Visualization and Prediction of US Visa Application by Using Machine Learning Techniques

Authors: Joshua Sackey, Blaise Ayirizia

University: University of Texas at El Paso

Date: May 6, 2024

Executive Summary

A permanent labour certification issued by the Department of Labor (DOL) allows an employer to hire a foreign worker to work permanently in the United States. In most instances, before the U.S. employer can submit an immigration petition to the Department of Homeland Security's U.S. Citizenship and Immigration Services (USCIS), the employer must obtain a certified labour certification application from the DOL's Employment and Training Administration (ETA).

In this study, our focus is on visualizing the US permanent visa applications dataset and to predict visa decisions based on some factors by using three different machine learning techniques.

The data analysed covers years 2012-2017 and includes information on employer, position, wage offered, job posting history, employee education and past visa history, associated lawyers, and final decision.

The original dataset contains 374,362 observations, characterized by the same 154 labelled variables, each of them representing "Certified", "Denied", "Certified-Expired" or "Withdrawn" decisions. Since the data contain some records with case status "Withdrawn", we remove them from our dataset and for cases where status is "Certified" or "Certified Expired" we used just one value "Certified" so that we will end up having only the desired categories namely "Certified" and "Denied". Thus, our new datasets reduce to a total 356,168 applications described by 153 attributes each representing either Certified or Denied of a visa applicants' status. Among the total 356,168 observations, about 93% of the applicants had certified applications and 7% denied. The data sets were obtained from Kaggle.

Three machine learning algorithms were applied to datasets. Among these techniques, the tree-based models, turned out to be the best and stable binary classifiers as they properly create split directions, thus keeping only the efficient information.

1. Introduction

Motivation and Objectives

A permanent labour certification issued by the Department of Labor (DOL) allows an employer to hire a foreign worker to work permanently in the United States. In most instances, before the U.S. employer can submit an immigration petition to the Department of Homeland Security's U.S. Citizenship and Immigration Services (USCIS), the employer must obtain a certified labour certification application from the DOL's Employment and Training Administration (ETA). The DOL must certify to the USCIS that there are not sufficient U.S. workers able, willing, qualified, and available to accept the job opportunity in intended employment and that employment of the foreign worker will not adversely affect the wages and working conditions of similarly employed U.S. workers.

The goal of this project is to visualize and analyse the US Visa Applications using the training data set and based on the testing data set predict which applicant gets certified or denied. Based on these predictions we will assess which of these machine learning techniques performs best. In addition, we will showcase the importance of the choice of machine learning algorithms, the selection of relevant variables, the role of the evaluation criteria and the importance of humans when it comes to making final decision. Based on the results, prospective foreign workers who want to live and work in the US can have an idea of the US Visa application system.

2. Methodology

2.1 Models

In this project, we analysed the data set using three machine learning algorithms namely.

- Logistic Regression: A traditional binary classifier used to model the probability of a binary outcome.
- Random Forest: An ensemble learning method that constructs multiple decision trees and combines their predictions to improve accuracy.
- K-Nearest Neighbours (KNN): A non-parametric method used for classification and regression tasks based on similarity measures.

Data Collection and Preprocessing

The dataset comprises 356,168 observations with 153 labelled variables, representing decisions such as "Certified," "Denied," "Certified-Expired," or "Withdrawn." It presents a binary classification problem, with 93% of applicants being certified and 7% denied. Features include employer information, job details, class of admission, and country of citizenship.

Exploratory Data Analysis

Descriptive statistics and visualizations are used to explore the distribution of variables and identify patterns within the data. Key visualizations include bar plots depicting the distribution of certified vs. denied applications and the relationship between various features and visa outcomes.

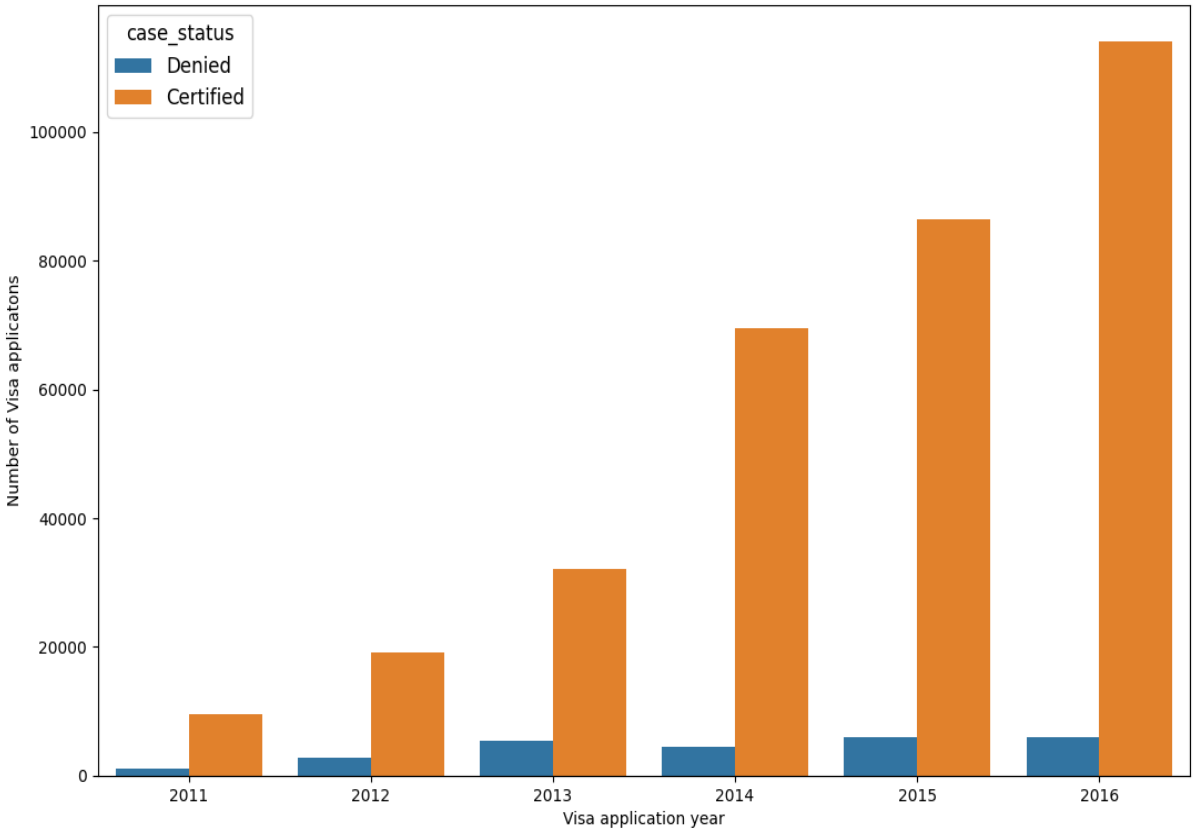

Figure 1: Plot of Certified vs Denied Based on Years

As we can observe, the number of submitted Visa applications increases every year. It's interesting that while the number of positively considered applications increases, the number of "Denied" ones seems to be similar from year 2013. As a next step, let's see, what where the most popular cities.

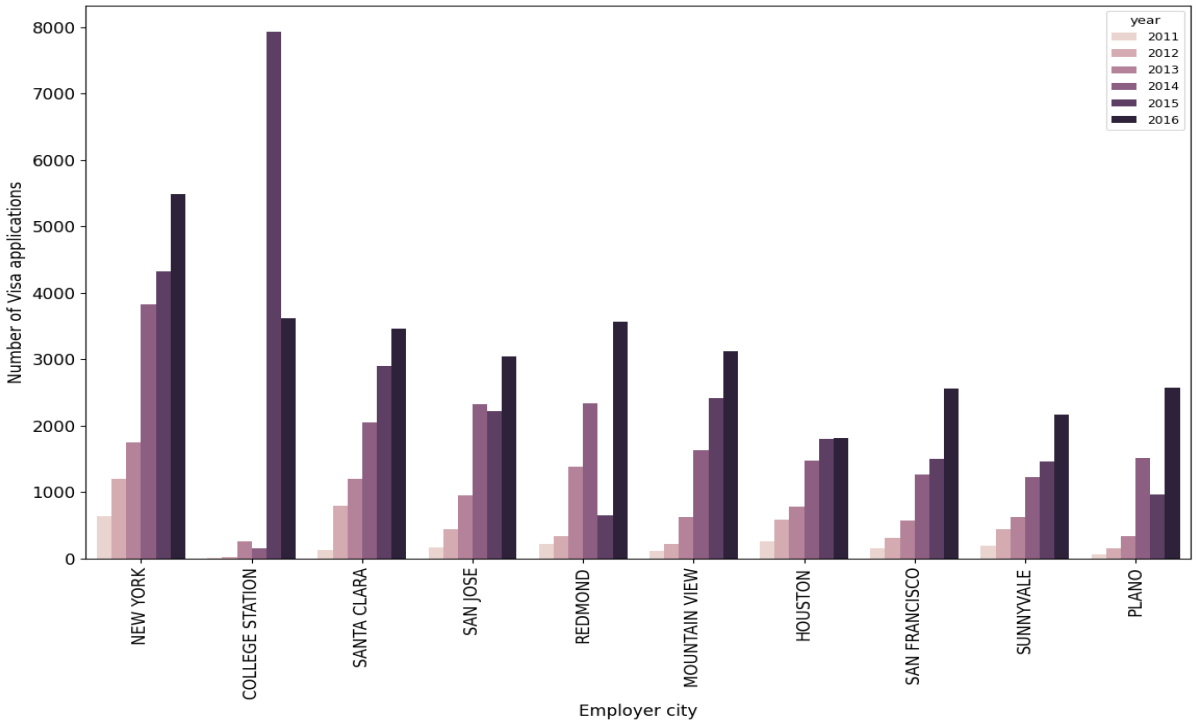

Figure 2: Employer City Vs Number of Visa Applications

From the diagram, we can see that the most popular destination cities were New York, College Station, Santa Clara, San Jose, Redmond, Mountain View, Houston, Sunnyvale, San Francisco, and Plano. In most of the cities there was a positive trend in Visa applications. A bizarre situation occurred in College Station in 2015 where the number of submitted Visa applications was twice large as in other cities.

Now, let's look what were the most hiring employers and economic sectors through these years. For "us_economic_sector" variable we have only 120 868 non-missing values, but this should give us an insight.

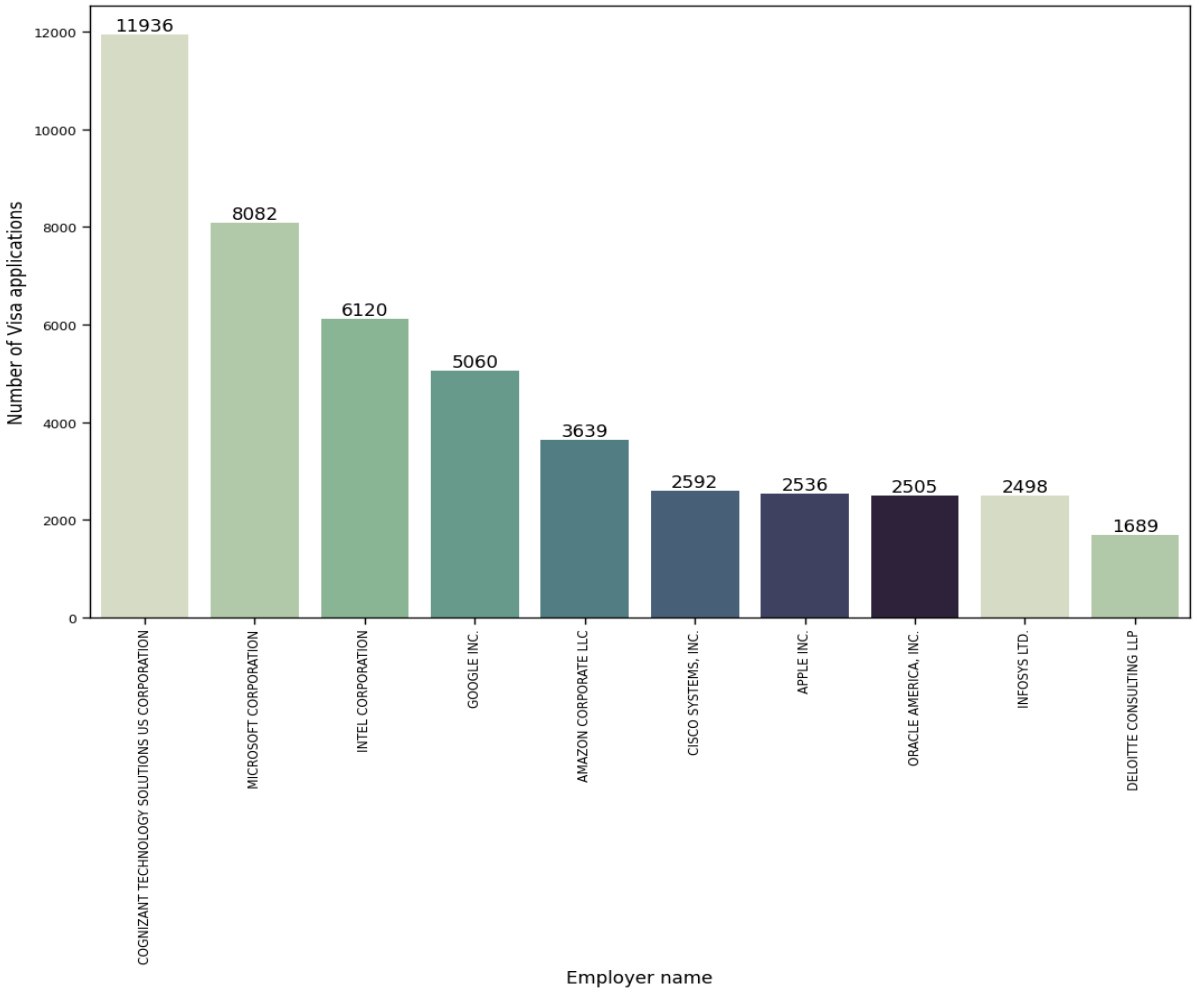

Figure 3: Employers Vs. Visas Certified

As we can see, 9 out of 10 most beneficial companies for Visa applicants are IT industry representatives. This leads to the assumption that IT sector is both most favourable and demanding one in United States. Let's check what is the distribution of industries across all Visa applications.

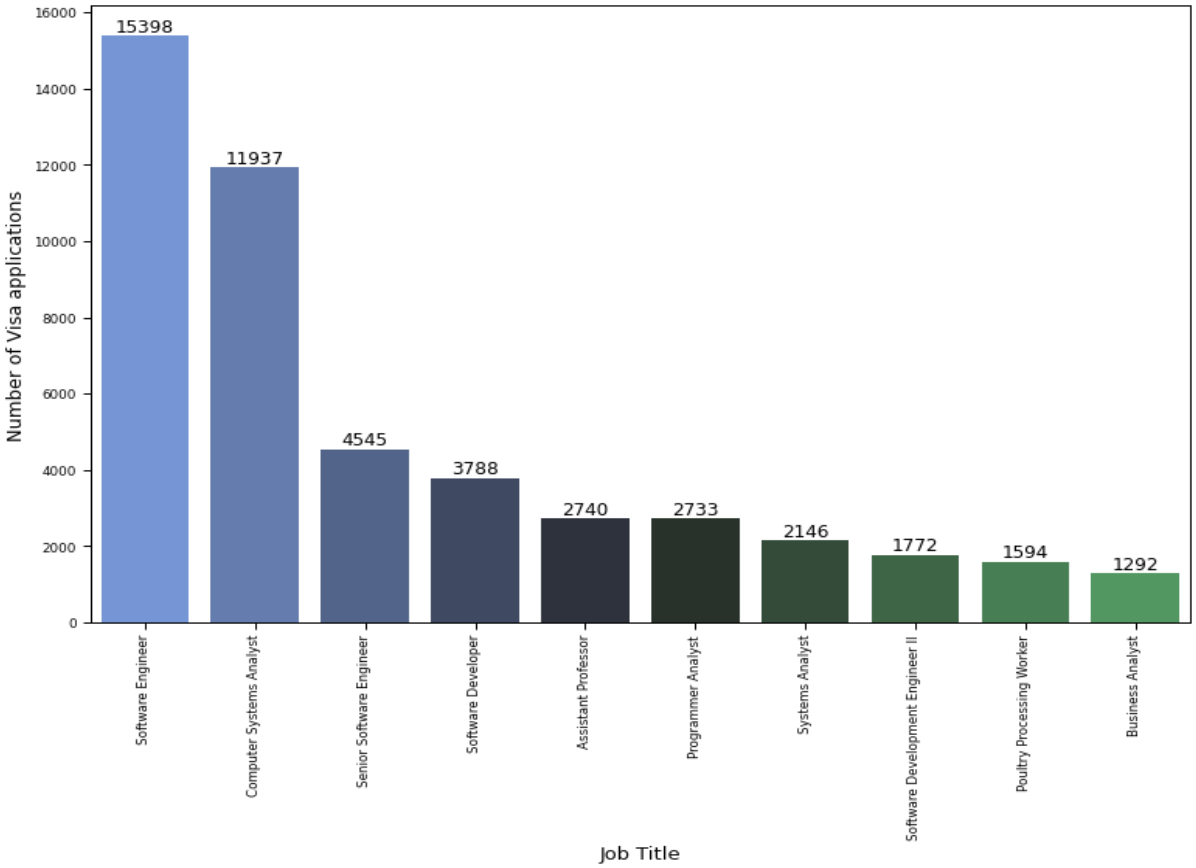

Figure 4: Job Title Vs Number of Visa Applications

Interestingly, most of the popular positions except "assistant professor" are derived from the IT industry. This is another confirmation that there is a huge demand for IT specialists in USA and being one of them increases our chances to obtain a permanent Visa.

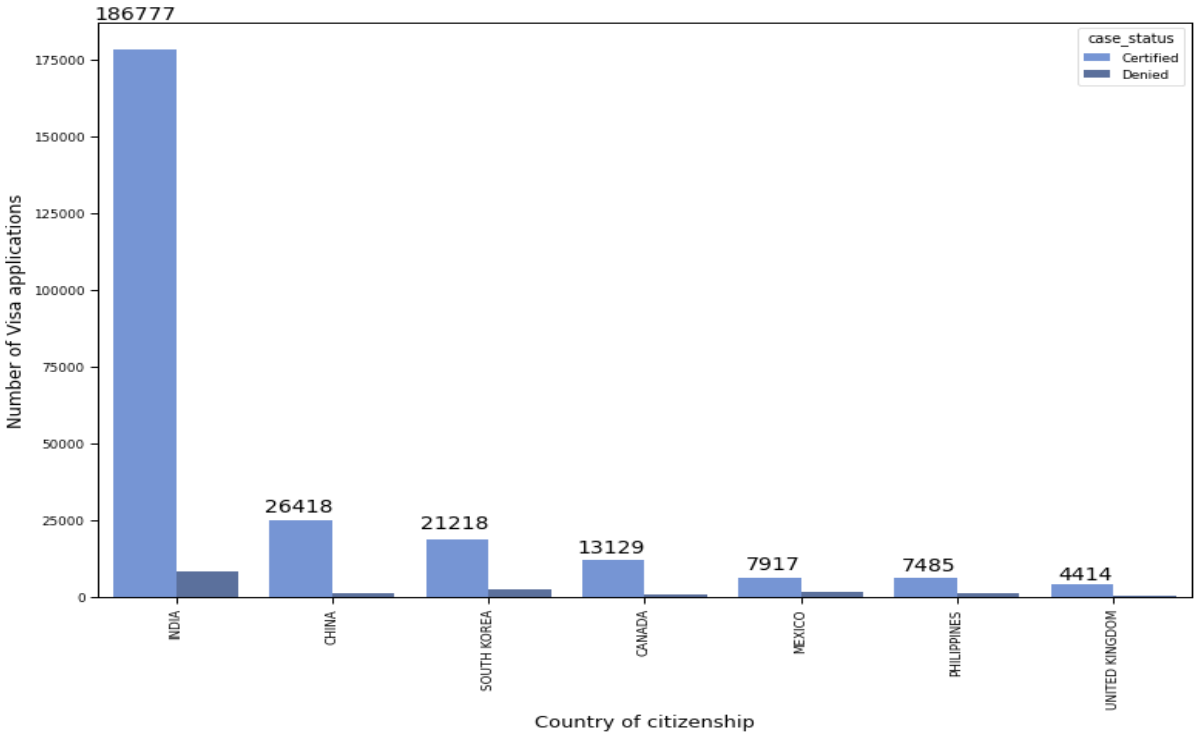

Figure 5: Country of Citizenship Vs Number of Visa Application

As we can see, most Visa applications has been submitted by Indian citizens. They constitute to more than half of our observations, we can assume that most of them are computer specialists.

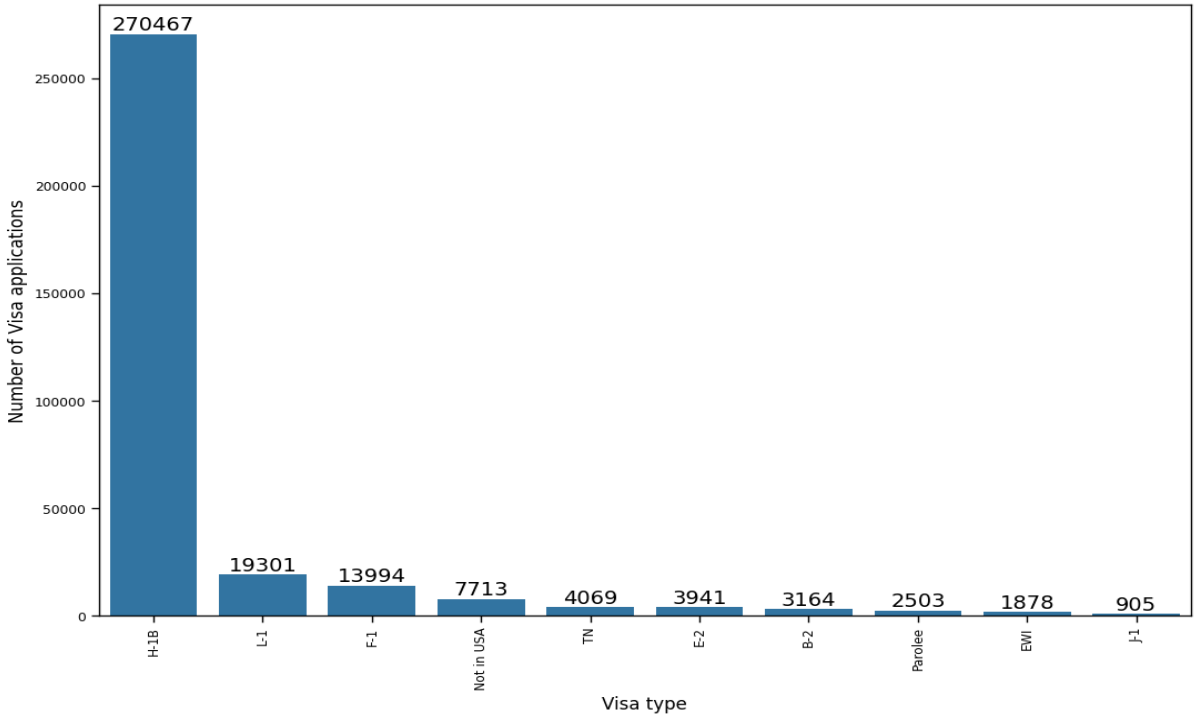

Figure 6: Visa Type Vs Number of Visa Applications

Most petitioners were applying for the H-1B Visa, which according to the Wikipedia, allows U.S. employers to employ foreign workers in specialty occupations.

Finally, let's try checking on the number and kind of application types.

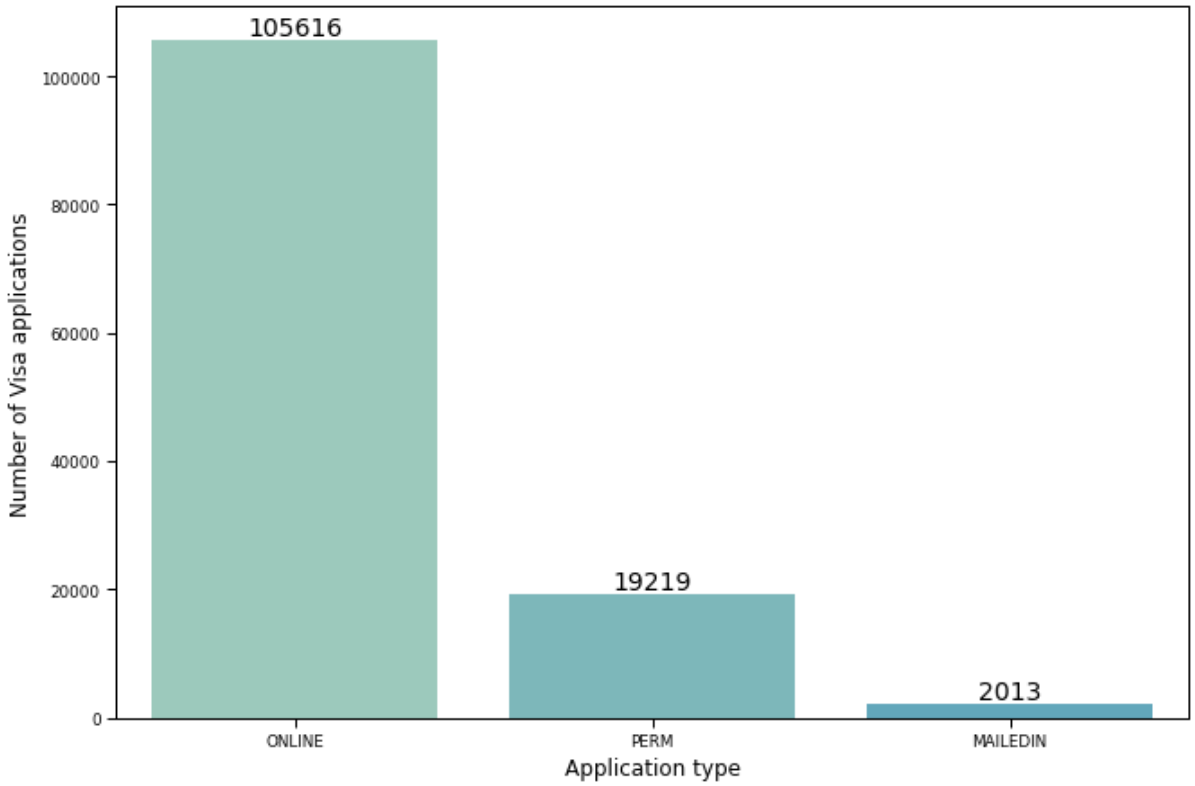

Figure 7: Application Type Vs Number of Visa Applications

Online submission was the most popular form of application type. The last plotting activity will be displaying the applicant's education level and remuneration.

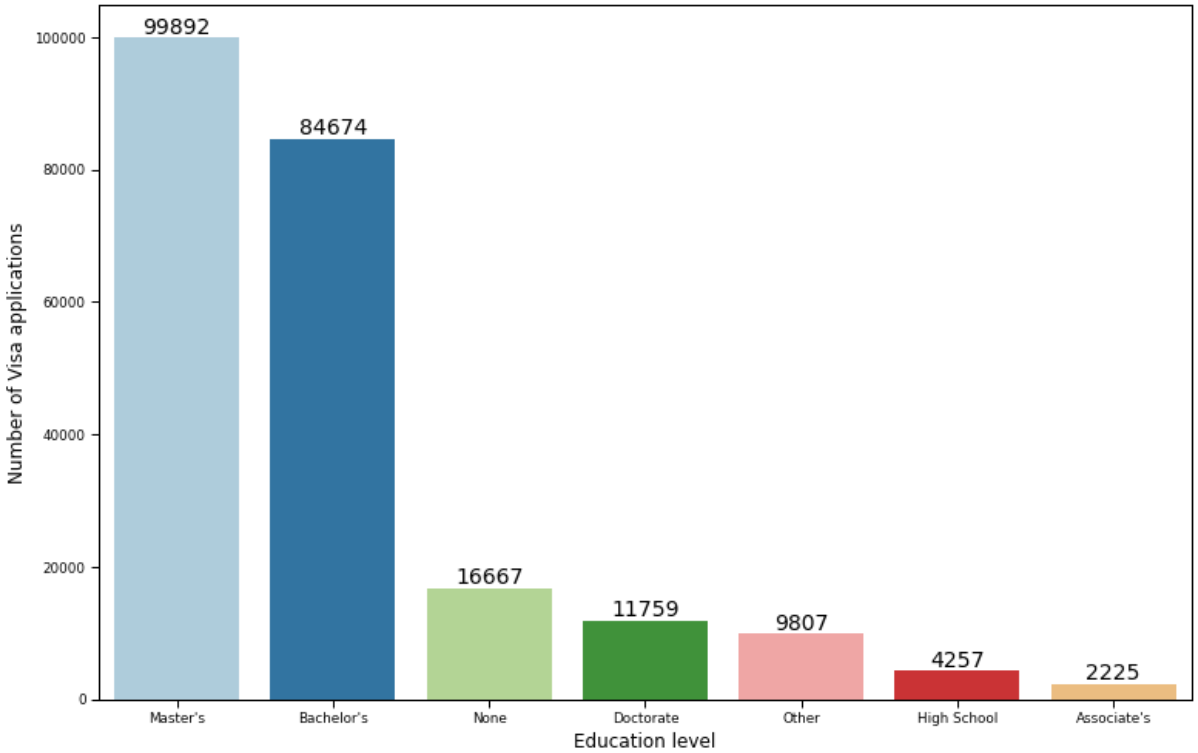

Figure 8: Education Level Vs Number of Visa Applications

As we can see, over 50% of applicants obtained a university degree.

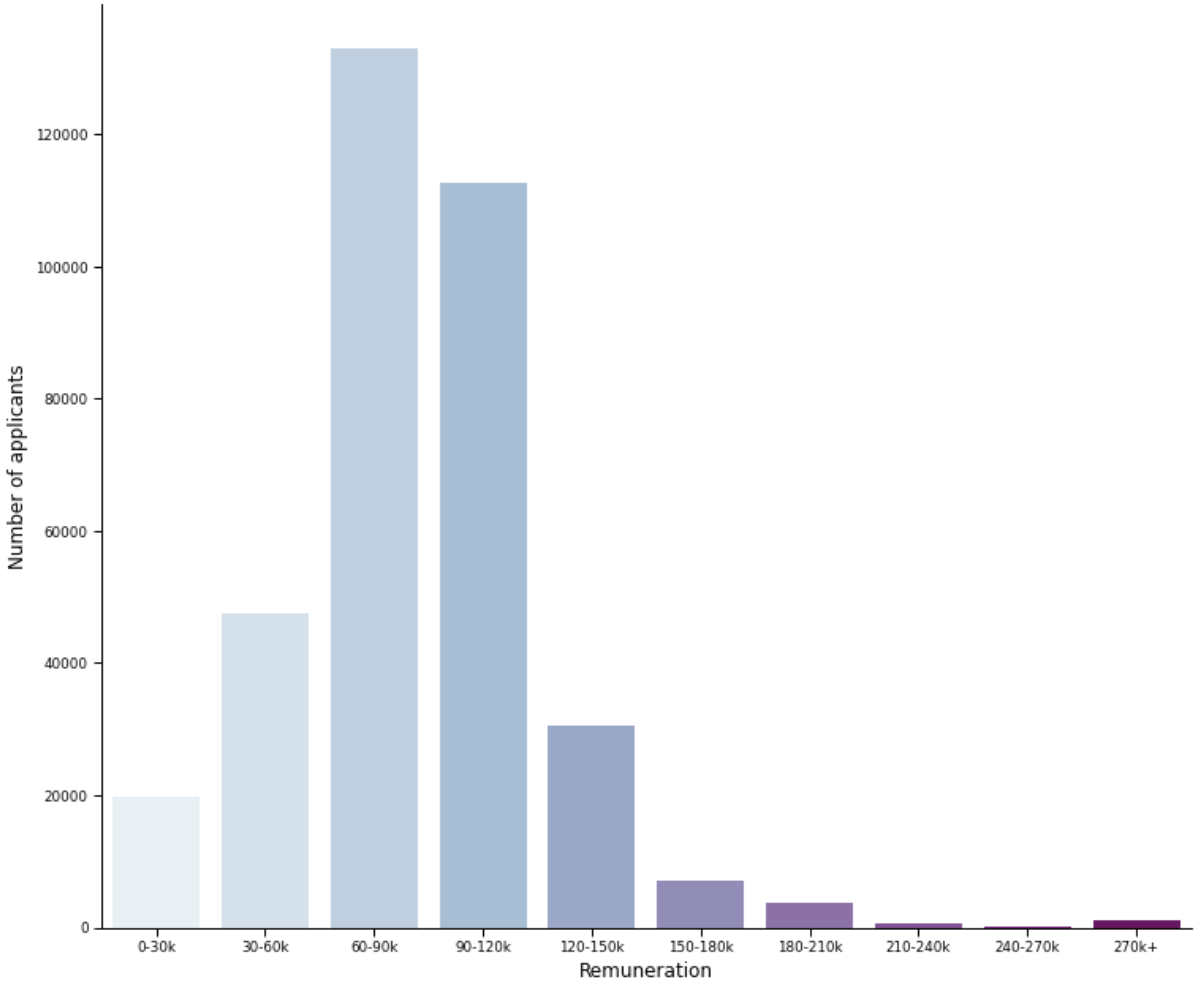

Figure 9: Remuneration Vs Number of Applications

Data Cleaning and Preparation

The dataset, consisting of 356,168 applications described by 154 attributes, underwent rigorous preprocessing to ensure data quality and model robustness. Attributes with at least 60 percent missing values were identified and subsequently removed from the dataset to mitigate potential biases. Missing values within the remaining parameters were then imputed using a two-step process: numeric missing column values were replaced with the respective column means, while character and factor missing column values were substituted with the column mode.

Given the computational constraints and the sheer volume of visa applicants, the analysis was focused on a subset comprising 26,418 applicants from China. Despite this reduction, the dataset still retained substantial representativeness. Within this subset, a notable class imbalance was observed in the distribution of the response feature (case status), with 24,931 applications certified and 1,487 denied. To address this imbalance, the dataset was randomly partitioned into training and test samples in a 2:1 ratio.

Furthermore, to rectify the class distribution imbalance, Random Over-Sampling (ROSE) was employed. This technique involved replicating instances of the minority class (denied applications) to match the frequency of the majority class (certified applications), thereby ensuring a balanced representation in the dataset.

Model Evaluation

The performance of each model is evaluated using metrics such as accuracy and Mean Squared Error (MSE). Both imbalanced and balanced datasets are considered to assess model robustness.

3. Results and Discussion

The data set was randomly divided into the training sample and the test sample with a ratio of 2:1. The training sample was used to build the three models and then the test sample was used to check the prediction accuracy of these models. Using the 8 features, the results was presented when the three models were applied to the test dataset.

There are several performance measures to compare the performance of the models. However, the results would be presented based on the accuracy of the models.

Due to the large number of visa applicants and less computer power, the analysis was focused on the 26418 visa applicants from China which is still a large number. Among this, the distribution of the response feature (case status) was 24931 Certified and 1487 Denied applications which is obviously a class imbalance problem. The class distribution of the data set was readjusted by performing Random Over-Sampling (ROSE). Random Over-Sampling (ROSE) solves the class imbalance problem.

The imbalanced data was used for the baseline model to predict the visa status and the corrected imbalance data was for the actual predictions and the performance of the model in each case was compared.

From the results presented in the table, random forest gave the best accuracy among the models built with unbalanced data set. This value is not significantly different from the other values obtained by the Logistic Regression (LR) and K-Nearest Neighbours (KNN) models. Among the models built by the balanced data set, again the random forest with the over sampling which falls under the random forest model had the best accuracy. The top four (4) topmost important features are the type of job, size of company, educational level and the salary of employees which happened to be the variables that predict certification or denial of permanent US Visa.

TYPE	ACCURACY	MSE
Imbalance	0.94	0.057
Balance	0.74	0.26

Table 1: Accuracies and MSE's of the LR model

TYPE	ACCURACY	MSE
Imbalance	0.95	0.0549
Balance	0.95	0.0495

Table 2: Accuracies and MSE's of the RF model

TYPE	ACCURACY	MSE
Imbalance	0.95	0.0649
Balance	0.87	0.1326

Table 3: Accuracies and MSE's of the KNN model

5. Conclusion

This study demonstrates the utility of machine learning techniques in predicting US visa outcomes. By leveraging a comprehensive dataset and employing state-of-the-art models, valuable insights are gained into the factors influencing visa certification or denial. The findings have implications for employers, applicants, and policymakers involved in immigration processes.