Job_info_major
- Job_info_alt_combo_ed which stands for the required degree
- Wage_offer_from_9089, the final offer which has been confirmed by USCIS

After removing the empty instances for all the above mentioned variables, the number of instances was reduced from 374,362 to 69,566. Next, we looked carefully into the existing values for each of these variables to recognize the need for any feature engineering tasks. Wage offers have different units consists of: biweekly, hourly, weekly, monthly, yearly. We decided to first replace the missing values in wage unit and then compute the yearly wage offer so the scaling is the same between all instances and minimize impacts of different scaling on our findings. To find the unit we compared the wage offer with maximum wage offer in each wage unit category and initialized it based on that. Then, different computations were done for each category to compute the yearly wage offer and it replaced the existing wage offer. At this step, we no longer needed the variable wage_offer_unit and excluded from data.

Employer size was a number between 0 and 263550614. This big range makes it hard for us to analyze the relationship between the size of company and the decision was made. Therefore, we decided to do feature engineering and map the number of employees into following groups based on the ranges we found in [1]:

- Unknown(instances with blank values)
- Micro (instances with employees between 0 to 10)
- Small (instances with employees between 10 to 49)
- Medium (instances with employees between 49 to 249)
- Large (instances with employees greater than 249)

The new variable was used in the analysis, and actual number of employees variable was excluded from data. In the next step, we wanted to acquire more knowledge about the visa process for employees in different states. Therefore, we grouped the employees based on the type of visa and state which they filed the request into the following groups:

- California Center
- Vermont Center
- Both California and Vermont
- Unknown

The preliminary information for this mapping was gathered from [2]. However, we included both the newly created variable Centers and actual state and class of admission columns in the data.
As foreign workers apply from different countries all over the world, there are many differences in the majors, which makes it impossible for us to use the majors in further analysis. Therefore, we gathered some information about different career clusters and formed keywords for the main categories of majors. We have the following main categories for majors which have been used to map the existing majors to one of this group:

- Agriculture
- Architecture
- Business
- Education
- Science majors
- Computer science and mathematics
- Health sciences
- Engineering
- Hospitality
- Human services
- Law and public safety

- Manufacturing
- Transportation

Using this grouping helped us to transform our information to a type which is usable and helpful in our analysis. This same mapping was used for job majors as it was so diverse and consisted of more than 210 levels. Therefore, instead of the variables indicating *actual foreign worker education* and *job required education*, we have used the information we acquired after performing mappings on these variables.

Through analyzing our variables, it was found that employer and foreign worker state were given in both full name and abbreviation. Therefore, we processed the state and convert all the states into abbreviation for sake of our analysis.

Methodology

Classification

To classify the case status of applicants as *certified*, *certified-expired*, *denied*, or *withdrawn*, for U.S. permanent visas, the following classification methods were performed: Decision Tree, Random Forest, and Naive Bayes. First, a 70-30 train and split was conducted on the data remaining after pre-processing. The training set was used to train each classification model.

Based on the training data, the Decision Tree model had an accuracy of 79.45% with a cp value of 0.005303992. The Naive Bayes model had an accuracy of 53.37% when the parameter for the kernel was true.

For Random Forest, the model was created using an ensemble of 1000 trees, which corresponds to 1000 bootstrap samples, and the parameter *importance* was set to true to compute a measure of importance for the variables using the out-of-bag samples. Training the model resulted in an out-of-bag error of 17.16% with majority of the error being contributed by the classes *denied* and *withdrawn*, which had class-specified errors of 71.94% and 96.88%, respectively. The variable importance plot for the Random Forest model, shown in Figure , indicates that the case_received_date was of the variable of most importance for based on both accuracy and the Gini Index. The variable wage_offer_from_9089 was of next importance based on accuracy, and the variables decision_date and wage_offer_from_9089 were next in importance based on the Gini Index. Based on this plot, it is shown that the date when an individual’s application is received and the salary offered to the applicant are major indicators of the case status of an application.

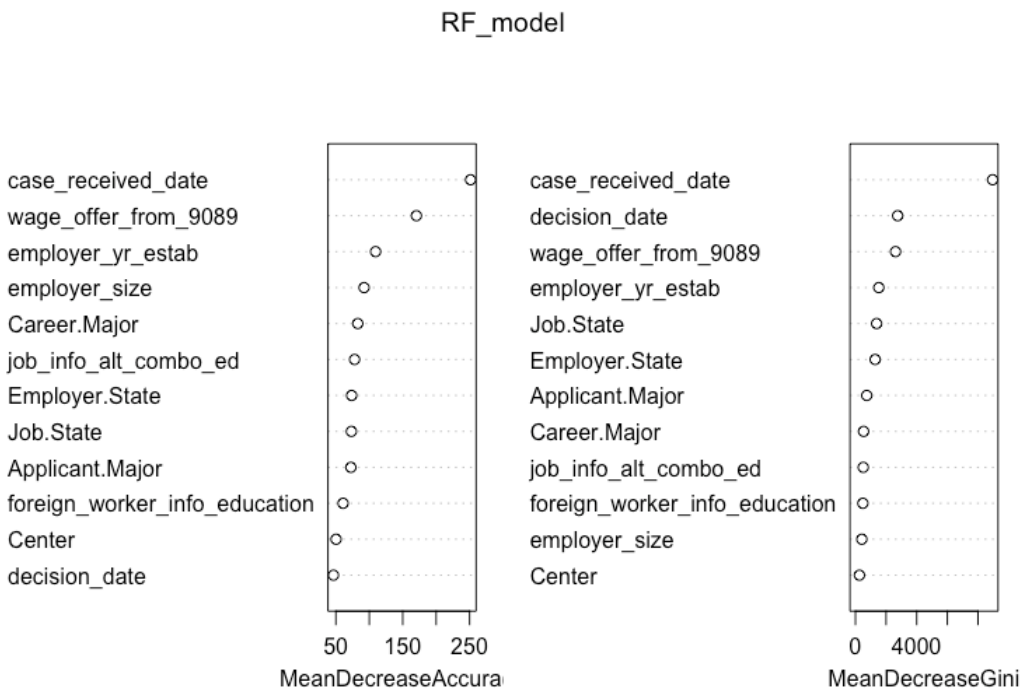


Figure 1 Variable Importance Plot of Random Forest

Clustering:

To address the second problem statement, clustering was conducted on a subsample of the data remaining after pre-processing. 10 percent of the data was subsampled to cluster. The following clustering techniques were employed: the hierarchical method using complete linkage, K-Means partitioning around medoids (PAM), and the divisive method Diana. For complete linkage and Diana, the dissimilarity matrix was used to cluster the data. The daisy function was used to compute the dissimilarity on the subsampled data using the Gower metric, which is able to combined variables of mixed type into a single dissimilarity matrix.

Results

Classification:

As mentioned in the methodology section, we have implemented Random Forest, Decision Tree and Naive Bayes. The results of each of these methods are given in the following subsections. The metrics, which are accuracy, precision and recall, used in the analysis are defined below.

Precision is a percentage of information which are relevant which is computed by the following equation:

Precision = True Positive / (True Positive + False Positive)

On the other hand, recall, which refers to the percentage of total relevant results correctly classified by your algorithm, is computed using the following equation:

Recall = True Positive /(False Negative + True Positive)

Group Members: Kiana Alikhademi, Mahdi Kouretchian, Diandra Prioleau

F-1, which is the harmonic mean of precision and recall, could give some metric to analyze them together and is computed as follow:

$$F\text{-}1 = 2*((precision*recall)/ (precision + recall))$$

Decision Tree:

Running test data through Decision Tree led to accuracy of 79.32. The confusion matrix of these results are shown in Table 1. The precision, recall and F-1 score for each method is shown in Table 2. Due to computational limitation, a subsample was taken from our dataset in order to run the decision tree and since the total number of denied or withdrawn cases are relatively very small in compare to other cases we see that in our subsample the total number of denied and withdrawn cases are zero.

	Certified	Certified-Expired	Denied	Withdrawn
Certified	7692	0	128	381
Certified-Expired	2988	8859	526	292
Denied	0	0	0	0
Withdrawn	0	0	0	0

	Certified	Certified-Expired	Denied	Withdrawn
Precision	93.79	69.98	NA	NA
Recall	72.02	1	0	0
F1	81.4	82.31	NA	NA

Random Forest:

Running test data through Random Forest led to accuracy of 82.5. The confusion matrix of these results is shown in Table 3. The precision, recall and F-1 score for each method is shown in Table 4.

	Certified	Certified-Expired	Denied	Withdrawn
Certified	8849	669	197	393
Certified-Expired	1766	8155	263	241
Denied	39	32	192	9
Withdrawn	26	3	2	30

Group Members: Kiana Alikhademi, Mahdi Kouretchian, Diandra Prioleau

	Certified	Certified-Expired	Denied	Withdrawn
Precision	87.54	78.22	70.58	49.18
Recall	82.85	92.05	29.35	4.45
F1	85.13	84.57	41.46	8.17

Naive Bayes:
Running test data through Naive Bayes led to accuracy of 59.41. The confusion matrix of these results are shown in Table 5. The precision, recall and F-1 score for each method is shown in Table 6.

	Certified	Certified-Expired	Denied	Withdrawn
Certified	7439	3901	278	318
Certified-Expired	3241	4958	376	355
Denied	0	0	0	0
Withdrawn	0	0	0	0

	Certified	Certified-Expired	Denied	Withdrawn
Precision	62.32	55.52	NA	NA
Recall	69.65	55.96	0	0
F1	65.78	55.74	NA	NA

Clustering:

In Figure 2, the silhouette plot is shown for the hierarchical method using complete linkage. It was decided to use two clusters as it resulted in the highest silhouette width, which was 0.23. One of the clusters has a silhouette width of 0.32, which somewhat effective but definitely needs improvement, and the other is close to 0 indicating that the observations in the cluster lie between 2 clusters.

Group Members: Kiana Alikhademi, Mahdi Kouretchian, Diandra Prioleau

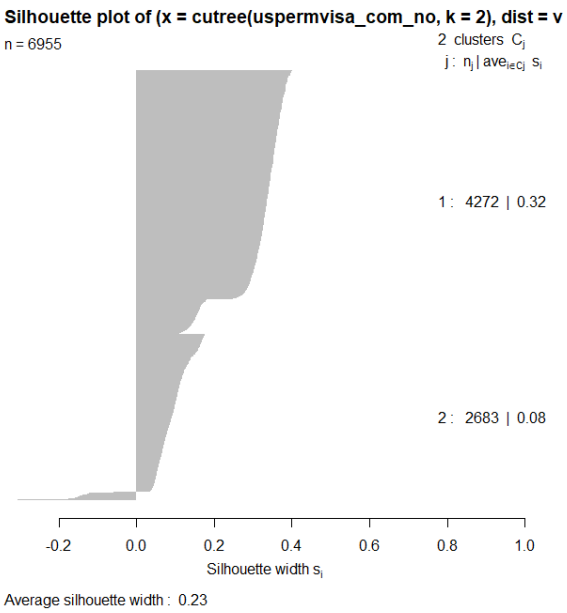


Figure 2 Silhouette Plot of Complete Linkage

Figure 3 shows the silhouette plot after clustering using the divisive method Diana. The best number of clusters was 2 using the Manhattan metric on the dissimilarity matrix. This resulted in an average silhouette width of 0.24, which is low and indicates that divisive method was ineffective in clustering the data. One of the clusters have a silhouette width of 0.31, which is still relatively low, and the other have a silhouette of 0.1, which is close to 0 and indicates that the instances in the cluster lies intermediate between 2 clusters.

Figure 4 shows the frequency based on the case status for each cluster. It is shown that cluster 1 has a higher number of certified applications in comparison to the other clusters.

Group Members: Kiana Alikhademi, Mahdi Kouretchian, Diandra Prioleau

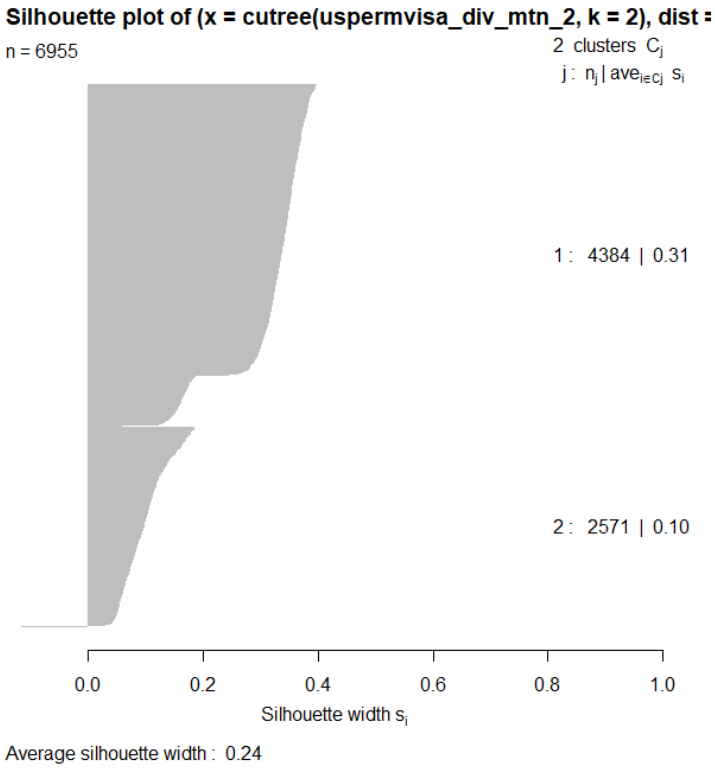


Figure 3 Silhouette Plot of Divisive Method Diana

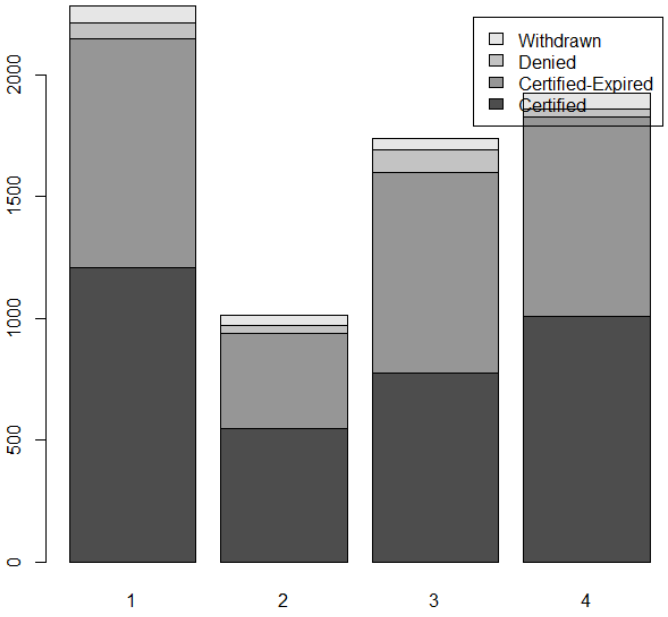
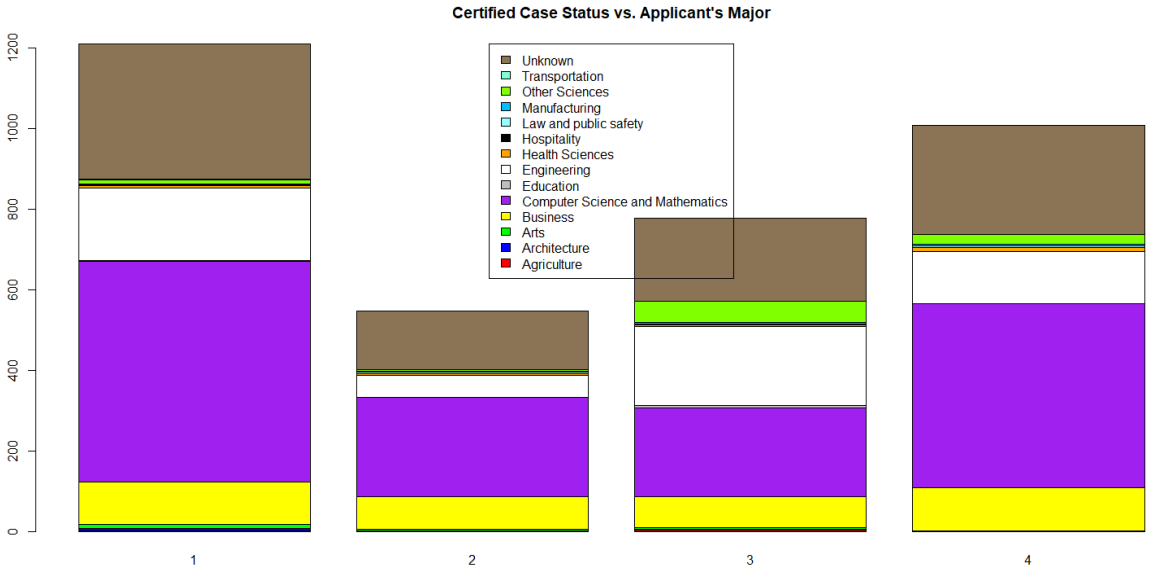
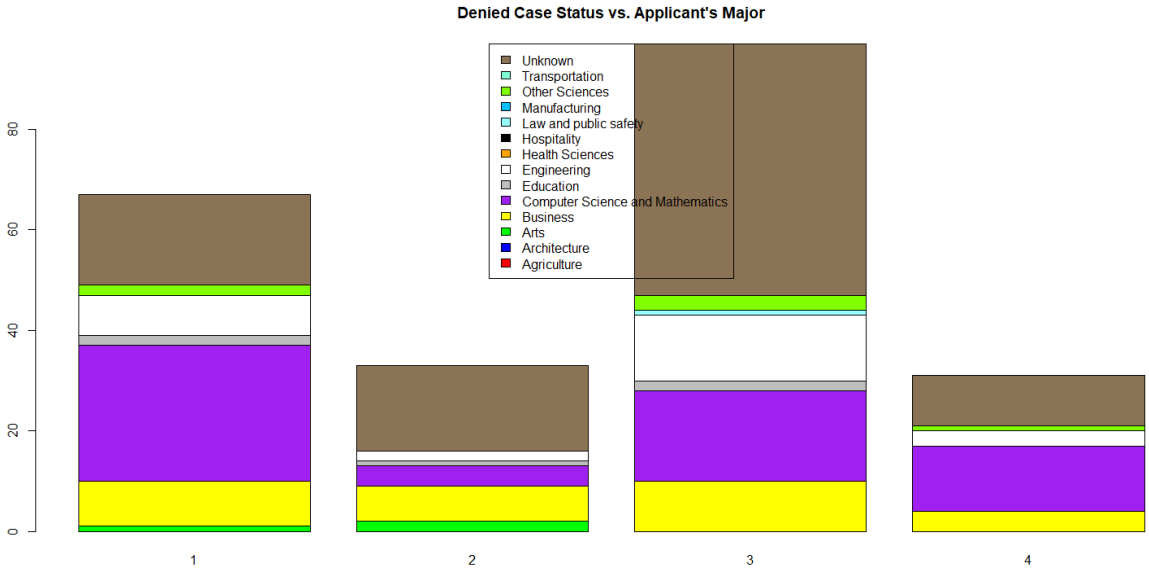


Figure 4 Case Status across Cluster Groups

Group Members: Kiana Alikhademi, Mahdi Kouretchian, Diandra Prioleau





Conclusion

This research analyzed the current patterns in U.S. working visa decisions by using different classification and clustering methods such as Decision Tree, Random Forest, Naive Bayes, PAM, complete linkage and the divisive method Diana. Although Random Forest acquired an accuracy around 82.5% with some potential misclassifications, it still may not be a viable option for predicting visa decisions, which affect the livelihood of people. One of the counter intuitive results of this study is that the qualification of each applicant like level of educational degree does not play an important role in making visa decision. Based on the variable of importance graph for random forest the variables case_received_date, decision_date, and wage offer are

important factors for approval of the worker visa, which is interesting considering that these variables are not directly related to qualities of the applicants. For clustering section, three methods namely hierarchical complete linkage, divisive method Diana and PAM are applied to the given data set. Low silhouette scores of clustering methods show that there is no strong structure between the variables which could help us in making meaningful clusters. Consequently, clustering methods perform a poor job in defining the underlying pattern of different classes. Lastly, one needs to pay attention to the correlation between the major of each applicant and the likelihood of the getting the visa. Based on an exploratory data analysis, as depicted in previous graphs, for an applicant with a major degree in mathematics, computer science and engineering, there is a higher chance of getting a working visa while this chance is nominal for an applicant with a degree in education.

Group Members: Kiana Alikhademi, Mahdi Kouretchian, Diandra Prioleau

References:

1. “Enterprises by business size” by OECD.org https://data.oecd.org/entrepreneur/enterprises-by-business-size.html
2. “Direct Filing Addresses for Form 1-129, Petition for a Nonimmigrant Worker” by U.S. Citizenship and Immigration Services https://www.uscis.gov/i-129-addresses