National College of Ireland

Analysing Different Machine Learning Algorithms to Study Visa Decisions in the United States

MSc Research Project
Data Analytics

Devika Kulkarni
Student ID: X19202865

School of Computing
National College of Ireland

Supervisor: Dr. Catherine Mulwa

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Devika Kulkarni
Student ID:	X19202865
Programme:	Data Analytics
Year:	2021
Module:	MSc Research Project
Supervisor:	Dr. Catherine Mulwa
Submission Due Date:	16/12/2021
Project Title:	Analysing Different Machine Learning Algorithms to Study Visa Decisions in the United States
Word Count:	5252
Page Count:	20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Devika Kulkarni
Date:	29th January 2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Analysing Different Machine Learning Algorithms to Study Visa Decisions in the United States

Devika Kulkarni
X19202865

Abstract

For those wishing to relocate, the United States is the most favored choice. A variety of visas are offered depending on the sort of profession. Obtaining a visa in the United States is difficult in most situations. The reason for this is because many people are unaware of the factors that determine whether or not a visa is granted. To figure this out, you'll need a powerful model that can forecast an individual's situation. Despite the fact that there are several established efforts on visa prediction in the United States. However, they all concentrate on the H1b visa, which is the most frequent and in-demand visa category. This study is concerned with the prediction of a wide range of visa kinds, not simply H1b. It will also provide the knowledge about the various sorts of visas available, the opportunities and organizations that can sponsor them as well as state wise distribution of the visa status. This project analyses the different kind of visas and to do this, five models were run and compared (Multilinear Logistic Regression, Random Forest Classifier, AdaBoost with LR, Radius Neighbor Classifier and XGBoost) to find the best fit model. It was found that among all the models, Random Forest gave the best results with around 93% accuracy. Other models also gave satisfactory results. The results of each model is discussed in chapter 4.
Keywords- Prediction, Employment, Visa, Data mining algorithms, Random Forest

1 Introduction

A visa is a stamp or permission that allows a foreign national to visit a country for a variety of reasons, such as travel, business, or employment. There are numerous types of visas available depending on the country to which one intends to travel. Similarly, in the United States of America, there are various classes or sorts of visas, such as H1, L1, and J1. The h1b visa, which is a form of non-immigrant visa in the United States that allows foreign nationals to work in occupations that require specialized knowledge and a bachelor's degree or higher in the specified branch, is the most prevalent. There are a lot of projects done on prediction of H1b type visa. However, not much on the different types of visa and predicting other visa types. To investigate this, five Machine learning models (Multinomial Logistic regression, Random Forest, AdaBoost with LR, Radius neighbor classifier and XGBoost) were trained by using the elements involved in the visa acceptance process. Many trial and error of the models was performed, in order to find the most successful model.

Several studies on the future of visas, specifically H1B visas, have been done. There are, however, other, more popular visas in the United States. The type of visa issued is determined on the nature of the employment. To be eligible for these visas, applicants must have a genuine work offer from a business. Due to a variety of obstacles, most people who get a job do not get a visa. To avoid this, visa prediction of the relevant employers who will be able to offer employment as well as the possibility of receiving a valid visa is important. In order to help this issue, this project was developed. There are, however, other factors to considered. This project not only teaches new data mining technologies, but it also focuses on predicting suitable employers.

1.1 Motivation and Purpose

Applicants face many issues when applying for jobs and then for visas. The purpose of this project was to assist in predicting the employers who will sponsor the visa based on the candidate's profile and other factors. In order to obtain a valid visa, one must also meet a number of additional requirements besides the company or the job. Thus, a vast analysis of the data set has been conducted to understand and know the elements found in the data set. It is a common misconception that obtaining required US visa is difficult. It is crucial to identify an employer who is willing to sponsor an overseas employee's visa. It is the first step in obtaining the visa you desire. To determine whether an employer is willing to hire a foreign worker, he must meet the employer's expectations. The project builds a model that will predict a candidate's outcome based on their status, so that they can know where to apply or what their prospects are for getting a job offer. Additionally, it examines how visas are approved and denied, as well as the criteria that are considered when applying for visa. The goal of this project is to help people looking for work in the United States find a respectable company that matches their skills. Also, this will make obtaining a visa easier for them.

1.2 Research Question, Objectives and Contributions

RQ: *What role can machine learning models play in studying visa outcomes using data from previous visa applications?*

Sub-RQ: *How can machine learning models enhance the prediction of visa status (denied/or approved) to help candidates applying for visas and to know if the visa might be denied/or approved?*

In order to address the aforementioned research question, the objectives described below are implemented, evaluated, and the appropriate results are provided in section 4.

Objectives

Objective 1: Analysis and prediction of the Visa trends in the US market.

Objective 2: Implementation and evaluation of visa outcome prediction models.

Objective 2.1: Implementation and evaluation of Multinomial Logistic Regression model.

Objective 2.2: Implementation and evaluation of Random Forest Classifier model.

Objective 2.3: Implementation and evaluation of AdaBoost with Logistic Regression model.

Objective 2.4: Implementation and evaluation of Radius Neighbor Classifier model.

Objective 2.5: Implementation and evaluation of XGBoost model.

Objective 3: Compare the models and determine which one is the most effective.

Contribution When applying for a US visa, it might be complex, and it also requires a lot of information. Furthermore, whether the visa will be issued or not is a significant unknown. To address this issue, this initiative was established to assist candidates in analyzing their status before to applying for a visa. It provides a more accurate image of an applicant before they apply for a visa. Through the analyzed study, it also assists aspiring applicants in learning more about different sorts of visas and gives them an idea of how to approach a US visa procedure. This contributes young people in making their visa application process a little easier.

The further report is divided into 5 chapters and its sub chapters. Following chapter 2 is based on the Literature review of the previous studies done on the prediction of the visa in the US and use of a different data mining models. Later, chapter 3 is based on the methodology and chapter 4 is based on implementation, evaluation, results and discussion of the methods used in the project followed by chapter 5, conclusion and future work.

2 Literature Review on the U.S Visa Prediction (2012-2021)

There has been a lot of research done on US visas, particularly H1b visas. However, we have looked at different visa kinds and their employment status in various cities or states of the United States as part of this research. In this section, we will look at previous projects on visa prediction and other data mining approaches. We'll also compare and contrast the benefits and downsides of the research, as well as how they can be improved in the future.

2.1 Study of Different Research Projects Based on the U.S Visa Prediction

Foreign workers are hired by a range of companies, the majority of them are hired for technical positions. States such as California, Washington, Illinois, Massachusetts, New York, New Jersey, and Texas all attract international workers. Numerous studies have been undertaken using various methodologies to forecast US visas. Based on the data set used, we used three different data mining techniques in this study. In this project we have discussed the difference between these three algorithms and why these are suitable to predict the visa status and we have also discussed the best fit algorithm for this study. In a similar work, Swain et al. (2018) used different data mining algorithms are used based on the candidate's profile, they used K-means, KNN, and Logistic regression algorithms to predict the H1B visa outcome. In their project, they considered three factors which are: Degree, experience (which should be minimum of 12 years in their own field and both of the above). However, sometimes taking into account these factors is insufficient

to ascertain the candidate's true status or whether he or she is qualified for any type of visa.Therefore, applying more filters for prediction can help to make more clear decisions and having more filters mean more expanded view.

There are too many articles and publications devoted to the study of US immigration procedures or H-1B visas, but there is basically no computer science research in this field according to Prateek and Shweta (2019) . The study intends to improve various categorization models in order to address the issue of H-1B visa eligibility. Another research was carried out using ANN algorithm by Khaterpal et al. (2020) in which the data was converted into the suitable format using one-hot encoding. After that, ANN algorithm was used to predict whether or not the petition would be accepted after the data had been trained. They used two data sets which was retrieved from kaggle. Later, the most relevant details in all of the petitions were examined, including the employee's name, job title, position, wage, caste status, and so on. Random forest approach was used by Sundararaman et al. (2017) to forecast if a visa petition in any US state will be approved or denied. According to their research, the companies identified as abusing these visas (negative) are quite similar to those shown in publications and news stories. The disparities in employment between two groups, namely 'non-college-educated natives' and 'native college-educated labor,' were discovered to be insignificant by a study carried out by Peri et al. (2015). These findings imply that STEM professionals, particularly those with a college education, help to boost economic growth through increasing productivity. In a research paper by Roy (2021), basic as well as the detailed analysis of the H1B visa is done. It provides us with the knowledge about the gaps in employment that are filled by foreign workers. Also it provides the information about which areas the U.S government should focus more for students. It also guides the international students for employment. Artificial Neural Network (ANN) was used to predict the success rate of the H1b visa in the research done by P. et al. (2021). They used the binary method to determine the status such as 0 means visa has been rejected and 1 for visa accepted. Their model achieved an accuracy of 94%.

2.2 Comparison of Different Data Mining Algorithms Used in the Prediction of U.S Visa

The visa status was predicted by Omar (2019) using a Random Forest algorithm. On two data sets with different attributes, they employed the Random Forest technique in their study. They developed an app for the candidates that forecasts the likely outcome of their visa application. Whether or whether it will be approved. This allows the candidate to receive immediate feedback on their status. In comparison to us, this is a similar study. Their technique, on the other hand, has an accuracy of roughly 83 percent, while our model, while using the same algorithm, was more efficient. The feature selection and data analysis are the reasons for this. Furthermore, the data set they used differed from the one we used for our experiment. On the other hand, three data mining algorithms were utilized in another comparable study by Vyas and Prakash (2018). Random forest algorithm was one of them, along with two other models: Naive Bayes and Logistic Regression. In comparison to the Random forest method, their results suggest that the other two worked effectively. Although various algorithms/models are used to analyse the H1b visa status it is necessary to note that the results are based on different criteria such as, feature selection, data sets, year, number of entries, etc. Another similar analysis was carried out by Thakur et al. (2018). In which they used six different models and that

was presented with the help of graphical representation. In a similar case study, multiple models were used to predict the H1b status. Beliz Gunel (2018) In this study they used in Naive Bayes, LR (Logistic Regression), SVM and Neural Network in order to predict the application case status in which they used employer success rate and prevailing wage as their variables which resulted in better accuracy of the models. In another research Kumar and Naresh (2018) the data set was split in the ratio of 80-20 and on the final dataset, three algorithms were applied. Among the three models (Naive Bayes, RF, XGBoost), XGBoost achieved the highest accuracy of 87%. A state wise distribution of the visa applications was focused in the project by Dhanasekar Sundararaman and Misraa (2018). They also introduced a part which features the company's priority as according to them just taking into consideration of common factors such as salaries would not help. According to them the company with a good name as well as which pays high carries higher priority. Below Table 1 shows the comparison between the models that were used for predicting the H1B visa..

Table 1: Comparison of models used in the prediction of US visa

Author	Accuracy Achieved (%)	Models used	Year
Aisulu Omar	82	LR, XGBoost, RF	2019
Hitesh Vyas et al.	99	LR, Naïve Bayes, RF	2019
Pooja et al.	95	Decision Trees, C5.0, SVM, NN, LM	2018
Beliz et al.	98	LR, SVM, NN, Naïve Bayes	2017
Madhana et al.	87	Naïve Bayes, RF, XGBoost	2018

2.3 Analysing the Features Used in the Prediction and Identified Gaps

In the years 2010 and 2011, it was reported that roughly 70 thousand average enterprises filed LCAs (Labour Condition Applications) for H1B applicants as per Ruiz et al. (2012). About half of these businesses requested only one person, and 94 percent requested fewer than ten. The majority of organizations apply for visas in the fields of information technology, consulting, and manufacturing. Results after running LR and ANN models showed that they got the greatest outcomes with 92 percent positive F1-score and 84 percent negative F1-score, respectively according to Pandya (2018). They claim that balancing the data was critical to achieving the desired outcomes. They discovered that characteristics from their data set, such as job title, employer or employee acceptance percentage, remuneration, and so on, played a significant effect in deciding case status.A study conducted by Kulkarni et al. (2019), looked at different firms over the course of a year and compared the city-by-city visa application status for a specific role over the course of several years. They presented the findings according to the company. i.e. for the top companies separately. They also identified the highest-paid positions in each company and analyzed their pay scales throughout the course of the year. A study based on migrants carried out by Parey et al. (2015) noted in their research report that, visa decisions are also based on the nation of residence. They further say that migrants from Germany are selected positively for immigration compared to the United States natives. Although, Grades, subjects, gender, quality, and other key variables are, nevertheless, taken into account. They also looked at data from the Community Survey

of America and found that the immigrants from Germany are more talented and have advanced degrees than native Americans. Interestingly, another study on the L1 visa was conducted by Dalmia et al. (2017) which was focused on the L1 visa type. Since there are a lot of studies done on H1b visa, in their perspective. Few, if any, are available on an L1 visa or any other type of visa. According to them, the dynamics for choosing an L1 visa differ from those for an H1B visa because L1 is limited to managers, executives, and other professionals with specialized knowledge. As a result, they thought it would be fascinating to concentrate on the qualities specific to the L1 visa. A similar study conducted by Tandon (2021) in which the fundamentals of US visas were presented and detailed examination of different sorts of visas was conducted. This research also revealed the latest visa patterns in the United States. They also look at the relationship between visa types and the applicant's home country. A random forest model was employed by Sampath et al. (2009) to forecast the chance of a binary response for freshmen visa applications, using freshmen data. For example, they employed yes/no responses in the enrollment process, as well as binary values for residency, race, sex, and other factors. There were over 5% missing values in their data collection. As a result, they erased columns like GPA, SAT, distance, and others that had a lot of missing numbers in order to deal with the missing values. The next section discusses about the methodology and design specifications used in this project.

3 Methodology Approach and Design Specification

The KDD or Crisp-DM methodology is utilized in data analytics. This project, on the other hand, does not deal with deployment and is purely educational. As a result, the existing KDD technique is the greatest fit for explaining the various stages of this project. Further part of this section explains details about the methodology used to predict the visa outcomes.

3.1 Visa Outcomes Methodology

In this visa prediction methodology there are five stages. In the first stage data was collected from kaggle as a .csv file. In the second stage this data was pre processed. In the following stages, this preprocessed data was used to draw the predictions with the help of machine learning models shown in the Figure 1.

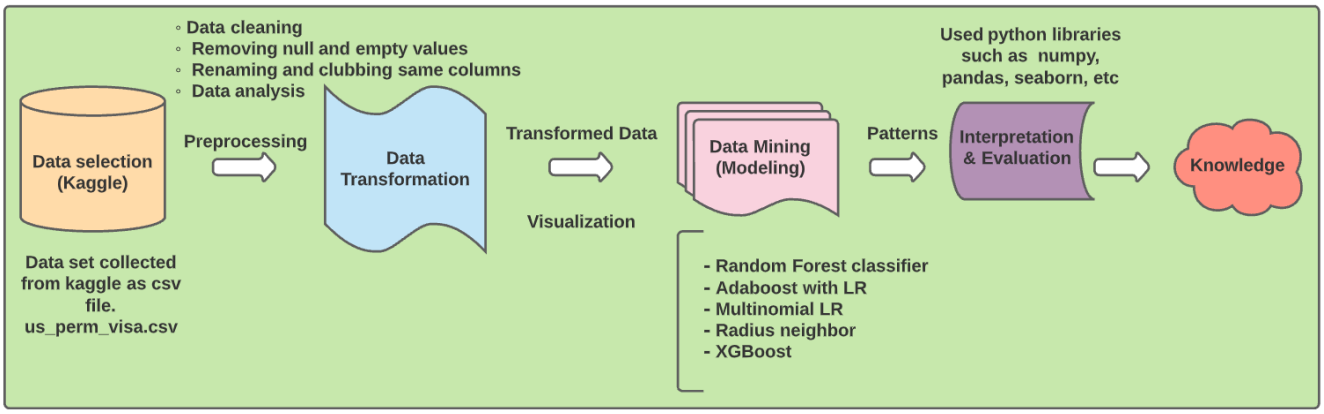

Figure 1: Visa Prediction Methodology

3.2 Design Specification

The project's process flow depicted in Figure 2, is divided into two layers: business logic and display. Dataset extraction as a.csv file was extracted from Kaggle, and data preprocessing and exploratory data analysis were performed in the first layer. Following that, machine learning models were used to obtain the best suited model.

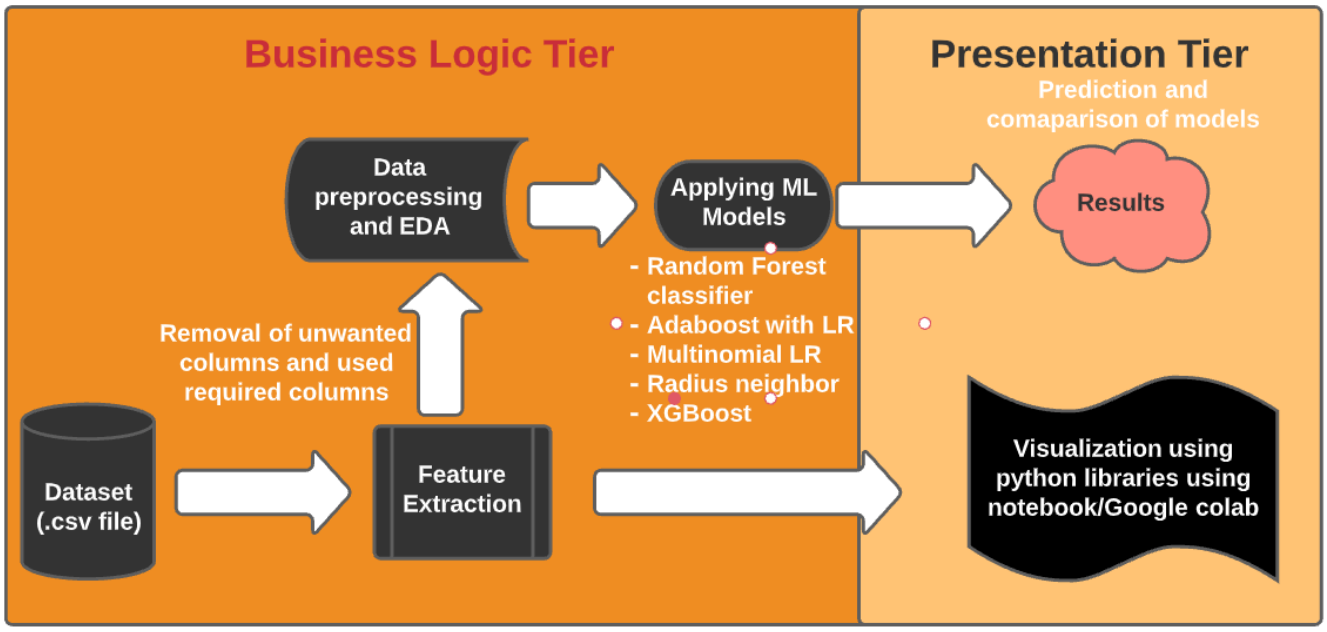

Figure 2: Design Process flow

The building and evaluation of the models took place after the KDD cycle and all of the stages were completed. Since KDD is cyclic, these steps was repeated until the predicted results were not obtained.

3.3 Data Pre-processing and Feature Selection

It is imperative that we pre-process data before putting it into our model because the quality of the data and the information it can provide directly impacts our learning model. Therefore, pre-processing data before feeding it into our model is crucial. To preprocess the data, the dataset was first downloaded as a csv file from kaggle and then converted into a dataframe in a Jupyter notebook. (Because it is a cloud-based platform, this project was also cloned on the Google colab to avoid disk space or ram limitations and compare the operating results.) However, the models produced similar findings and took the same amount of time to execute on both systems.) The raw data set has a total of 374362 rows and 153 columns. As can be seen, there were allegedly comparable columns and rows that were later eliminated as part of the data pre-processing procedure. Consider the 'case number' and 'case no' columns in the data set, which contain similar information. The data contained several comparable columns with null values, and the data was inconsistent. To address the issue of missing values, only columns with less than 12% missing values were chosen for further analysis. Figure 3 shows the results of the non null values. Feature variables were then converted into categorical variables.
The process of reducing the number of input variables in a predictive model is known as feature selection. In some cases, it may be advantageous to minimize the number of model input variables to reduce computational costs. In other words, To determine the relative impact of attribute values on final decision, a feature elimination approach was implemented called Recursive Feature Elimination (RFE). In this approach, features which are need are used and the rest are removed from the data set until the required amount of features are obtained. In this data set, the main variables to select were the case status. Two columns 'certified' and 'Denied were' defined as the 'case_status' with input 0 and 1 against them. Such as 0 means Denied and 1 means certified.

```
0 Column: 'add_these_pw_job_title_9089' contains  10.99 % non-null values
1 Column: 'agent_city' contains  56.92 % non-null values
2 Column: 'agent_firm_name' contains  55.74 % non-null values
3 Column: 'agent_state' contains  56.05 % non-null values
4 Column: 'application_type' contains  35.61 % non-null values
5 Column: 'case_received_date' contains  64.39 % non-null values
6 Column: 'case_status' contains  100.0 % non-null values
7 Column: 'class_of_admission' contains  94.08 % non-null values
8 Column: 'country_of_citizenship' contains  94.59 % non-null values
9 Column: 'country_of_citzenship' contains  5.4 % non-null values
10 Column: 'decision_date' contains  100.0 % non-null values
11 Column: 'employer_address_1' contains  99.99 % non-null values
12 Column: 'employer_address_2' contains  60.4 % non-null values
13 Column: 'employer_city' contains  100.0 % non-null values
14 Column: 'employer_country' contains  64.37 % non-null values
15 Column: 'employer_decl_info_title' contains  64.37 % non-null values
16 Column: 'employer_name' contains  100.0 % non-null values
17 Column: 'employer_num_employees' contains  64.36 % non-null values
18 Column: 'employer_phone' contains  64.38 % non-null values
19 Column: 'employer_phone_ext' contains  6.32 % non-null values
20 Column: 'employer_postal_code' contains  99.99 % non-null values
```

Figure 3: Displaying first 20 non null values in the data set

3.4 Exploratory Data Analysis

An integral part of every Data Analytics study is exploratory data analysis (EDA). It involves examining a dataset to discover patterns, inconsistencies, and anomalies, and making hypotheses about the data based on what is known about it. EDA entails constructing various graphical representations for numerical data and generating summary statistics for the dataset to better comprehend the data. Because the data collection was so large, there were a lot of things to look at. A few observations were taken in order to arrive at the intended result. It was tough to reach every point in the dataset due to the time constraint. The observations that were made is discused in this section while showcasing the results achieved from EDA.
Figure 4 shows the total number of visa applications per year. Interestingly the visa applications were increased as the years passed. Also, number of denial cases was seen similarly same since 2013.

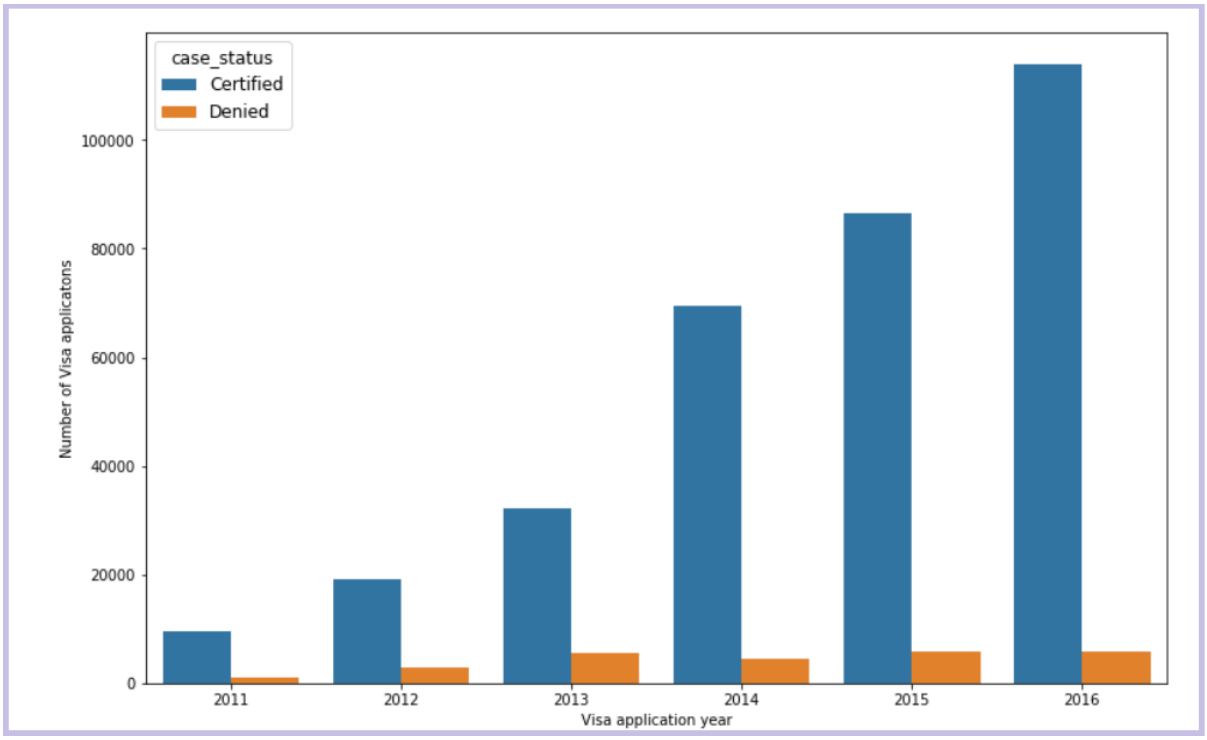

Figure 4: Total number of visa applications from year 2011 to 2016

As illustrated in Figure 5, visa applications have been on the rise in recent years in locations such as New York, Santa Clara, Mountain View, San Jose, and others. These cities were labeled as the most popular cities. In addition, there was a positive trend in visa applications in numerous cities. According to the bar graph, the number of submitted Visa applications at College station was more than twice as high as in other cities in 2015.

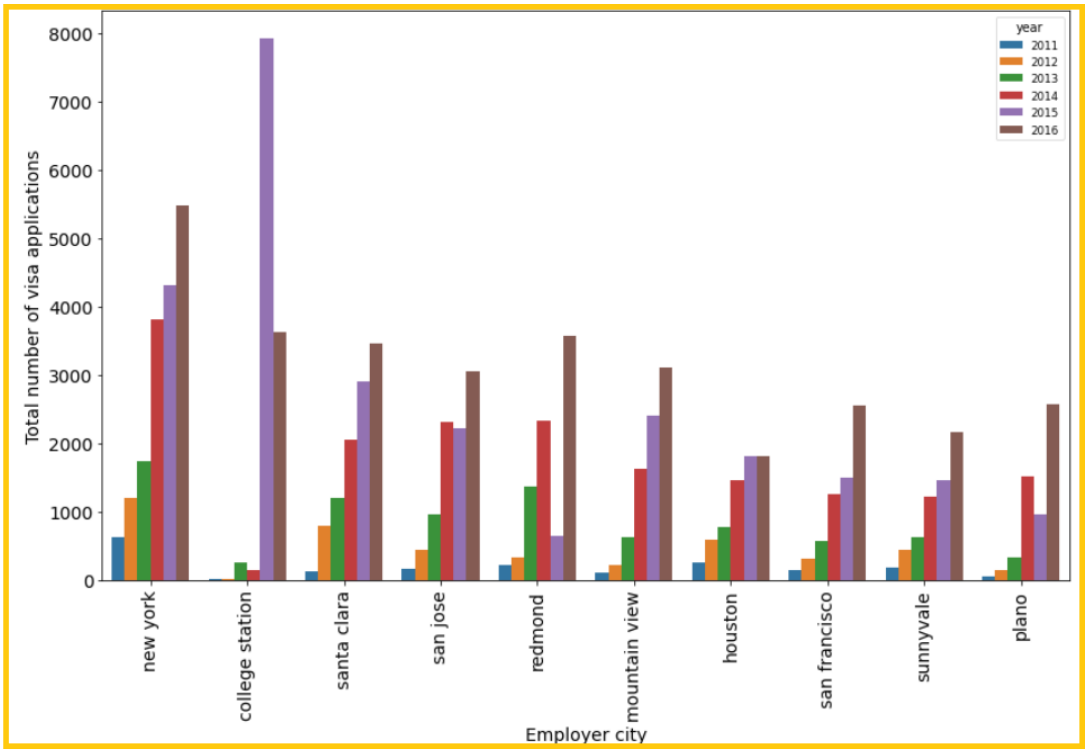

Figure 5: The most popular cities in the United States with a high number of visa applications

It has been discovered that 90 percent of IT organizations are the most useful to visa applicants. According to Figure 6, the IT industry has the highest certified count of visa applicants affiliated with these organizations. People apply for H1B visas more than any other visa category because to the high demand in the IT sector. According to the pie chart below, the IT sector accounts for roughly 45 percent of visa applications, with advanced manufacturing and economic sectors accounting for a significant portion of the remaining applications.

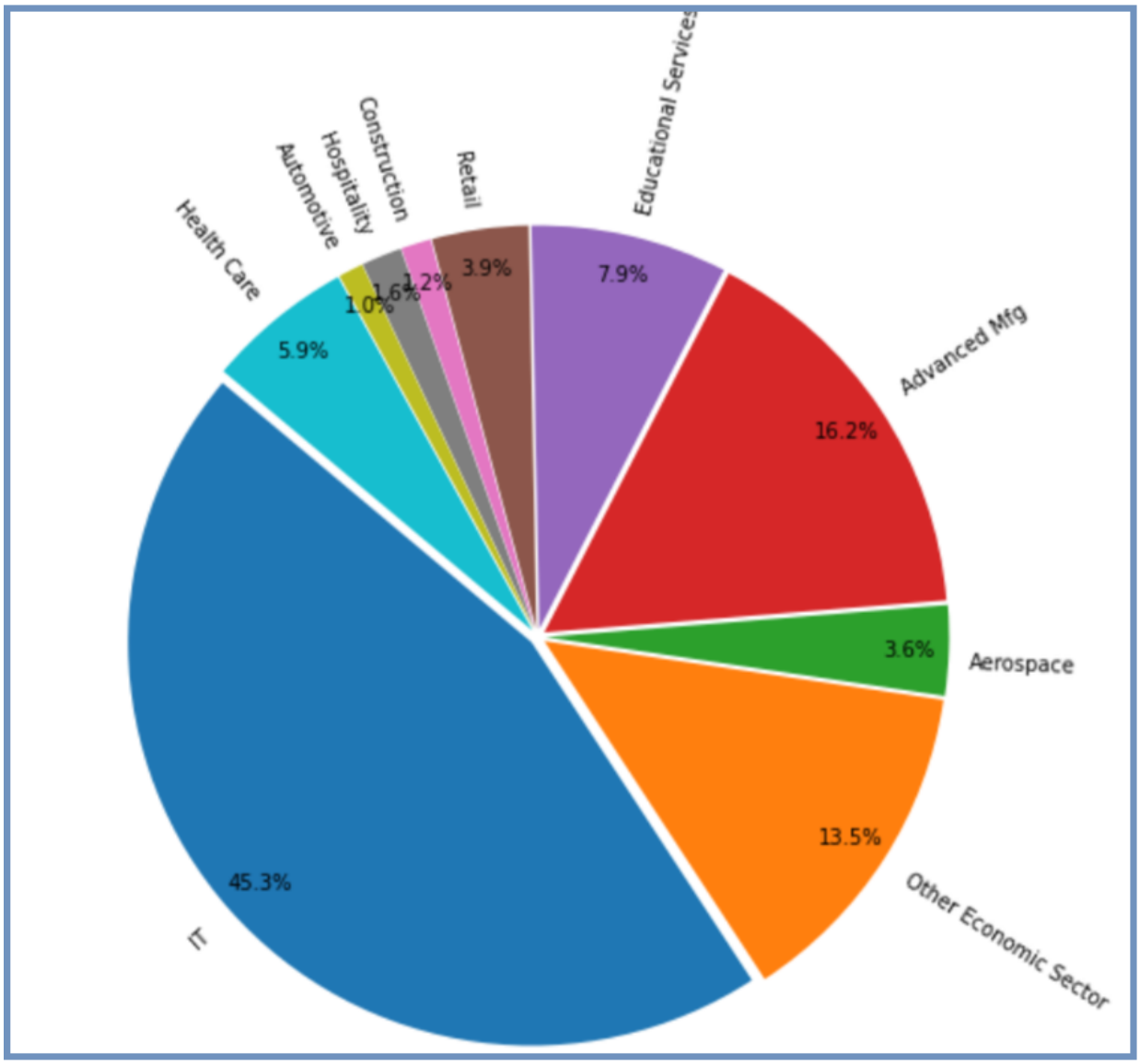

Figure 6: Sector wise distribution of visa applications

Another related observation was that IT companies account for approximately 45 percent of the US economy. Figure 7 depicts the distribution of companies and applications in the United States by sector. The majority of applicants were also found to be applying for the H-1B Visa, which allows U.S. companies to hire foreign workers in specific occupations, according to Wikipedia. If an H-1B visa holder departs the United States or is fired from his or her sponsoring company, he or she must either apply for and be granted another status, find another employer ready to accept an application for adjustment of status or change of visa, or leave the country. Another observation was that, online submission of visa application was most popular among the applicants. More than half of the applicants have a bachelor's or master's degree, indicating that the majority of those who applied were well compensated.

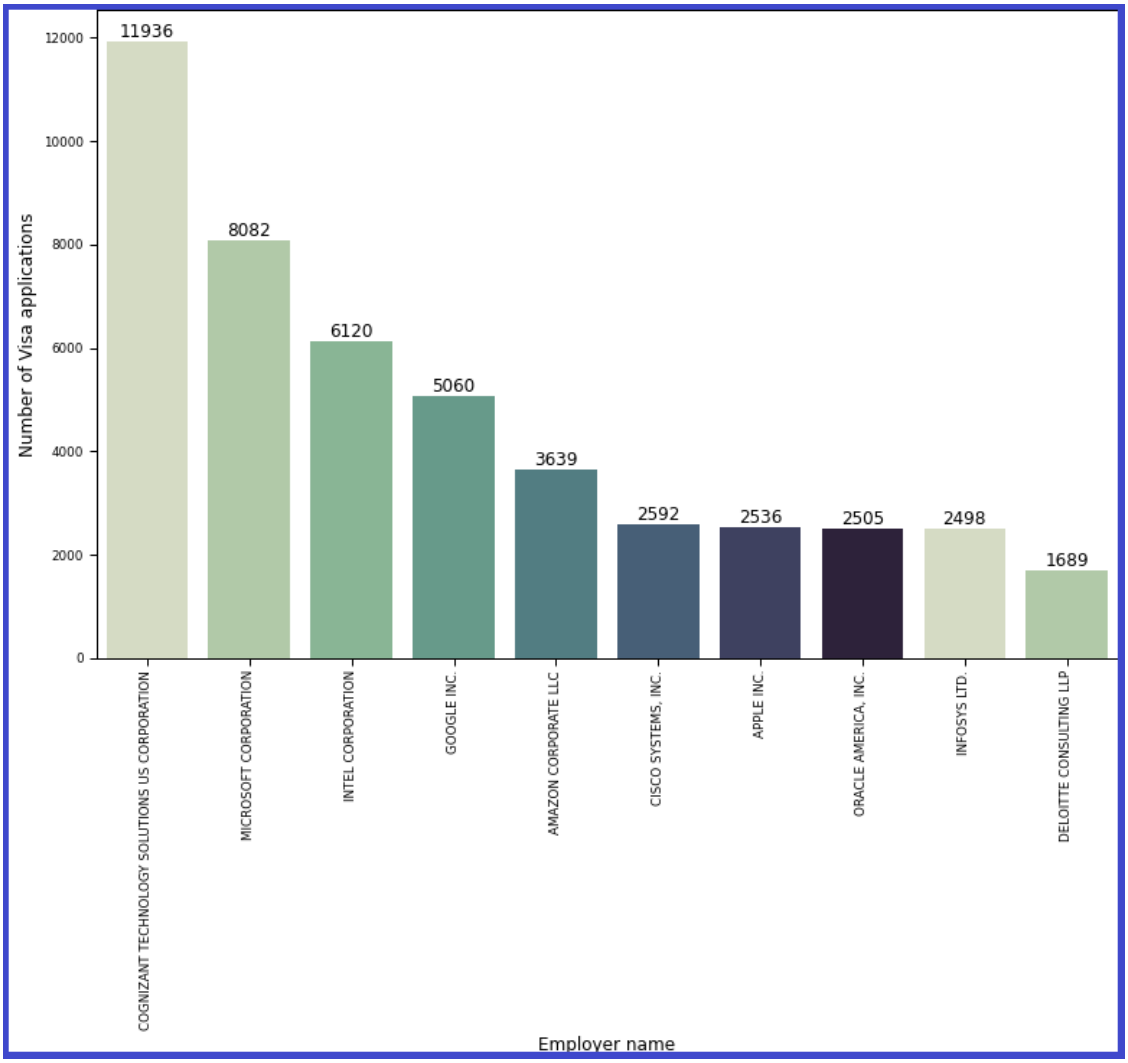

Figure 7: Company wise distribution of visa applications

The data set consists of a variety of visas that are available in the US. Visa type depends on the type of the employment. Hence, it was observed that number of H1B visas was the highest as there is a huge demand of STEM workers in the United States. Figure 8 shows the types of visas and the number of applications for the particular visa. According to Wikipedia,

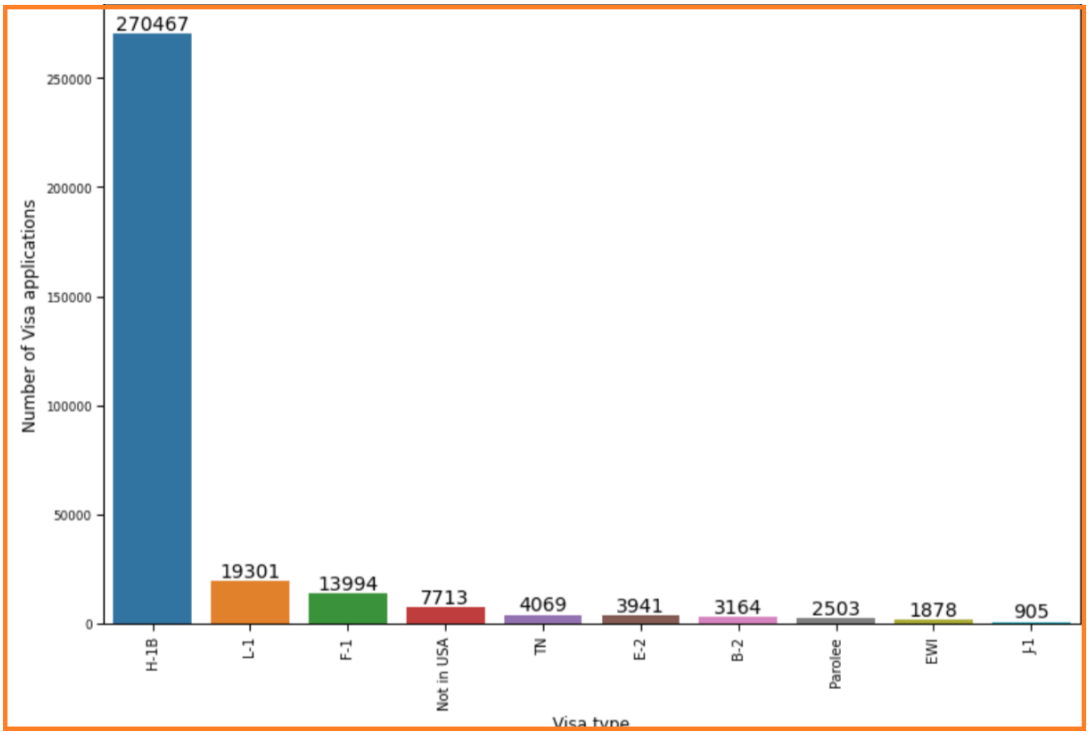

Figure 8: Visa Type and the number of applications

As previously noted, the data collection had identical types of positions, making analysis challenging due to differences in naming practices for the same post. As a result, similar positions were grouped together for easier comprehension. The number of applications for each type of post in the United States is depicted in Figure 9.

```
software engineer                    18582
computer systems analyst             12054
senior software engineer              5802
software developer                    4501
programmer analyst                    3763
assistant professor                   2869
software development engineer         2766
systems analyst                       2587
senior programmer analyst             1884
senior software developer             1625
Name: job_info_job_title, dtype: int64
```

Figure 9: Position wise visa application distribution

We all know how important the IT industry is in the United States. As it can be seen in the graph below, the majority of the applications are for the position of "Software Engineer." Figure 10 depicts all of the IT career positions for which a visa application was submitted.

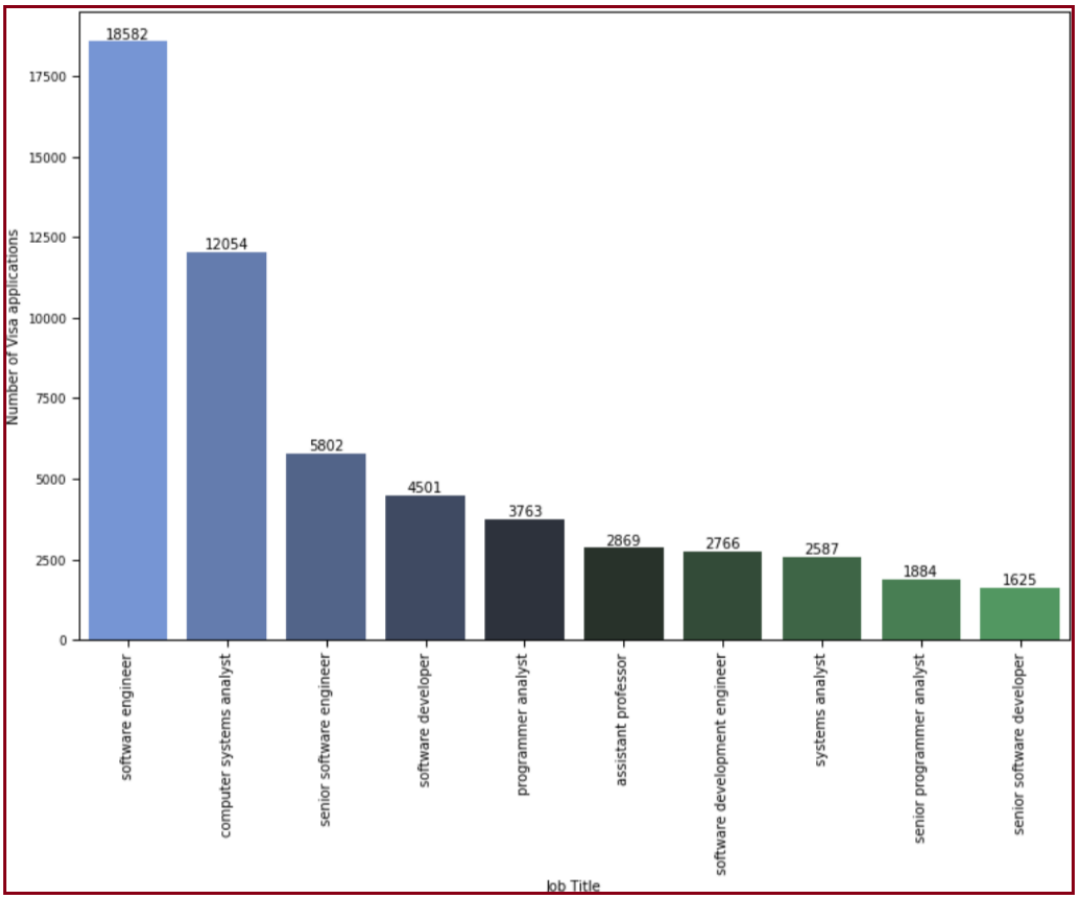

Figure 10: Position wise visa application distribution

There are several combinations that can be used to analyze data, and the data can be visualized in a variety of ways. In this part, only a few of the observations were displayed. The implementation of machine learning models is discussed in the next section.

4 Implementation, Evaluation and Results of Prediction Models of US Visas Outcome Decision

This section describes the implementation, evaluation and discussion of the obtained results. Comparison of 5 models were done and which inclues Multinomial logistic regression, Random forest classifier, Adaboost with logistic regression, Radius neighbors classifier model and XGBoost model.

4.1 Multinomial Logistic Regression

Multinomial Logistic Regression which is an extension of Logistic regression adds buit-in support for 'multiclass' classification problems.

Implementation It uses python library called sklearn and RepeatedStratifiedKFold() function. To collect the scores cross_val_score library was used.

Evaluation and Results

A logistic regression model, also known as Multinomial Logistic regression, is used to adapt and learn a probability distribution. MLR is a more advanced kind of logistic regression that forecasts a multinomial probability of more than two classes for each input. The data set contained a large number of categorical variables and a large number of categories from which specific observations might be derived. It achieved an accuracy of 68%. Figure 11 shows the box plot representation of L2 penalty vs Accuracy.

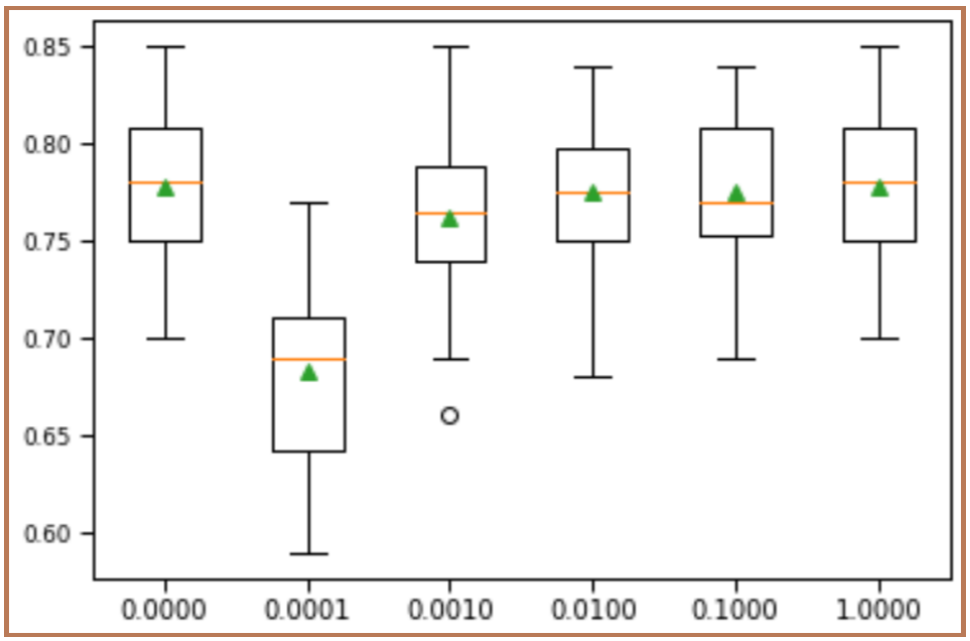

Figure 11: L2 penalty configuration vs Accuracy

4.2 Random Forest Classifier

Random Forest Classifier yet another ensemble ML model. To validate the unvariate data walk_forward_validation library was used.

Implementation It was implemented using python library called sklearn and RandomForestRegressor. To graph the plot pyplot pyhton library was used. Data was splitted using train_test_split python library.

Evaluation and Results

Random forest was the best fit for the data set since it is an ensemble learning method that is used to obtain improved performance for prediction and classification of tasks. It achieved the highest frequency of all the five models, at 93%, with a minimum samples leaf node of 24. The minimal number of samples required at a leaf node is specified by min samples leaf. For example, if min samples split = 5 and an internal node has 7 samples, the split is permitted. It can be mathematically stated as:

$$\sigma = \sqrt{\frac{\sum_{b=1}^{B}(f_b(x') - \hat{f})^2}{B-1}}$$

The Figure 11 below shows the shape of the objective function, sample and a red line showing the best results.

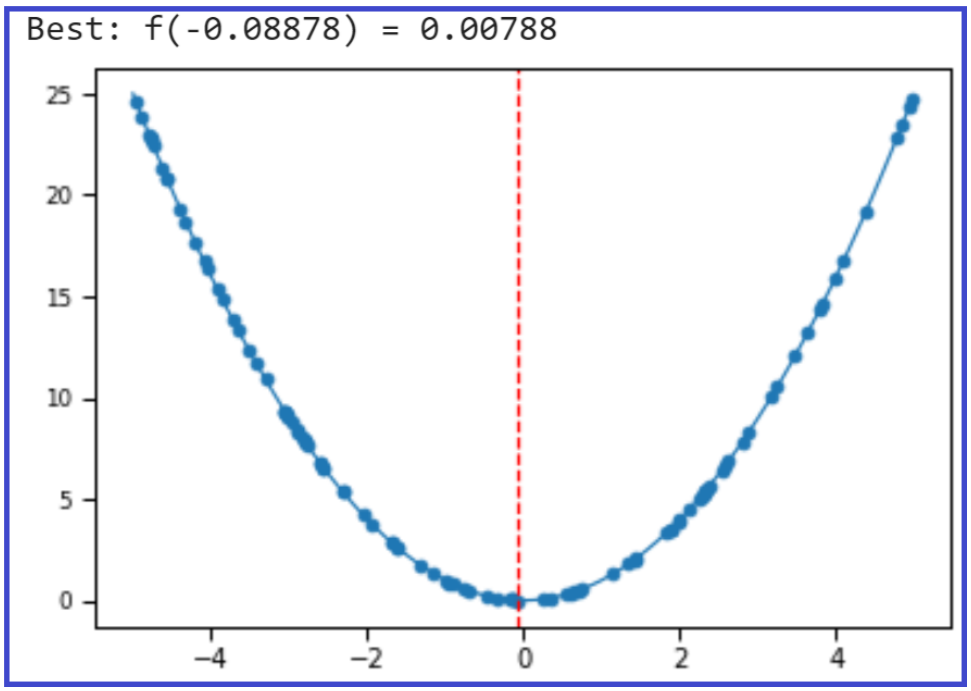

Figure 12: One-Dimensional objective function

4.3 AdaBoost with Logistic Regression

AdaBoost is used from the ensemble decision tree. It is a boosting algorithm and can be used with another algorithm. To collect the scores cross_val_score library was used.
Implementation AdaBoost is implemented using Logistic regression model. It uses python library called sklearn and RepeatedStratifiedKFold() function.
Evaluation and Results
In this model make_classification() function was used to create a synthetic binary classification with 1000 examples and 20 input variables. Then it was evaluated using repeated stratified k-fold cross-validation method using 3 repeats and ten folds. Adaboost with LR achieved an accuracy of 79% which was less than the Multinomial LR model. It achieved 79% accuracy. It can be mathematically represented as:

$$\sum_i w_i y_i \log(p_i) + w_i(1 - y_i)\log(1 - p_i)$$

This model was then evaluated by using k-fold cross validation by using scikit-learn library. Scikit-learn library gives negative values of MAE. Hence it was maximized instead

of minimized. In other words, as compared to MAE score 0, the models are better with maximum scores. Adaboost achived MAE about 72.

```
MAE: -72.899 (4.457)
```

4.4 Radius Neighbor Classifier

Radius Neighbor Classifier is one of the machine learning algorithm that used classification. To create a synthetic binary classification, make_classification() was used with 1,000 examples and 20 input features. To collect the scores cross_val_score library was used.
Implementation This model was implemented using Scikit-Learn library in python. It was implemented by creating a pipeline.It was evaluated using RepeatedStratifiedKFold () method with 10 splits, 3 repeats and 1 random state.
Evaluation and Results
The Radius Neighbors Classifier is a machine learning technique for classification. It achieved an accuracy of 75% which was good than expected. Although, problem arrived while tuning the algorithm. Also, to achieve better results radius of 0.8 was used which improved the accuracy to 87%.

```
Mean Accuracy: 0.872
Config: {'model__radius': 0.8}
```

4.5 XGBoost

XGBoost also knows as Extreme Gradient Boosting algorithm. The xgboost library version in 1.1.1. To create a synthetic binary classification, make_classification() was used with 1,000 examples and 20 input features. To collect the scores cross_val_score library was used.
Implementation The XGBoost model was created using the Scikit-learn API. A synthetic binary classification issue with 1,000 instances and 20 input features was created using the make classification() method.
Evaluation and Results
Using XGBoost, the most powerful technique for building predictive models, was a solid option because it attained 89% accuracy, which was significantly higher than the other five models. It can be represented mathematically as:

$$\hat{f}(x) = \hat{f}_{(M)}(x) = \sum_{m=0}^{M} \hat{f}_m(x)$$

As XGBoost ensemble is fit for all the data. hence predict() function was used on the data to make predictions. Also, model was evaluated using k-fold cross validation using 3 repeats and 10 folds to get the mean absolute error value (MAE) which was about 62.

```
MAE: -62.762 (3.219)
```

Thus to conclude the implementation, Random Forest was the best fit with an accuracy score of 93%. XGBoost performed well after RF with an accuracy score of 89%. Adaboost and RN Classifier gave similar performance with 79% and 75% accuracy scores respectively. Multinomial Logistic regression was on the lowest among all the models with accuracy score of 68%.

Model	Accuracy (%)
Multinomial Logistic Regression	68
Random Forest Classifier	93
AdaBoost with Logistic Regression	79
Radius Neighbor Classifier	75
XGBoost	89

5 Conclusion and Future Work

The research question and objectives specified in section 1.2 of chapter 1 were effectively answered, and the objectives were implemented, evaluated, and the results were presented in chapter 4. This initiative is critical for US visa applicants and anybody considering applying for a visa of any sort. The analysis and models that are conducted on this specific collection of data produce the best possible results and are not used anywhere else. The models utilized were not overly complicated, but they produced the satisfactory results. A thorough model comparison was carried out between five different models (Multinomial Logistic regression, Random Forest, AdaBoost with LR, Radius neighbor classifier and XGBoost). They achieved an accuracy of 68%, 93%, 79%, 75% and 89% respectively with Random forest standing to be the best fit model with the achived accuracy. The data set had concerns since it was disorganized and required a lot of work and time to analyze. Because the H1B visa is the most popular, there have been a lot of research papers and studies done on it. The main focus of this research was on learning outcomes and data analysis, as well as anticipating specific visas. If there was an interface for entering the applicants' inputs and then receiving the visa projections or choices as an output, it might have been more complicated.

In order to commercialize, it would be beneficial to create an interface that takes client input and returns anticipated outcomes relating to their visa status. Artificial intelligence can also be used to provide supplementary feedback. This idea could perhaps be expanded for other countries in the future. The United States was chosen for this study because it receives the most visa applications. Also, because many projects have been completed before, data for US visas is readily available everywhere.

Because there was a lot of trial and error with the models during the project's execution to discover the best model, it was a great learning experience. From the data set, this project provided knowledge about numerous employers, firms, and visas, as well as decision methods for implementing a wide range of models.

Acknowledgement

I would like to express my gratitude to National College of Ireland and the Computing department for their time to time assistance and constant support. My supervisor, Dr. Catherine Mulwa, whose guidance was vital over the two semesters. Your valuable time beyond your work schedule, informative remarks and unwavering support throughout my research endeavor encouraged me to improve my thinking and elevate my work. Finally I'd want to express my gratitude to my parents and friends back home for their constant support throughout my masters journey and beyond.

References

Beliz Gunel, O. C. M. (2018). Predicting the outcome of h-1b visa applications.
URL: *https://cs229.stanford.edu/proj2017/final-reports/5208701.pdf*

Dalmia, N., Murthy, M. and Vivekanandan, N. (2017). Understanding factors that influence l1-visa outcomes in us.
URL: *https://www.ischool.berkeley.edu/sites/default/files/sproject$_a$ttachments/aml$_p$roject$_r$eport.pd*

Dhanasekar Sundararaman, N. P. and Misraa, A. K. (2018). An analysis of nonimmigrant work visas in the usa using machine learning.
URL: *https://dhanasekar-s.github.io/research/3paper.pdf*

Khaterpal, R., Ahuja, H., Goel, J., Singh, K. and Manoj, R. (2020). *Predicting the outcome of H-1B visa using ANN algorithm*, IJRTE.

Kulkarni, T., Nayak, G., Devasthali, R., Dhuri, N., Godbole, S., Sardinha, R. and Nazareth, D. (2019). Data preprocessing and analysis for h-1b visa petitions.
URL: *http://www.iosrjen.org/Papers/Conf.19017-2019/Volume-3/9.%2049-54.pdf*

Kumar, M. S. and Naresh, N. (2018). A predictive model for h1-b visa petition approval.
URL: *https://bit.ly/3GmcYki*

Omar, A. (2019). Predicting h-1b status using random forest.
URL: *https://towardsdatascience.com/predicting-h-1b-status-using-random-forest-dc199a6d254c*

P., C., M.S., V. and T, J. (2021). Success of h1-b visa using ann.
URL: *https://doi.org/10.1007/978-981-33-4859-2*$_4$8

Pandya, D. A. (2018). Predicting filed h1-b visa petitions' status.
URL: *https://www.irjet.net/archives/V5/i8/IRJET-V5I859.pdf*

Parey, M., Ruhose, J., Waldinger, F. and Netz, N. (2015). The selection of high-skilled migrants.
URL: *https://ftp.iza.org/dp9164.pdf*

Peri, G., Shih, K. and Sparber, C. (2015). Stem workers, h-1b visas, and productivity in us cities, *Journal of Labor Economics* **33**(S1): S225–S255.
URL: *http://www.jstor.org/stable/10.1086/679061*

Prateek and Shweta, K. (2019). *Predicting the outcome of H-1B visa using ANN algorithm*, IJRTE.

Roy, R. (2021). Data analysis of h1b visa applications.
URL: *https://lib.dr.iastate.edu/creativecomponents/797*

Ruiz, N. G., Wilson, J. H., and Choudhury, S. (2012). The search for skills: Demand for h-1b immigrant workers in u.s. metropolitan areas.
URL: *https://www.brookings.edu/wp-content/uploads/2016/06/18-h1b-visas-labor-immigration.pdf*

Sampath, V. S., Flagel, A. and Figueroa, C. (2009). A logistic regression model to predict freshmen enrollments.

Sundararaman, D., Pal, N. and Misraa, A. K. (2017). *An analysis of non-immigrant work visas in the USA using Machine Learning*, Vol. abs/1711.09737, ArXiv.
URL: *https://bit.ly/3DfFKRv*

Swain, D., Chakraborty, K., Dombe, A., Ashture, A. and Valakunde, N. (2018). *2018 International Conference on Advanced Computation and Telecommunication (ICACAT)*, IEEE.

Tandon, A. (2021). Analysis of immigration trends in the u.s. to discover patterns and make better policy decisions.
URL: *https://scholarworks.lib.csusb.edu/cgi/viewcontent.cgi?article=2387context=etd*

Thakur, P., Singh, M., Singh, H. and Rana, P. S. (2018). An allotment of h1b work visa in usa using machine learning.
URL: *https://bit.ly/3oEukTj*

Vyas, H. and Prakash, S. (2018). Likelihood of a work visa approval.
URL: *https://cs229.stanford.edu/proj2019aut/data/assignment*$_3$08832$_r$*aw*/26601658.*pdf*