Understanding factors that influence L1-visa outcomes in US

By Nihar Dalmia, Meghana Murthy and Nianthrini Vivekanandan

Link to online course gallery:
https://www.ischool.berkeley.edu/projects/2017/understanding-factors-influence-l1-work-visa-us

1. PROBLEM STATEMENT

With increasing number of visa applications to US, USCIS has intensified their processes to approve candidates work visas. It is important for an applicant and their employer to know if the visa application would be approved or not. On one hand, applicants make a life changing decision of leaving their home country and starting a new life in the US. On the other, employers spend a lot of money looking to recruit the right candidate and applying for their visa. Thus, knowing about the chances of visa approval and the levers that can be adjusted (to an impact) will help them put their best foot forward. In this paper, we choose to restrict our scope of analysis to the L1 visa category.

When an employer makes a decision to sponsor L1 visa for their employee, they are impressed by the skills of the employee and want to make use of the employee's talents in their US office. From USCIS perspective these visas have to be granted to an applicant with high qualifications without affecting the employment opportunity of a US citizen with similar qualifications. So, these applications are thoroughly vetted to make sure that there are no qualified US employees to fill that job and also to make sure that the foreign applicants are paid wages as per the requirements for that position. Given that L1 visa applicants have higher chance to qualify for an EB-1 category Green card application, they also have to perform careful background checks to verify the credentials of the applicant and any associated risks pertaining to their country of origin. Our goal is to identify the factors that have an influence in the application decision as it would help the applicants and employers to identify drawbacks and make better informed decisions.

2. DATASET AND FEATURE ENGINEERING

We procured the dataset from US Department of Labor. This dataset (Link) has about 374k instances with 154 features and we are considering only L1 visa type which has about 19k instances. One row in this dataset represents data about one visa applicant with information about the case status, employer name, wage, type of visa, education,

experience, date of application, country of citizenship, type of job, job title, lawyer information, previous application history etc (Metadata). We appended information about religious diversity (Link), rise of GDP (Link) of the country of citizenship of the applicant, education rate (Link) and unemployment information for the employer state (Link). These will capture information about the home country which, we hypothesize, could have a role to play in visa decisions. Further, we appended data regarding unemployment and education rates in the employer's state as it could possibly affect a candidate's visa approval outcome.

This dataset required extensive cleaning due to the presence of a number of empty columns, columns with more than 50-80% of missing data, presence of a number of categorical columns with 1000's of categories (leading to sparse data), duplicate columns and same data stored in multiple columns. The columns with more than 50% missing values were dropped as there was high uncertainty in filling these values with techniques like mean/mode. Dates were separated to get information about year, month and day. The months were grouped into half-yearly and quarterly seasons. Certain columns which had very few missing values still had data in multiple formats. For example, the employer state column had states written both in abbreviated form and in full forms. We designed a mapping to change the abbreviations to full forms before appending state level data pertaining to unemployment and education rates. There were also instances of data which were split between multiple columns. For example, applicant's country of citizenship was split between two columns which we then had to put together. Religion and GDP data was then appended based on the citizenship data. Agency information was converted into a binary with presence (or absence) of information indicating the use of agency.

When dealing with missing values, for continuous variables, we chose to group by fields that are correlated with those variables, and then take the mean of the grouped values to fill the missing variable. We repeated the same process for categorical variables as well, however, we chose to group by and insert the mode into the missing values. Continuous variables were tested for normality using Shapiro test and for all the selected continuous variables, the null that the distribution is normal was rejected, so these values were standardized to z-scores. Categorical variables like job title and major of the applicant were very sparse with 1000's of possible values, to handle this, major and job title were grouped to more abstract groups of about 25-30 categories. After reducing the number of possible values of these categorical variables, they were one-hot encoded. We also examined each of the features carefully with the value they add and decided to drop a number of them as they were not contributing to the final outcome.

To summarize, we started with 19938 rows and 154 columns. After appending new data, and cleaning up the existing data, we ended up with 19301 rows and 38 columns.

3. PREVIOUS WORK ON THIS DATASET

Extensive amount of work has been done on predicting H1-B visa decisions on a previous limited dataset including exploratory data analysis and predictions based on a number of models. The major features influencing H1-B visa outcomes were found to be the wage and job title. Exploratory data analysis was done for H1-B dataset to determine the top companies, job title, top cities for the applicants, how salaries of these applicants differ from that of a US employee for the same job, how the decisions vary by year, how the salaries differ among the various cities, states, job titles etc.,[1,2,3]

Limited work has been done on our current dataset with exploratory analysis like histogram for top 10 employers, country of citizenship, job title etc., for all visa types and a model to predict the decision based on 10 features like decision year, pay, employer name, employer state, wage and source from 9089, pay unit, standard occupation classification code and title using Xboost model[4]. This work, however, has been done on the entire dataset, and not particularly focussed on L1 visas.

We decided on working with L1 visa application outcomes as the dynamics for selecting L1 visa are different than those for H1-B visas. L1 is restricted to managers, executives or personnel belonging to specialized knowledge category who are looking to move from their home country to USA to work for a specific employer. Thus, we found the L1 visa space interesting and research worthy.

4. METHODS USED

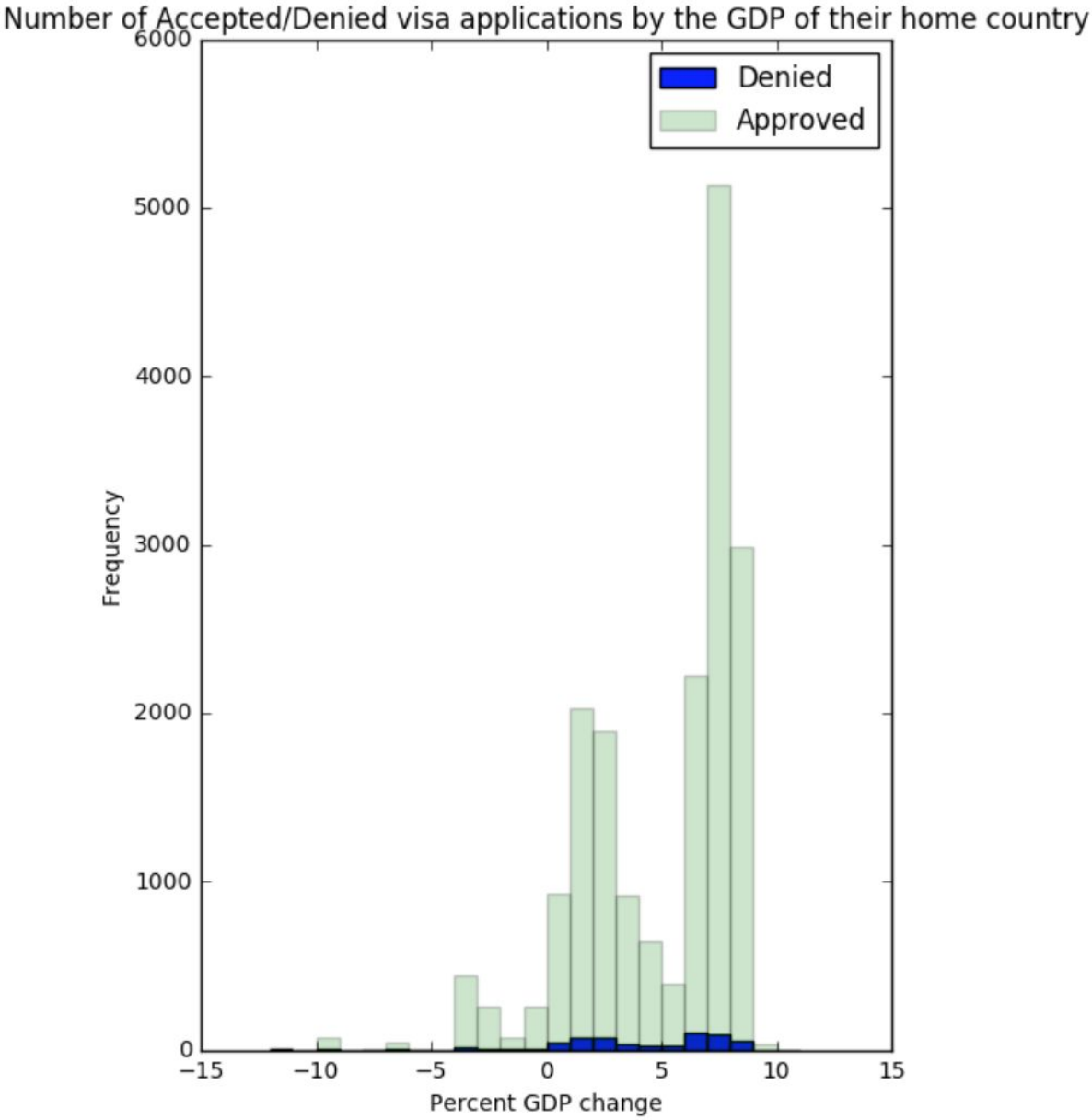
4.1 Statistical Significance Tests:

We begin our analysis by running significance tests between select few variables (both numerical and categorical) to determine any significant differences in visa outcomes. Following are the tests that we run along with their results:

1. Does the applicant's country's GDP affect Visa approval rate?

Null Hypothesis: The mean of GDP increase of the home country of the applicant who is denied the visa is the same as that for the applicant whose visa is approved.

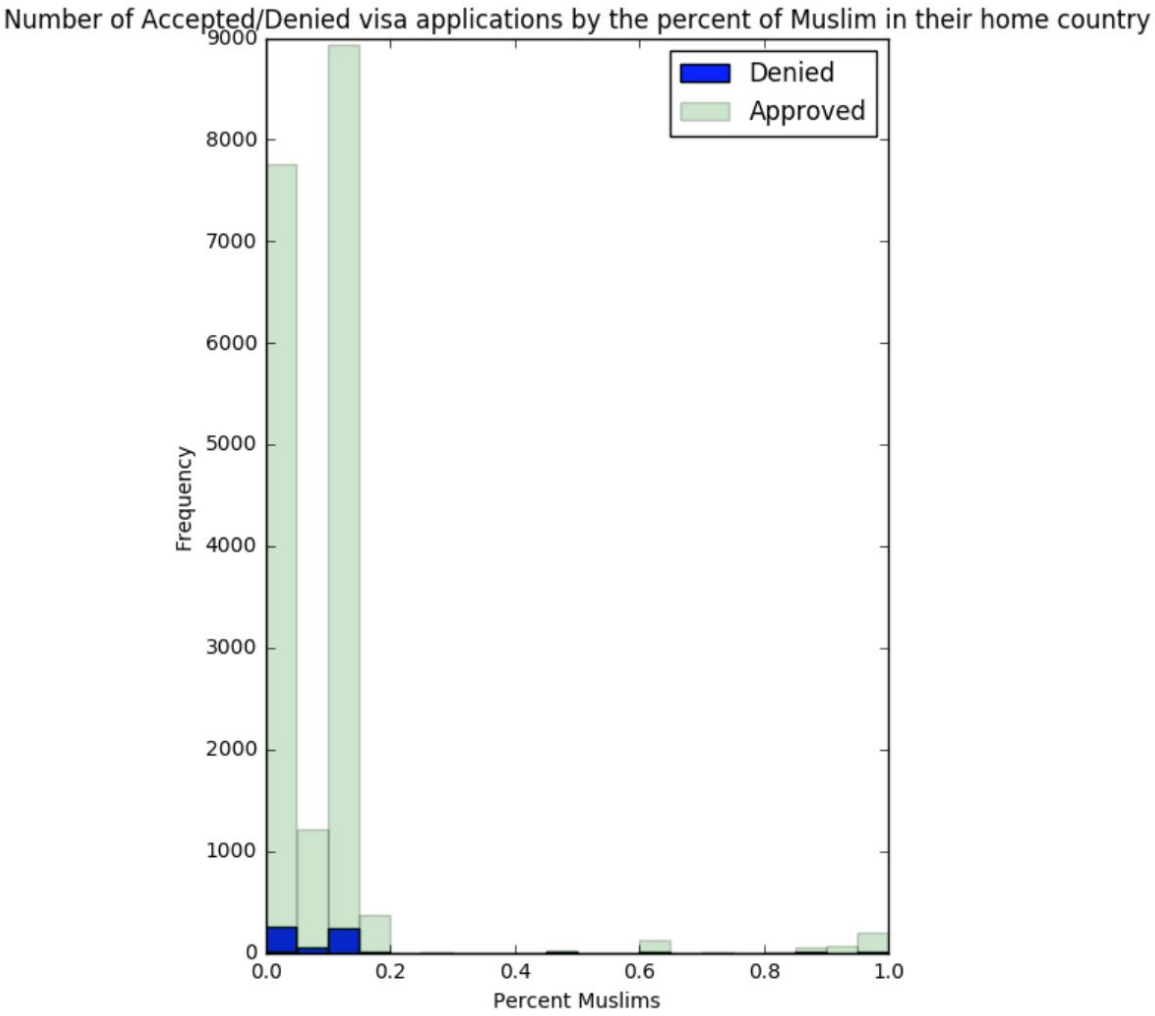
T-Test: Statistic: -5.283
P-value = 1.28e-07
Result: **Significant**
Mean of GDP increase for Denied: 4.135
Mean of GDP increase for Approved: 4.922
Conclusion: We reject the null hypothesis and claim that there is a statistically significant difference in the mean GDP increase of the home countries of those who were approved and that of those who were denied.



2. Does the percentage of Muslims in applicant's home country affect his/her visa approval chances?

Null Hypothesis: The mean of percentage of Muslims in the home country of the applicant who is denied the visa is the same as that for the applicant whose visa is accepted.

T-Test: Statistic: -2.325
P-value = 0.0201
Result: **Significant**
Mean of % Muslims for Denied: 0.0896
Mean of % Muslims for Approved: 0.1033
Conclusion: We reject the null hypothesis and claim that there is a statistically significant difference in the mean percentage of Muslims in the countries of those who were approved and that of those who were denied.

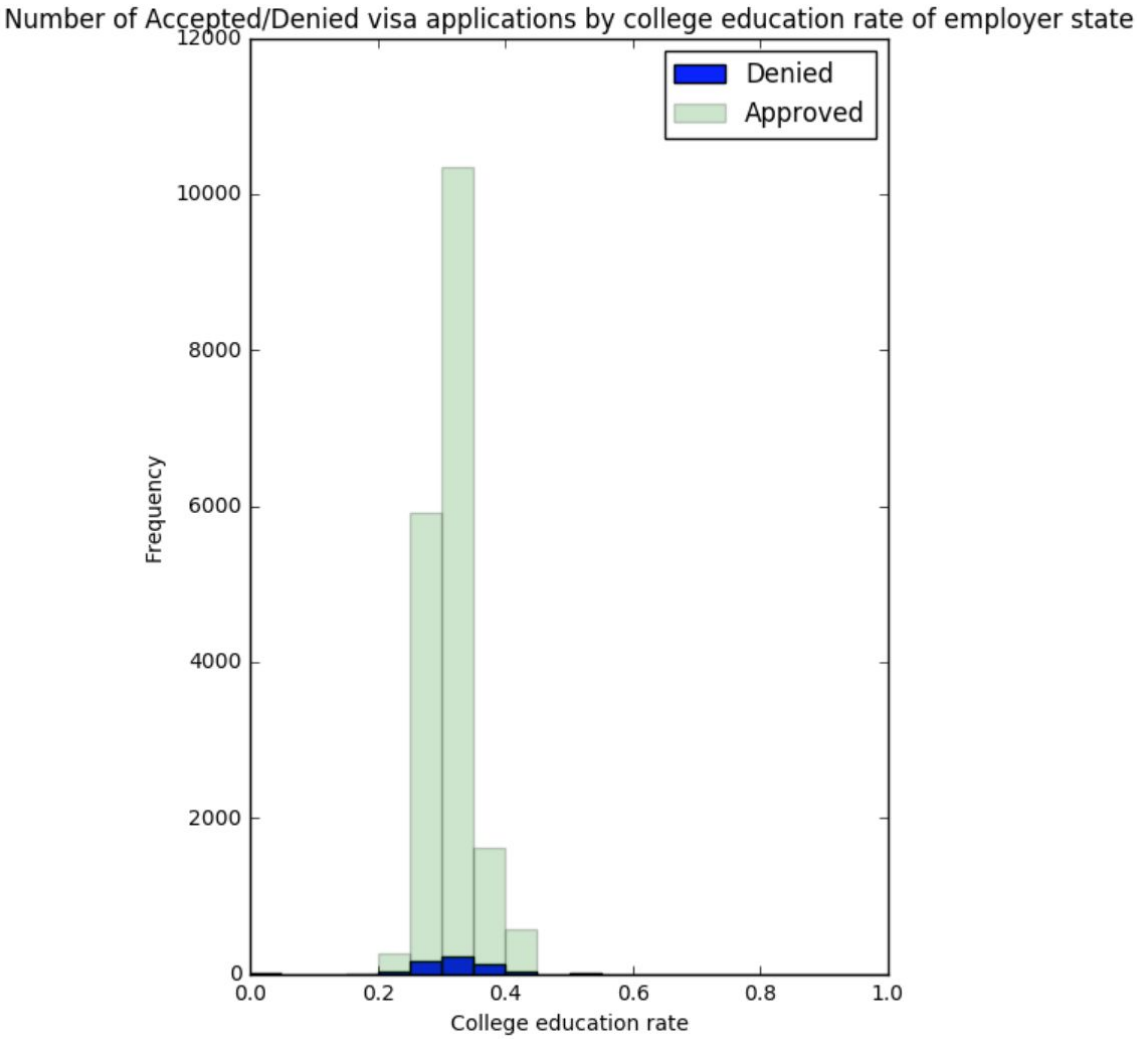


3. Does the college education rate in the employer State affect a candidate's visa approval chances?

Null Hypothesis: The mean of job state college education rate is the same for candidates whose visa has been approved or denied.

T-Test: Statistic: 4.82
P-value = 1.439e-06
Result: **Significant**
Mean of education rate for Denied: 0.32027
Mean of education rate for Approved: 0.3130

Conclusion: We reject the null hypothesis and claim that there is a statistically significant difference in the education rate in the employer state of those who were approved and those who were denied.

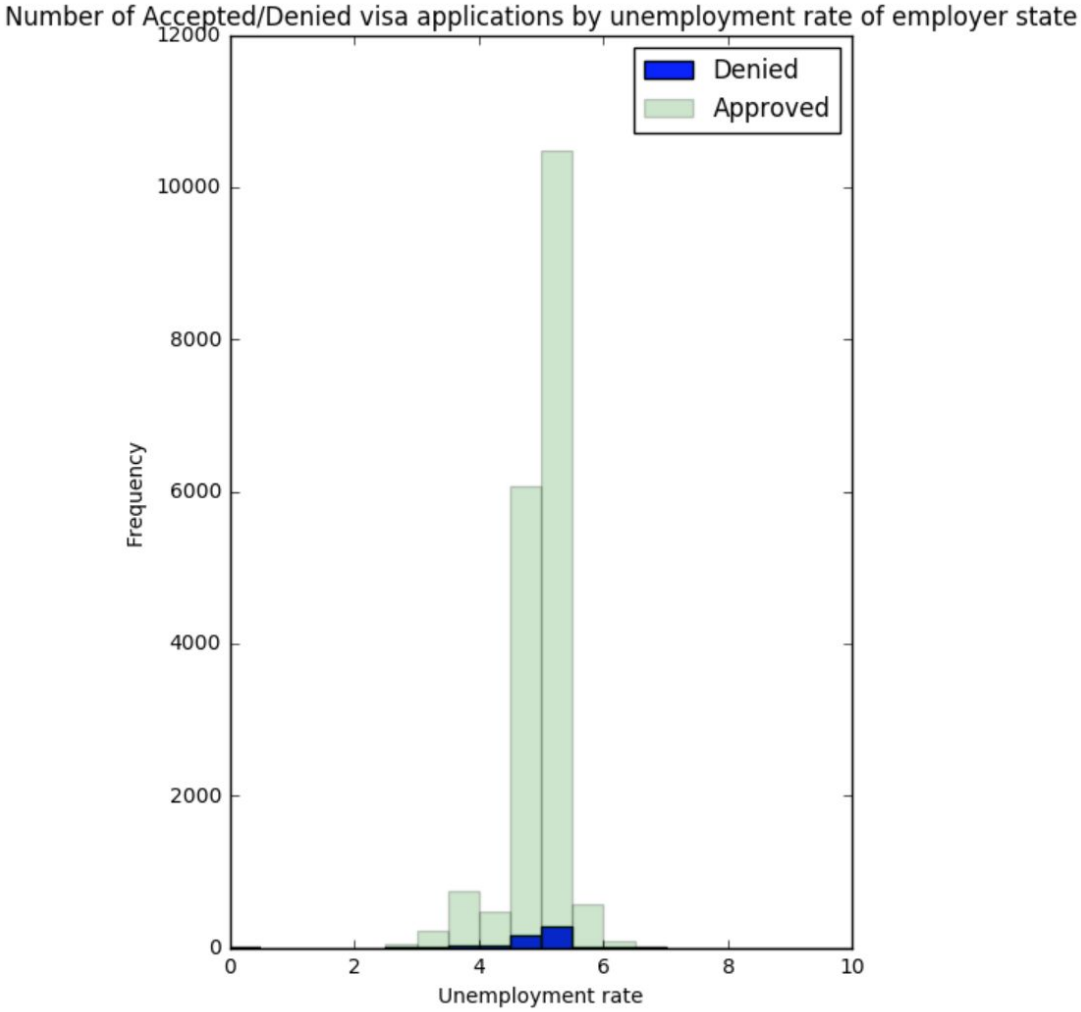


4. Does unemployment rate in employer State affect a candidate's visa approval chances?

Null Hypothesis: The mean of state unemployment rate is the same for candidates whose visa has been approved or denied.

T-Test: Statistic: -4.728
P-value = 2.286e-06
Result: **Significant**
Mean of unemployment rate for Denied: 4.922
Mean of unemployment rate for Approved: 5.029

Conclusion: We reject the null hypothesis and claim that there is a statistically significant difference in the unemployment rate in the employer state of those who were approved and that of those who were denied.

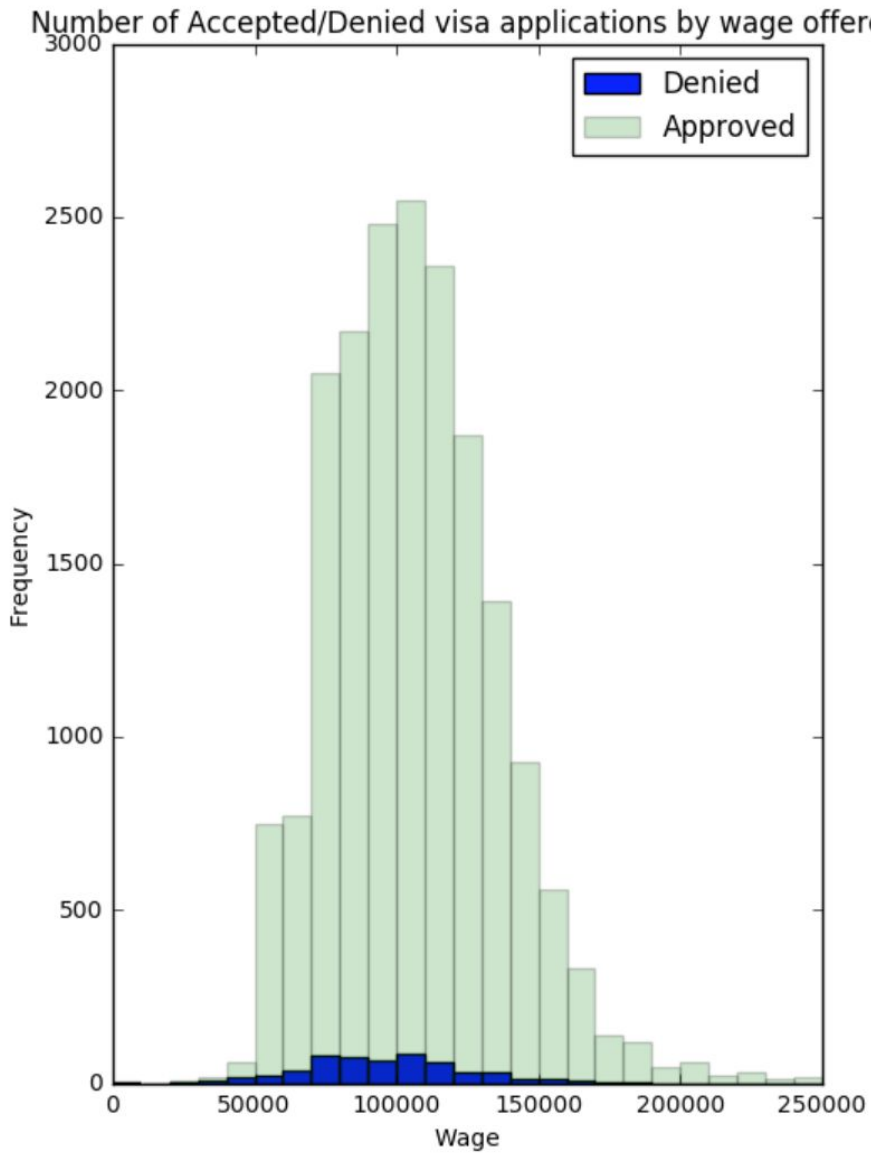


5. Does wage offered to candidate affect visa approval rate?

Null Hypothesis: The mean of wage offered to candidates whose visa was approved is the same as that for candidates whose visa was denied

T-Test: Statistic: 1.561
P-value = 0.212
Result: **Insignificant**
Mean of wage offered for Denied: 109321.077
Mean of wage offered for Approved: 106068.458

Conclusion: We do not have sufficient evidence to reject the null hypothesis



6. Does the education level of the applicant affect visa approval rate?

Null Hypothesis: The outcome of the visa is independent of the education level of the applicant

Chi Square Test: Statistic: 370.141
P-value = 5.99e-76
Result: **Significant**

Conclusion: The outcome of the visa is dependent on the education level of the applicant.

education_level	0	1	2	3	4	5	6	7
status								
Approved	105	8123	273	90	4491	409	474	4777
Denied	6	113	1	16	68	37	23	295

Legend:
0: Associate's
1: Bachelor's
2: Doctorate
3: High School
4: Master's
5: None
6: Other
7: unknown

7. Does the years of experience of the applicant affect visa approval rate?

Null Hypothesis: The outcome of the visa is independent of the years of experience of the applicant.

Chi Square Test: Statistic: 370.141
P-value = 5.99e-76
Result: **Significant**

Conclusion: The outcome of the visa is dependent on the experience level of the applicant.

experience_level	0-2years	2-3years	3-5years	5years_above	unknown
status					
Approved	2541	3849	2005	5570	4777
Denied	41	90	30	103	295

4.2 Machine Learning:

We use machine learning models to determine the importance of various features and also to predict the binary class outcome of visa approval or denial. Since our focus is on understanding what features contribute most to the outcome of visa application process, our choice of model should allow us to explore feature importances along with their interpretations. Due to this reason, we primarily conduct our analysis using Logistic Regression and Decision Tree Classifier models.

Further, from our initial assessment, it was evident that the output classes were extremely imbalanced. The majority class ("Approved") was **97.3%** and hence the classes had to be balanced before running any machine learning models on it. To balance the classes we used two approaches - Oversampling with SMOTE and regular downsampling.

Prior to exploring the sampling methods or running the models, the dataset was split into 80% training and 20% test. The 80% training would be used to train to find the best model, and then, the accuracy will be checked on the 20% test set.

Sampling Methods

a. SMOTE (Synthetic Minority Over-sampling Technique):

SMOTE is an over-sampling approach in which the minority class is over-sampled by creating "synthetic" examples rather than by over-sampling with replacement. The minority class is over-sampled by taking each minority class sample and introducing synthetic examples. Depending upon the amount of over-sampling required, neighbors from the k nearest neighbors are randomly chosen [5].

While training our models, we applied SMOTE to the training data within each fold, without impacting the holdout fold so that we only train on the synthetic data and test on the actual data.

b. Downsampling

Downsampling is implemented by randomly removing samples from the dataset which belong to the majority class, in order to have a more even distribution of the output classes. In our case, we downsample the visa “Approved” class to make the number of samples comparable to the visa “Denied” class.

While training our models, we downsampled the training data within each fold, without impacting the holdout fold. A downside of this approach was that by removing samples, the dataset may not remain representative for some features and thus, it could impact the overall learning of the model.

Models:

a. Decision Trees:

Once we had the data cleaned, sampled and ready, we moved on to the Decision Tree classifier. In order to get the best prediction model, we used our sampled data to tune the hyperparameters - max_depth and min_samples_leaf. The range we chose for max_depth was (1,15) and for min_samples_leaf was (1,15). The hyperparameter tuning was done using stratified 10-fold cross-validation. Following are the results, and important features for different sampling techniques.

SMOTE sampling: The best values obtained were **max_depth = 14** and **min_samples_leaf = 14**. These provided us the best accuracy using decision trees with a training accuracy of **94.01%** and testing accuracy of **89.04%**. For the same model, we also evaluated using the following scores:
Precision: 14.14%
Recall: 55.12%
F1 score: 22.51%

Since the tree plot with 14 levels is really complicated to analyze, we are showing a decision tree with max_depth of 4 in Appendix 3.

Following are the most and least important features of the model, along with its impact on the class outcome.

Decision Trees			
Top 10		Bottom 10	
Features	Impact on class outcome	Features	Impact on class outcome
education_level_Bachelor's	Approved	clean_state_WEST VIRGINIA	N/A
education_level_Master's	Approved	clean_state_WISCONSIN	N/A
clean_state_WASHINGTON	Approved	clean_state_WYOMING	N/A
GDP	Approved	case_month_9	N/A
Percent_Folk_Religions	Approved	experience_level_2-3years	N/A
education_level_None	Denied	experience_level_unknown	N/A
education_rate	Denied	education_level_Associate's	N/A
pw_amount_9089	Denied	education_level_High School	N/A
Percent_Buddhist	Approved	education_level_None	N/A
education_level_Doctorate	Approved	education_level_unknown	N/A

Downsampling: The best values obtained were **max_depth = 14** and **min_samples_leaf = 14**. These provided us the best accuracy using decision trees with a training accuracy of **80.06%** and testing accuracy of **79.75%**. For the same model, we also evaluated using the following scores:
Precision: 8.64%
Recall: 62.82%
F1 score: 15.19%

Following are the most and least important features of the model with downsampling, along with its impact on the class outcome.

Decision Trees			
Top 10		Bottom 10	
Features	Impact on class outcome	Features	Impact on class outcome
pw_amount_9089	Approved	educarion_level_Associate's	N/A
GDP	Approved	education_level_Doctorate	N/A
education_level_unknown	Approved	education_level_High School	N/A
education_rate	Approved	education_level_Master's	N/A
clean_state_WASHINGTON	Approved	education_level_None	N/A
unemployment_rate	Denied	education_level_Other	N/A
Percent_Unaffiliated	Denied	Percent_Christian	N/A
Percent_Folk_Regions	Denied	Percent_Hindu	N/A
case_month_7	Approved	Percent_Buddhist	N/A
Percent_Jewish	Approved	Percent_Other_Religions	N/A

Since the tree plot with 14 levels is really complicated to analyze, we are showing a decision tree with max_depth of 4 in Appendix 3.

b. Logistic Regression:

Moving on the the next model that is logistic regression, we tuned the hyperparameters using stratified 10-fold cross-validation for both sampling methods. Following are the results that we got.

SMOTE sampling: The best values obtained for C (inverse of regularization strength) is **1000** with **Penalty = L2** (ridge regression). The accuracy obtained with these hyperparameter values was **79.97%** on the training and **75.20%** on the test set. For the same model, we also evaluated using the following scores:

Precision: 5.29%
Recall: 44.87%
F1 score: 9.46%

Following are the most and least important features of our model with SMOTE sampling, along with its impact on the class outcome.

Logistic Regression			
Top 10		Bottom 10	
Features	Impact on class outcome	Features	Impact on class outcome
citizen_COLOMBIA	Denied	agency_use_N	N/A
citizen_TRINIDAD AND TOBAGO	Denied	citizen_VIETNAM	N/A
clean_state_OREGON	Approved	citizen_ZIMBABWE	N/A
clean_state_TENNESSEE	Denied	citizen_SLOVAKIA	N/A
clean_state_WASHINGTON	Approved	citizen_SLOVENIA	N/A
education_level_Bachelor's	Approved	citizen_SERBIA AND MONTENEGRO	N/A
education_level_Doctorate	Approved	citizen_SENEGAL	N/A
education_level_Master's	Approved	citizen_UZBEKISTAN	N/A
clean_state_NORTH CAROLINA	Approved	citizen_SUDAN	N/A
clean_state_ALASKA	Denied	citizen_SERBIA	N/A

Downsampling: The best values obtained for C (inverse of regularization strength) is **0.01** with **Penalty = L1** (lasso regression). The accuracy obtained with these hyperparameter values was **76.93%** on the training and **85.75%** on the test set. For the same model, we also evaluated using the following scores:

Precision: 70.00%
Recall: 46.55%
F1 score: 55.92%

Following are the most and least important features of our model with downsampling, along with its impact on the class outcome.

Logistic Regression			
Top 10		Bottom 10	
Features	Impact on class outcome	Features	Impact on class outcome
GDP	Approved	clean_state_MISSOURI	N/A
Percent_Buddhist	Approved	clean_state_KENTUCKY	N/A
Percent_Folk_Religions	Approved	clean_state_MISSISSIPPI	N/A
Percent_Hindu	Approved	clean_state_LOUISIANA	N/A
Percent_Jewish	Approved	clean_state_MINNESOTA	N/A
Percent_Muslim	Approved	clean_state_MAINE	N/A
Percent_Other_Religions	Approved	clean_state_MICHIGAN	N/A
Percent_Unaffiliated	Denied	clean_state_MARYLAND	N/A
agency_use_Y	Approved	clean_state_MASSACHUSETTS	N/A
citizen_AUSTRIA	Denied	agency_use_N	N/A

A note about downsampling: As it is evident above, the downsampling for Decision Trees performed really poorly, as most of the important features (like education level) were not considered as important. On the other hand, downsampling for Logistic Regression performed really well, and gave positive results as well. This variance is due to random downsampling of the data and its inability in truly representing the actual dataset. The down sampled data set may not be a true representation of the actual dataset, thus, we cannot completely trust it's validity.

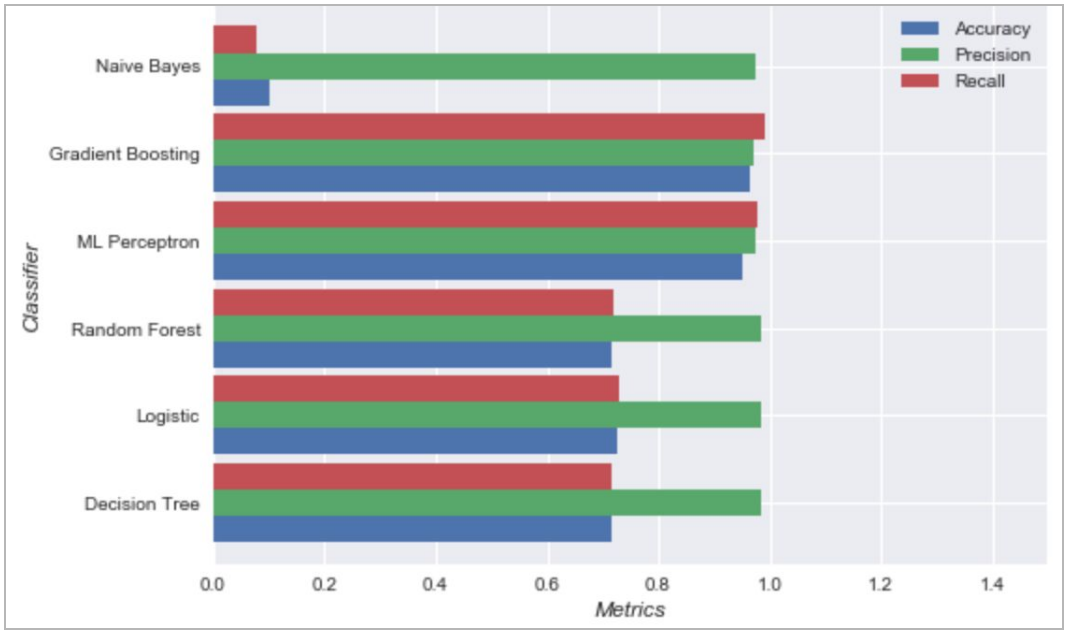
c. Horseraces:

We also implemented various models to understand which model provided the best results for accuracy, precision, recall and F1 score. We implemented the following models:

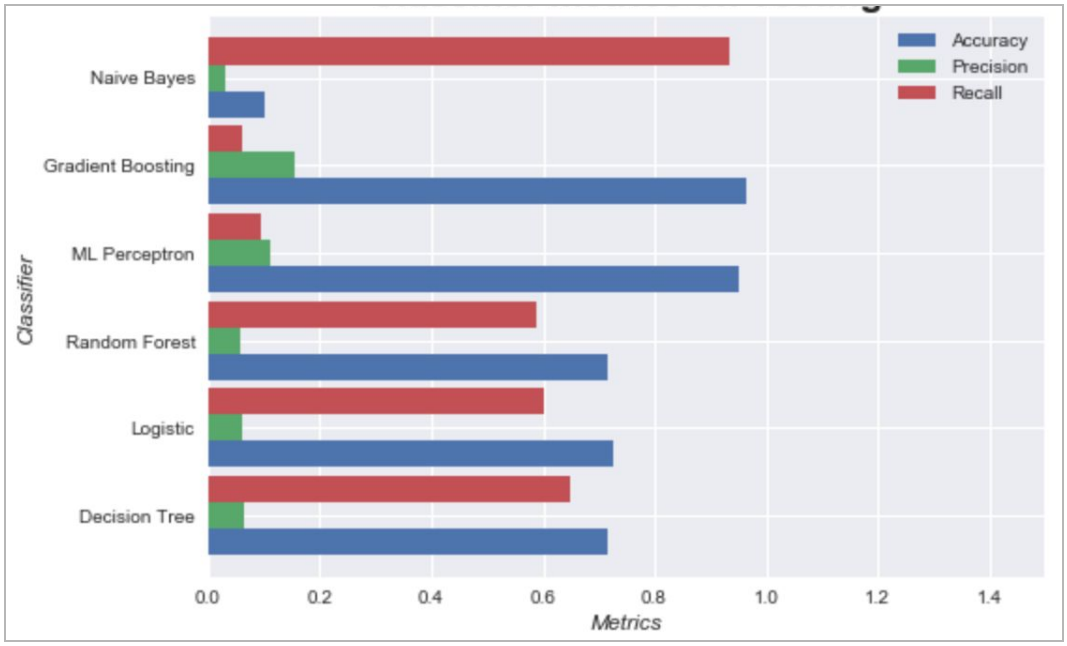
- Decision Tree
- Logistic Regression
- Random Forest

- ML Perceptron
- **Gradient Boosting - Best performing model (Accuracy)**
- Naive Bayes

Precision Recall results for majority class (Visa Approved)



Precision Recall results for minority class (Visa Denied)



5. CONCLUSION:

Following are our recommendations for the most important features:

1. Quantitative Features:

We evaluated the ranking of all the continuous variables based on both sampling methods using logistic regression and decision trees and also evaluated the same for statistical significance using t-test. The following table shows the rankings:

Features	Statistical Significance	Logistic Regression ranking		Decision Tree ranking	
		SMOTE	Downsampling	SMOTE	Downsampling
pw_amount_9089	N	245	1	21	6
unemployment_rate	Y	150	10	19	220
education_rate	Y	154	12	23	2
Percent_Christian	Y	155	14	51	13
Percent_Muslim	Y	156	16	45	19
Percent_Unaffiliated	N	128	18	48	221
Percent_Hindu	Y	188	20	54	15
Percent_Buddhist	N	241	22	6	8
Percent_Folk_Religions	Y	103	24	4	14
Percent_Other_Religions	Y	121	26	33	11
Percent_Jewish	N	137	28	222	222
GDP	Y	162	30	3	1

We found some interesting insights from the above table. Firstly, for downsampling, most of the features that were statistically insignificant were ranked fairly high for Logistic Regression (Eg: pw_amount_9089 was ranked 1) whereas for decision trees, a few were ranked high and a few were at the bottom. This again shows that the models running on downsampled data has erratic behavior.

Secondly, for Logistic regression, the continuous variables have a very low ranking and hence less influence on the outcome, whereas for decision trees, they are ranked fairly high and have considerable influence on the outcome.

GDP is highly influential for decision trees with both the sampling methods but the same does not apply to logistic regression.

Comparing the ranking of the features for SMOTE and downsampling, many features that were highly significant in one was not in the other and vice versa (Eg: Percent Buddhist was ranked 22 by downsampling whereas 241 with SMOTE for logistic regression). This is applicable to both decision tree and logistic regression.

2. <u>Categorical Features</u>

2.1 State - Using Chi Square Test, we were able to establish a significant result proving that the visa outcome does depend on the state that the employer is based in. Applicants to the states of Washington, Oregon and North Carolina have the highest chance of getting their visa approved. However, applicants to the states of South Carolina, Tennessee and Alaska have the least chance of visa approval.

2.2 Education level - Using Chi Square Test, we were able to establish a significant result proving that the visa outcome does depend on the education level of the applicant. Applicants with Doctorate, Masters and Bachelors degrees have the highest approval rates whereas those with high school or no education have the least chance.

2.3 Experience level - Experience level is also significantly proven to be important in determining visa outcomes. Candidates whose experience level is missing in the dataset have the lowest chance of getting the visa approved. Further, generally speaking, the chance of visa approval increases with the increase in experience levels.

2.4 Country - Using Chi Square Test, we were able to establish a significant result proving that the visa outcome does depend on the home country of the applicant. Applicants from the countries of Colombia, Trinidad & Tobago, Austria have a lower chance of getting approval. However, most countries are not important as features to our models.

2.5 Agency_Use - It is statistically significant that those applicants who use an agency have a higher chance of success than those who do not.

6. REFERENCES:

1. https://www.kaggle.com/tonyislamaj/analyzing-only-certified-h-1b-cases
2. https://www.kaggle.com/gskhurana/h1b-visa-prediction
3. https://www.kaggle.com/analystamit/h1b-data-exploration
4. https://www.kaggle.com/ambarish/eda-us-permanent-visas-with-feature-analysis
5. https://www.jair.org/media/953/live-953-2037-jair.pdf

Appendix 1

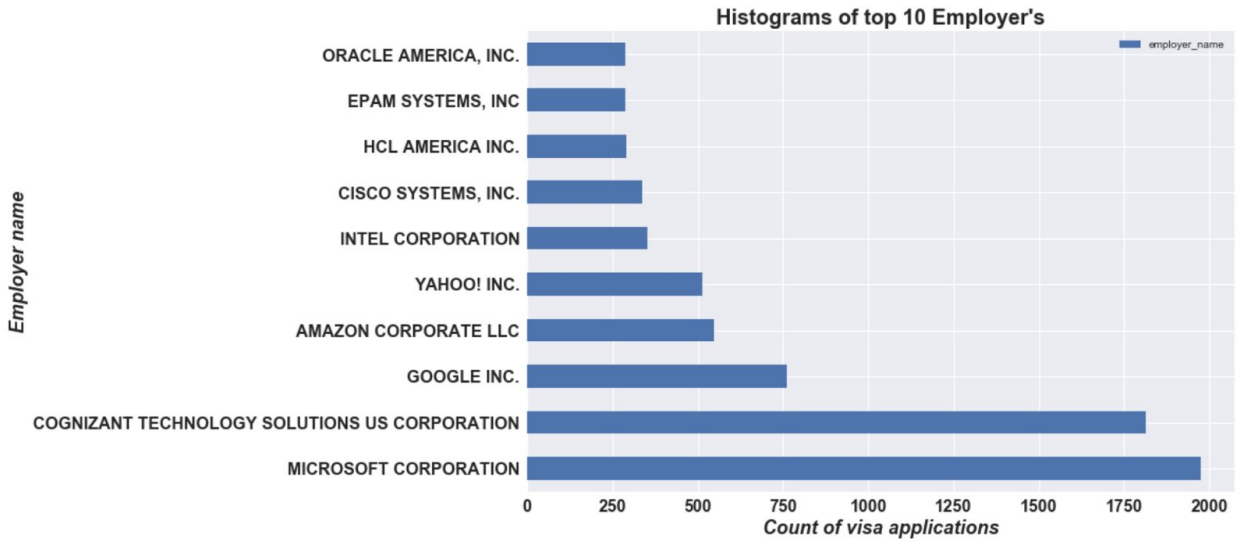
SUMMARY STATISTICS:

	Variable name	Mean	SD	Median	Min	Max	Count
12	GDP	4.898896	3.397687e+00	6.638364	-36.699952	2.627606e+01	19301
8	Percent_Buddhist	0.039348	8.647195e-02	0.008000	0.001000	9.320000e-01	19301
4	Percent_Christian	0.304623	3.650892e-01	0.051000	0.001000	9.950000e-01	19301
9	Percent_Folk_Religions	0.031735	7.916037e-02	0.005000	0.001000	5.890000e-01	19301
7	Percent_Hindu	0.362228	3.931786e-01	0.014000	0.001000	8.070000e-01	19301
11	Percent_Jewish	0.016281	1.037211e-01	0.001000	0.001000	7.560000e-01	19301
5	Percent_Muslim	0.102877	1.362473e-01	0.143000	0.001000	9.990000e-01	19301
10	Percent_Other_Religions	0.015329	2.043228e-02	0.009000	0.001000	1.620000e-01	19301
6	Percent_Unaffiliated	0.129382	1.786060e-01	0.031000	0.001000	7.640000e-01	19301
3	education_rate	0.313244	3.483365e-02	0.314000	0.192000	5.460000e-01	19301
0	pw_amount_9089	136696.889869	2.763146e+06	96762.000000	17389.000000	2.210354e+08	19301
2	unemployment_rate	5.025820	5.250043e-01	5.300000	2.800000	6.700000e+00	19301
1	wage_offer_from_9089	106162.661503	6.065809e+04	103626.000000	20904.000000	7.310000e+06	19301

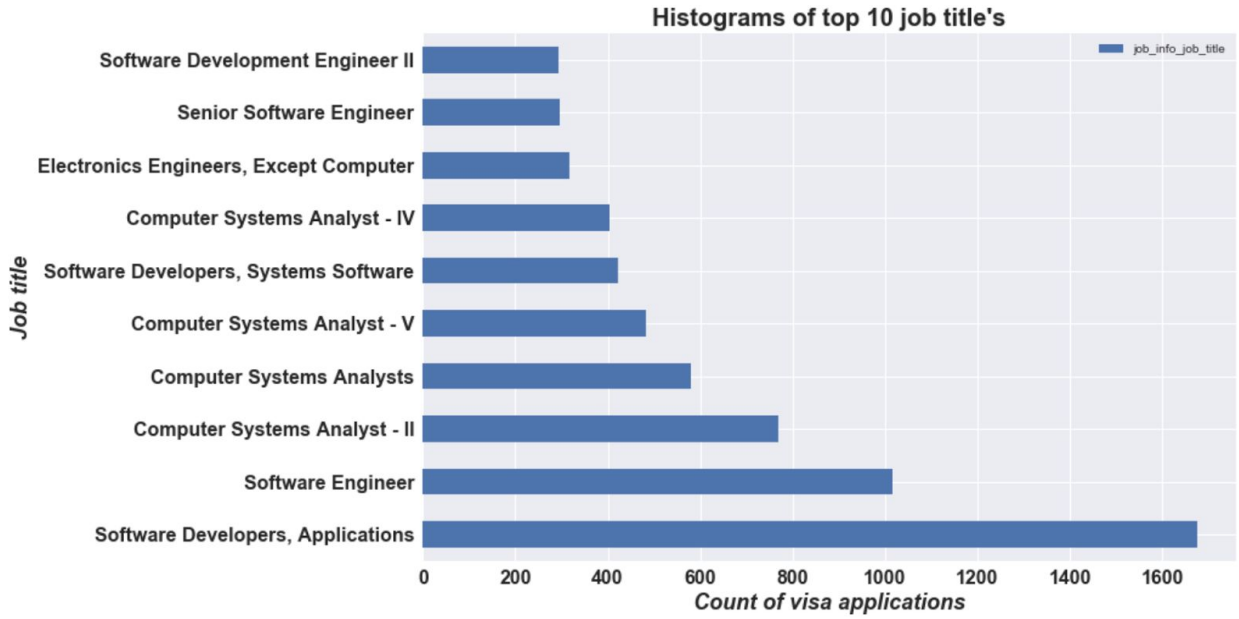
Appendix 2

Who applies for L1 Visa?

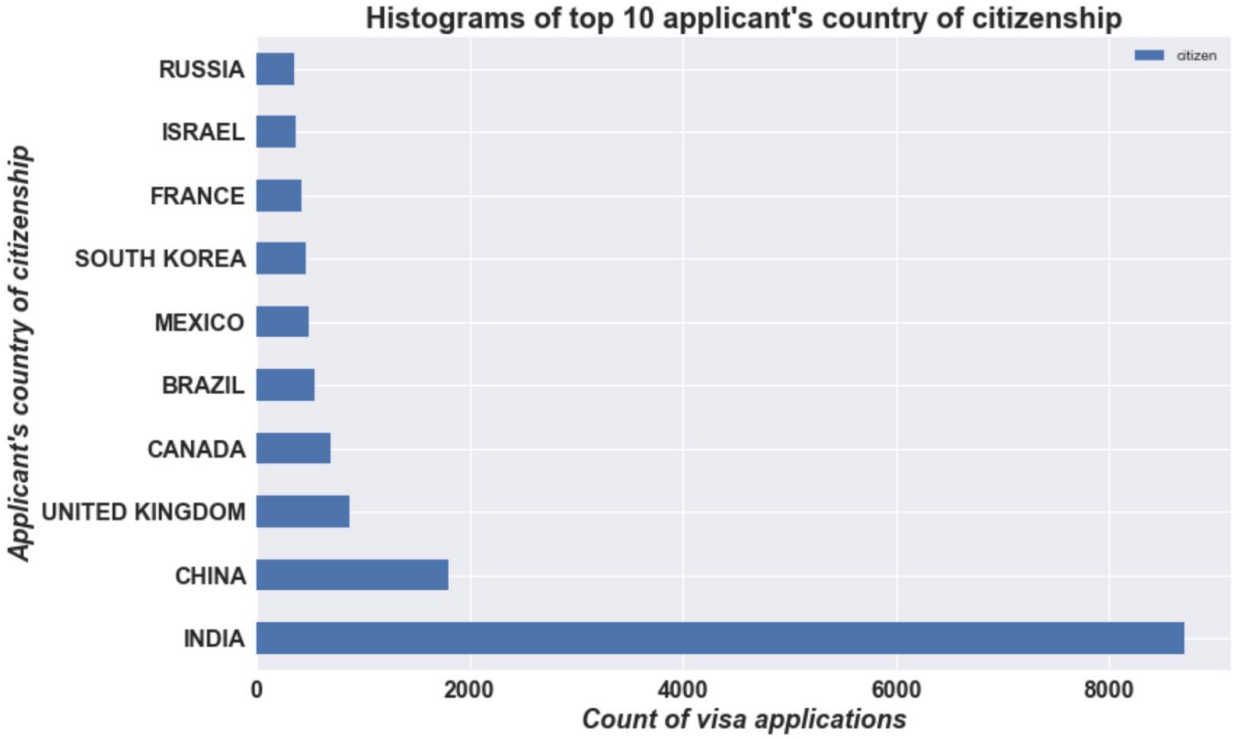
We did some EDA on the dataset to get more insights about the applicants and the employers. The histogram of top 10 employer's shows that Microsoft is the company which has sponsored the most applicants for L1 visa, whereas H1-B's were sponsored mostly by Infosys(Ref).



Almost all the jobs are associated with computer/software technology which is not surprising as most jobs available in US are pertaining to this industry.

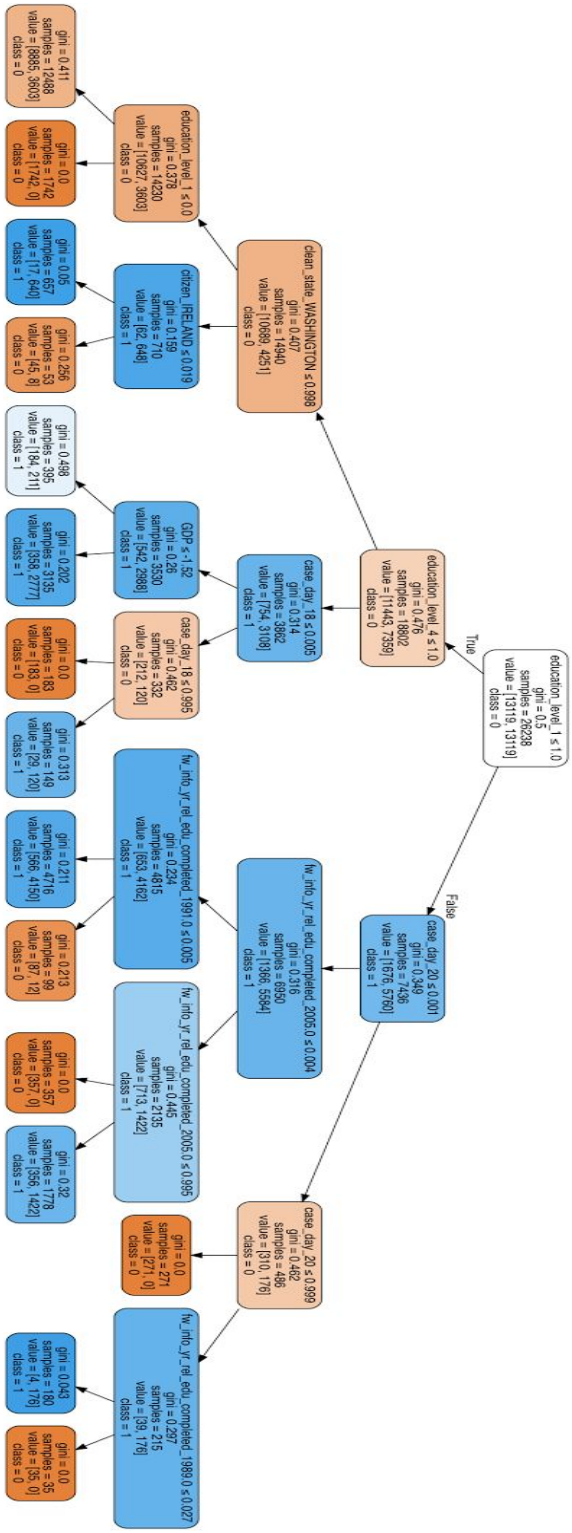


As expected, majority of these applicants are from India.



Appendix 3

Decision tree post SMOTE visualization - max_depth = 4, min_samples_leaf = 14



Decision tree post SMOTE visualization - max_depth = 4, min_samples_leaf = 14