How Different Factors Affect U.S. PERM Application Approval

Data Mining Final Project - CSCI5502

Yushuo Ruan
yushuo.ruan@colorado.edu
102272546
In-Class Section

Xiao Xiao
xiao.xiao@colorado.edu
105966261
Distance Section

Dongyao Wang
dongyao.wang@colorado.edu
103129526
In-Class Section

Tianshu Pang
tianshu.pang@colorado.edu
109258729
In-Class Section

ABSTRACT

The paper focused on the different factors th at may affect the approval rate of the application for U.S. permanent labor certification (or PERM, for short), which may be interested to those who are willing to or plan to legally work and live in the United States permanently. During the latest decades, the number of population to apply the U.S. permanent visa, including students, engineers, businessmen and families, has been remarkably increasing year by year. This makes the U.S. to be a biggest immigrant country. However, the visa system is the gateway that controls the entrance to the country, since people who enter foreign countries, including the U.S., must hold a valid visa. Moreover, the application for permanent labor certification is an even longer process which may take much time to prepare for the paperwork and wait for the final decision. Many prior works regarding this problem provide us insight toward problem of impact brought by this immigration boom, but without providing a solution or suggestion to the applicants, which is the approach of our work. We evaluate and visualize the important attributes in the dataset which can provide applicants with practical suggestions while applying. We finally build a predictor to predict applicants' final case status by an 84% accuracy. The result of this project might not work for too long since the immigration policies are changing all the time, but the model can be trained over and over again using the new coming in datasets. Our process is valueable for future pridiction.

KEYWORDS

Permanent labor certification, approval rate, immigration control, data mining, data visualization, wage, position

1 INTRODUCTION

The percentage of the number of international students studying computer science has been increasing rapidly in the past few years. Finding jobs and applying for H-1B visa has been the most popular topic among our computer science international students. According to the data[10] from last year, the approval rate of U.S. H-1B working Visa application is less than 60%, which is really unstable and seems to decrease year after year. At the same time based on rumor about Trump's future immigrant policy the filtering process of validating applicants will be more strict and hard to be approved. The same goes for the F-1 visa which is for foreigners who want to study in the U.S.. For almost all types of visa applications, there are always different kinds of factors that might make the certification hard to made. Therefore a very serious question has been raised, which is for people who want to go to America, how they could increase the possibility and probability of getting their Visa or PERM application approved.

Additionally, as for international students who hold the F-1 visa, they do care about the process of applying for the H-1B visa as well as the approval rate in order to be eligible to work in the U.S. after the expiration of OPT (Optional Practical Training). The main purpose of this paper is to find some patterns and relationship between different factors, as well as extracting useful information related to the PERM application approval rate from the dataset through the methods of data mining and data analysis.

For the PERM application with respect to different types of original visa, there are many different kinds of factors that we need to concern about and these factors weigh very differently for different candidates of course. For example, the financial condition of the candidates, the credits of the university the candidates applied for or the competition of the company that is going to sponsor the employee. Sometimes even the nationality of the candidate matters in this process which we do not like. There are so many variables that could affect the final results of the PERM application analysis. But among all of the factors, there are definitely some factors that always have a huge impact on whether the application will be approved or not. Therefore, in our study, we are going to find the hidden correlation between these factors and the certification results for U.S. PERM application using data mining methods and tools.

The main contributions of this project include:

- We first analyzed the situation of U.S. application for permanent labor certification, such as the H-1B visa application. Inspired by this, we found some dataset from the U.S. Department of Labor, which contains the U.S. PERM application cases between 2013-2018, and we'd like to find some patterns and get useful information related to the PERM application approval rate from the data. A lot of related work, referring to section 2, has been explored to understand the U.S. PERM application cases (including certified, certified-expired, denied, and withdrawn) from different aspects.
- The most important thing at first is to mine the dataset, and we did a variety of data processing on the dataset to make the analysis easier, including data reduction, data cleaning and

data integration. Then, we analyzed the attribute features in order to provide key attributes to SVM model for further usage.

- To evaluate the factors that may cause the approval rate or cases' decision of PERM application, we used data visualization to find the outcomes more directly including boxplot, histogram, pie chart and spatial mapping. For instance, the higher average income from a field (e.g. computer and mathematical occupations) is usually related to a larger number of PERM applicants. In addition, the higher wage offer may cause higher probability of the approval rate. Overall, we interpreted these patterns in real life, did prediction, visualized the relationship between different factors.
- In the construction of SVM model, we modified the data set in order to fit them into our classifier well. We transformed data type to make them readable. We also transformed numerical data into categorical data using one-hot-encoding technique. Then, we split out data into train and data set and fed them into our classifier as well as comparing the influence from each individual features we picked and chose proper candidates.

The rest of the paper is organized as follows. Section 2 includes review of related works categorized by different issue. Section 3 describes the methodology we used for this project, including dataset, tools, data pre-processing, and attribute selection process. Section 4 demonstrates the evaluation of out data including visualization and classifier. Section 5 discusses the result we obtain from our work. Section 6 is the conclusion.

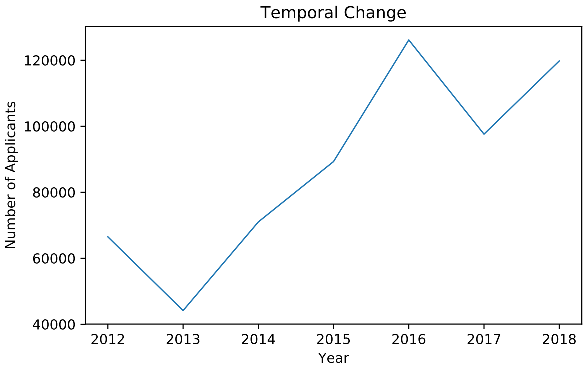

Figure 1: Temporal change of number of PERM Visa application from 2012 to 2017

2 RELATED WORK

2.1 Visa Issue

Visa issues always interest non-Americans who try to find opportunities in the U.S.. As a result, there are many prior works regarding the application approval issue, including trends for different types of visa in different states. The deeper analysis paper regarding the similar issue also include topics such as the equality of Visa approval for different races, the relationship between PERM application approval and economy, and the process of imposing and limiting the application approval.

The paper "Unequal access to foreign spaces: how state use visa restrictions to regulate mobility in a globalized world" written by Eric Newmayer[7] discussed issues derived from visa restrictions and immigration control. As Eric said, "there can be no doubt that cross-national movement, both short-term and long-term, has increased dramatically over the last three decades", the product of this phenomenon is the passport system, which is widely used among many countries nowadays. Passport system bring countries more security while it controls the quality of entering foreigners, but it also limits the freedom of international movements. Eric discussed two questions, why states impose visa restriction and why states refrain from imposing visa restrictions. In the paper, Eric did some Empirical analysis to further discover the pattern of visa restriction statistically. Eric provided a table which includes some statistical values about visa restriction, restrictions to political conflict, number of terrorist attacks, bilateral trade, and so on. And he also provided the estimation results and predicted probabilities of visa restrictions. As a conclusion, Eric mentioned, "states remain willing and eager to systematically keep out passport holders from certain nations". Through the analysis, we can see that the visa restriction to OECD countries is obviously looser than the restriction to some countries which are poor and have history of violent political conflict. This conclusion provides us a foundation of what we can further analyze in our data mining project.

Another paper "Controlling access to territory: economic interdependence, transnational terrorism, and visa policies" written by Nazli Avdan[1] also discussed visa restrictions, but from the perspective of its relationship with external factors such as economy and security. In the paper, Nazli tries to compare the impact on visa restrictions brought by economy interdependence and security concern. He argues that "economic ties affect visa policies though a reconfiguration of preferences and the opportunity costs of economic loss and by tempering the impact of terrorism". In the paper, Nazli also provided tables demonstrating relationships between transnational terrorism, economic interdependence, and visa restrictions, and the tables analyze it in two respects, interaction effects, and additive effects. He also provides plot about marginal effect of standard deviation increase in terrorism on visas as trade changes. As the conclusion, this article provides us a way of understanding the relationship and balance between "satisfying material objectives and maintaining security". It give us a great insight into the visa restriction through the view of globalization.

2.2 High skill immigrants

There were also some studies related to the immigrants with high skills in technology fields such as STEM field. There was one paper "Analysis of US Census and permanent Resident Data for High Skill Employment", written by Rajesh Kumar [5], analyze the impact to the United States economy and the native workers brought by the high skill immigrants holding H-1B visas or permanent labor certificate, or so-called green card. Rajesh states that although high skill immigrants might have certain negative influence to the low skill native workers due to the job opportunity equality, the impact on the economy of the United States brought by these immigrants is so great in a positive way such that it could make the negative impact ignorable.

Education wise, “More than 11 percent of foreign -born workers have advance degrees - slightly above the fraction of Americans with post-college degrees”[5]. Kumar’s paper’s dataset involves job types, work location region, and the level of skill, which are also included in this paper. At the end of the paper, Rajesh concluded that, visa and PERM certification are likely to favor those who are well educated, he conclude that “Roughly 28 percent are allocated to persons with extraordinary ability: outstanding professors and researchers and multinational managers or executives. Another 28 percent are assigned to workers with advanced academic credentials or workers with exceptional ability in business, science, or arts. This result gives us a great insight of the outcome of job type analysis. Other factors that could influence the application results will be examined in later part of this paper.

2.3 Impact on wage

Several studies focus on the impact on wages brought by immigration in the past decades. Ottaviano and Peri state that the real wages of U.S. were influenced negatively due to the immigration of low skill workers, who were often uneducated and undocumented.[12] In 2002, U.S. Census showed that the wage impact in the U.S. due to immigrated workers was about 3 percent.[2] However, due to the massive growth of immigration, especially illegally immigration, in the past few years, this percentage must also grew significantly. The number of PERM application from different types of jobs and their average wage will be analyzed in later part of this paper. A paper “The Impact Of Experience on Wage Premiums for Permanent Employment-Based Visa Applicants in the Information Technology Sector” written by Scott Schonberger[14] focused on the wage premium, which indicates the gap between two kinds of wages for foreign workers, one is called prevailing wage, which is the average wage for a specific position determined by the department of labor, and one is the wages offered by sponsoring companies for the same position. In this paper, we will also calculate the average wage for different type of jobs and also their approval rate. In our trained classifier, the wage will also be a very important attribute.

Although we can get a more well-rounded view of the current phenomenon and issues regarding immigration process and impact by getting a glance of the prior studies, we also find something that is not covered in these works. They developed nice insight of problems, but they didn’t provide a solution to the actual immigration applicants, and this is what we will do in our project. We are going to put the datasets into use and develop a trained model to predict if a applicant is likely to be denied, and we also provide information gain of attribute, which can tell the applicants what matters the most while they apply to the Perm program. These studies can lead our group to a more valuable topic which we will focus on later.

3 METHODOLOGY

After early preparation, we will start digging out valuable information which we think mostly reveal several phenomenons that interest us behind the dataset. From historical perspective, knowing the trend towards the applications of each Visa type and the trend towards the approval rate of each Visa type might help us to approximately trace what the immigration policies and background were back then. From political perspective, analyzing the PERM application approval rates for applicants from different countries/regions may disclose federal government’s attitude towards each country/region and bilateral relations with each country/region at certain time. From international students perspective, we care about each year’s employment Visa/PERM statistics. This could include reports such as employment application approval rates, the amount of employment applications from each professional field, total employment Visas or permanent labor certification issued for each professional field in a certain year, and etc. Furthermore, we expect to build a model that can functionally forecast employment PERM application approval rate and expected waiting time for a virtual applicant with appropriate background setting.

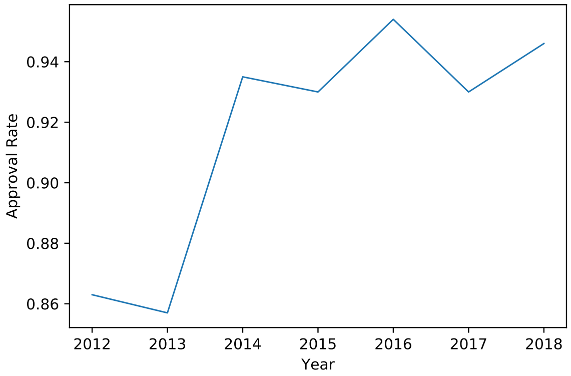

Figure 2: Temporal change of pass rate of PERM Visa application from 2012 to 2017

The model we are going to build varies depend on what kind of data information we are going to build and what attributes we are going to use. The most basic model we could take a try is that a model predicting whether a candidate could pass the PERM application or not based on some of the attributes listed in our data set that we considered to be important. During this process the training process and the picking of a proper classifier algorithm could be tricky case the final result we are looking for is a binary value. Another model we could build is the classifier to predict the trend of application passing rate within a certain time of period or a certain set of bounded contexts defined by some of the attributes in out source data. For example, we could predict between 2014 and 2015 how the trend of the application passing would be based on all the mod values entered as all of its correlated attributes.

A very big problem we are going to have during our classifier training process is that we have too many attributes in our data set which is more than 100. So, we need to find out what value is the most important attributes to out model. Also, at the same time some attributes might have a too big influence on our classifier that our classifier does not really care about the other attributes and the same goes for the attributes that have big influence on our classifier but not being valued enough. Therefore, we might need to do some feature engineering to our data sets. For example, we are going to assign weights to our attributes in our classifier to make it unbiased toward different type of situations. Not mention a lot of attributes are not valuable at all to our classifier therefore for this type of attributes we are just going to ignore them by assigning 0 to their weights. At the same time, we should not be able to confirm the relation between our attributes and our results is basic linear relations. Therefore, we might need to configure the relation used in our classifier to make it more precise for our model. Which means we need to use Support Vector machines algorithm if possible and we might need to do some kernels to our training process. And in the end, We might need to try different classifier algorithms to find the best one suitable for our prediction models.

3.1 Dataset

Our data is collected and distributed by U.S. Department of Labor. It includes the PERM applications' case NO., case status, class of visa, the applicants' country of citizenship, the case decision date, final decision. It also includes applicant's job and past job information such as employer's address, employer's name, wage offered, economic sector, job title, job training field and job posting history. It also includes more personal information like past education institute and associated lawyers. Through this dataset, we can have deep understanding in the situation of U.S. PERM application. Based on the analysis of these data we can get the trend of decision of PERM applications especially for year 2017 and 2018. Through our team's data analysis and trend charts, we can roughly predict the final decision of a U.S. PERM application.

3.2 Tools

The whole data is all included in one csv file, and the most frequent tool used to handle this kind of file is Microsoft Excel, because Excel can show data in a form and can provide many functions that can help get some high level ideas through the whole data. Because of its powerful data processing functions, we can grab some general knowledge about this "U.S. PERM application" data set in a short time. We also need other data analysis tools to mine the data, because a big amount of data usually means more interesting knowledge can be mined, so only one tool is not enough. An open source data analysis language: Python along with its a wide variety of libraries, can be a powerful tool for mining the data set. With this, we can analysis the relationship between any attributes in the data set. For example, we will use a curve to find the trend of the number of different types of Visa while applying for the permanent labor certification over time. We also use the histogram to see the relationship between the approval rate of different types of Visa and nationality, the type of work and the place of work. The histogram can also show the relationship between the number of PERM applications and the application area. On the other side, data visualization is also considered to address any interesting information found in the data.

3.3 Data Reduction

The data file we find online contains huge amount of data. The time span of the data is from 2013 to 2018, including PERM visa application cases from every day during the four years, and that end up with more than three hundred thousand rows in the data file. And also, the data file has 157 attributes, and most of them are trivial or useless. This tremendous amount of data can cause the data mining program to run slowly, so some data cleaning process such as reduction is necessary. Among the 157 attributes, most of them are sparse. Instead of spreading everywhere, many of the attributes start to have data since certain date and the data disappears on another date. When the data is sorted by date, we can see kind of data chunks under those attributes. So we conclude that this is caused by some temporary system update of the U.S. embassy. Some attribute are required for the system to have under certain policy, and when the policy expires, the attribute stop to have data. In this case, since the time spans of many attribute are not long enough, and there is no way to fill in the blanks, they become not very useful for the long-term analysis of the data, so we decided to perform data reduction using attribute subset selection. It can not only make date look cleaner, but also speed up the data mining programs.

3.4 Data Cleaning

We also find some outliers under some attributes. For example, under wage attribute, we encountered some outliers, such as $400000/hr. These numbers that are greater than IQR/2 are recognized as outliers and are smoothed with a class mean, where the class is the time range within a year. However, we found that, in many cases, these unbelievably high wage could end up with a denial in visa application approval. There are also some blanks in numeric attributes such as wage, for these blanks, we also fill them with class means. However, there are blanks in some nominal attributes that are hard to fill in, for example, there are 57 empty cells under nationality. For these attributes, we simply ignore the missing values.

3.5 Data Integration

There are also some redundant attributes with absolutely same or correlated values such as wage per hour and wage per year. These attributes are also arranged or deleted manually. There are two special attributes named "citizenship" and "citzenship". After we sort the data by date, we found that "citzenship" starts at the beginning of the data file, and it ends at a date when "citizenship" start to have value. So we conclude that this is simply caused by the spelling error in the application system, and the error is fixed with the system being upgraded at that date, which end up with two attribute with same data starting from different dates. To deal with it, data integration comes into the play. We simply integrate the two attributes and make them one complete attribute.

In our data sets, different attributes are not really appearing at the same time. Some attributes were added to the form very earlier like 2000 and some other attributes are very new since we are only able to see them in late 2015 or even late 2016. Therefore, for us to load our attributes data into our classifier, we need to take care of whether the attributes have enough data and whether different attributes have valid data in the same position or time period. So, what we did is that we wrote a Python script to get the location of valid data for each attributes and compare this data between different attributes. Then, we grouped attributes with valid data in same position since they could have decent affect to our prediction results in a proper way. Also, some attributes only appear in a very short time period which means they were only recorded in a short timeframe. For this type of attributes, we need to ignore them by removing them in our dataset our set their weights to zero in our classifier.

3.6 Attribute Feature Analysis

The dataset contains many columns, which is described in an official structure file. It contains the data from the Applications for Permanent Employment Certification (ETA form 9089) by the employer and other administrative data such as certification determinations and the date of the determination was issued. etc.

Among those columns, not all of them are useful, we can select some of them by empirical method and data mining methods. For those data which are unique for each application, we simply omit them based on the thought that these data won't contain any data that will be useful for any other applications. These columns contain 'case number' 'decision date', 'case received date', 'original case number'.

3.6.1 Refiling and Schedule A or Sheepherder Information. There are other columns which will affect the status of the case heavily. These columns include 'refile', which indicates if the application was previously filed. And 'schedule a or sheepherder' which indicates if the application is in support of Schedule A or a Sheepherder Occupation. Compare to a total amount of more than 210,000 applications, although there are only 228 applications which has a "Yes" in their refile columns, nearly 40% of those applications was denied. This can tell us if the application is previously filed, it is more likely to be denied. As for if the applications are in support of Schedule A and Sheepherder Occupation, things get more interesting. According to the U.S. government, Schedule A is comprised of certain occupation, for which DOL (Department of Labor) has determined there are not sufficient U.S. workers who are able, willing, qualified and available. Sheepherder Occupation is at the same place. According to our experience, if there are not enough worker in the U.S. in this field, then the application is very likely to be certified. However, in our dataset, none of the applications is in support of the Schedule A or sheepherder occupation, which mean this column is totally nonsense. When we referred to the instruction to the form 9089, we noticed that if the answer to this question is yes, then the application will not be sent to the DOL, but directly to the appropriate department of homeland security office, which can explain why answer are all no to this question in our dataset.

3.6.2 Employer Information. The columns come next in our dataset are some basic information about the employer, such as the name and address, city; state and the postal code of the employer. Also, the establishment year of the employer and number of employees employed by the employer are included.

In the column of the number of employees, the values vary heavily. For example, the minimum value is 0, which means there are some company which doesn't have any employee at all! The maximum value is over 600,000,000, which means this company hired one tenth of the total population in the world, unbelievable! However, the third quartile is slightly over 10,000, which is still lower than the mean value by 50%. Then we can know most of companies are hiring a reasonable number of employees. As for other information about the employer, they are neither useful for the determination nor interesting. For example, all the employers are from the U.S.A.. And employer phone numbers are private information. Besides, there are other information about the agent such as the company or law firm that employs the agent, where does the agent come from, which seems no use at all.

Now comes the information of the work. As we all know, the information of the work is very important. If the work type is very popular and attract a lot of foreigners, then the approval rate will be very high. In our dataset, the job title of each applicant's job is given. However, the job titles are written in very different ways. Even the same job will be described in the way that they are totally different of each other. Hopefully in our dataset, there is another column called SOC code, which is related to the Standard Occupational Classification System (SOC)[11]. The SOC system is used to classify the job being requested for permanent certification. According to the standard occupational classification manual given by the executive office of the president, SOC is a 6-digit code. the SOC code classify the work performed into 23 major groups by its first two digits, and then the major groups are broken into minor groups, and then divided into other broad occupations, and then one or more detailed occupations. In the 6-digit code, we only take the first two digit for our information because too detailed information will block our insight into the dataset and even the major groups are detailed enough.

3.6.3 Prevailing Wage Information. The dataset includes another column which includes the level prevailing wage determination. This Prevailing wage is used to protect the foreigner workers as is often the case that foreign labor is low paid. The prevailing wage level are used to indicate the level of the prevailing wage, are divided into 4 levels. But as we can know the latter part of the paper, with several mining tools used on this dataset, the level attribute has very little impact on the case determination. There is information about the wage such as the lower range and the upper range of the wage offer and their pay unit. Since this information are numerical, we will discuss them in other section.

We have other data columns about the job opportunity. This information includes the fact the in which information the foreign worker is going to be employed and the detailed address. The number of the states here is beyond the total number of states in the U.S.A.. After a deep look, we can find that some works are going to work in the island outside the mainland of the America, while other districts in other countries, are considered, such as the British Columbia in Canada. We will see that the state where the employer is going to work affect the determination of the case a lot.

3.6.4 Job Opportunity Information. Next, we encounter one of the most powerful that will affect the decision of the case status. This column contains the minimum level of education required to perform the duties of the job being offered. Indeed, all the columns considering the educations is very important, including the class of the admission, the highest educations of the foreign worker. We mined the relation between the education required by the job and the education accomplished by the employee. First, we map different level of the education into integers. And then we calculate the difference between the two columns, and we can know the if the employee gets the acquired education by the job or not. For example, the positive value denotes that the employee's education level is equal to or higher than the education level required by the job. And the negative value indicates the opposite thing. If the value is zero, then we can know that the job has no education requirement. To our surprise, the fact whether or not the employee accomplishes the education required by the job seems has no effect on the determination of the case. But if the job does not require and education, then the ratio of being denied increases. But if we look at the two columns independently, not considering the relation between, we can find the items with none education, associate degree or high school degree have the higher possibility of being denied, which indicates the marketplace of US like foreign workers with higher education. There is another attribute indicate whether the applicant is being sponsored. If not, then the case will definitely be denied.

3.6.5 Alien Information and work experience. Then we come to the part of information of the foreign worker. Some columns are stating where the employee is living, including the state the worker is in, the exact address and the postal code. The value in the state column is the same as those in the column of the state of the job. However, the employee's living state is not that important considering the case status. And we have other information such as the citizenship of the employee, and the education level of the applicant, which are analyzed in detail in the previous section. Other attribute includes the major of the study of the applicants and whether the applicant accomplish the training required by the job, which are not that important.

We have another important information about the employer, called NAICS code. According to North American Industry Classification System[9], NAICS code is a hierarchical 6-digit code, whose first indicates general categories of economic activities of the employer, just like SOC code. So, we use the same technique, just take out the first 2-digit of the NAICS code and leave out some too detailed information. We can still find the similar pattern as that of SOC code. Also, if an applicant didn't provide a NAICS code, then he is very likely to be denied.

3.6.6 Information Gain and Gain Ratio. Besides the empirical methods mentioned above, we also use some data mining techniques to analyze the attribute features, including information gain and gain ratio. Then, we sort the feature according to their information gain value or gain ratio value. There are some interesting results in the sorted attribute list, whose top 10 rows is shown in table 1. We can find some attributes that we didn't expect to be a very important factor that can affect the determination of the case status very much. For example, the unit of pay. If we look into this attribute deeper, we can find that if the employee is paid by hour or by week, then he is more likely to be denied. This indicates that these units of payment may show a sign of instability of the work status.

4 EVALUATION

4.1 Data Visualization

4.1.1 Box Plots for Wage of Each Unit. This section is about the box plots for wage offer with respect to different case status. In the dataset, we are given the information of "wage offered from 9089" (lower range of the wage offer, and this attribute is what we used for plotting) and "wage offered to 9089" (upper range of the wage offer) by U.S. Department of Labor, for each applicant from different fields of work, which includes 5 valid units of wage offer: hour, week, bi-weekly, month, and year. With regard to the visa approval rate, there are 4 kinds of case status to be considered: certified, certified-expired, denied, and withdrawn (note that certified-expired means one doesn't file 140 within 6 months after the PERM is certified). In this case, our group of students got really interested in whether or not the larger amount of wage may cause higher probability of the approval rate for the permanent labor certification, and our guess would be positive.

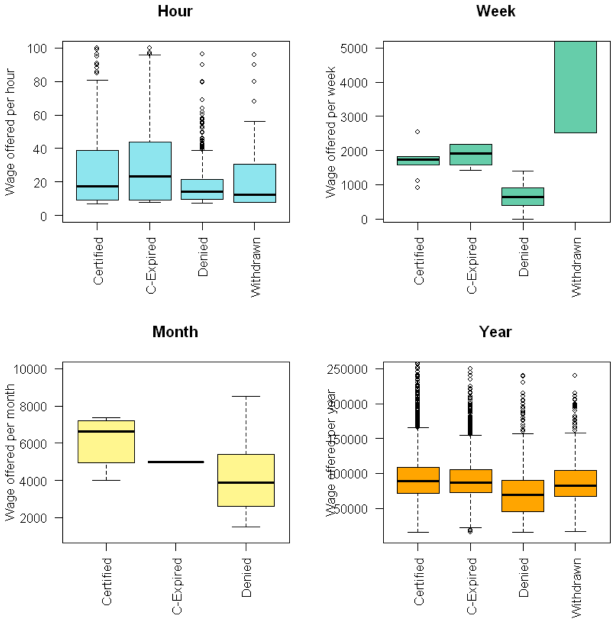

Figure 3: Wage of different units

To implement the relationship from the real data, we did the box plot, which is a standardized way of displaying the distribution of those data based on 5 numbers of summary (min, Q1, median, Q3, and max), and it can also tell us the outliers information. To start with the box plot, we have to make some changes of the data. The first step is to separate the original data by groups of wage units into 5 subsets. Then, for each subset, we created the dummy variables (or categories) based on 4 kinds of case status, for example, defining "certified == 1", "certified-expired == 2", "denied == 3", and "withdrawn == 4". Next, we plotted the box plot of each group as Figure 3 from above.

We may probably say that the larger amount or wider range of wage can cause a higher probability of the visa approval rate. It may be obviously true for the "wage per week" and "wage per month" in our plots above (see Figure 4), however, in the "wage per week" plot, the values and the range of wage for "withdrawn" is much larger than other case status. But it's not true with respect to the "wage per hour" group, since the median values of "certified" and "denied" are really close to each other. In addition, the ranges of wage offer

Attribute Name	Information Gain	Gain Ratio
CASE_STATUS	0.36673872118962114	1.0
JI_OFFERED_TO_SEC_J_FW	0.0007626576080062564	0.25504959903339397
RECR_INFO_EMPLOYER_REC_PAYMENT	0.000239920165837626	0.18032592617499624
SCHD_A_SHEEPHERDER	0.00013395375308294932	0.16846840599414814
EMPLOYER_COUNTRY	0.0007055259876602715	0.12597596119837962
JI_LIVE_IN_DOM_SVC_CONTRACT	0.00860906585661253	0.11268434683542324
PW_UNIT_OF_PAY_9089	0.015855408079182842	0.11261106852130848
JI_FW_LIVE_ON_PREMISES	0.0017997458051501325	0.09851144399703898
JI_LIVE_IN_DOMESTIC_SERVICE	0.000779392985416405	0.07032001088690451
REFILE	0.0007324616277396001	0.05999418320605409

Table 1: Top 10 Factors According to Gain Ratio

per year for 4 case status are really close as well. Moreover, we obtained an opposite example for the "wage bi-weekly" as Figure 4, where the median of "certified" is much smaller than that of "denied". However, it is hard to tell that the approval rate and the wage offer are unrelated, since the data may not be guaranteed to be precise enough every time. So, we still need more exploration and analysis.

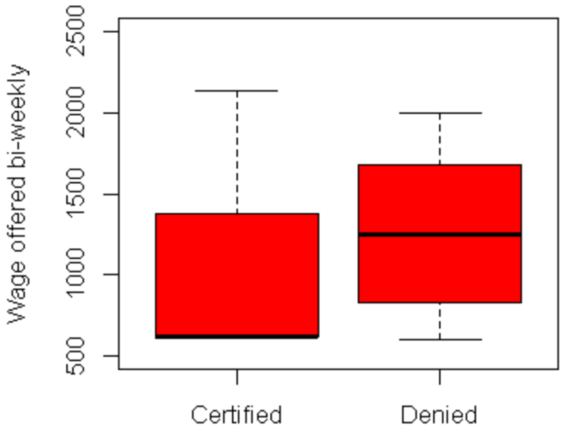

Figure 4: Wage biweekly

4.1.2 Box Plots for Yearly Wage from 2015-2018. Based on section 7.1 for different unit of wage offer, we'd also like to see the outcomes if unifying all the units of wage offered from 9089 to unit of "year" only. To do this, we just deal with the year from 2015 to 2018. That is, assuming the regular working time is 40 hours per week, then we multiplied the "wage offered from 9089" by 12 if the unit is "Month"; multiplied by 26 if the unit is "Bi-Weekly"; multiplied by 52 if the unit is "Week", and by (40×52) if the unit is "Hour". This could give us the rough results for yearly wage offer, and the box plots are as Figure 5.

Roughly speaking, from the above boxplot, one can see that the median wage of "Certified" status is much larger than that of "Denied", and the difference value of them is close to 50,000. In this case, we might conclude that the higher amount of wage will cause the higher approval rate for U.S. permanent application, the majority of which is actually H-1B visa application. In addition, one can also see that the median wage values of "Certified" status have increased year by year, which is reasonable to us since the average per capita income also increases year after year.

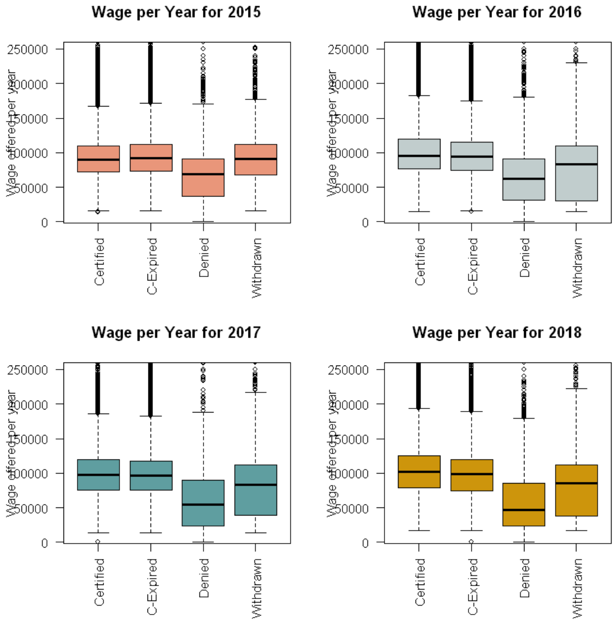

Figure 5: Yearly wage for 2015-2018

4.1.3 Cases by Countries. We want to find out the relationship between the PERM application cases and the nationalities of the applicants. In the dataset, there are two attributes which can show the nationality of applicants directly. They are "citizenship" and "citzenship". Apparently, "citzenship" is a typo of "citizenship", but both columns are not empty, which is to say, we need both of them if we want to analyze the relation mentioned above. When digging into the two columns, an interesting patter can be found. The cases in the dataset is arranged in the time order. At a given point, the data in the column "citzenship" no longer exists, and data begins to appear in the column “citizenship”, which has no data before this point at all. We can also notice that other columns, such as “previous job title”, “naics code”, “naics title”, and “wage offered” A reasonable guess is that at this point in time, the database system has undergone some upgrades and fixed this spelling mistake. We import these two columns, move all the data from column “citzenship” to column “citizenship” before that particular point, and discard the typo column. We can do this using entity integration, a popular data integration method in data preprocessing.

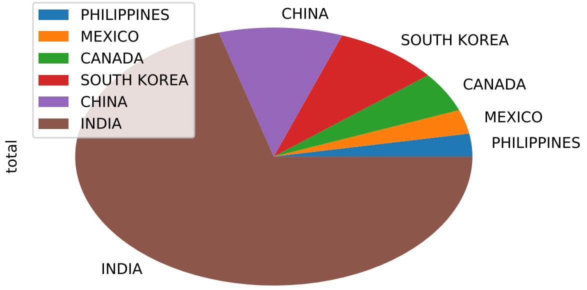

Figure 6: Number of application cases in difference countries

According to the dataset, there are people from 202 countries who applied for the U.S. permanent certification, nearly as many as the total number of countries in the world. However, not every country has quite a few applicants. In fact, most of the countries contribute very little to the application cases in these 5 years. We select those countries whose number of application cases is over 5 thousand in these years and make a pie chart to show the relationship between them, as shown in figure 6. Just as we can expect, the first two countries which have the most applicants are India and China, followed by South Korea, Canada, Mexico, and the Philippines. India has over 19,000 application cases, which is about half of the whole dataset. China and South Korea both have over 20,000 applicants, Canada has over 10,000 cases, Mexico and Philippines both have less than 10,000 application cases. The sum of number of the applications from those 6 countries exceeds 270,000, which means they make up over 70 percent of the whole data. As we can see, these countries are either border the U.S., such as Canada and Mexico, or in good relationship with the U.S. such as the Philippines and South Korea.

Even in those countries, the approval rate is not nearly the same. The countries whose has the highest passing rate are still India and China, as well as Canada. These three countries have their over 80 percent of application cases passed. We can roughly say that the order of passing rate is almost the same as the order of the number of cases, which probably means people from a country who has more applicants will get more chance to be certified the PERM. However, the difference is not that much, in these 6 countries, more than 70 percent of the applicants can get their application passed. But we can still notice that although South Korea has a lot more applicants than Canada, its pass rate is still lower than it of Canada pretty much. This may partly because above a half of people from Canada can speak English well. India and China have the most applications and highest pass rate, which is a very interesting phenomenon. We know that both India and China are huge countries. They have a lot of students who come to the U.S. to seek higher education every year, so they can have more chance to stay in the U.S. and get a work Visa (H-1B). As time goes on, these people can also attract much more people from their own countries to study and work abroad.

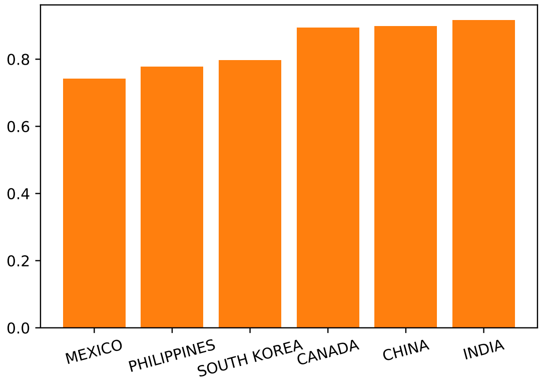

Figure 7: Pass rate of application cases in difference countries

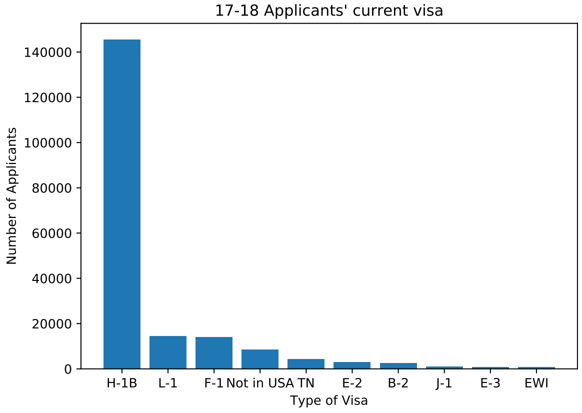

Figure 8: Histogram of Type of Visa holding by applicants while applying to Perm

4.1.4 Type of Visa While Applying to PERM. There are nearly 100,000 PERM application case received each year; these applications are received from people who currently hold different kind of visa, and some of them were trying to get the PERM from outside of the U.S.. We count the number of applicants with different type of visa and made a histogram of it. From figure 8, it is not hard to see that, from 2017 to 2018, most of the applicants were in the U.S. with H-1B Visa at the time they applied for PERM. The number of applicants holding H-1B overwhelmed the number of applicants holding other type of Visa. This result verifies the paper written by Ron Hira, "Bridge to Permanent Immigration or Temporary Labor? The H-1B Visa Program is a Source of Both".[4] The full name of PERM is the permanent labor certification. According to the definition of H-1B and PERM, an immigrant holding PERM can apply to any kind of jobs just like any other U.S. citizens, and he have the right to stay in the U.S. as long as he wishes even he quit the U.S. job market. However, H-1B is a temporary pass to U.S. sponsored by the company which the applicant works for, and it is limited to 6 years. After it expires after the 6 years, the worker can't remain in U.S. unless his current sponsor wish to extend the visa or he finds a new sponsor company. Although they provides different status to the foreign workers, they are both related to workers and labors, and clearly, PERM provides better degree of freedom to the ones who want to work in the U.S. and at the mean while want to visit their home town regularly.

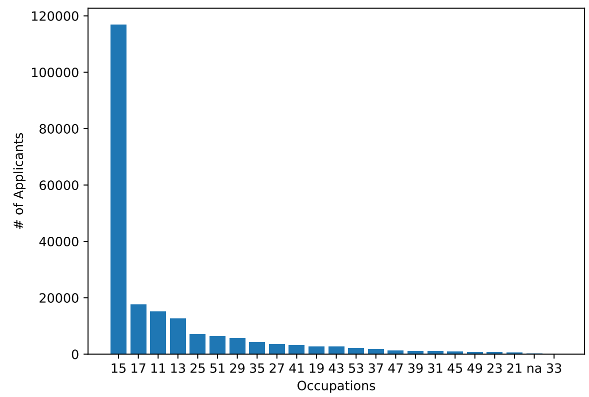

Figure 9: Histogram of applicants' jobs

4.1.5 Number of Applicants from Different Job Fields.

- 11 - Management Occupations
- 13 - Business and Financial Operations Occupations
- 15 - Computer and Mathematical Occupations
- 17 - Architecture and Engineering Occupations
- 19 - Life, Physical, and Social Science Occupations
- 21 - Community and Social Service Occupations
- 23 - Legal Occupations
- 25 - Educational Instruction and Library Occupations
- 27 - Arts, Design, Entertainment, Sports, and Media Occupations
- 29 - Healthcare Practitioners and Technical Occupations
- 31 - Healthcare Support Occupations
- 33 - Protective Service Occupations
- 35 - Food Preparation and Serving Related Occupations
- 37 - Building and Grounds Cleaning and Maintenance Occupations
- 39 - Personal Care and Service Occupations
- 41 - Sales and Related Occupations
- 43 - Office and Administrative Support Occupations
- 45 - Farming, Fishing, and Forestry Occupations
- 47 - Construction and Extraction Occupations
- 49 - Installation, Maintenance, and Repair Occupations
- 51 - Production Occupations
- 53 - Transportation and Material Moving Occupations
- 55 - Military Specific Occupations [11]

There was no doubt that Information Technology has become so popular in the past few decades. As shown in figure 9, most PERM applicants have IT and math background. The number of applicants from Computer and Mathematical occupations is way larger than any other job fields, and it is nearly six times of the second one, which is architecture and engineering occupations. According to U.S.News software developer was ranked #1 among the 100 best jobs in 2018, and was ranked #1 in best technology jobs. One reason that IT job is so popular is because it involves in nearly all kinds of job fields. According to U.S. News, "Software developers are in high demand right now, they're employed in a range of industries, including computer systems design, manufacturing and finance"[6]. Also, according to the Bureau of Labor Statistics, there will be a 30 percent growth on the demand of software developers up to 2026, and it is faster than any other occupations.

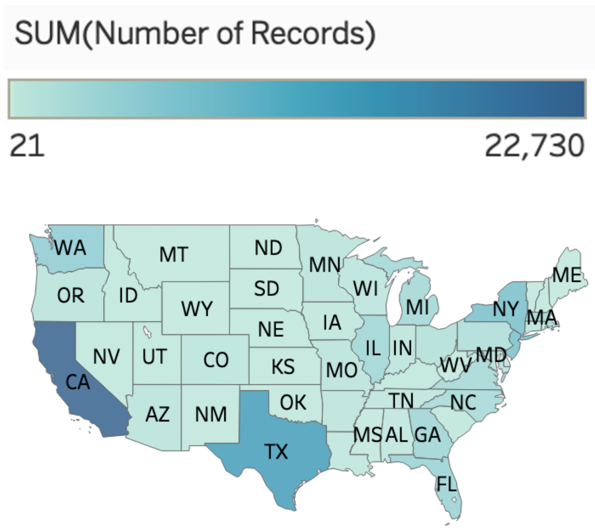

Figure 10: Number of applicants in 2017 in different states

4.1.6 Job Locations Analysis. Related to the occupation, the number of states of application is also impact by the IT fever. As shown in the figure, among the top 10 states where the applicants work, about a quarter of the applicant have their jobs located in California. It makes sense because California is one of the "IT center" of the U.S.. On the other hand, as a coastal city, California has more immigrants than any other states, there are 10 million immigrants living in California, which is about a quarter of all immigrants in the U.S.. According to Public Policy Institute of California, "in 2016, the most current year of data, 27 percent of California's population was foreign born, about twice the U.S. percentage"[8].

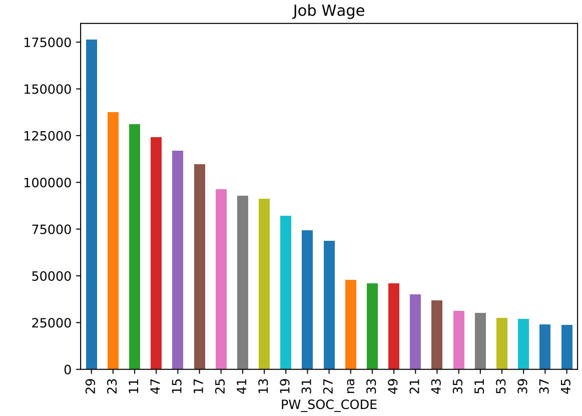

Figure 11: Histogram of wage from different occupations

4.1.7 Job Wage Analysis. In addition to applicant analysis, we did a wage analysis for different occupations. According to the bar graph, the occupation which has the highest wage are health-care practitioner and law occupations, in other words, doctors and lawyers. Computer and mathematical occupations are ranked number 5. We also made a scatter plot demonstrating the correlation between the average wage and number of applicants of occupation. We can see that, there are some correlation between the two data sets, at least at the low wage range, the number of applicants is smaller. The correlation coefficient is 0.47.

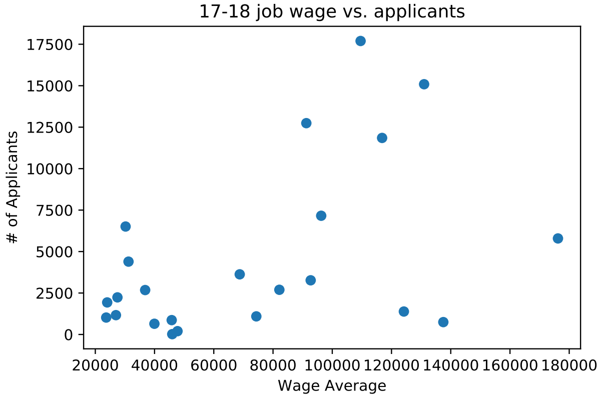

Figure 12: Scatter plot of correlation between average wage of the job and it's number of applicants

4.2 Classifier Using Support Vector Machine Algorithm

4.2.1 Model Selection. Although we have many attributes in our data set and these attributes have many different type of data, we consider CNN or some other classifier algorithms not fitted for our need. The reason is that we are certain that most of our attributes are not really associated with each other from neither mathematical aspect nor logical aspect. All of our attributes are kind of independent of each other and we should not relate them with more complicated algorithm. In this case Support Vector Machine fits our needs decently.

4.2.2 SVM Data Process. Data processing is the first necessary step for us to build our model. A very obvious problem of our data set is that the label we are going to predict, which is pass/reject attribute is not balanced. Number of people who passed the application is significantly higher than the application that failed. This will make our classifier biased since it will have an willing to predict a applicants to pass compared to be rejected. Therefore we duplicate rows where applicant failed the process to make our label unbiased for our predicting model. Normalization was also implemented in our data processing stage since it will also reduce conditions where our model could make biased predictions.

Unfortunately the algorithm we chose, SVM only works for numerical data while we have a lot of categorical data and strings in our data set. Which means that almost all of our data need to be transformed into some type of format to meet the needs for our classifier. Label is the first attribute that we modified. We mark rejected as -1, pass as 1 and another other status as 0. We also transformed binary string attributes into binary int attributes. There are two aspects of numerical data in our model context. One type is that when the scale of the value matters in its underlying meaning. For example salaries have a tendency to help the application if it is in a large scale compared to lower numbers. In this case we leave the value as it is in its position. Processing action against it was just removing some null values and removing commas from digits to make it into an int type. Another type of numerical attribute is that when scale does not matter. For example the code of the state of the applicants and the company is recorded as numeric but the value does not have an underlying meaning of this attributes. In other words it is actually categorical data instead of numerical data although it is written in digits. In this circumstances we are going to consider them as separate attributes. Therefore the technique of one-hot-encoding was used in order to transform one attribute into different attributes. One-hot-encoding took one attribute, classified values into sets and generate the corresponding columns. For example if we have 50 states in our state column we would have 50 columns generated by using one-hot-encoding technique. And we are going to feed these 50 columns as attributes into our SVM classifier model training stage. Also, some values that are categorical values, might actually could transformed into numerical value and we keep them in one column cause the scale difference has meaning for our classifier. Education level is very good example to demonstrate this scenario. Education level is recorded as âĂŸmiddle school, bachelor, master and PhD etc'. It seems like they are categorical string values but actually there is a hierarchical meaning implying by those values. We knew that they could be ordered by difficulty and this difficulty might have a big contribution to our classifier model. Therefore we transformed this column assigning digit values to education level from low to high in an ascending order. Another issue is that salary information of candidates is recorded with two column in our data sets. One column for value and one column for unit(hour, week, month, year). Therefore we multiplied these two columns and got a new column recording the salary in a unified unit which is year.

Attribute Name	Accuracy
PW_SOC_CODE	0.1186
JOB_INFO_EDUCATION	0.1280
CLASS_OF_ADMISSION	0.1230
FOREIGN_WORKER_INFO_EDUCATION	0.1094
SALARY_YEAR	0.3123

Table 2: Important Attributes and Accuracy

5 DISCUSSION

In the data visualization part, most graphs seems to demonstrate the reality. We visualized the number of applicants who are from different countries, different occupations, holding different type of visa while applying, and different states of their jobs. The result are obvious and close to our intuition. For example, we know that IT occupations are popular, we know that doctor and lawyer have higher wage, and we know that California is a big immigration state. These knowledge are clearly demonstrated and verified in our graphs. Undoubtedly, most of our data visualization succeed in providing correct information to ones who need it.

However, there is an uncertain factor from the box plots with respect to the wage distribution for different kinds of case status. Due to inconsistent unit of wage offer and in order to compute the values of wage with unit of “Year" only, we just assumed that all the employees work 40 hours per week. But it’s not always true in the real working situation, since some positions often need extra working time, such as engineers, computer programmers, medical workers and etc.. In addition, the form of wage is flexible according to different job titles. There are many job which the workers earn their incomes outside their wage, service fee, or tip, is a most common example. Some occupations such as product managers get cut from the product. If a product successes, the product managers might get much more money from the product. Therefore, there a some uncertain factors regarding wage which we couldn’t do anything to it.

When analyzing the attributes, the empirical methods and other data mining techniques like information gain and gain ratio are used. The limitation lies in these data mining concepts is that only one attributes can be analyzing the for its impact on the case determination at one time. Due to the amount of the columns in the dataset, the authors did not consider combining two or more attributes when leveraging the empirical methods. The frequent pattern mining, which can find the value combinations in different columns that frequently appear in the dataset, is a kind of valuable future work. This method is not implemented because of a lack of knowledge on tools and short of time.

For our classifier, we only used data from 2017 and 2018, because we want to set our classifier to a specific generation of government since policy is changing in different political period. We used 70% of our data as training set and 30% as testing set. The accuracy we have in the end is 89.8% using 16 attributes selected from our data set. Due to the long training time of our model we only use a subset of our original dataset but the result we got is good enough to compare importance of different attributes. By plugging each attribute individually into our model we can see the difference of feature in one-feature model condition. We had the following results indicating increment of accuracy by using some outstanding features in table 2. From this we can tell that the most important attribute to our model is SALARY_YEAR which could increase the prediction accuracy by 31%.

From the model we can also see that not all of our features contribute to enhancement of our predictions accuracy, like REFILE/EMPLOYER_NUM_EMPLOYEE etc. Some of the features even decrease our accuracy. But these features were actually observed that they might have some potential influence over our label. Which means that influence of feature over label predictions observed manually does not really indicates any underlying mathematical relationship in some cases. Which perfectly emphasized the importance of mathematical model in data mining science.

There might be some better algorithms for our prediction for the data set. SVM might not be able to dig out more complicated information. But based on our understanding of our dataset we believe SVM did a good job of analyzing the relation between our features and labels cause in our data set features reside in different field that can hardly related to each other according to our common sense. Therefore this model worked significantly for our prediction as we expected.

Some limitation occurred in our training stages. We fitted each features into our model and then we fitted all 16 features into our model to get results and compare differences. But a more practical way of feature engineering would be picking all combinations from 16 features and train all of them to get many models and then we compare the results to choose the best one. Because there might be some relations between features that is hard to be mined without mathematical models. But that would be 16! = 2E13 models to train and test which requires much more computing resource and time that we do not possess. If we had the resource to execute this procedure there might be some really interesting fact to be discovered.

We have done plenty of analysis and prediction in this project, and it should help ones success their Perm application processes to some extent. However, for those who tend to apply for green cards, this study can’t help them to the last step. Green card involves many step, and Perm is just one step. And also, green card applicants can’t get the green card once they are certified because of the long line, and the waiting process usually cost years even if the applicants pass the Perm step with no problem.

In the future, since the U.S. president may change after years, the new policies with regard to the application for permanent labor certification will also be regulated and implemented. At that time, we may have new data sets which could cause different relationship and factors as we discussed in this paper. For example, we found that one has higher probability to be denied again if the first application is already filed. However, in the future, the failure or a denied decision may not affect the further application. We’ll never know!

6 CONCLUSION

There is no doubt that there might be a lot of factors we are not going to use or even have access to. And the factors we are going to use might not be fully capable for representing or predicting the result of PERM application. In addition, we admit it possible that not all of the important factors that affect officers' decision could be recorded by numerical values, or even recorded at all. The uncertainty of measuring for person's capability of making values to society is always a hard topic for both scientists and government. But we believe that by digging and analyzing the recorded data from past few years with all of the tools and methods we focused in the previous section. we can at least have a basic and clear understanding of some of the important factors within the decision making procedure. And this could always help our international students to make the path to achieve American dreams wiser and easier.

We successfully fitted our data into the SVM model and the accuracy generated from this model did imply a decent understanding of the application approval process. But the difference in contribution of different features in our model is still noticeable, which means there might be more underlying relationship between features making a decent impact on these decision process that requires more complicated data analyzing in the future.

REFERENCES

[1] Nazli Avdan. 2014. Controlling Access to Territory: Economic Interdependence, Transnational Terrorism, and Visa Policies. *The Journal of Conflict Resolution* 58, 4 (2014), 592–624. http://www.jstor.org/stable/24545655

[2] Douglas Gollin. 2002. Getting Income Shares Right. *Journal of Political Economy* 110, 2 (2002), 458–474. http://www.jstor.org/stable/10.1086/338747

[3] Jiawei Han, Jian Pei, and Micheline Kamber. 2011. *Data mining: concepts and techniques*. Elsevier.

[4] Ron Hira. 2018. Bridge to Permanent Immigration or Temporary Labor? *US Engineering in a Global Economy* (2018), 263.

[5] Rajesh Kumar. 2017. *Analysis of US Census and Permanent Resident Data for High Skill Employment*. Ph.D. Dissertation.

[6] U.S. News & World Report L.P. 2018. Software Developer Overview. https://money.usnews.com/careers/best-jobs/software-developer

[7] Eric Neumayer. 2006. Unequal Access to Foreign Spaces: How States Use Visa Restrictions to Regulate Mobility in a Globalized World. *Transactions of the Institute of British Geographers* 31, 1 (2006), 72–84. http://www.jstor.org/stable/3804420

[8] Public Policy Institute of California. 2018. Immigrants in California. https://www.ppic.org/publication/immigrants-in-california/

[9] U.S. Department of Commerce. 2018. North American Industry Classification System. https://www.census.gov/eos/www/naics

[10] U.S. Department of Labor. 2018. OFLC Performance Data. https://www.foreignlaborcert.doleta.gov/performancedata.cfm

[11] U.S. Bureau of Labor Statistics. 2018. Standard Occupational Classification. https://www.bls.gov/soc/2018/home.htm

[12] Gianmarco I. P. Ottaviano and Giovanni Peri. 2012. RETHINKING THE EFFECT OF IMMIGRATION ON WAGES. *Journal of the European Economic Association* 10, 1 (2012), 152–197. https://doi.org/10.1111/j.1542-4774.2011.01052.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1542-4774.2011.01052.x

[13] Ben A. Rissing and Emilio J. Castilla. 2014. House of Green Cards: Statistical or Preference-Based Inequality in the Employment of Foreign Nationals. *American Sociological Review* 79, 6 (2014), 1226–1255. https://doi.org/10.1177/0003122414553656 arXiv:https://doi.org/10.1177/0003122414553656

[14] Scott Schonberger. [n. d.]. The Impact of Experience on Wage Premiums for Permanent Employment-Based Visa Applicants in the Information Technology Sector. https://repository.library.georgetown.edu/bitstream/handle/10822/1043973/Schonberger_georgetown_0076M_13627.pdf

A HONOR CODE PLEDGE

On my honor, as a University of Colorado Boulder student, I have neither given nor received unauthorized assistance.

B WORK DONE BY INDIVIDUAL GROUP MEMBERS

All members have equal contribution to this project.