H-1B Visa Data Analysis and Prediction by using K-means Clustering and Decision Tree Algorithms

Team: Renchi Liu (rxl161430), Jinglin Li (jxl163530)

1. Introduction

The H-1B is an employment-based visa in the United States, which allows U.S. employers to temporarily employ foreign workers in specialty occupations. To apply for H-1B visa, an U.S employer must offer an job and petition for H-1B visa with the U.S. immigration department. This is the most common and legal visa status and for international students who complete their college / higher education (Master, PhD) and work in a full-time position. This year, the U.S. Citizenship and Immigration Services (USCIS) will officially begin accepting petitions for the H-1B Visa on April 1st, 2017 for the next fiscal year. The status of H-1B visa will definitely influence the life and work, and even the career of the international students. So this project try to use algorithm learned in machine learning class, analyze historical H-1B data to produce helpful information. Briefly, In this project, we apply machine learning algorithms including Decision Tree and K-means analyze the conditions (or attributes) of the foreign workers, such as EMPLOYE_NAME, JOB_TITLE, WAGE, etc. We utilized the 2011-2015 H-1B petition disclosure data to analyze the employers with the most applications, data science related job positions and relationship between salaries offered and cost of living index. In our project, K-means clustering algorithm was utilized to separate the locations of H-1B applications.

In addition, our analysis will also provide some statistic data to answer some questions. Such as: What is the top companies that have apply to the H-1B for employees? What is the trend of total number of H-1B application is? What is the top popular Job Title and Worksites for H-1B Visa holders? What is the salary mean values of respective Job Titles? As H-1B visa is the most common and legal status for the international student, these data might help to guard them to choose the most easier way to work in the United State and accomplish their American Dream.

2. Problem Definition and Algorithm

2.1 Task Definition

Our model and analysis will provide a whole picture of the different approval rates by comparing different conditions based on previous data. Therefore, it will help us to predict the approval and deny rate of H-1B visa of the current year or the application in 2016. In addition, our analysis will also provide some statistic data to visualize the characteristics of the application case and trends.

2.2 Algorithm Definition

2.2.1 K-Means Clustering to separate the h1b location

K-Means Clustering is an algorithm among several that attempt to find groups in the data. In pseudo code, it is shown to follow this procedure:

```
Initialize $\mathbf{m}_i$,  $i = 1,\ldots,k$, for example, to k random $\mathbf{x}^t$
Repeat
  For all $\mathbf{x}^t$  in X
          $b_i^t \leftarrow 1$ if $\| \mathbf{x}^t - \mathbf{m}_i \| = \min_j \| \mathbf{x}^t - \mathbf{m}_j \|$
      $b_i^t \leftarrow 0$ otherwise
  For all $\mathbf{m}_i$,  $i = 1,\ldots,k$
      $\mathbf{m}_i \leftarrow$ sum over t ($b_i^t \mathbf{x}^t$) / sum over t ($b_i^t$ )
Until $\mathbf{m}_i$ converge
```

The vector **m** contains a reference to the sample mean of each cluster. **x** refers to each of our examples, and **b** contains our "estimated [class] labels".

2.2.2 Decision Tree

Decision trees model sequential decision problems under uncertainty. A decision tree describes graphically the decisions to be made, the events that may occur, and the outcomes associated with combinations of decisions and events. Probabilities are assigned to the events, and values are determined for each outcome. A major goal of the analysis is to determine the best decisions. The model of the decision tree were illustrated in the Figure 1.

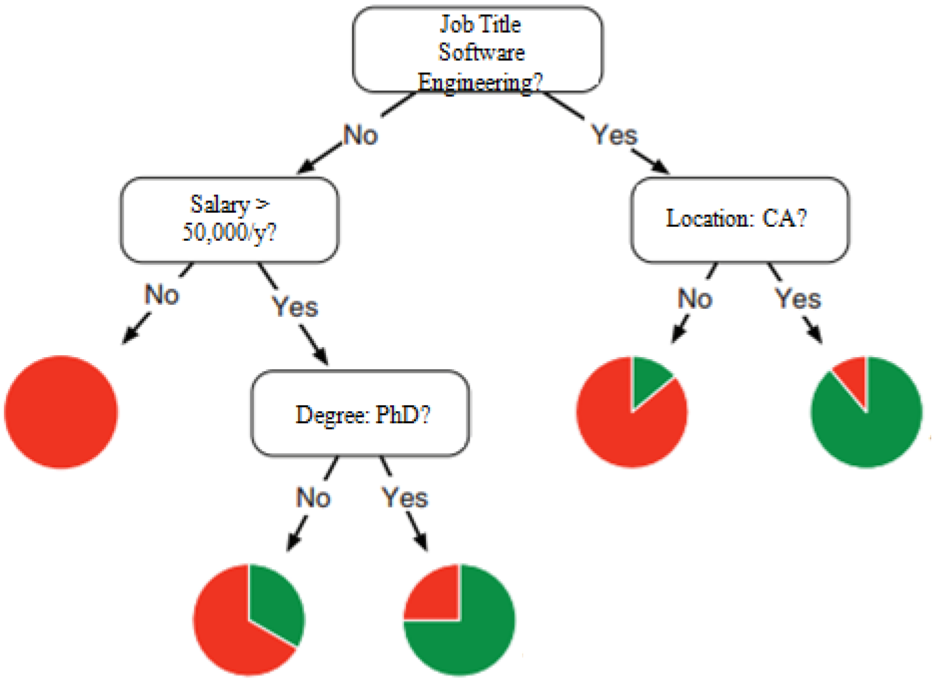

Figure 1. A hypothetical decision tree in which each node contains a yes/no question asking the training example about a single feature of the data item. An example arrives at a leaf according to the answers to the questions. Pie charts indicate the percentage of attributes from the training examples.

3. Experimental Evaluation

3.1 Methodology

1） The project will use Python 2.7 to program, Pycharm and Python-Notebook as the IDE. The 3rd library includes pandas (Data processing), matplotlib (plot display), numpy (match calculation) and some machine learning algorithms.

2） The analysis will utilize the algorithm learned from the Machine Learning class, such as K-means clustering and Decision Tree algorithm to analyze the conditions (or attributes) of the foreign workers, such as EMPLOYE_NAME, JOB_TITLE, WAGE, etc.

3）The dataset H-1B Visa Petitions 2011-2015 were cited from UNITED STATES DEPARTMENT OF LABOR.

4） As the H-1B Visa Petitions 2017 is not available by the time of project deadline, dataset of H-1B Visa Petitions 2011-2015 were utilized as the training dataset, the dataset of 2016 will be the test dataset based on the data of the previous years.

3.2 Results

3.2.1 Analysis of the H-1B by the status of their visa applications

The function showCASE_STATUS(self, H1Infor) of the source code shows the percentage of the status of all application, which is illustrated in the Figure 2. The panel shows that the majority (87%) of the visa application is certified, followed by 7% Certified-Withdrawn. It is noticed that there are 3% visa application is denied.

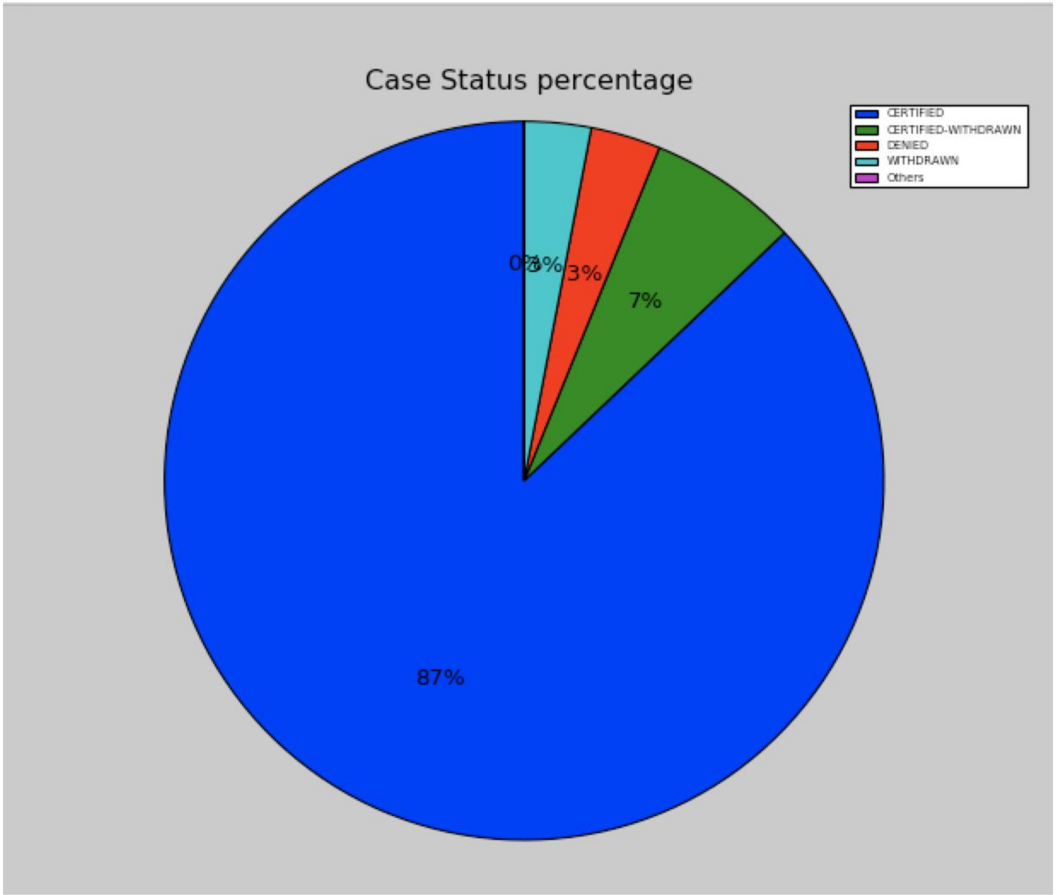

Figure 2. Analysis of the H-1B by the status of their visa applications. The panel shows the percentage of the visa status, certified, certified-withdraw, denied, withdraw and others. The majority of the application is certified.

3.2.2 K-Means Clustering to see where H-1Bs are

The function def K_meansAnlyze(self, H1Infor) of the source code implements the K-Means algorithm. It classifies the application locations based on the longitude and latitude of the sites. The density of the applications were shown in Figure 3. It suggested that the majority of the applications were located in the California and Northeast states of US.

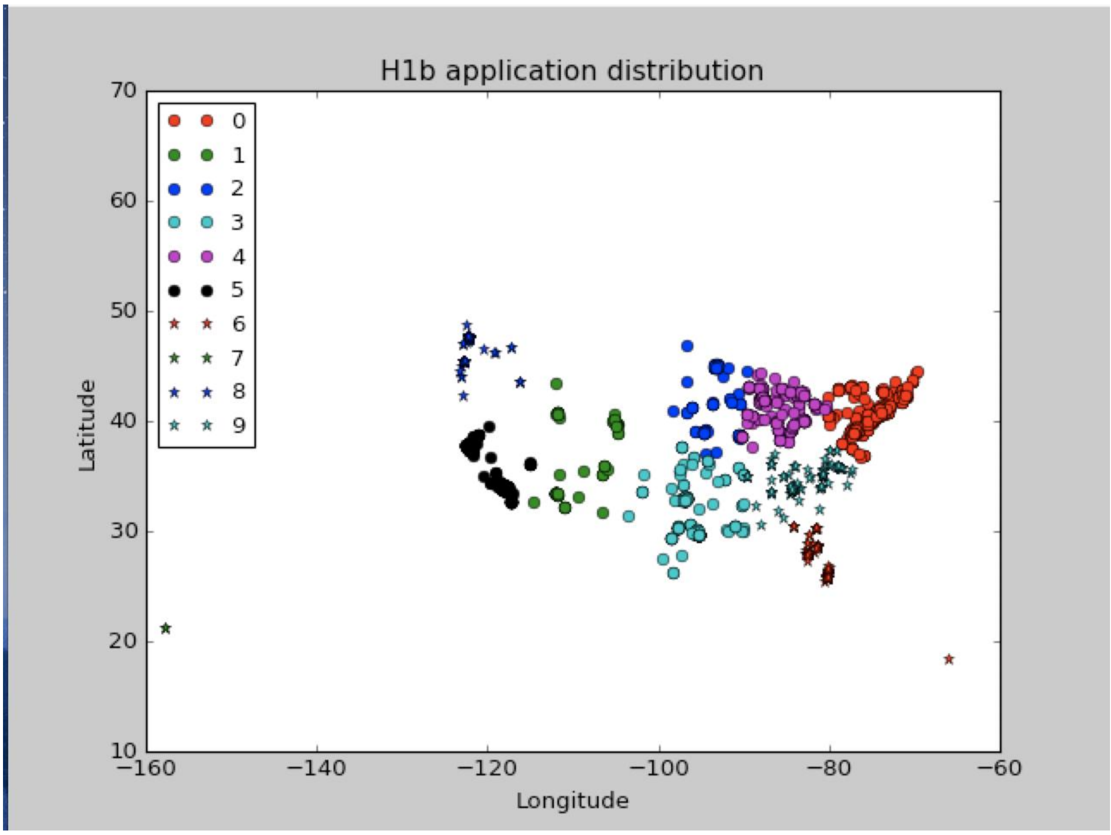

Figure 3. K-Means Clustering implementation to see the location and density of the applications.

3.2.3 K-means clustering algorithm to analyze the top 10 cities that apply the H-1B visa.

The function showWORKSITE(self, dense, H1LatLong) of the source code utilized the K-Means clustering algorithm to get the 10 clustering, which suggested the top 10 cities that apply the H-1B visa for the employees (Figure 3).

As shown in Figure 3, the most H-1B were applied from New York (about 190,000 cases). It is noticed that there are three cities, San Francisco, San Jose, Sunnyvale in the state California were list in the top 10 cities of the application. Considering these two phenomena, the New York and California have the most applications, which is consistent with our previous finding that plotting the locations of the application in map (Figure 3).

In addition, the Houston, TX has the 2nd application, which about 90,000 cases. Another city, Dallas, also from Texas, was also listed in this location plotting. This suggested that the cities in Texas also has many applications in US.

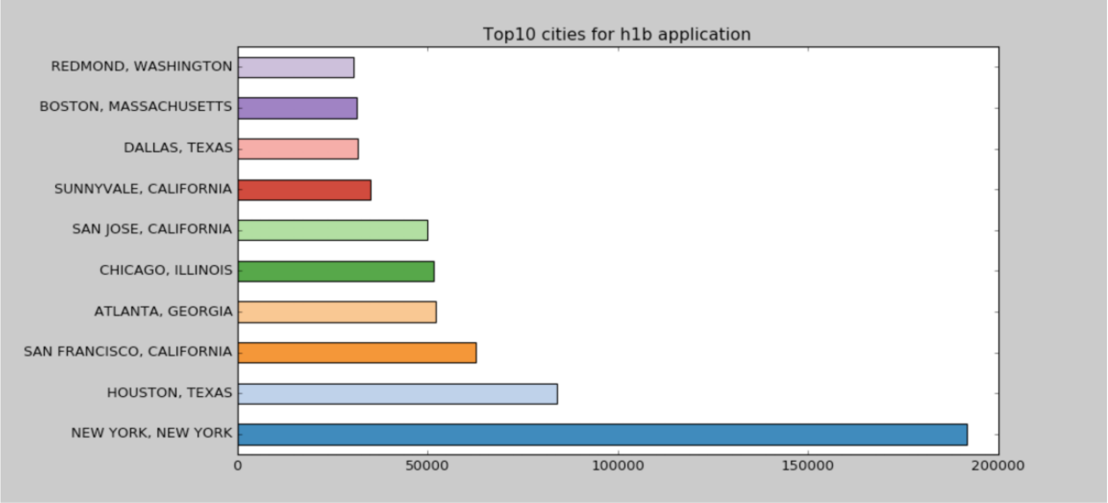

Figure 4. K-Means clustering algorithm to analyze the top 10 cities that have applied the H-1B visa. The bar shows the number of the application based on the training data. The most cases of application was proposed in New York, followed by Houston in Texas. There are three cities in California and Two cities in Texas were listed here.

3.2.4 Analysis of average salary for these top 10 cities

The function salaryAnalyze(self, H1Info, H1LatLong) implements the salary analysis of these 10 top cities, which is shown in Figure 5. The highest median salary of these 10 cities are Houston in Texas. The 2nd and most similar high salary is found in San Jose in CA. They show that about 80,000 USD is the median salary of all the applications. However, San Francisco in CA has the lowest salary in these top 10 cities, which about 55,000 USD.

Figure 5 implied that among these 10 cities, the cities in Texas, for example, Dallas and Houston, have the highest median salary. Three cities in California, have the 2nd top highest median salary. Considering the living expense compared with Texas and California, the savings of the people in Texas might be more than those of the people in California. However, because the living expense is not available in our dataset, it is better if we could analyze other data to draw a conclusion.

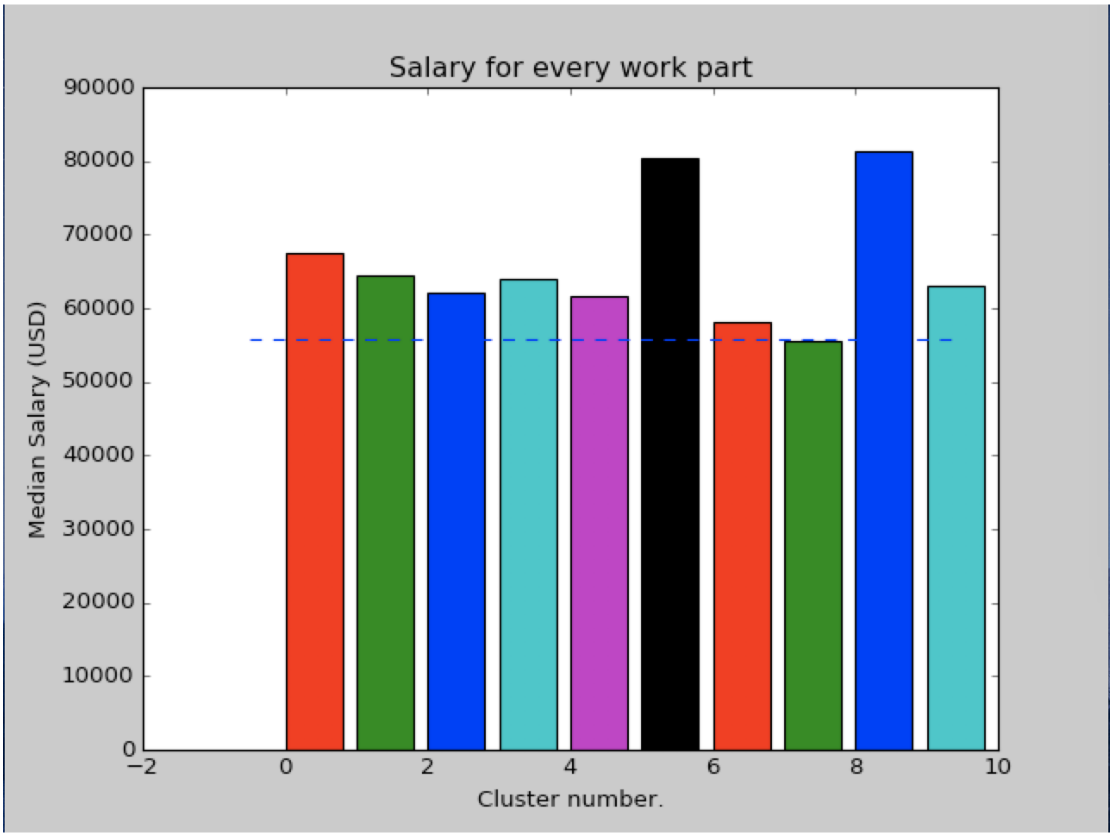

Figure 5. Median salary of these top 10 cities. K-Means algorithm plotting of the Median salary of these top 10 cities. The greatest salary is from the 5th city, Houston.

3.2.5 The top 6 company that have the most application number and the application trend analysis

The function showTOP6com_table(self, H1Info) of the source code was implemented and it is used to analyze the top 6 companies that have the most applications. The H-1B application number of the top 6 companies was shown in Figure 6.

The trend of the application of these top 6 companies were also illustrated in Figure 6B. It is interesting to find that from 2011 to 2015, almost all companies have an increasing application, while in 2016, there is slightly decreased among these companies.

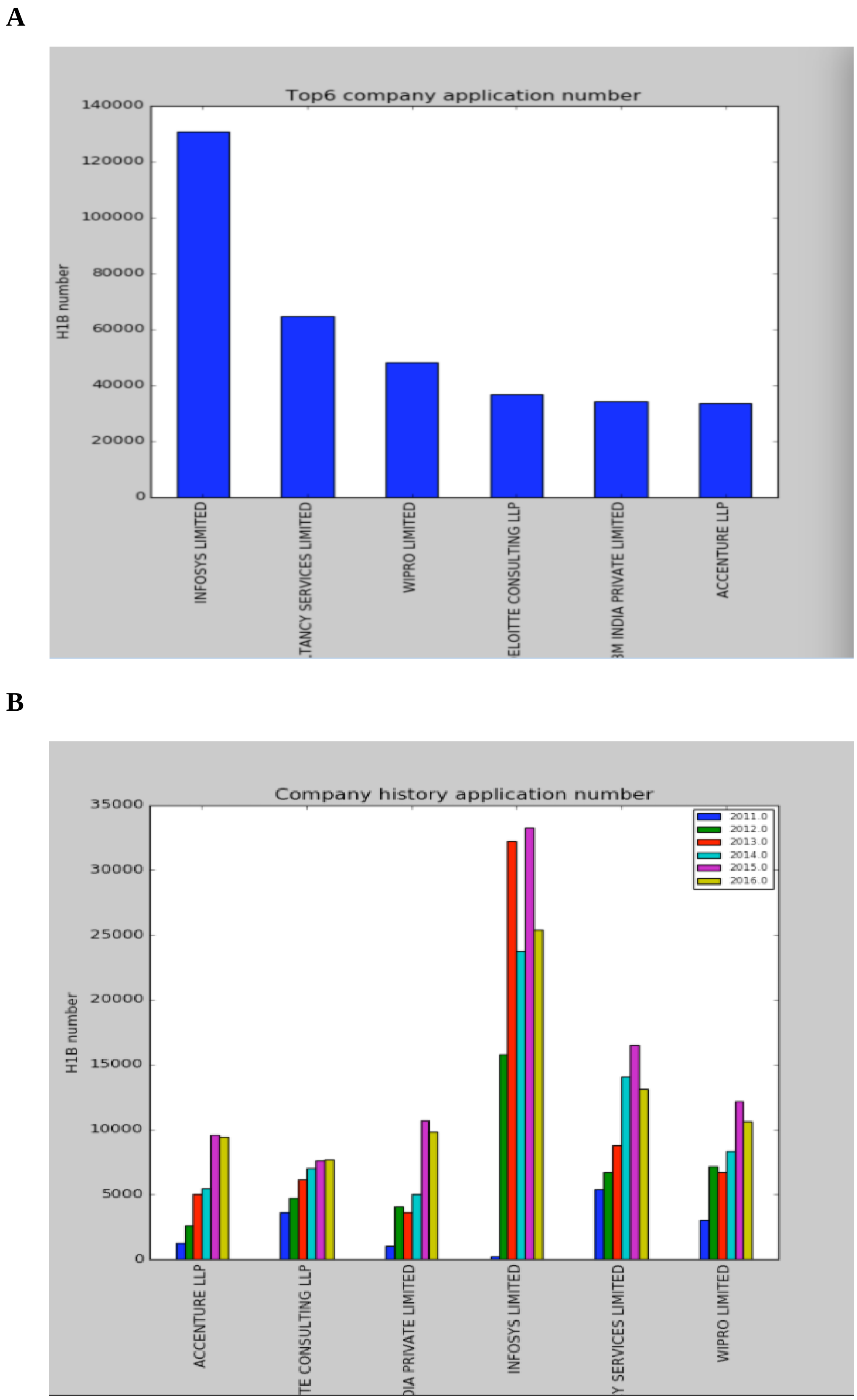

Figure 6. The top 6 company that have the most application number and the application trend analysis. This figure shows the top 6 companies that have the greatest application numbers (A) and the trend of the application number in the recent

years 2011 – 2016 (B). The trend shows that almost the application in 2015 has a obviously increased application and 2016 shows a slightly decrease of the application numbers.

3.2.6 Analysis of the job title for the most application

The function showJOBTITLE_plot(self, H1Info) of the source code analyze the top 20 jobs that get the most application proposed, which is shown in Figure 7.

The top job title was demonstrated to be the Programmer Analyst, with about 250,000 cases. The 2nd and 3rd top job titles were suggested to be Software Engineer and Computer Programmer. It is noticeable that there among the top 20 job titles, Computer Science-related job titles occupied most in this list. It might encourage more IT company and start-up company to focus on IT and technique-related area.

Other title jobs, such as Business analyst, consultant, physical therapist, accountant have also shown very high application numbers compared to other non-CS-related job titles. It could also provide some insight into the career choice for the students or the self-employed individuals.

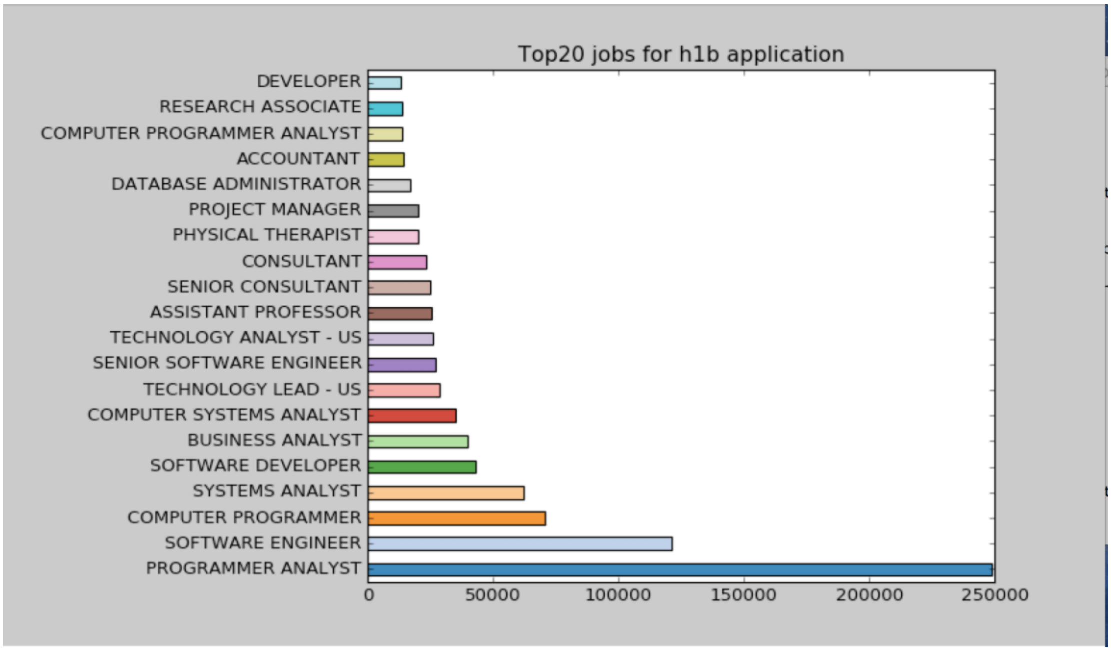

Figure 7. Analysis of the job title for the most application. The top 20 job titles show that the majority of the jobs that have the application number is computer science-related jobs. The most number of the application of the top 1 job title, Programmer Analyst, if about 250,000.

3.2.7 Analysis of the top salary of as per different job titles

The function showAVGSalary_plot(self, H1Info) of the source code provides the plotting of the top 20 average salary of different job titles, which is shown in Figure 8. Some senior job, such as director, vice president, manager, has no doubt high salary among the applications. Others like quality test engineer, developer also have very high salary. However, compared with the data from Figure 6, which is shown the large amount of application is for CS-related job, it shows that although there are large demand for the CS-related occupation, many other jobs, such as assistant professor, consultant, hospitalist, health economics, *etc*. have quiet high salary compared with the CS-related occupation.

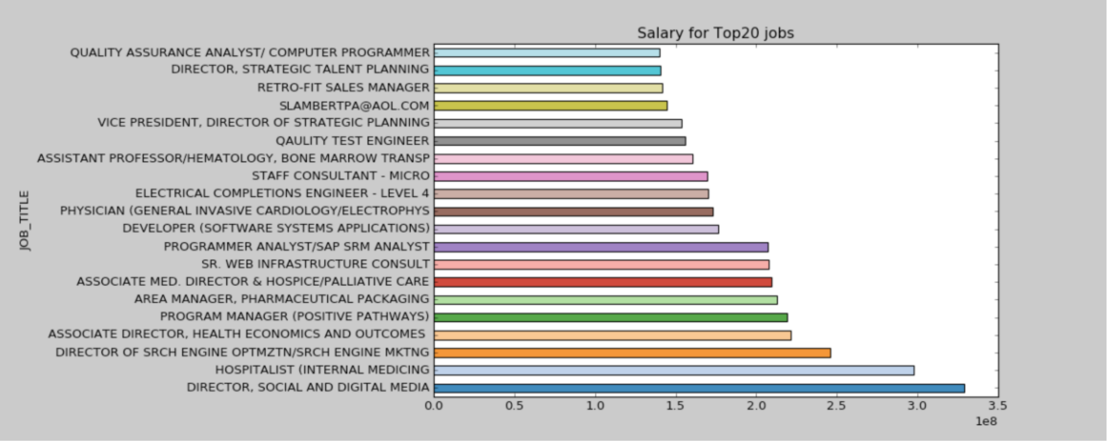

Figure 8. Analysis of the top salary based on different job titles. Many senior level jobs earn highest salary among other occupation. Some CS-related job, such as developer, quality test engineer, also have high salary.

3.2.8 Comparison of the application numbers of Full-time V.S. Part-time jobs

The function showFullvsPart_plot(self, H1Info) of the source code plots the H-1B application number comparison between the Full-time and Part-time jobs, which is showns in Figure 9. It is shown that the majority of the cases if full-time jobs.

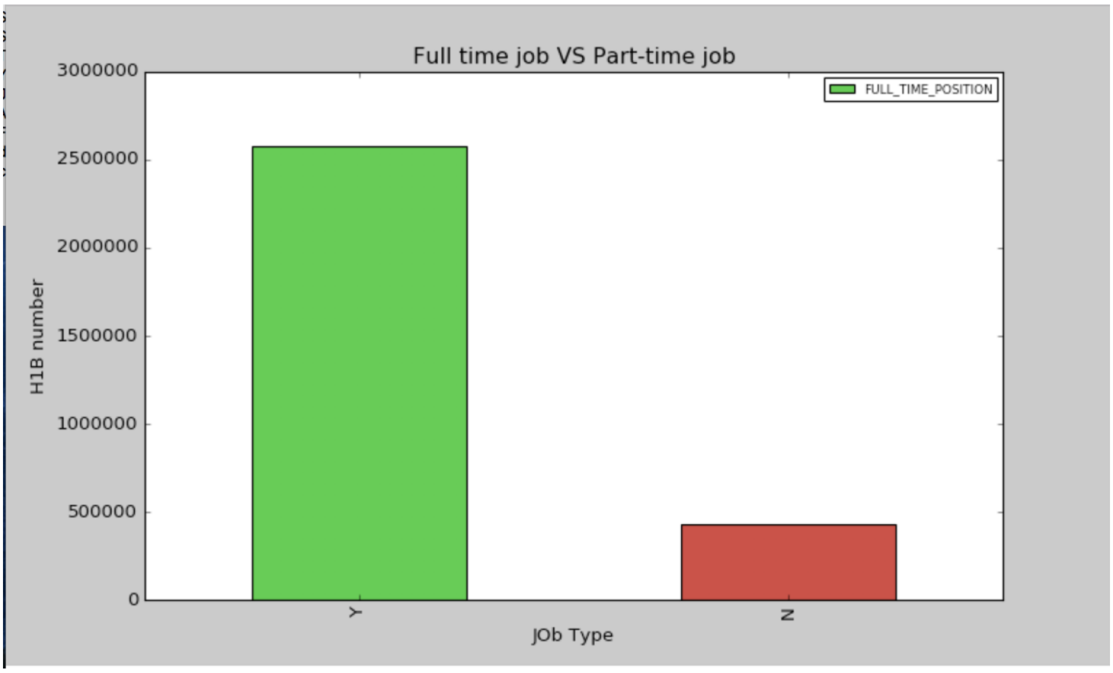

Figure 9. Comparison of the application numbers of Full-time V.S. Part-time jobs. The majority of the application is for full-time jobs.

3.2.9 Analysis of the trend of application number for each year

Figure 10 shows the numbers of application for each year, by the function showYearTrend_plot. It provides an overall picture of the trend of application along with the year. It is shown that from 2011 to 2016, the application number is increasing.

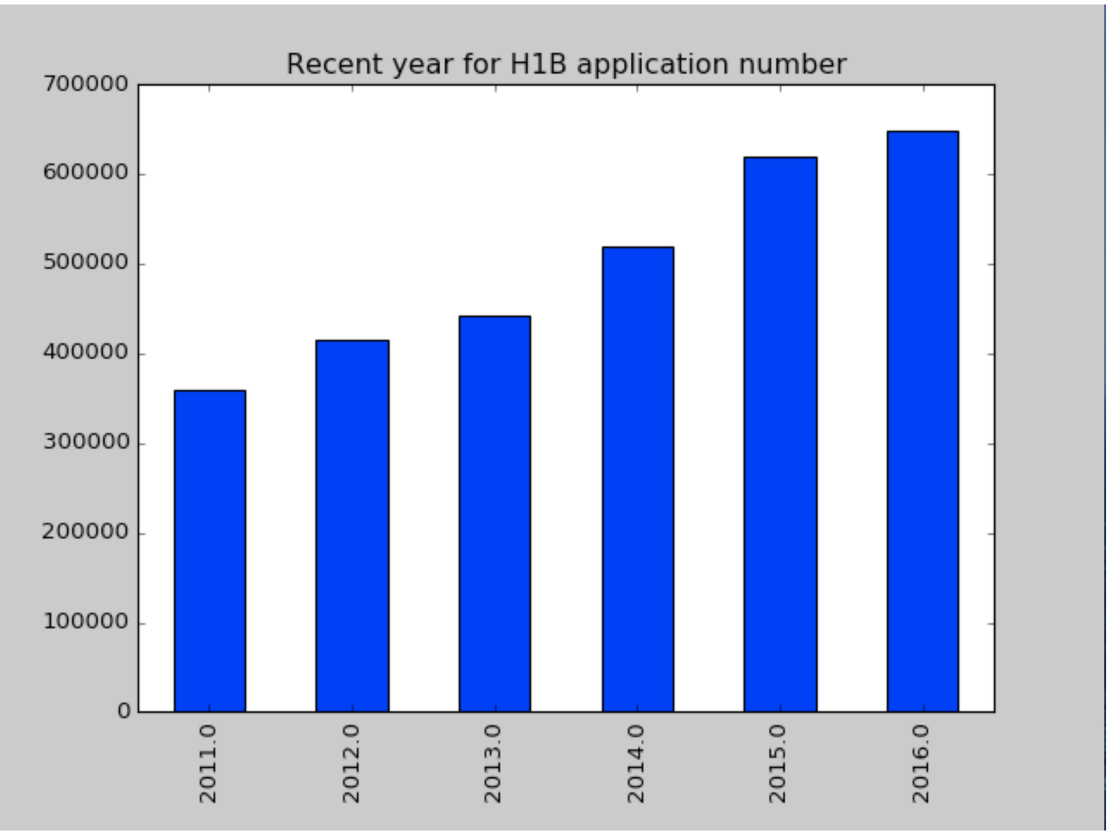

Figure 10. Total application numbers of the H-1B visa for each year. The application number is increasing.

3.2.10 Analysis of accuracy of testing dataset for different occupations according to the training dataset by using Decision Tree

Data from 2011 to 2015 was utilized as training dataset and data from 2016 was utilized as testing dataset. The Top 10 accuracy of case status (approved or not) for each job titles in 2016 dataset was plotting in Table 1.

Among the job titles listed in Table 1, technology lead has the greatest accuracy (99.8%). It implied that the trend of this occupation did not change in 2016. Because the decision tree utilized the data from 2011 to 2015 as the training dataset, the higher accuracy implied that this job situation remains the same in 2016, while a lower accuracy implied that the situation of this occupation might has some change in 2016 compared with the previous years.

Table 1. Top 10 application accuracy for different job title was plotted. Accuracies of the result of application (approved or not) was plotted and analyzed as per different job titles. It provides some insight into the trend of each occupation. Greater accuracy implied that the application status remain same within recent years, while less accuracy implied that the application status for that occupation has some change in the testing dataset (data from 2016).

	Job Title	Accuracy
1	BUSINESS ANALYST	0.7603359877822624
2	SYSTEMS ANALYST	0.8153321422770831
3	COMPUTER PROGRAMMER	0.8596160438806993
4	COMPUTER SYSTEMS ANALYST	0.8420289855072463
5	TECHNOLOGY ANALYST - US	0.9967974379503602
6	SENIOR CONSULTANT	0.8848737549223997
7	ASSISTANT PROFESSOR	0.5853952926976463
8	SOFTWARE DEVELOPER	0.7762979844740403
9	SENIOR SOFTWARE ENGINEER	0.7268209349277838
10	TECHNOLOGY LEAD - US	0.9981515711645101

3.2.11 Case prediction based on training dataset

Datasets from 2011 to 2015 were trained and a decision tree was implemented. For the next step, we could input some attributes to predict the status (approved or not) with high accuracy. For example, Programmer, California, 10,000 will be input as Job Tile, Location, Salary into our program and based on the decision tree, the result Denied might be predicted.

3.3 Discussion

This program presented interesting analysis of the H-1B visa application. The top 10 jobs, top 20 companies which proposed large amount of applications were analyzed. Some Computer Science-related occupations, such as developer, quality assurance and tester, *etc.* have show the greatest demand of these jobs in the US market. This data could also guide the students or self-employed individuals to have their career plans.

Machine Leaning is a powerful technique to deal with data. Decision Tree was implemented and the accuracy of each job title in 2016 was analyzed. This decision tree was built and provide an opportunity to predict the result of cases by inputting some attributes. In addition, K-Means clustering algorithm was implemented to plot the locations of application, which provide an overall picture of the application cases and some job demands.

4. Related Work

Organizing data into sensible groupings is one of the most important modes of understanding and learning [1]. Cluster analysis is one of the classic approach and algorithm for grouping, or clustering, objects according to their attributes. Among clustering formulations, the most widely used is k-means clustering [2]. The aim of the K-means clustering algorithm is to divide M points in N dimensions into K clusters, so that minimizing the mean square distance from each data point to its nearest center [3]. Thus, in our project, K-means clustering algorithm was utilized to separate the locations of H-1B applications, which is an ideal model to analyze the locations-related data.

The Decision Trees were derived with techniques based on classification and regression trees (CART) and Chi-squared Automatic Induction (CHAID) [4], which are classifiers that predict class labels from data items [5]. In our project, the Decision tress algorithm is utilized to predict whether the case of H-1B Visa application will be approved or not based on the previous data. This decision tree approach is a fairly simple type of classifier compared to CART and CHAID, and this is one of their advantages. Thus, in our project, decision tree was implemented based on different attributes and could provide a prediction if some new attributes are given.

5. Future Work

This dataset presented interesting challenges for dealing with complex dataset. The data still needed to cleaned, and more Machine Learning algorithm, such as Native Bayesian and SVM, *etc*. could exploited and implemented.

For the statistic data analysis, some model could be more detailed. For example, with K-Means clustering algorithm to plot the application cases location, a detailed map of US will be better, which could provide the exact location for visualization. In addition, more type of data could be added to provide a more detailed data. For example, if the living expense of each states of US is available, the data could be combined with the salary of the application and provide and predict which cities have more comfortable life and people there could save more compared with other places. This could be more meaningful to relocate for the individuals.

By the time for the project submission, the H-1B visa application cases in 2017 is not available online. Therefore we split the dataset of 2011 to 2015 as training and 2016 as testing datasets. In the future, the cases in 2017 could utilized as the testing datasets to see the trend for different occupation or different salary range.

6. Conclusion

In sum, our model and analysis provide a whole picture of the different approval rates compared with different conditions based on previous data. The utilization of machine learning related algorithm enables us to analyze the data based on training and learning steps. Therefore, it will help us to predict the approval and deny rate of H-1B visa of the testing year by Decision Tree algorithm. In addition, our analysis also provide some statistic data plotting, such as the top 10 occupations of applications, the salary of each job *titles*, etc. These data provides the situation of the H-1B visa during the recent years. The analysis of H-1B visa application might help to guard students, individuals to accomplish their American Dreams.

7. Reference

1. Jain, A.K., *Data clustering: 50 years beyond K-means.* Pattern recognition letters, 2010. **31**(8): p. 651-666.
2. Kanungo, T., et al., *An efficient k-means clustering algorithm: Analysis and implementation.* IEEE transactions on pattern analysis and machine intelligence, 2002. **24**(7): p. 881-892.
3. Hartigan, J.A. and M.A. Wong, *Algorithm AS 136: A k-means clustering algorithm.* Journal of the Royal Statistical Society. Series C (Applied Statistics), 1979. **28**(1): p. 100-108.
4. Yuan, M., et al., *Geospatial data mining and knowledge discovery.* A research agenda for geographic information science, 2004. **3**: p. 365.
5. Kingsford, C. and S.L. Salzberg, *What are decision trees?* Nature biotechnology, 2008. **26**(9): p. 1011-1013.