ENPM808W - Data Science - Final Project Report

Purpose Of Our Project

It can be a time-consuming and expensive process to apply for a Work Visa in the United States. Furthermore, an application does not guarantee acceptance. In order to solve this problem, we created a classifier that would determine whether a work visa application can be certified or denied. Before applying, anyone who wants to save some time and money could potentially try our classifier to see if it helps.

Summary Of Our Work

We took the dataset from Kaggle and performed exploratory research, analysis and visualizations of unprocessed data to see the quality of the dataset and how the results vary with different features. We then did an Initial clean-up(This was more subjective to mid-term deliverable). Further, we cleaned the data and eliminated the fields which we did not care about. Finally, we used the clean data for transformations and normalization and fed it to our classification models.

Clean-Up Involved

The raw dataset we chose can be seen below:
https://www.kaggle.com/jboysen/us-perm-visas

This dataset has 374,362 observations distributed from the US Department of Labor website between the years 2012 to 2017. The raw data contained 154 columns with 300 MB size .As a result of the large dataset, we had to decide which fields we wanted to keep and which to discard.

Following were our steps to clean the data:

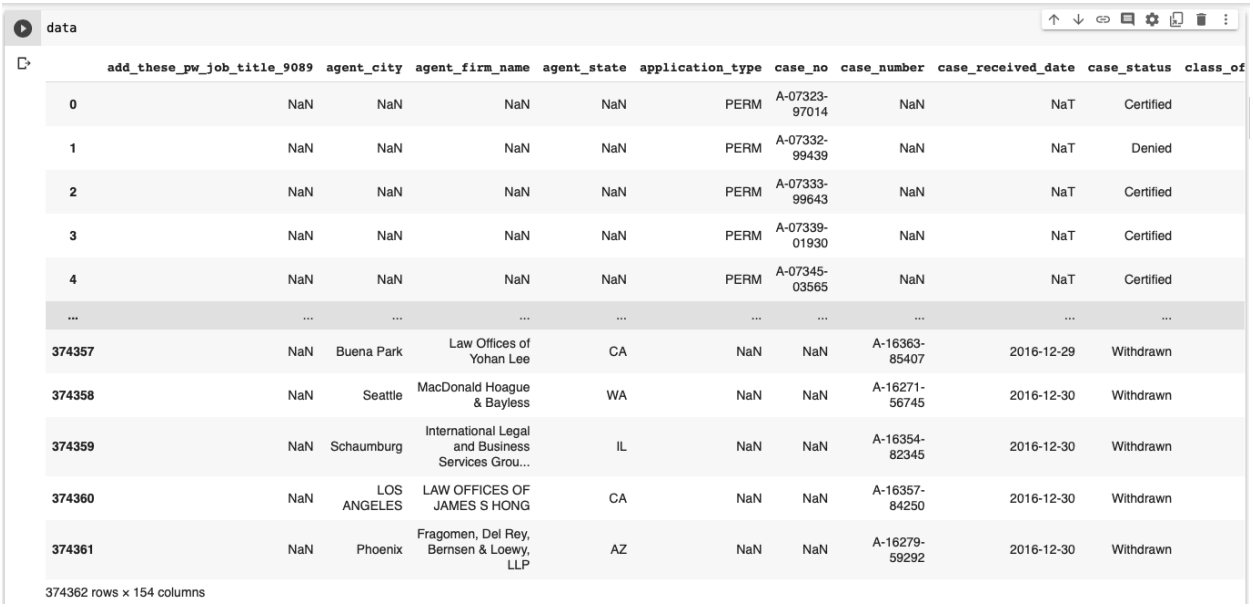
- The first step was to throw out features that mainly contained nulls (this cleaned up a lot of fields).

- It was also necessary to combine a few fields that were repetitive and had no overlaps so that they could be combined into one.

  Eg 1: (Columns 'country_of_citzenship' and 'country_of_citizenship' combined into a new column 'countryofcitizenship').
  Eg 2: (Columns 'case_no' and 'case_number' combined into a new column 'casenumber') .
- We performed dimensional reduction analysis to remove the anomalies and null values in the dataset.

- Removing the unnecessary columns and anomalies resulted in a smaller dataset to analyze when compared with the raw data. This helped us with a significant choice selection, we selected features that we thought would make good features which would help in our classifier.

Raw dataset



data

	add_these_pw_job_title_9089	agent_city	agent_firm_name	agent_state	application_type	case_no	case_number	case_received_date	case_status	class_of
0	NaN	NaN	NaN	NaN	PERM	A-07323-97014	NaN	NaT	Certified	
1	NaN	NaN	NaN	NaN	PERM	A-07332-99439	NaN	NaT	Denied	
2	NaN	NaN	NaN	NaN	PERM	A-07333-99643	NaN	NaT	Certified	
3	NaN	NaN	NaN	NaN	PERM	A-07339-01930	NaN	NaT	Certified	
4	NaN	NaN	NaN	NaN	PERM	A-07345-03565	NaN	NaT	Certified	
...	...	...	...	...	...	...	...	...	...	
374357	NaN	Buena Park	Law Offices of Yohan Lee	CA	NaN	NaN	A-16363-85407	2016-12-29	Withdrawn	
374358	NaN	Seattle	MacDonald Hoague & Bayless	WA	NaN	NaN	A-16271-56745	2016-12-30	Withdrawn	
374359	NaN	Schaumburg	International Legal and Business Services Grou...	IL	NaN	NaN	A-16354-82345	2016-12-30	Withdrawn	
374360	NaN	LOS ANGELES	LAW OFFICES OF JAMES S HONG	CA	NaN	NaN	A-16357-84250	2016-12-30	Withdrawn	
374361	NaN	Phoenix	Fragomen, Del Rey, Bernsen & Loewy, LLP	AZ	NaN	NaN	A-16279-59292	2016-12-30	Withdrawn	

374362 rows × 154 columns

Cleaned dataset

```
Data columns (total 16 columns):
 #   Column                  Non-Null Count   Dtype
---  ------                  --------------   -----
 0   agent_state             356113 non-null  object
 1   case_status             356113 non-null  int64
 2   class_of_admission      356113 non-null  object
 3   decision_date           356113 non-null  int64
 4   employer_country        356113 non-null  object
 5   employer_name           356113 non-null  object
 6   employer_num_employees  356113 non-null  float64
 7   employer_state          356113 non-null  object
 8   employer_yr_estab       356113 non-null  int64
 9   job_info_work_state     356113 non-null  object
 10  pw_soc_code             356113 non-null  int64
 11  pw_source_name_9089     356113 non-null  object
 12  casenumber              356113 non-null  object
 13  countryofcitizenship    356113 non-null  object
 14  pw_amount_9089_new      356113 non-null  float64
 15  wage                    356113 non-null  category
dtypes: category(1), float64(2), int64(4), object(9)
memory usage: 51.9+ MB
```

Features we used for Analysis or Classification

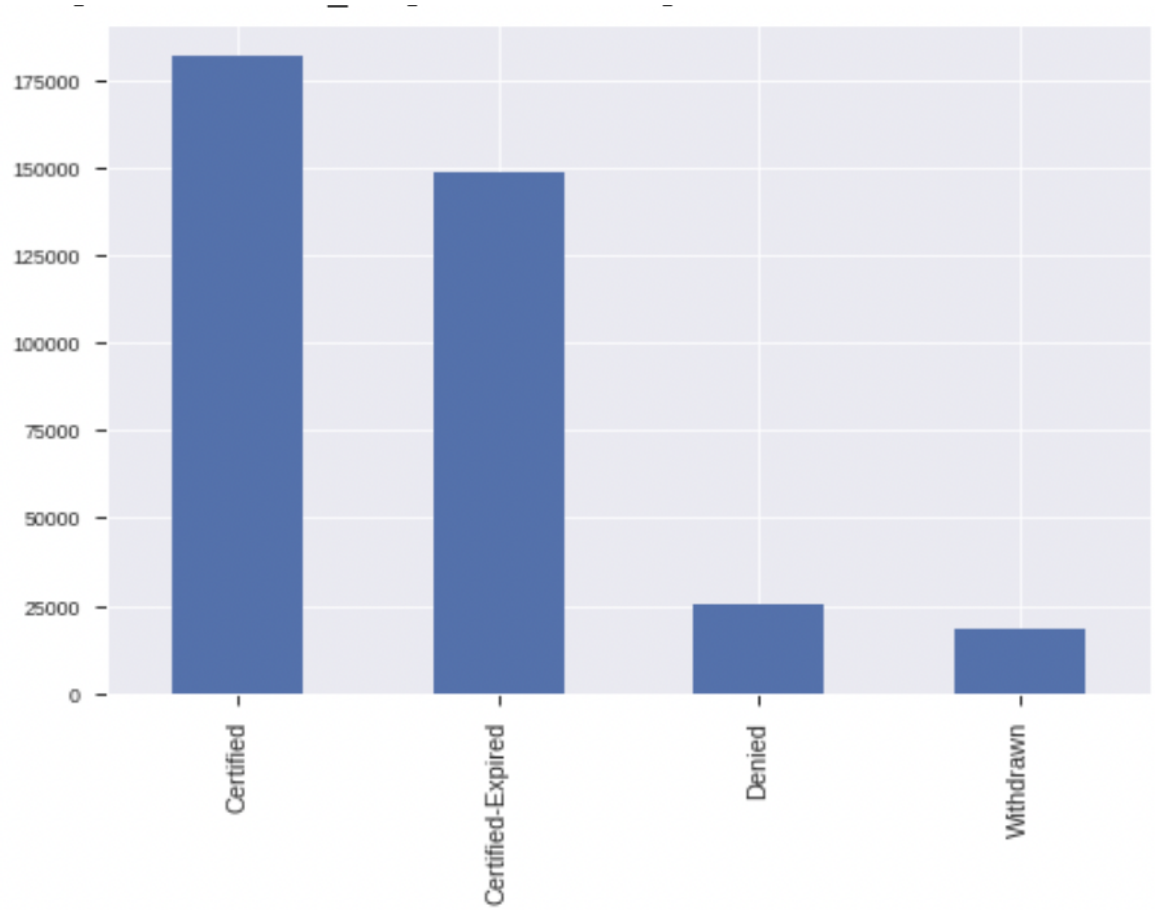
Features	Description	Notes
"agent_state"	State information for the Agent or Attorney requesting a permanent labor certification on behalf of the employer.	State data was not all in the same format (Maryland Vs. MD), data was formatted such that state data was homogenous in format (all upper case abbreviations).
"case_status"	Status associated with the last significant event or decision. Valid values include “Certified,” “Certified-Expired,” “Denied,” and “Withdrawn”	This field is the information we wanted to try to guess correctly using a classifier.
"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)	There were 58 unique entries here. More than 70% of the data was H-1B visas.
"decision_date"	Date on which the last significant event or decision was recorded by the ETA National Processing Center (Year only).	We extracted only the YEAR part from the column decision_date as exploratory analysis.
"employer_country"	Country of employer	Majority of the companies are from the United States of America.
"employer_name"	Name of the employer	Used for data Analysis and as a feature classification model.
"employer_num_employees"	Number of employees for an employer	Used as a feature in classification models.

"employer_state"	State of the employer requesting permanent labor certification	Same problem as in agent_state. Converted employer state data into homogeneous format.
"employer_yr_estab"	Year of Establishment	Used as a feature in classification models.
"job_info_work_state"	Indicates the work state of employment for the applicants	Used as a feature in classification models.
"pw_soc_code"	Occupation code associated with the job being requested for permanent labor certification.	Used as a feature in classification models.
"pw_source_name_9089"	Name of the entity providing prevailing wage information for the job. Valid values include : OES, CBA and others.	Used as a feature in classification models.
"casenumber"	Unique identifier per entry	Combined column for (case_number) and (case_no).
"countryofcitizenship"	Country of citizenship of the foreign worker	Combined column.
"pw_amount_9089_new"	Prevailing wage for the job being requested for permanent labor certification	This value was scaled to be annual salary using the associated fields pw_amount_9089 and pw_unit_of_pay_9089.
"wage"	Calculated column	Group the applicants with different wage ranges and used in exploratory analysis & classification models.

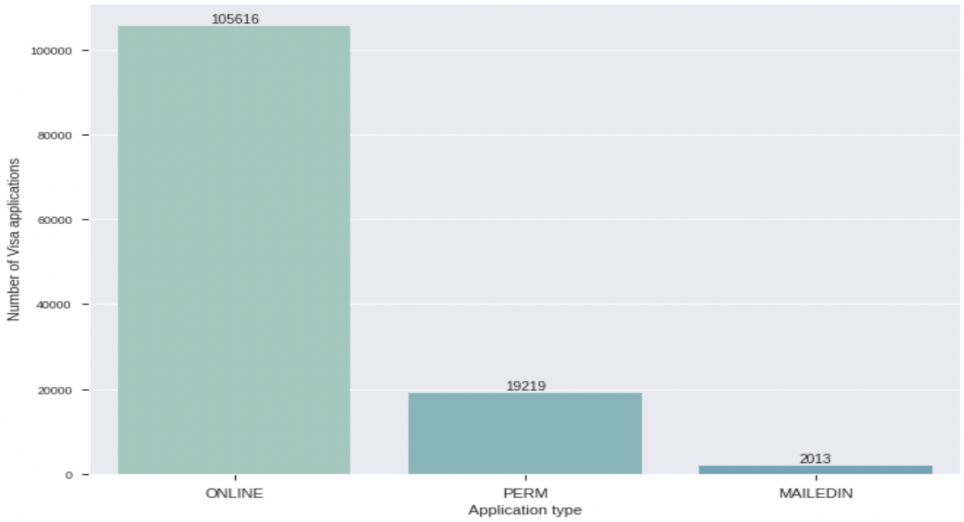
Exploratory Analysis

After removing unnecessary fields and keeping only the features we wanted to use, we next plotted some graphs to visualize the data we had.

First, we wanted to check the distribution of the status of the application. Majority of the application status entries in our data were Certified/Certified-Expired (case_status field) and comparatively least entries had withdrawn status. So we merged Certified and Certified-Expired as the 'Accepted' field. Also, the ratio of Accepted to Denied was 12:1.

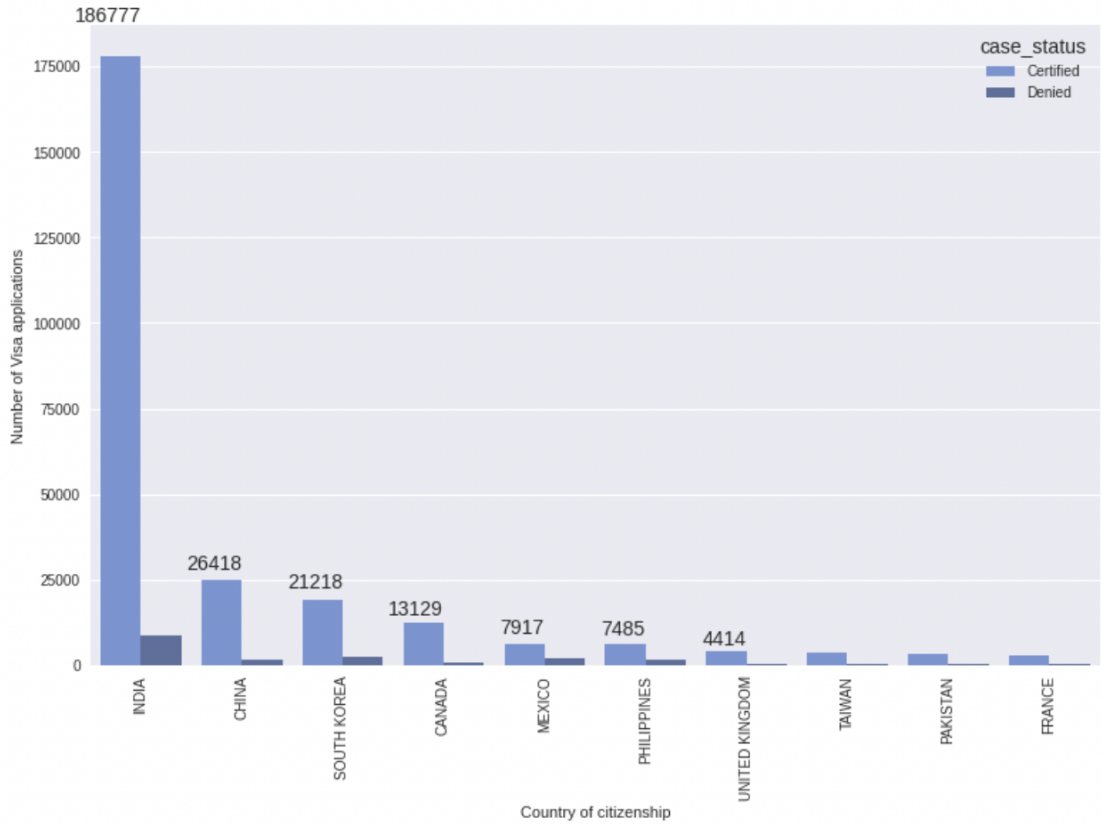


We also checked the distribution of the mode of application submissions and below graph shows Online mode is the most opted mode by the applicants.



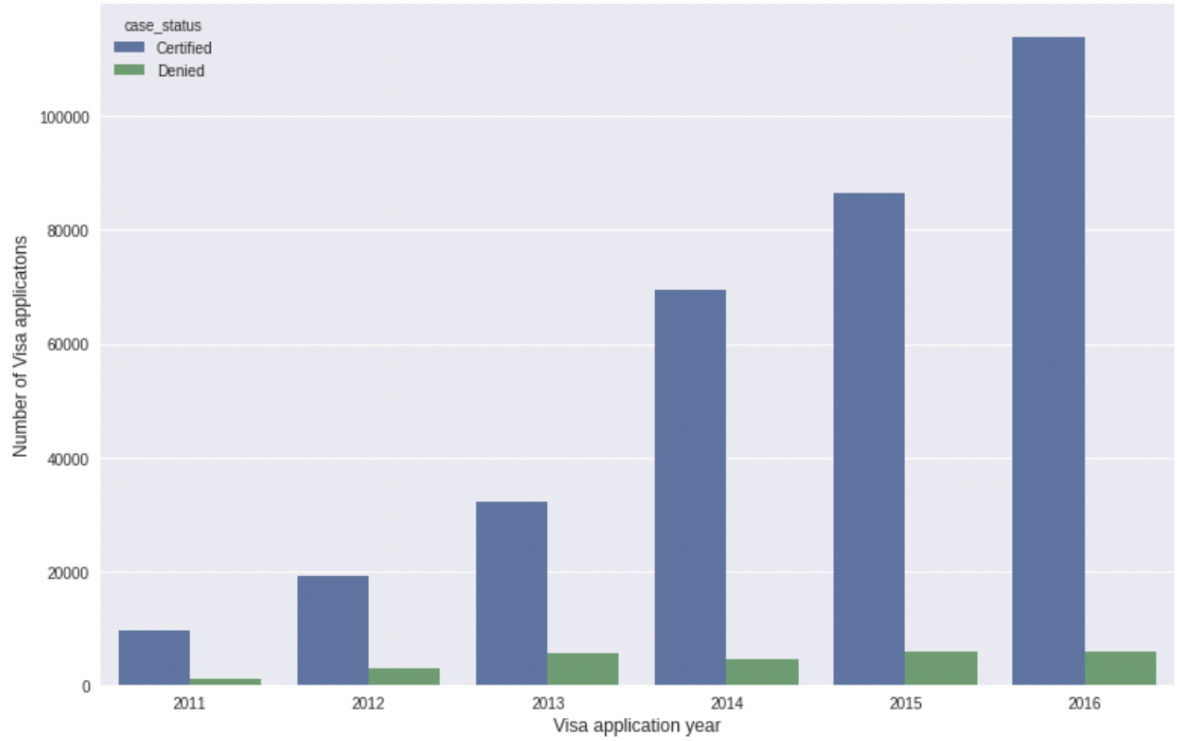
Next we plotted the top 9 countries that were applying for Visa. Surprisingly, many of the applicants received in our data were from India. Among these top 9 countries applying, China has the largest Accepted vs Denied ratio while Mexico has the lowest.

```
There are 201 different countries of citizenship in this dataset
[Text(0, 0.5, 'Number of Visa applications'),
 Text(0.5, 0, 'Country of citizenship')]
```



As we can observe below, the number of submitted Visa applications increases every year. It's interesting that while the number of positively considered applications increases, the number of "Denied" ones seems to be similar from 2013.

```
[Text(0, 0.5, 'Number of Visa applicatons'),
 Text(0.5, 0, 'Visa application year')]
```

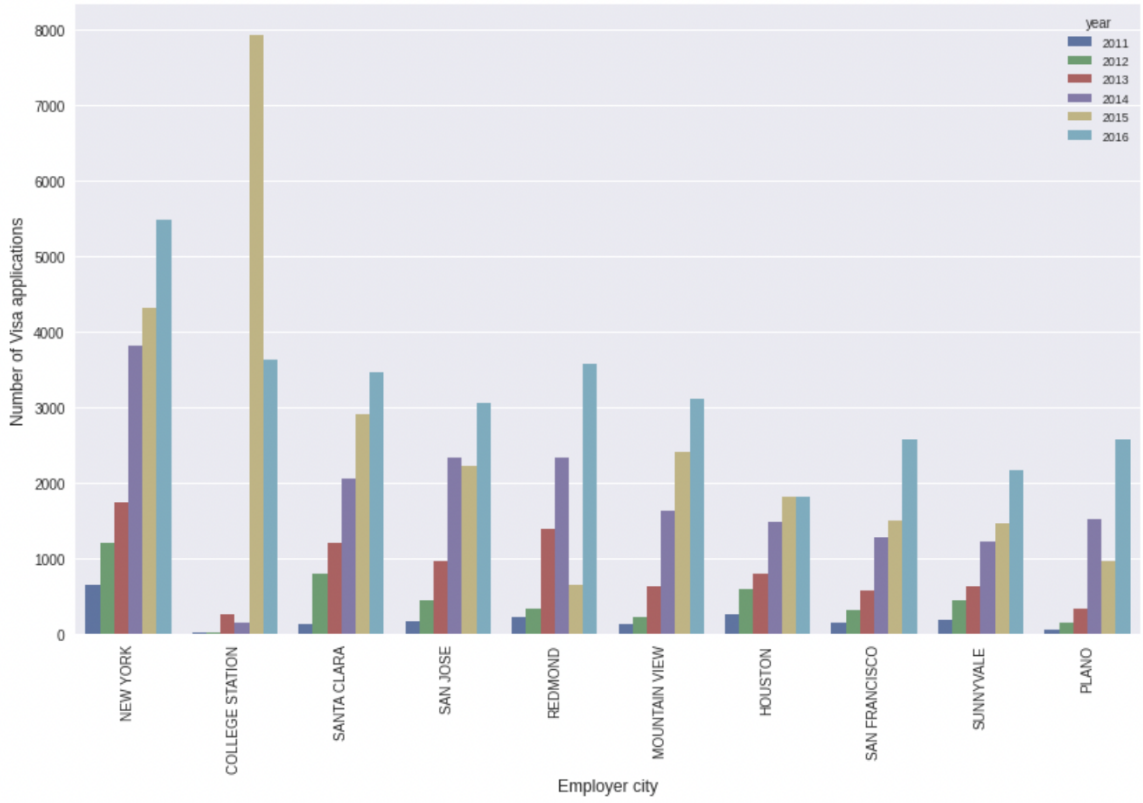


As a next step, let's see, what where the most popular cities

In the last few years, the most popular destination cities were: New York, College Station, Santa Clara, San Jose, Redmond, Mountain View, Houston, SunnyVale, San Francisco and Plano. We saw this above (in the beginning) as well.

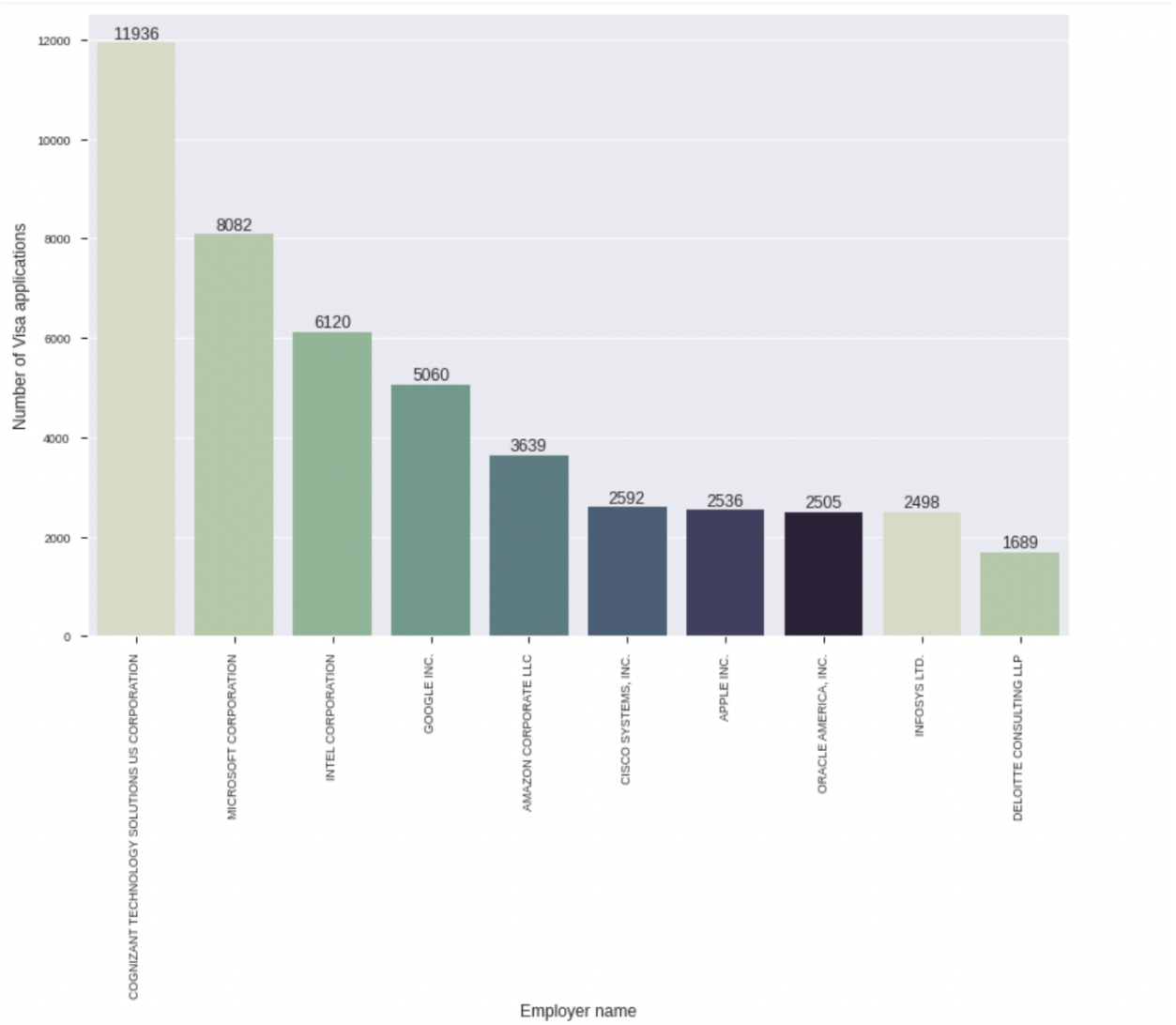
In most of the cities there was a positive trend in Visa applications. A bizarre situation occurred in College Station in 2015 where the number of submitted Visa applications was more or less twice large as in other cities.

```
[Text(0, 0.5, 'Number of Visa applications'), Text(0.5, 0, 'Employer city')]
```

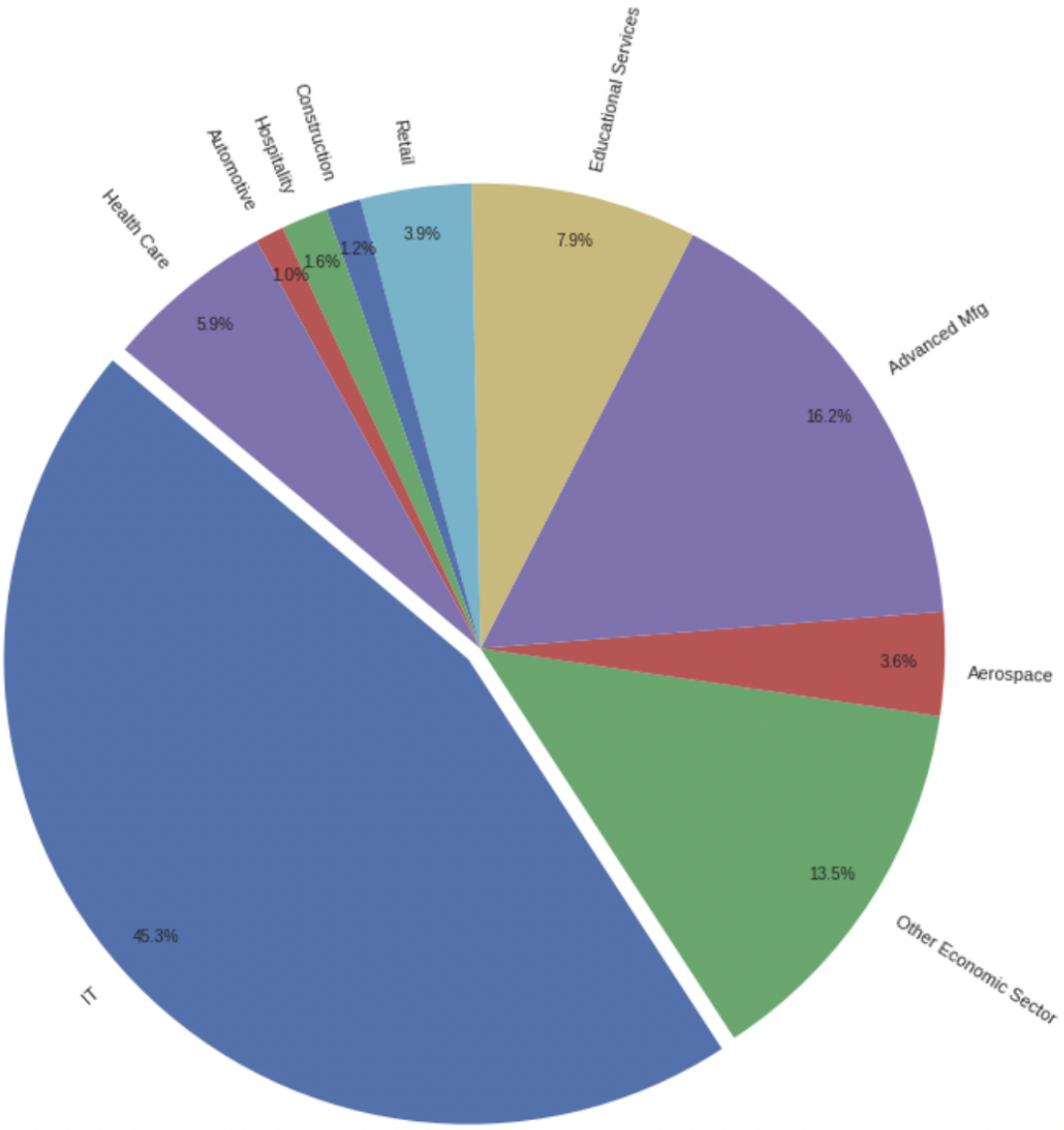


Now, let's take a look at what were the most hiring employers and economic sectors through these years. For the "us_economic_sector" variable we have only 120 868 non-missing values, but this should give us an insight. As we can see, 9 out of 10 most beneficial companies for Visa applicants are IT industry representatives. This leads to the assumption that the IT sector is both the most favorable and demanding one in the United States.

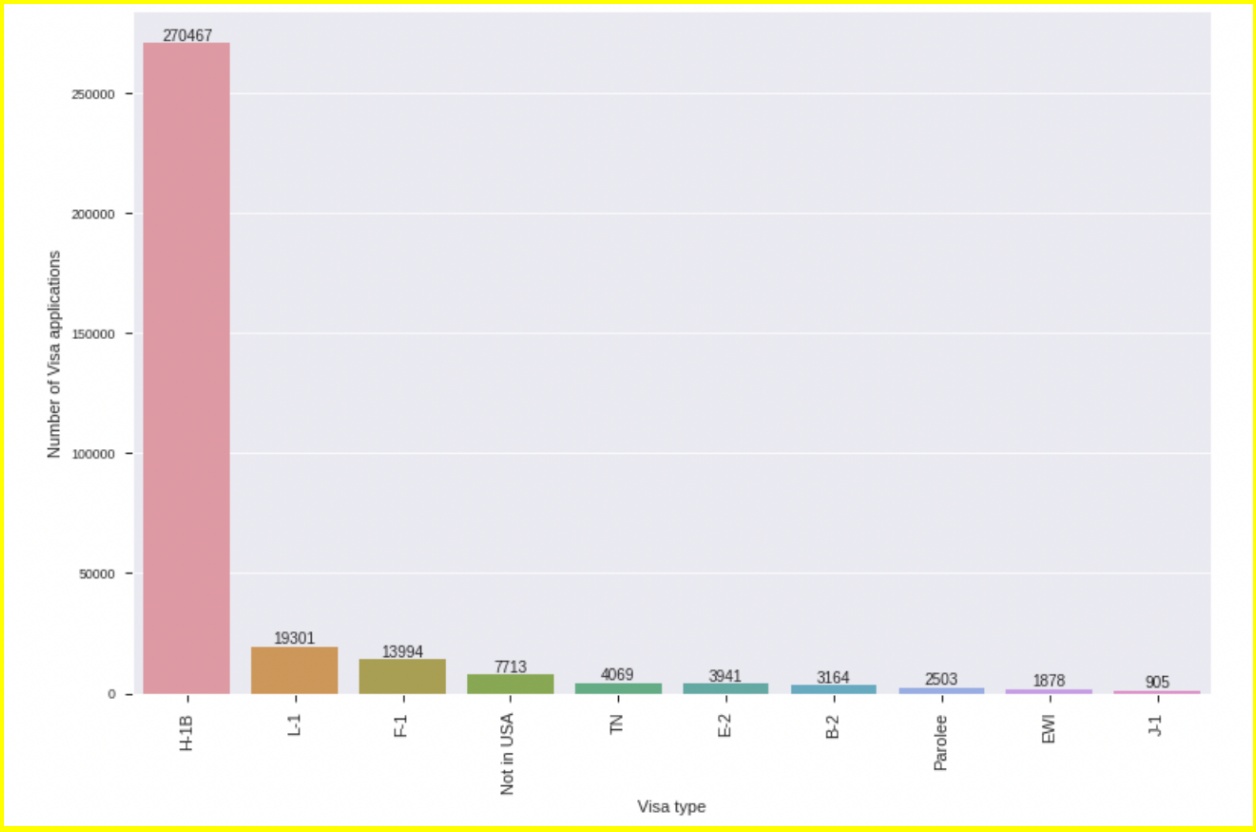
Cognizant Technology Solution US Corporation had the most visa applicants, although their accepted to denied ratio was rather low. Microsoft is the second most desired company by visa applicants.



Let's check what is the distribution of industries across all Visa applications. Even though the US economic sector sample contained only 120 868 non-missing values, this somehow confirms that IT (not surprisingly) and Advanced Manufacturing are the most convenient sectors for applying-foreigners. Advanced Manufacturing seemed to have an extremely high ratio of accepted vs denied.

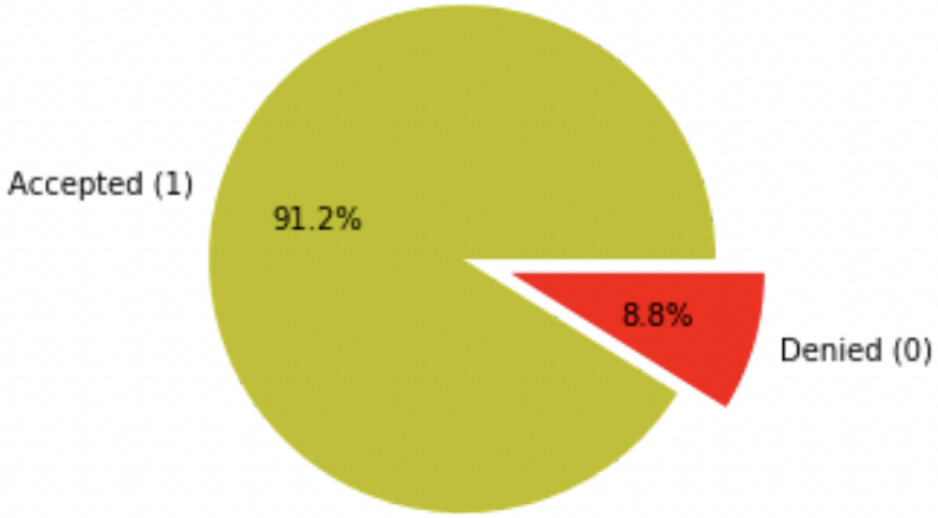


We then checked the type of visa the applicants were interested in. The vast majority of petitioners were applying for the H-1B Visa, which according to Wikipedia, allows U.S. employers to employ foreign workers in specialty occupations. If a foreign worker in H-1B status quits or is dismissed from the sponsoring employer, the worker must either apply for and be granted a change of status, find another employer (subject to application for adjustment of status and/or change of visa), or leave the United States



Classifier Creation and Results

The case status field, which will be our prediction field, consists of three types of status: 1) Accepted 2) Denied 3) Withdrawn. Since, Withdrawn status only accounts for 5% of the total data, we dropped all the columns with that case. Rest of them had Case Status as accepted or denied. After performing Label Encoding, the Accepted(Certified) cases were represented as 1 and the denied cases were represented as 1. The case status of around 90% of the total dataset accounts for the certified cases whereas only 10% accounts for denied cases as shown in the chart below. This shows that the dataset is highly imbalanced and we have to find a mechanism for dealing with it. The dataset also had a lot of missing value: despite having a large number of data features, only few of them could be counted as useful for visa status prediction.



After analyzing the dataset, we decided to use 15 columns from which we believed that our classifier would be able to extract most of the features. The table below provides percentage of missing data for the 15 chosen data features.

Data Features	**Percentage of missing data**
1. Case received date YEAR	65.8%
2. Application type	34.2%
3. Class of admission	5%
4. Education level	65%
5. Country of citizenship	0%
6. Economic sector	37.2%
7. Employer state	0%
8. CTC (Cost to Company)	0%
9. Agent State	0%
10. Employer Name	0%
11. Employer Number of Employees	0%
12. Employer Year Established	0%
13. Decision Date	0%
14. PW_SOC_Code	0%
15. PW_Source_Name_9089	0%

Many machine learning and data science algorithms can not operate directly on the categorical values. These values must be transformed into numerical values. For the categorical data features, we decided to extract features by performing two types of encoding depending on the data: Label Encoding and One-Hot Encoding. Since columns like class of admissions and country

of citizenship have around 57 and 201 unique values respectively, we decided to normalize the data using One Hot Encoding. In One-Hot encoding, a new column is added for each unique category as shown in the figure below:

...	class_of_admission_47	class_of_admission_48	class_of_admission_49	class_of_admission_50	class_of_admission_51
...	0	0	0	0	0
...	0	0	0	0	0
...	0	0	0	0	0
...	0	0	0	0	0
...	0	0	0	0	0
...	...	...	...	...	...
...	0	0	0	0	0
...	0	0	0	0	0
...	0	0	0	0	0
...	0	0	0	0	0
...	0	0	0	0	0

After normalizing the data, the dataset was splitted into the standard 70-30 ratio for training and testing dataset. A decision tree classifier was trained and the following output of accuracy was obtained from the trained classifier:

Percentage of denied cases predicted correctly.	**Percentage of cases predicted correct. (Accuracy Score)**
7.06%	93.72%

Although the accuracy of around 94% seems to be very high, we realized that our classifier was not able to predict the denied cases accurately. To be precise, it could only predict the denied cases 1 out of 14 times in the test dataset. In this way, the classifier can predict the accepted case status for almost all of the data rows in the test set, and since more than 90% of the data-set consists of accepted case status values (and the test data set also contains the majority of data rows with accepted case status values), the accuracy is high.

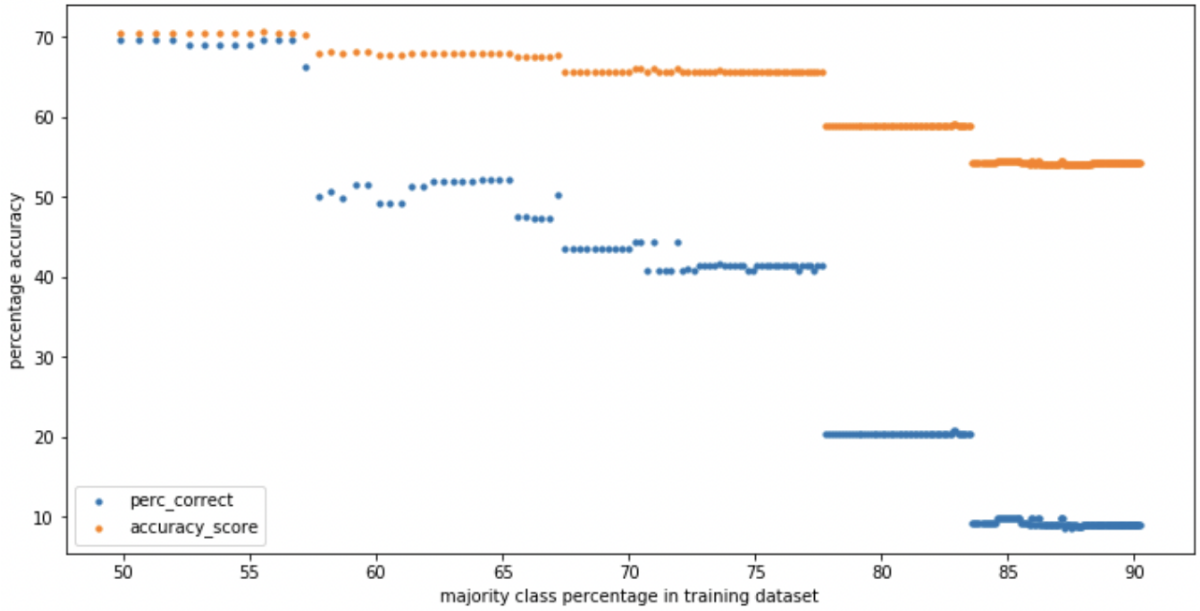
To get a more clearer picture of how our dataset is performing, we decided to test our dataset into evenly distributed cases of Accepted data and Denied data. Specifically, we tested our dataset into an even split of around 3000-3000 accepted and denied cases. The table below shows the accuracy of the results obtained from the classifier. Among the 3000 denied cases, only around 8% of the cases were accurately predicted by the classifier.

Percentage of denied cases predicted correctly.	Percentage of cases predicted correctly. (Accuracy Score)
8.51%	54.03%

	Predicted Denied	Predicted Accepted
Actual Denied	257	2770
Actual Accepted	6	2994

The confusion matrix for the predictions on the 6000 cases are shown. As we can see, for 2770 datasets where the results should have been denied are being predicted as accepted which only justifies the oversampling of our data.

Finally, for our classifier to give a better representation of the predictions, we decided to train our datasets evenly. Since we have around 15000 Denied cases dataset, we decided to train our classifier with approximately 15000 Accepted and 15000 Denied cases. Reducing the dataset to balancing the minority class with the majority class is called undersampling which is the technique we have used. The rationale behind the selection of evenly distributed dataset is shown in the figure below. The accuracy score decreases overall with the increase in the majority class percentage. The perc_correct field (accuracy of only the denied cases) decreases significantly as the majority class percentage increases.



After settling on the number of datasets, we trained four classifiers with four different algorithms: Logistic Regression, Support Vector Classifier, Random Forest Classifier and Decision Tree Classifier. The denied cases accuracy and overall accuracy of each of the four classifiers are presented in the table below:

Classifier	Accuracy	
	Denied Cases Accuracy	Overall Accuracy
Logistic Regression	82%	73%
Support Vector Classifier	77%	71%
Random Forest Classifier	**86%**	**82%**
Decision Tree Classifier	80%	77%

The result shows that the Random Forest Classifier has the best accuracy for both: the accuracy of only the denied cases and the accuracy of the overall dataset. The accuracy provided by the Decision Tree Classifier is also good. Tree based algorithms have better accuracy than Logistic Regression and the Support Vector Machine classifier.

Did Our Technique Work?

When using the initial data set, where most accepted cases were represented, our classifier had a lower accuracy than if you were to guess accepted visas randomly for every single data point. Consequently, we were unable to do better than the typical baseline used in many classification problems, since a human could achieve a "guess only achieved" accuracy of over 90%. Due to extremely skewed data, we would rather consider how accurately we were able to classify denied cases. The naive classifier would be 100% accurate if all cases were accepted. We were able to increase both our accuracy for denied cases and accepted cases up to 70% each using the best classifier we were able to train. If we were to classify visa applicants according to whether they would be denied, we would be able to predict whether 70% of applications would be denied. As a result, it would save both applicants and companies a great deal of money when deciding whether to apply or not.