A report on

Prediction of US Visa Application Decision

under the guidance of

Mr. Venkatesan Sundaram

Submitted by

Asavari Limaye

14CO108

VII Sem B.Tech (CSE)

Kavya Atmakuri

14CO123

VII Sem B.Tech (CSE)

in partial fulfilment for the award of the degree of

BACHELOR OF TECHNOLOGY

in

COMPUTER ENGINEERING

Department of Computer Science & Engineering Technology

National Institute of Technology Karnataka, Surathkal.

November 2017

ABOUT THE PROBLEM

This dataset was collected and distributed by the US Department of Labor from 2012-2016.
The data includes information on employer, position, wage offered, job posting history, employee education and past visa history, associated lawyers, and final decision.
The motivation behind studying this dataset is to help US visa applicants gauge their chances of getting their visa approved based on various factors that have been taken into account in the dataset. We have drawn inferences about which factors play a major role in deciding the status of the applicant's decision.
However, with the recent changes (2017) in the US government, the prediction model that has been built will have to be trained with more recent data to provide realistic predictions.

DATASET

The dataset was available in CSV format with **374362 entries**, **152 features**.
Here, one of the features is the case status, which must be predicted based on other features. It can have 4 values:

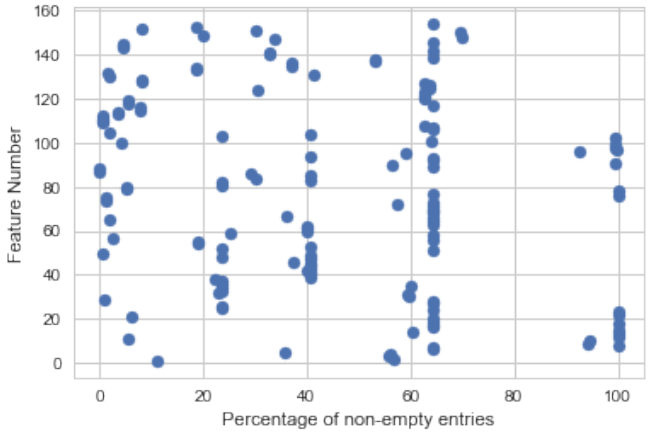
1. Certified
2. Certified-expired
3. Withdrawn
4. Denied

Since the Certified-expired class consists of applications that were Certified, we combine the 2 classes.
Those applications which were withdrawn were not included in the study.

Thus, there are 2 classes- **Certified and Denied**- which boils down this problem to one of **binary classification**.

FEATURE SELECTION

The following plot illustrates how filled the features in the raw data are:



The feature number is a unique ID given to each feature. This shows that a lot of the features have **less than 60%** meaningful data.

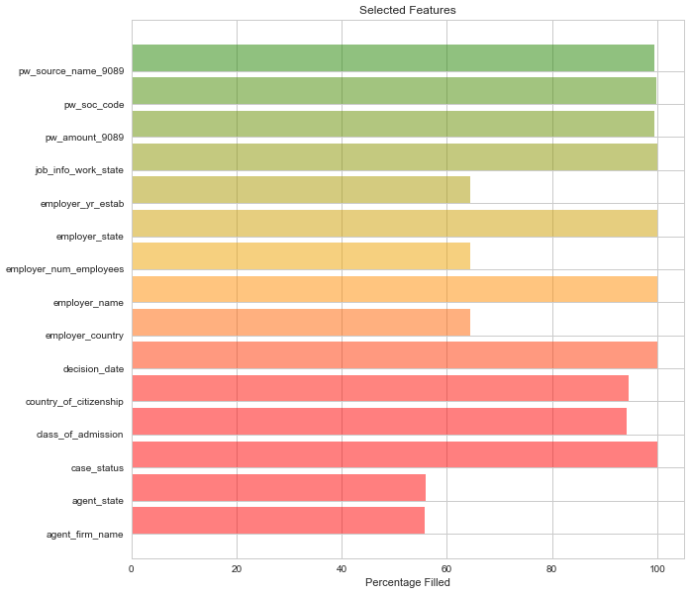
As there wasn't any meta-data provided along with the dataset, the exact significance of certain features was not evident. The features which are known to not contribute to the final visa decision, like Case Number were removed. Also, redundant features like Postal Code, State and City were eliminated.

Firstly, all features with less than 50% non-empty values were dropped.
Top **15 features** were selected using their correlation values (Pearson's standard correlation coefficient).
Each row (application) is identified by a unique Case Number (string).

The chosen features are:

1. agent_firm_name: Firm name of the agent handling the applicant's case
2. agent_state: State to which the agent belongs
3. class_of_admission: Type of visa being applied for
4. country_of_citizenship: Country which the applicant is a citizen of
5. decision_date: Date the application's decision was released
6. employer_state: State which the employer(company) is based out of
7. employer_country: Country in which the company was established
8. employer_name: Name of the employer (company)
9. employer_num_employees: Size of the employer's organisation
10. employer_yr_estab: Year the company was established
11. job_info_work_state: Location of the applicant's job
12. wage_source_name: Different pay grades specified by the government
13. soc_code: Profession of the applicant
14. yearly_wage: Yearly pay being offered to the applicant

The following figure shows how filled the selected features are:



We have given more importance to features with lesser empty values during selection, as predicting too many values is not recommended. As is seen, only 5 features out of 15 are less than 90% filled. These were chosen as they are intuitively important to the final decision.

The features which were not chosen to train the model either had more than 50% missing values or the information it conveyed was irrelevant to the problem.

DATA TRANSFORMATION

The dataset used cannot be used directly to train a classification model. A very large number of values are missing for some features, and many discrepancies in the data make data cleaning necessary.

Here are some of the challenges presented by the data:

1. For the features with less than 10% **missing values**, either the mode or the mean was used to fill up the empty values. In features like class_of_admission (Visa Type) missing values were replaced with the mode, whereas in features like decision_year, the mean of the data is used.
2. For the features with a lot of missing data, we created a new category for the missing data, 'Unknown' for eg. agent_state.
3. Some of the features had a type mismatch in its values. Appropriate **type conversion** was used.
4. In some features like year_established, some of the entries were too small to represent a **valid** year. These were rounded up to the year 1700.
5. The details of the salary had been given as two separate features, salary per unit time, and the unit of time. We **combined** these into a single feature which is uniform over all the rows and removed the old features.
6. All the features were converted to **categorical** numerical data.

PREDICTION MODELS

StratifiedKFold algorithm was used to repeatedly divide it into test set and training set. This ensures that the proportion of samples belonging to a particular class are about the same in the training and test set.

Decision Tree classifier was **implemented** in Python. Due to memory constraints, it was run only on a subset of the data (6,000 rows). This subset was chosen in a way that both classes were present in the same proportion as they are in the original dataset.

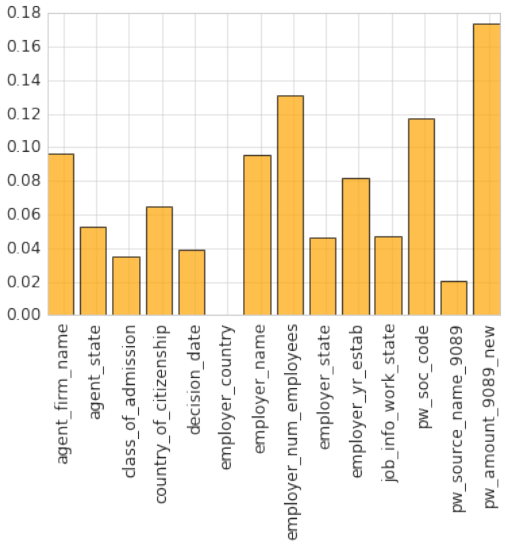
The following scores (**ROC_AUC**) were obtained: [74.395329441201, 75.0625521267723, 74.72894078398666, 75.81317764804002, 73.97831526271894]
Mean Accuracy: **74.796%**

Random Forest (using Python sklearn library) classifier is run on the train data for 5 different splits of the data. Then, the model which gives the maximum value for the **AUC** measure which was chosen.
The following results were obtained:

TRAIN SIZE	TEST SIZE	TRAIN ACCURACY	TEST ACCURACY
284934	71234	0.99724607629377848	0.71399796150858286
284934	71234	0.99791252225886806	0.62166131565618499
284934	71234	0.99750874227041253	**0.71467811132979886**
284934	71234	0.99809385703148046	0.65871122448907371
284936	71232	0.99802234670929657	0.70161322531428061

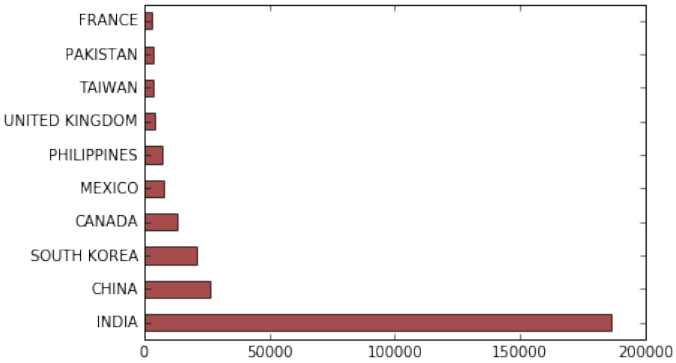
The following graph shows the relative importance of the features used during classification.
The top four features are:

1. Salary
2. Number of Employees in the Company
3. Type of Job
4. Name of the Company



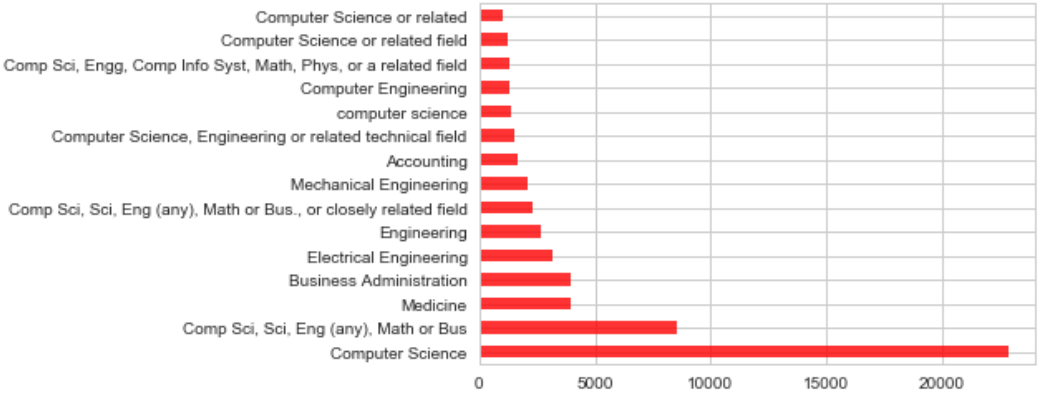
ANALYSIS

The following graph shows the Top **10 Countries** in terms of number of applications:

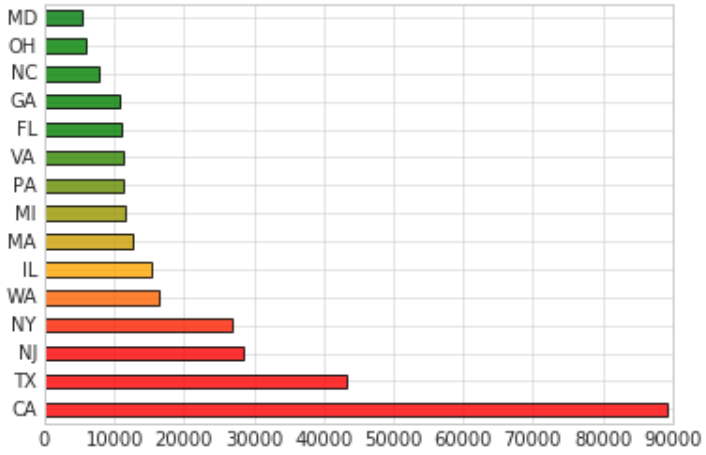


India is the source of maximum number of applicants for US visas.

Job-major based (top 15) distribution of applicants

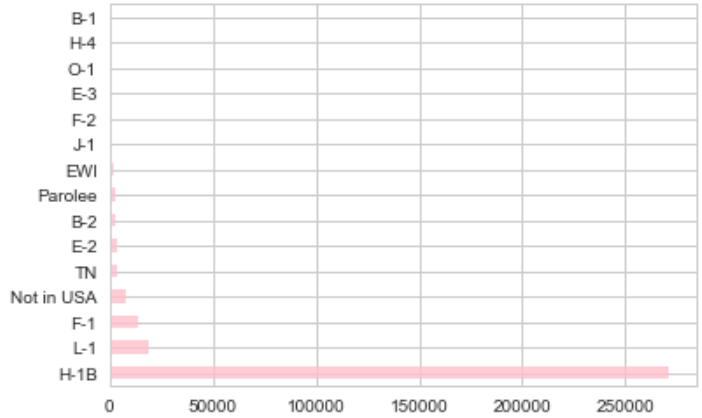


State-wise (top 15) distribution of applicants



Most applicants had job offers for computer applications/engineering jobs. As a result, California which houses Silicon Valley receives maximum number of applications.

Distribution of Visa Types Applied For



Maximum applicants are for H-1B visa which is issued by an employer on acceptance for a position at their firm. A visa was more likely to be obtained for a higher paying job.

CONCLUSION

The random forest and the single decision tree models have been used to predict the approval of a USA visa application. The dataset under consideration has many features, which were reduced to 14 important and complete features. Stratified K-fold cross validation was used along with the random forest model to find the best parameters for a random forest. Using these parameters, a random forest is trained, and an accuracy of 71% is got on the test set. A decision tree was implemented from scratch and trained on a balanced subset of the dataset giving an accuracy of 74.796%. It was observed that most of the applications for visa are of the H-1B type, and most of the jobs are in the computer industry.

REFERENCES

The presentation can be found at:
https://drive.google.com/drive/folders/1rHmurp3mQtpW77EVu8__cRrv47_ZebCD?usp=sharing

1. Dataset: https://www.kaggle.com/jboysen/us-perm-visas/data
2. Kaggle Blog: http://blog.kaggle.com/2016/07/21/approaching-almost-any-machine-learning-problem-abhishek-thakur/
3. Pandas Documentation: https://pandas.pydata.org/pandas-docs/stable/index.html
4. Sklearn Documentation: http://scikit-learn.org/stable/documentation.html