US Permanent Visa Applications

Justin Dagenhart
Muhammad Adil Sohail
Shivam Mishra
Corbyn Yhap

Overview

- Purpose of the project
- Data overview
- Exploratory analysis
- Classification attempts
- Classification results
- Issues faced
- Summary

Purpose of The Project

What we wanted to know:

> Is it possible to determine if an applicant pursuing a permanent work visa in the United States will be accepted or denied using applicant's available information

What data will be used:

> Data found from the US Department of Labor (already collated into a Kaggle dataset)

Data Overview

- 205,533 observations
- Initially 154 fields

# Observations	205,533
# Fields	154

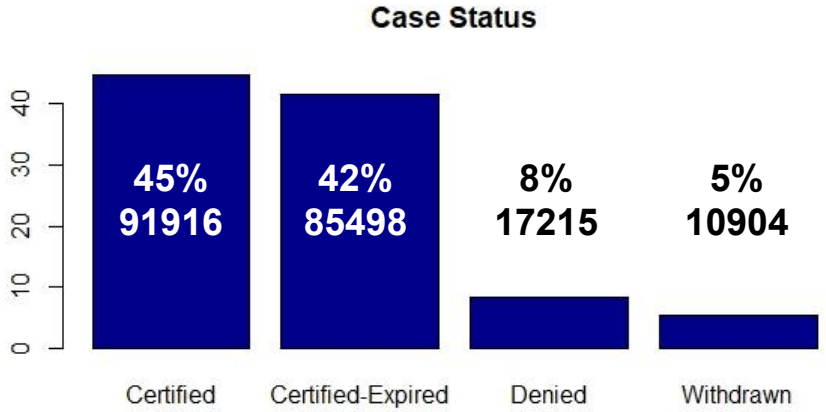
Removed

# Observations	~70,000
# Fields	19

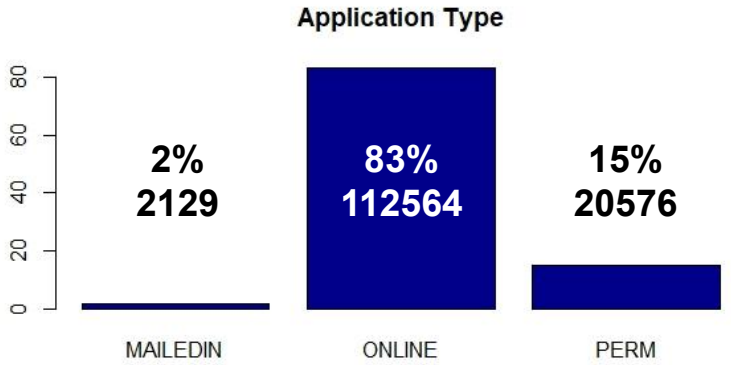
Final Fields Kept

1. case_no
2. Case_status (what we want to predict)
3. agent_state
4. case_received_date_EPOCH
5. case_received_date_YEAR
6. class_of_admission
7. country_of_citizenship
8. employer_state
9. application_type
10. decision_date
11. employer_postal_code
12. foreign_worker_ownership_interest
13. foreign_worker_info_education
14. job_info_major
15. pw_amount_9089
16. pw_unit_of_pay_9089
17. us_economic_sector
18. wage_offer_from_9089
19. wage_offer_unit_of_pay_9089

Exploratory Analysis



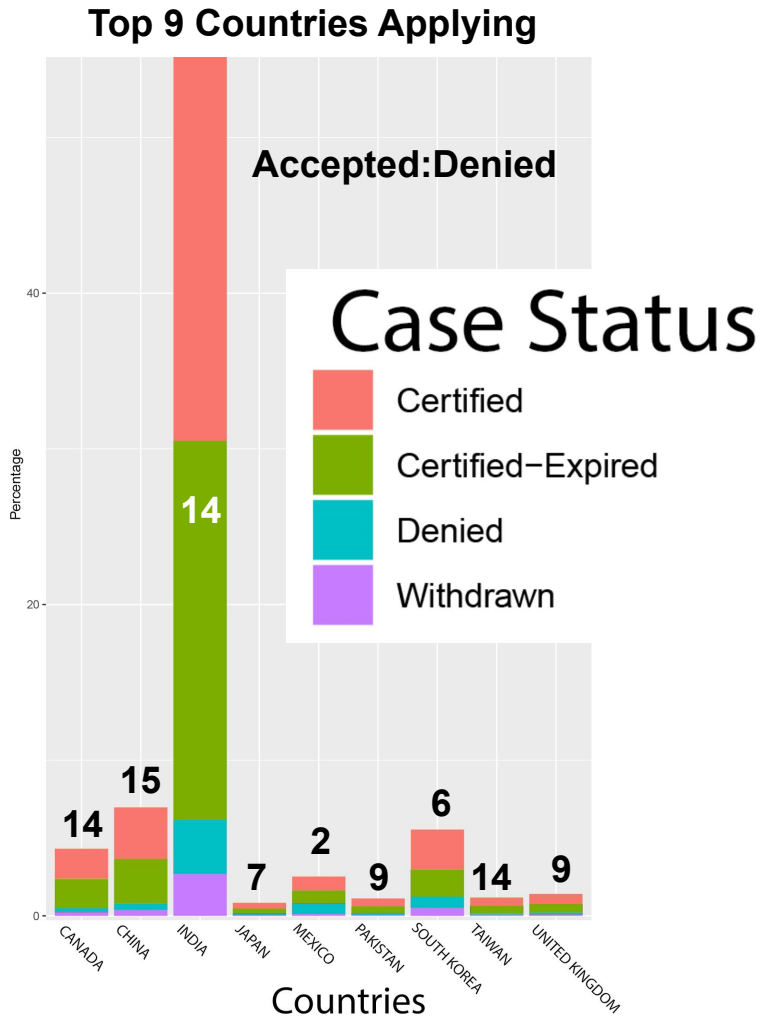
- Most of our data contains certified data
  - We took Certified-Expired and Certified to mean Accepted Visas (expired were accepted at one time)
  - The ratio of Accepted vs Denied for the entire data was ~12



- Most Applications are submitted Online
- Data found may be biased or most applicants are actually certified

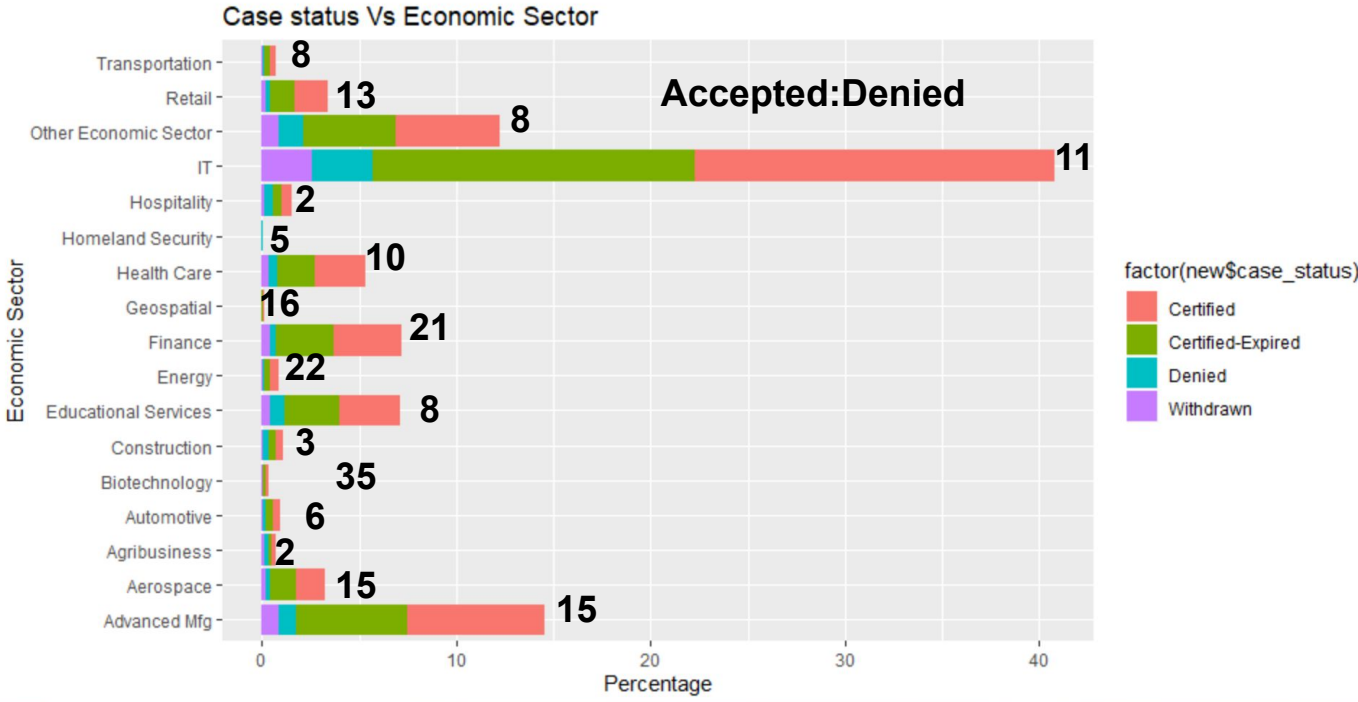
Exploratory Analysis

- Most applicants processed are from India
  - Second most is from China (with South Korea close behind)
- The ratio of accepted vs. denied across countries is relatively the same
  - (# of accepted / # denied per country is shown on the graph)



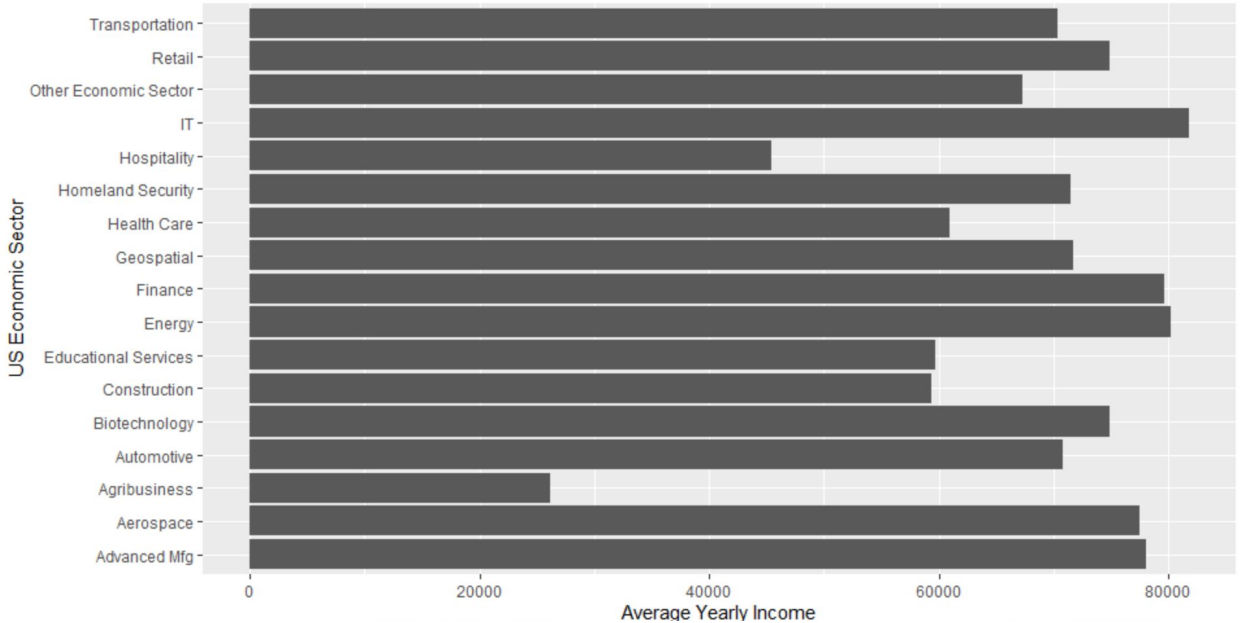
Exploratory Analysis

- Aside from IT, there are also a number of economic sectors that contribute significantly to the number of Certified Visa applications.
    - Biotechnology seems to have a extremely high ratio of accepted Vs. denied



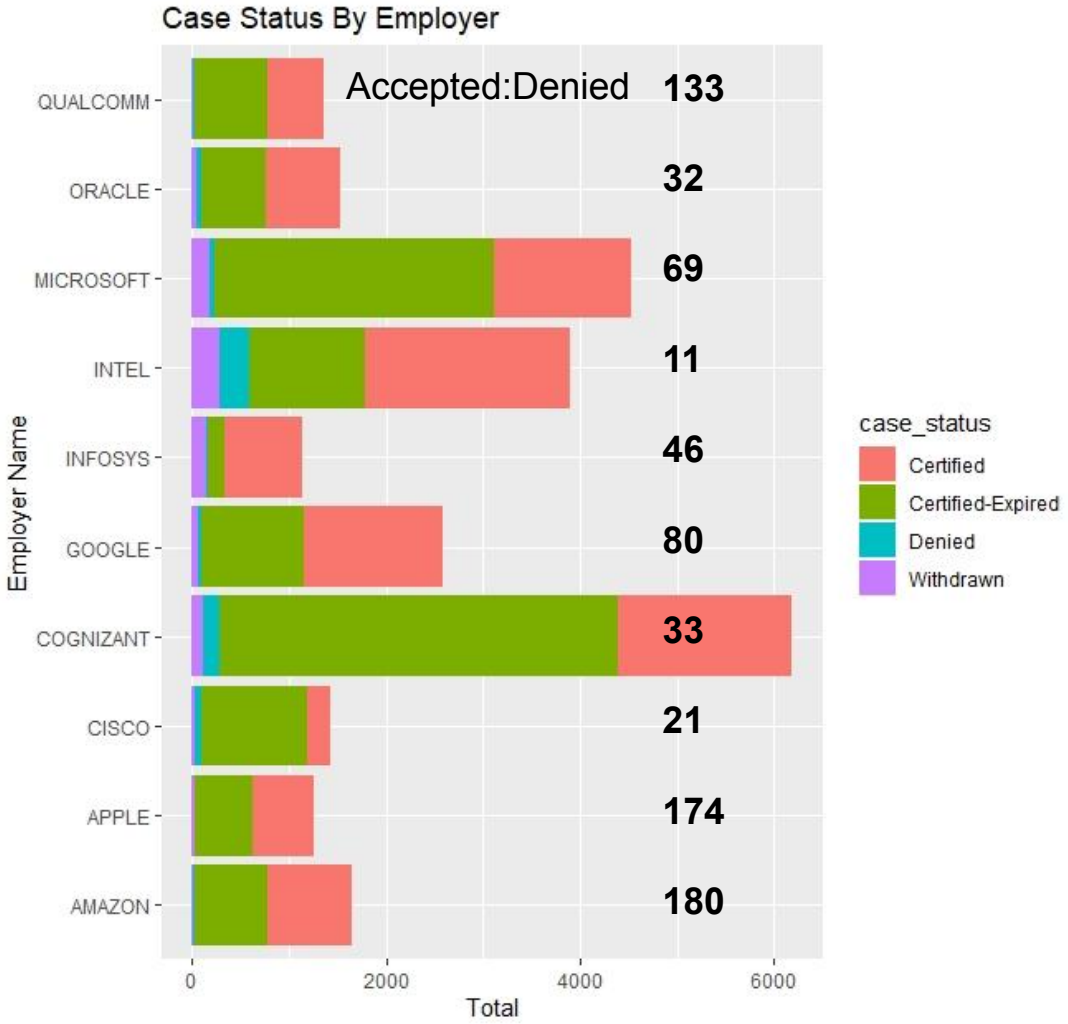
Exploratory Analysis

Average annual wage for visa applicants in IT, Finance, Energy, Advanced Manufacturing, and Aerospace are all fairly close. Lowest average annual wage was in Agribusiness and Hospitality

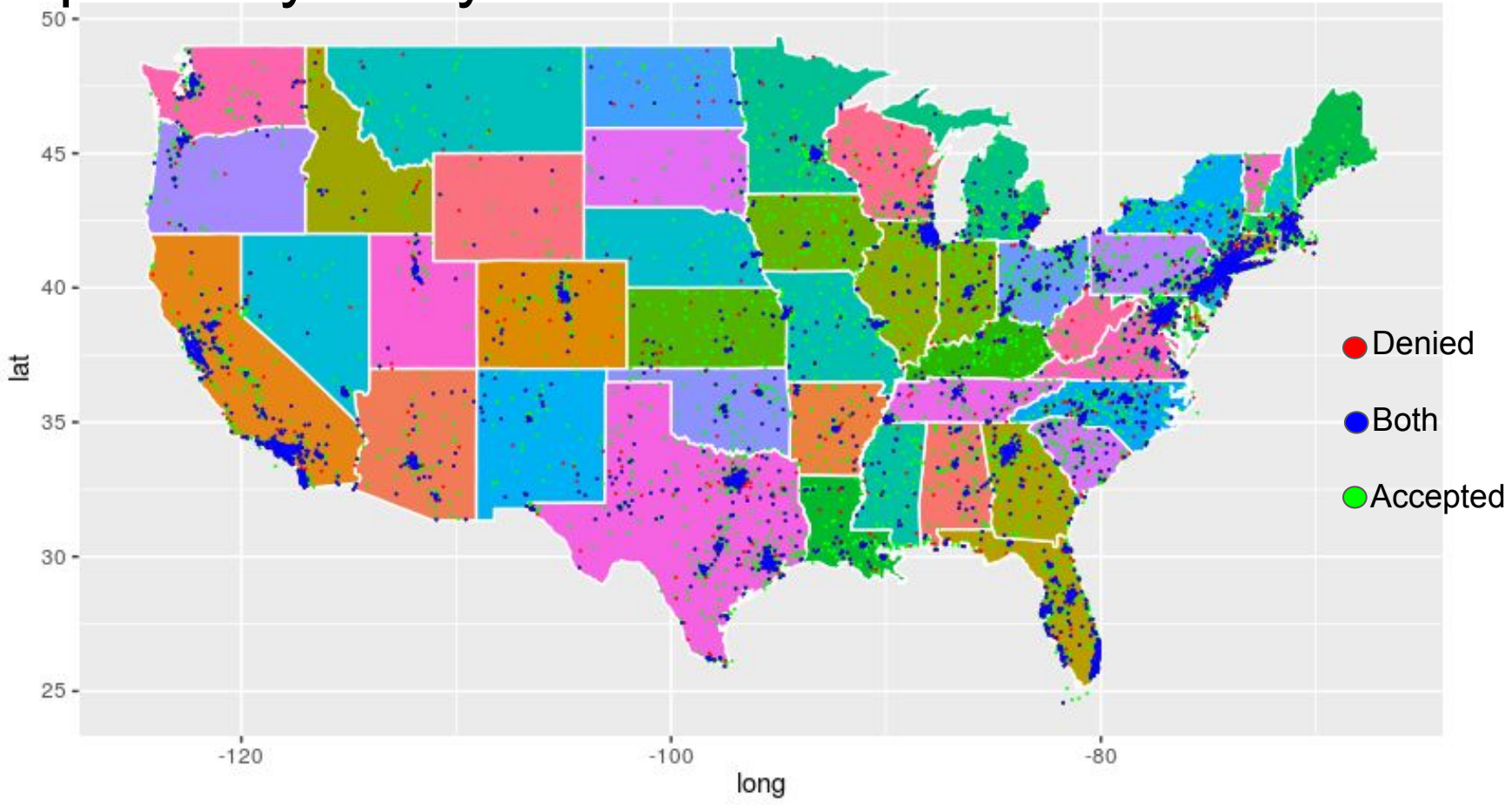


Exploratory Analysis

- Cognizant Technology Solutions Corporation had the most visa applicants by a fair margin though their accepted to denied ratio is rather low
- Certain companies appear to have slightly different Accepted Vs. Denied ratios (See Numbers On Graph)
    - Amazon, Apple, and Qualcomm all had fairly high accepted vs denied ratios



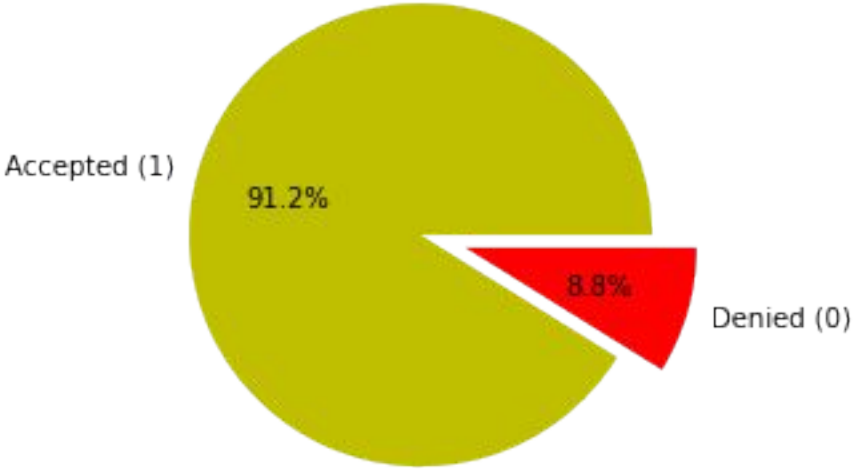
Exploratory Analysis



Classification Attempts

- Doing a naive classifier, and predicting 100% of the cases are accepted, we would get a 91.2% accuracy (pretty good for any classifier)
- This however would incorrectly label ALL denied cases
- Instead, we went with a label of how well did we classify the denied cases?

$$Accuracy_{Denied} = \frac{Number\ of\ Denied\ Guessed\ Denied}{Total\ Number\ of\ Denied\ Cases}$$

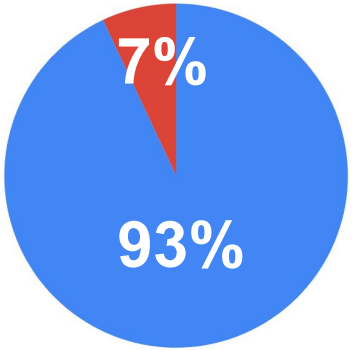


Data Features	Percentage of missing data
Case received date YEAR	65.8%
Application type	34.2%
Class of admission	5%
Education level	65%
Country of citizenship	0%
Economic sector	37.2%
Employer state	0%
CTC (Cost to Company)	0%

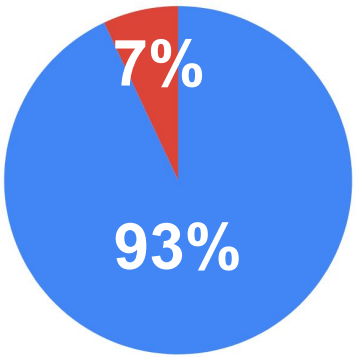
Classification Attempts (Initial Attempt)

- Trained a classifier (Decision Tree) on a random 70% of our data
- Tested on the remaining 30%
- Results are highly skewed towards labeling an applicant accepted
- We only get a Denied Accuracy of 7.06% (better than our naive classifier of predicting everything accepted but we can do better)

Data Trained On



Data Tested On

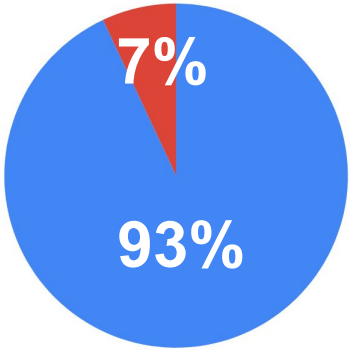


Percentage of denied cases predicted correctly.	7.06%
Percentage of cases predicted correct. (Accuracy Score)	93.72%

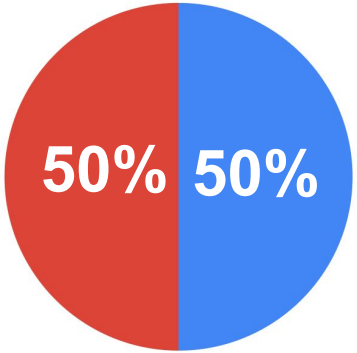
Classification Attempts (Continued)

- Trained a classifier (Decision Tree) on a random 70% of our data
- Tested on the remaining 30%
- Tested the same classifier from previous slide on an evenly sampled set of data, We can see that we're actually doing a little better on Percentage of denied cases predicted correctly

Data Trained On



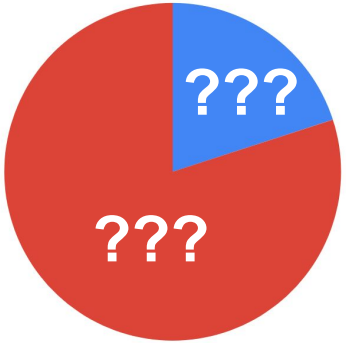
Data Tested On



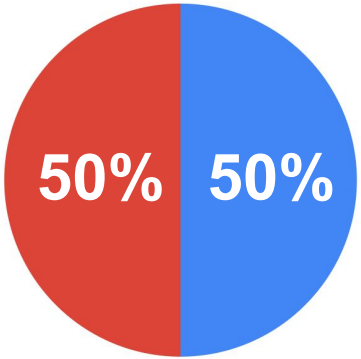
Percentage of denied cases predicted correctly.	8.51%
Percentage of cases predicted correct. (Accuracy Score)	54.03%

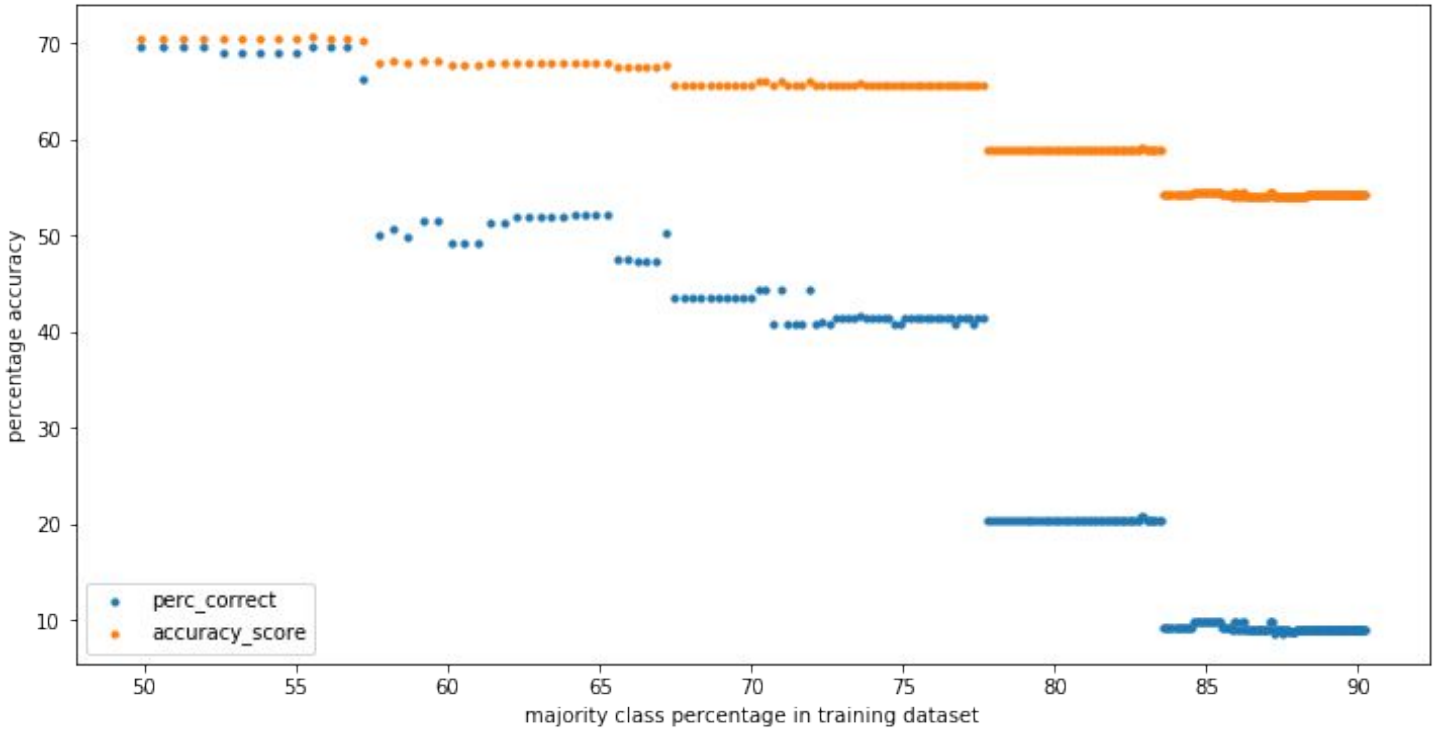
Classification (Sampling Problem)

Data Trained On



Data Tested On



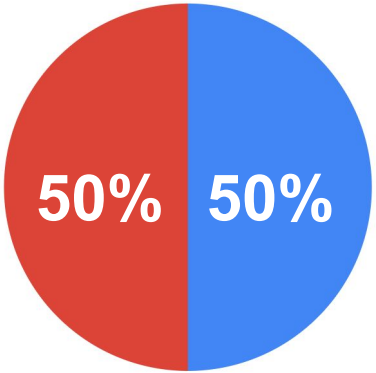


Classification Results (Continued)

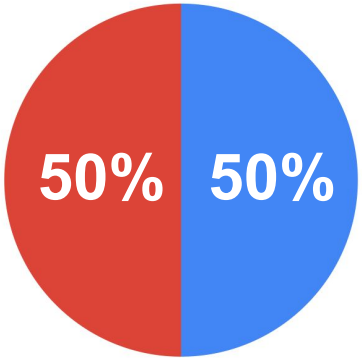
- How we got around our skewed data problem:
  - We trained on a random 50% Accepted vs. 50% Denied data set
  - Tested on another random 50% Accepted vs. 50% Denied data set
- Our Results can be seen below for various different classifiers
  - The Decision Tree Classifier ended up being the best classifier for our job (Perhaps due to our problem being entirely binary)

Classifier	Accuracy	
	Perc accuracy	accuracy score
Logistic Regression	48.62	50.24
Support Vector Classifier	49.12	50.74
Random Forest Classifier	**71.67**	**71.74**
Decision Tree Classifier	**73.88**	**70.12**

Data Trained On



Data Tested On



Issues Faced

Issue	Solution Implemented
Large amount of data (~291 MB) due to many fields	Removed fields we didn't think would help us
Many fields in the data were actually redundant fields that already existed (field name of "country_of_citizenship" and another field name of "country_of_citzenship")	Insured there were no conflicts between the two redundant fields and merged into one field
Large amount of the data was empty/NA/0	Empty numeric fields were populated with the average value across the entire dataset and empty categorical data was populated with NONE
Had many more data points for accepted visa applicants (87% of the data) and not many cases of denied applicants (8% of the data)	When training the classifiers, we tried differing ratios of Accepted to Denied and ended up using a 1:1 ratio that gave the best accuracy

Summary

Overall we were able to create a classifier that could accurately predict whether or not a visa application would be denied ~ 70% of the time. With this information both applicants and employers would be better equipt to make decisions regarding the visa application process. This could save them both time and money when deciding to apply, discouraging applicants that are unlikely to be selected from applying. Reducing the number of applicants applying would also save the US government time and money processing excess applications.

A higher accuracy classifier could potentially replace a large portion of that part of the government and improve the time between application submission and notification of application status.

Backup

Link to Kaggle Dataset used:

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