Prediction of Work Visa or H1b Approval using various classification techniques

Gopichand Kommineni
UFID : 03055523
gopichandk@ufl.edu

***Abstract*—Most of the United States Workforce has high-skilled foreign workers who rely on H1b Visa for working in the US legally. A job by a valid employer is required to submit a petition for H-1 Visa to the US immigration department. Most of the international students once they complete higher education in United States apply for this visa and begin working in a full time position in their specialised fields. Every year 85000 new petitions are picked up. If a person's petition didn't get picked up, he/she should go through the lottery. Although this process is done by a lottery system, approving these petitions are subjected to attributes like employer, salary etc,. A Person with good profile and salary has more chances of getting approved.In this project we try to develop and apply some machine learning predictors for predicting likelihood and risk of the new and continuing petitions.**

***Index Terms*—H-1B, Visa, Prediction, Machine Learning, Classifiers.**

I. Introduction

Every year thousands of immigrants across the world make their way into the United States. These immigrants require a Visa approved and issued by the united states in order to enter the country, One such Visa which his highly prominent and highly seeked across the immigrant community id H-!B visa. Every year there will be new allocation of 85000 new visas in H-!B category. These 85000 visas are picked up by a lottery system which works slightly in favour of students with higher degrees in United States colleges and Universities.
This entire process is done by an independent entity US immigration Service (USCIS). A certifies U.S employer should file the petition on behalf of the employee with a heavy visa fee. Even though the selection process is done by lottery, the final approval of the visa is subjected to the attributes of the profile of the employer and employees. These attributes are salary, employer reputation, location of the job, conditions of the year,type of job and designation. Even though the approval is done by the immigration officer they try to analyse a petitions potential to stay in united states based on the aforementioned attributes.

In this project we are trying to predict whether a person's H-1B visa can be certified or denied based on the attributes. For this project we are using data set collected from the kaggle which has over 3 million data points with 11 attributes. We are exploring different classification techniques and trying to gain some insights from the data we presented.

II. Description

For this problem we are taking an approach of applying different classification techniques on the Data Set we obtained. Before applying the classification techniques we need to get an understanding of the data and what need to process. Since most of the data present in the Data set is categorical and has strings as input values we need to change them into numerical values for the model to perform smoothly without any issues.The detailed list of techniques and preprocessing is given the upcoming sections. After preprocessing we need to split the data set and apply classification models on the training data set. Since the outcomes of the data set are mapped to 0,1 what we are performing is simply binary classification.
After training the model we use the model to predict the results of test data set and compile the accuracy. Since we are using classification techniques we can compute the confusion matrix for generating precision and recall.

III. Related Work

This problem statement is not well explored since there isn't a well defined structure for the process of predicting the Visas. All the visa approvals and rejections are done in person by USCIS officers and one cant give a structured definition on what parameters they choose for making the decision. However based on the existing data one can try to gain a insight on what parameters could have how much impact on the decision.
One report showed performing kmeans clustering and decision tree for classifying the decisions[16]. This report gave a good picture of hoe the data is distributed and this is performed on a sample of data instead of whole data set. Report[17] has performed all the given classifier techniques, they performed preprocessing which is different from that I performed. Their techniques has sounded more reliable than which I performed. There are some united studies that performed predictions on data set taken from single year with 500000 data points which used many scoring techniques to get the rating the each company,state,job title and Wage of the petition
Even though these methods could give some insights of the data we could not rely on the techniques since at the end every decision is made manually. What we could do is to gain some knowledge and prepare ourselves for the day.

IV. Data Set

The Data set we are using for this project is taken from kaggle.This Data set has over 3 million records of petitions. Each petition contains the following data

- **Unamed: 0:** This contains the petition id of the employee.
- **CASE_STATUS :** This gives the information of the final outcome of the petition.
- **EMPLOYER_NAME :** The Name of the company which is filing the petition on behalf of the Employee.
- **SOC_NAME :** Occupation name that is categorised into occupation codes.
- **JOB_TITLE :** The Designation of the Employee.
- **FULL_TIME_POSITION :** Information whether the job is full time time or part time.
- **PREVAILING_WAGE :** The yearly salary of the employee in dollars.
- **YEAR :**The year in which petition is filed.
- **WORKSITE :** The city and state where the job is located.
- **lon and lan :** The longitude and latitude of the location.

Since the data set has 3 million records there is a lot of noise and missing values in the data set.
Our output or target value is the "CASE_STATUS" in which there are 7 classified.

- **Certified :** If the visa is approved then it is classified as certified.
- **Certified-withdrawn :** If the visa is approved and it is withdrawn from consideration by the Employee or Employer.
- **Denied :** If the Visa pitetion ios rejected by USCIS
- **Withdrawn :** If the pitetion is withdrawn from consideration.
- **Unassigned :** If there has been a hold on the decision due to lack of documents.
- **Rejected and Invalidated :** This is also categorised as denied.

There are a lot of unique value ranges for other parameters in the Data set.

V. Data Pre Processing

Since the Data set has so many missing values and noise, we need to perform a lot of Data cleaning and fine tuning the data.
In this part we will explore how we can convert certain attributes from text values to numerical without loosing the necessary information.

A. Preprcoessing Case Status

Since the Case status is the target, we need to clean the attribute. The attribute has 7 values and our goal is to reduce them to 2. Among these values we have "CERTIFIED_WITHDRAWN" and "WITHDRAWN". These two values are not useful to us since they are voluntarily removed by the employee or Employer. We can also convert "CERTIFIED_WITHDRAWN" to "CERTIFIED" since they are both approved, but we are having a lot of data and we try to balance the data as much as we could. So, we drop those records which are "WITHDRAWN" and "CERTIFIED_WITHDRAWN".
After dropping these records we have Rejected. invalidated and unassigned. We can safely convert them into denied since there are very small number of these petitions and they could do have any direct impact on the model. After making all these conversion, we finally map the "CERTIFIED" to 1 and "DENIED" to 0, because we only have two outputs in the classification models.

B. Preprcoessing Employer Name

This is a tricky part since we are dealing with the text values instead of numerical values. The Employer reputation significantly effects the approval of the petition. So for processing this column we first collected data of fortune 500 companies which are listed in the stock exchange. Exploratory Data Analysis showed most of the fortune 500 companies filed for most of the visa petitions and these pitetions have a higher success rate than other companies. So We mapped all the companies in the 500 list to 1 classifying them as one category.
EDA also showed that University sponsored H-!B visas are all almost approved except some which are withdrawn. Hence we need to classify universities also into category of the fortune 500. For doing this we found the Employer name that has university in it and assigned a value 1 to it. Rest of all the values in Employer name is mapped to 0.
Even though this is not the ideal case to process the employers since we need to consider all the employers it is exhausting as there are236013 unique companies sponsoring the visa. Another alternative processing is finding the number of application for each company and determining their success rate based on that.

C. Preprcoessing Worksite

Worksite is one of the important aspects that effect the approval of the application. Most of the worksites with enough companies and amenities appeared in the data set has high approval rating. The "WORKSITE" in the Data set has location city and state in it separated by a delimiter. Since city and state has almost same weightage as city is located in the state, we can discard city and keep the state.
After keeping the state in the column we need to give weightage to states that have high success rates. For attaining this first we will find out how many application each state has filed.
After finding the number of applications we clustered them into categories such that States with high petitions and high success rates are classified as one cluster. we clustered them into 5 clusters and assigned corresponding values to the state

in the worksite column. We performed one hot encoding since it is a categorical column and discarded the original column since we don't need it anymore.

As we have 1 dimension data so we don't use traditional kmeans to cluster the data. So we use a special library called kmeans1d which uses mean shift and kernel utilisation to cluster the data.

D. Preprcoessing Prevailing wage

Salary is the basic commodity needed to survive at a workplace. Salary and visa approval rate often tends to be proportionate to each other. EDA showed there are is a high chances of getting approved if the Employer is providing high compensation. The mean salary of the prevailing wage is 65000 usd
The 98^{th} percentile of the prevailing wage is capped at 138000 usd. After removing some outliers which have hundreds of millions as compensation we capped everything that is above 138k to 138000 and everything below 34000 to 34000 to fit the data into 2^{nd} and 98^{th} percentile in the distribution.

E. Preprcoessing YEAR

The year in which the petition was filed is given in this column. The year of filing petition also could effect the procedure since in some years some constraints in the government and administration have some underlying impact in the approval rate. The years are simply one hot encoded.

F. Preprcoessing Full time position

There are only two values in this column namely "Y" or "N". Full time position means the worker is working only on this job full time and there is no time limit on his employment until fired. Workers who are working for a limited amount of period doesn't come under this category.

G. Preprcoessing Job title

As the name implicates it gives the name fo the job the person is given. There are 238175 unique jobs in the column and some of these columns have high applications and high approval ratings. For job titles we did the same preprocessing as we did for the states. We performed 1 dimension data clustering for clustering the jobs based on their number of petitions. we used 10 clusters for this and mapped each job to its corresponding cluster.

H. Preprcoessing SOC NAME

Soc name is the name associated with e corresponding soc code the job title has. After performing EDA we attained that all the Soc names could be categorised into 10 different categories. We manually mapped the soc names into these categories independently.

I. Preprcoessing lat and lon

These correspond to the latitude and longitude coordinates the job is located in. lat and has over 170000 missing values each. Moreover worksite has all the necessary information regarding the latitude and longitude so we can safely discard these columns.

After cleaning and preprocessing each attribute we can finally drop unnamed 0: and missing NAN values from the data set.

VI. Models

After completing all the preprocessing the main problem is applying classification models to the cleaned data, For now we have Decision Tree, Random Forest, Neural Network, Neural network with regression, AdaBoost, Naive Bayes, Linear SVM, Logistic regression, Logistic regression with elastic net regression.

A. Naive Bayes

In Naive Bayes all the features are considered independent to each other. Naive Bayes simply calculates the feature importance with respect to the class by taking the probability of the feature in the class. it calculates $p(x|y == 1|\theta), p(x|y == 0|\theta)$ and $p(y)$. Naive Bayes finds the Maximum likelihood estimation or Maximum a Posterior for the features using the following formula for $p(y == 1|x)$ = .

$$\frac{\prod_{i=1}^{m} P(x_i|y==1|\theta)p(y==1)}{\prod_{i=1}^{m} P(x_i|y==1|\theta)p(y==1)+\prod_{i=0}^{m} P(x_i|y==1|\theta)p(y==0)}$$

Naive Bayes is very scalable and can be used to deal with very large data sets with less number of outcome types. It is simple to implement and also very powerful enough to exceed many sophisticated algorithms.

B. Support vector machines

Support vector machine tries to classify the problems by maximizing the distance between planes of each class. First svm finds the line or hyper plane that separates two classes and them the maximum likelihood estimation tries to maximize this distance .

$$min_{\alpha,\gamma,bi}\ \frac{1}{2}\ ||w||^2$$

$$\text{s.t. } y^{(i)}(A^T x^{(i)} + bi) \geq 1, \text{ i = 1,2,,,,,,m}$$

Support vector machines are easy to implement and code yet powerful enough to increase the accuracy when a right kernel is selected. Even though svm is not suited for large data sets. It significantly performs well on data sets with high dimensionality. If we add the regularization term to $l(\theta)$ we get "linear SVM with regularization. Ussually we add l2 reg norm which is $\frac{1}{m}||\theta||^2$.

C. Logistic regression

One of the most important and most used classifier is logistic regression. Logistic regression is similar to linear regression but it only gives discreet values. i.e, logistic regression is build

to deal with data that has discreet output values by performing regression.

$$h_\theta(x) = sign(\theta^T x) = \frac{1}{1+e^{-\theta^T x}}$$

$$\text{P(y=1;x;}\theta\text{)} = h_\theta(x)$$

$$\text{P(y=0;x;}\theta\text{)} = 1 - h_\theta(x)$$

$$l(\theta) = \sum_{i=1}^{m} y^{(i)} log(h_\theta(x^{(i)})) + (1 - y^{(i)} log(h_\theta(1 - x^{(i)}))$$

Where $l(\theta)$ is the likelihood estimate of the data.

If we add the regularization term to $l(\theta)$ we get "logistic regression with regularization. Ussually we add l2 reg norm which is $\frac{1}{m}||\theta||^2$. Regularization is done to prevent over fitting the model.

D. Neural Networks

Neural Networks are the most powerful and robust model ever developed in Machine Learning. Neural networks consists of a node which is often called perceptron with input weights. These weights are processed by an activation function which resides in the perceptron.The output of the perceptron is transmitted as input t o nodes in the next layer. Neural Networks form complex structures with hidden layers and nodes. We perform forward propagation and then back propagation to update weights and bias in the Neural Networks.

if we have an activation function of logistic regression then typical functions in the Neural Networks are as follows.

$$L(\hat{y}, y) = -[(1 - y) log(1 - \hat{y} + y log(\hat{y})]$$

$$Z^{[1]} = W^{[1]} X + b^{[1]}$$

$$A^1 = g(Z^{[1]})$$

$$Z^{[2]} = W^{[2]} A^{[1]} + b^{[2]}$$

$$\hat{y} = g(Z^{[2]})$$

We update the Weights and bias based on the type of optimizer used here we are using SGD hence we update single weight and bias individually.

$$W^{[i]} = W^{[i]} - \alpha \frac{\partial L}{\partial W^{[i]}}$$

$$b^{[i]} = b^{[i]} - \alpha \frac{\partial L}{\partial b^{[i]}}$$

For getting regularized Neural Networks we add the regularized weights while updating the weights in backpropagation. the updating functions for regularized neural networks looks like this.

$$W^{[i]} = W^{[i]} - \alpha \frac{\partial L}{\partial W^{[i]}} - \frac{1}{m}||W||^2_{2:m}$$

$$b^{[i]} = b^{[i]} - \alpha \frac{\partial L}{\partial b^{[i]}} - \frac{1}{m}||W||^2_{b:m}$$

E. Decision Tree

Decision tree is a classification technique that follows tree type data structure. At each node Decision tree makes a decision based on the mathematical function and then makes a decision to go Left or right based in the function output.
The leaf nodes in the decision tree often represent the final classes of the data. Decision trees can be used for classifying both categorical and numerical data.

Loss : loss function is a mathematical function used to evaluate the loss or mismatch occurred in each iteration. The final goal; of any ,model is to reduce the loss as much as possible. Here we are using **hinge loss** for computing the loss in all classifiers other types of losses include mean square error, Huber loss etc.,.

Optimizer Optimizer is the technique used to update weights or parameters during the training of the models. There are many optimizers like proximal gradient decent, ADAM etc., Here we are using Stochastic Gradient Descent. In SGD we update the parameters after iterating a data point.

VII. Experiments and Evaluation

While implementing the project we downloaded the data set from kaggle and performed the data cleaning mentioned in preprocessing section. After performing the preprocessing we divided the data set into 60 40 for training and testing the dataset.
Before splitting the data we tried to make a balanced data set by keeping number of certified petitions equal to number of denied petitions.
We imported the Naive Bayes, Decision Tree, AdaBoost, Random Forest, Neural Networks, Logistic Regression, Support Vector Machines classifiers from the scikit learn library and applied them to the training data.
After attaining the model we predicted the outcomes on the test data and calculated the accuracies using the metric scores from scikit-learn. We tabulated accuracy, precision and recall for both the balanced and unbalanced data set

Classifier	Accuracy	Precision	Recall
Decision Tree	0.96678	0.96690	0.99987
Random Forest	0.96742	0.96689	1.0
Neural Net	0.96689	0.96689	1.0
Neural Net with reg	0.96742	0.96689	1.0
Adaboost	0.96742	0.8923	1.0
Naive Bayes	0.89274	0.97268	0.91432
SVM	0.967220	0.96689	0.99975
SVM with reg	0.96742	0.96689	1.0
logistic	0.96716	0.96690	0.99969
logistic with reg	0.96742	0.96689	1.0

Table for Balanced Data set

Classifier	Accuracy	Precision	Recall
Decision Tree	0.70495	0.73299	0.84544
Random Forest	0.70587	0.70416	0.91447
Neural Net	0.70078	0.70473	0.91420
Neural Net with reg	0.71058	0.72913	0.86520
Adaboost	0.70746	0.73110	0.85264
Naive Bayes	0.68507	0.72879	0.80472
SVM	0.65383	0.66002	0.94290
SVM with reg	0.67804	0.70996	0.83539
logistic	0.46216	0.708357	0.26627
logistic with reg	0.69038	0.69012	0.93295

Table for Balanced Dataset

VIII. Conclusion

By conducting this project we can conclude that we can predict the outcomes of the H-1B visa petitions to some extent. From the obtained results we can see that the difference between accuracies is significantly very small for unbalanced data set. Although the accuracies are high for unbalanced data set is about 96 to 97 percent the difference between accuracies of different models is of 10^{-5} scale. In these accuracies Neural net has the most accuracy since it is the most sophisticated and powerful algorithm. The precision and recall rates are also very high for the unbalanced data set since there are hundreds of thousands of data points.

For the unbalanced data set we created positive and negative data points with size proportion i.e, we have taken positive samples twice the size of negative samples. Since the data set is Small and negative samples play equal role in the model we have the model accuracies at about 60 to 70 percent in the classifiers and we have significant difference in accuracy scores between the classifiers. Of all the classifiers Neural Net with regression has best perform for obvious reasons and logistic regression has the least accuracy score. The precision and recall rates are also significantly lower when compare to the unbalanced data-set.

References

[1] https://www.datacamp.com/community/tutorials/predicting-H-1B-visa-status-python.

[2] https://www.kaggle.com/nsharan/h-1b-visa

[3] https://datahub.io/core/s-and-p-500-companiesdata

[4] "High-skilled visa applications hit record high," CNNMoney. [Online]. Available: http: //money.cnn.com/2016/04/12/technology/h1b-cap-visa-fy-2017/index.html. [Accessed: 20-Oct-2017].

[5] "Predicting Case Status of H-1B Visa Petitions." [Online]. Available: https://cseweb. ucsd.edu/classes/wi17/cse258-a/reports/a054.pdf.

[6] H-1B Visa Petitions 2011-2016 — Kaggle. [Online]. Available: https://www.kaggle. com/nsharan/h-1b-visa/data. [Accessed: 20-Oct-2017].

[7] "Using Text Analysis To Predict H-1B Wages," The Official Blog of BigML.com, 01-Oct2013. [Online]. Available: https://blog.bigml.com/2013/10/01/using-text-analysisto-predict-h1-b-wages/ . [Accessed: 20-Oct-2017].

[8] J.Ross Quinlan, "C4.5: Programs for machine learning", Elsevier, 2014.

[9] https://github.com/belizgunel/cs229proj