Predicting US Permanent Visa Application Decisions using Machine Learning

Aneesh Phatak, Sankalp Singh, Ying Lin

School of Information Studies, Syracuse University
aaphatak@syr.edu, ssingh56@syr.edu, ylin65@syr.edu
https://data-analytics-ist707-project.herokuapp.com

Abstract

The United States of America is known as the land of opportunities. A large number of international students and workers come to the US every year to pursue higher studies or for work related purposes. But, for a foreign citizen to work in the US, their employer must obtain a permanent labor certification. They then need to submit an immigration petition as well for their worker. There are several nonimmigration visas like H1-B, F1, J1, L1 etc. that lets a non-US citizen work in the US. Every year, tech giants like Google, Apple, Amazon etc. file for thousands of immigration applications for their foreign workers. Visa approval is a strenuous process which can take up to years to get approved. There are a lot of times when these long due applications are denied and that leads to major problems for the employers and their foreign workers. In this paper, we have developed a system that leverages machine learning models to figure out the key factors that contribute to the approval and denial of a visa application.

1 Introduction

Visa application decisions have a lot of significance for a variety of employers and their international workers across the US. A denied visa application decision can lead to severe consequences for both the international workers and their employers. This is a major societal problem that needs to be addressed. To tackle this problem, we have designed a system using machine learning algorithms to uncover the crucial features that are helpful in predicting whether a particular visa application will be approved or denied.

We have experimented with various machine learning models for our research work. We have implemented several classification-based models including Logistic Regression, Random Forest, Gradient Boosting Machine, Linear Support Vector Machine and Artificial Neural Networks to tackle the visa application decision problem. We have evaluated our models using the F1-score evaluation metric.

Section 2 defines our problem statement and the dataset that we have used for our research work. Section 3 introduces the approach and algorithm that we have implemented. Section 4 shows the results obtained for our research work. Section 5 and 6 summarizes the discussion, conclusion and future scope of our research.

2 Problem and Data Description

For our experiments, we chose to work with the visa application domain. Being

international students in the US ourselves, we thought of implementing our research in this domain to gather key insights on the visa application decision process. The objective of our research work is to design a system that can predict US permanent visa application decisions based on various input features. We have also tried to identify the most important factors that help in predicting if a certain visa application will be approved.

For our research work, we have used the 'US Permanent Visa Applications' dataset. This data has been collected and distributed by the US Department of Labor. This dataset is available on Kaggle and contains detailed information on 374000 visa application decisions between 2011 and 2016. The dataset consists of input features like employer name, employer city, employer state, employee education, class of admission, case number, case status (target attribute) etc.

3 Approach & Algorithm

For our research work, we have followed the CRISP-DM approach which is the 'Cross Industry Standard Process for Data Mining'. It consists of following steps:

- Business & Data Understanding
- Data Pre-processing & Exploratory Data Analysis
- Modeling
- Evaluation
- Deployment

Below we have showcased the end-to-end architecture of our visa application decision prediction system. First step is business and data understanding. Next step is data preprocessing which consists of techniques like data transformations, exploratory data analysis, feature selection, handling missing values, feature engineering, normalization and scaling. This step is followed by the modeling step. Next is the model evaluation part followed by the final step of system deployment.

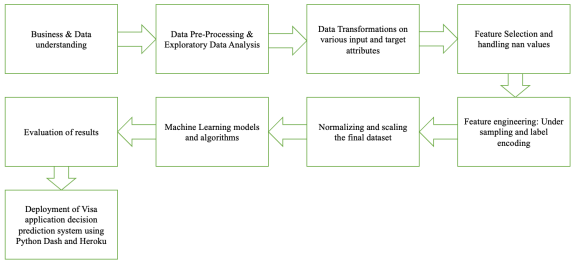

Figure1. Overall system architecture

3.1 Business & Data Understanding

For this step, we have tried to understand the overall impact our system would have on society. For the data understanding part, we initially went through each attribute of the dataset to try and understand what information each of them conveys.

3.2 Data Pre-Processing & Exploratory Data Analysis

For this part, we have performed several data cleaning and preprocessing techniques. Initially, we have performed data transformations on the target feature. Our target attribute 'case_status' initially consisted of four distinct values - 'Certified-Expired', 'Certified', 'Withdrawn' and 'Denied'. For our research purposes, we have treated this problem as a binary classification problem. We have transformed our target attribute 'case_status' into a binary label. We have merged all the observations for 'Certified-Expired' and 'Certified' into a

single ‘case_status’ representing ‘Certified’. We have removed all the data points for ‘Withdrawn’ status as these are the applications which have been withdrawn by the employers. Our final data frame contains a target feature ‘case_status’ with binary output labels - ‘Certified’ and ‘Denied’. We have also replaced the output labels with 0 for ‘Denied’ and 1 for ‘Certified’ for machine learning purposes. We have also merged ‘case_number’ and ‘case_no’ features into a single feature ‘casenumber’ as both of these features consisted of similar values. We have created a new feature ‘year’ from the ‘decision_date’ attribute and dropped the latter.

For feature selection, as our original dataset consisted of 154 columns, we have dropped all the columns consisting more than 20% missing values as imputing such a large portion of our dataset would be impractical. We have imputed all the missing values in our remaining attributes with their respective mode values as most of the missing features are categorical. We have also performed some feature engineering steps like under sampling and label encoding on our dataset. Our initial dataset was highly imbalanced with respect to the target attribute. Below, we have shown the distribution of the target attribute for the original dataset:

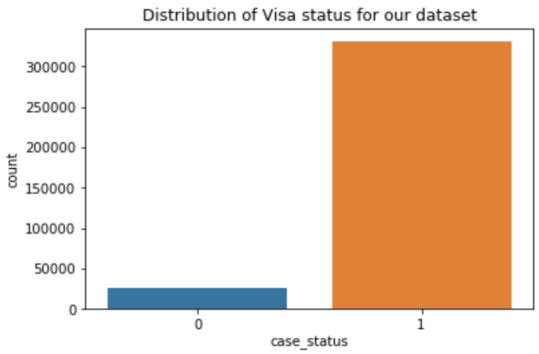

Figure2. Distribution of target feature before performing under sampling

To tackle the imbalanced dataset problem, we have leveraged the under-sampling technique. We have brought down the count of overrepresented classes to match the count of underrepresented classes using a random sample which has been extracted from the original dataset. Below is a plot of distribution of target attribute after implementing the under-sampling technique.

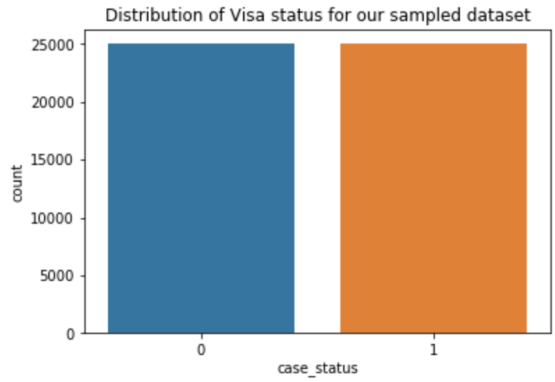

Figure3. Distribution of target feature after performing under sampling

Our final under sampled dataset contains 50000 observations and 15 features. We have label encoded all the categorical features in order to feed them to the machine learning models. For the final data preprocessing step, we have normalized and scaled all the input

features using the standard scalar technique so that each feature contributes to the target attribute.

Below, we have shown the exploratory data analysis and visualizations that we performed on our original dataset to gain more insights on some of the features. The first plot represents the total visa applications for each year. We can see from the below plot that the total visa applications that are certified have been increasing in the 5-year span. The denied visa applications count remains almost constant from 2011 to 2016.

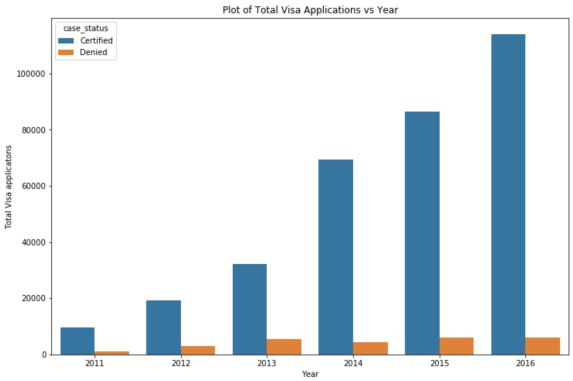

Figure4. Plot of total visa applications for each year

Next plot represents the class of admission for the visa applications. The plot shows that H1-B, L-1 and F-1 are the three major classes of admission for which most of the visa applications are filed.

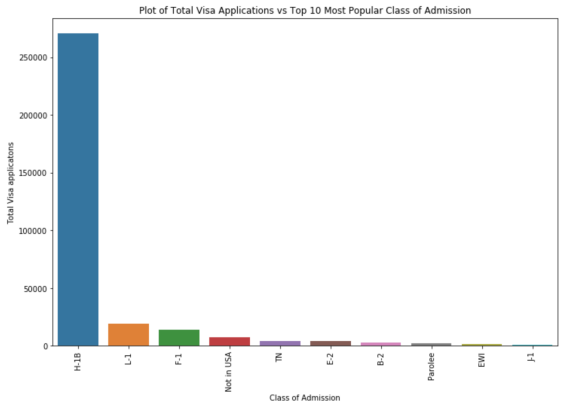

Figure5. Plot of total visa applications against the class of admission

Third plot represents that the most common mode of visa applications is the online mode.

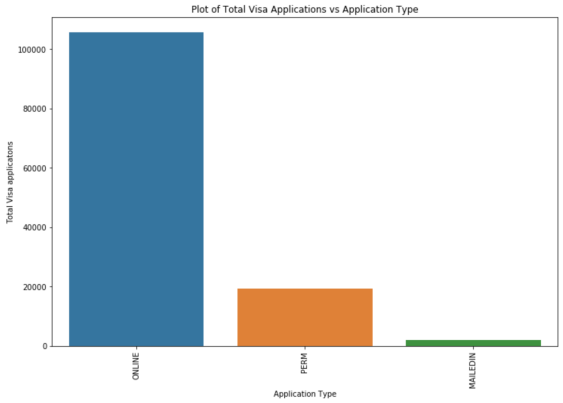

Figure6. Plot of total visa applications against the type of application

Next, we have a plot representing the education level of employees. Most of the visa applications are filed for the employees who have obtained a master's or bachelor's degree.

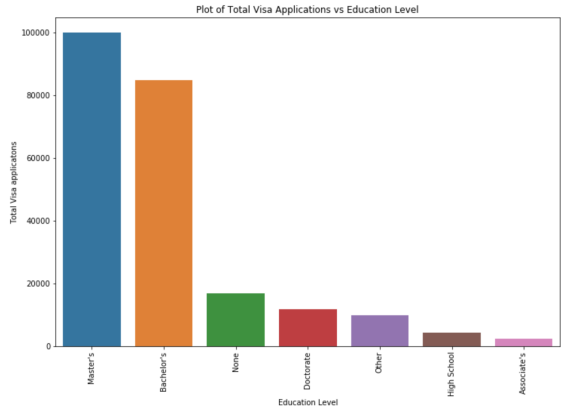

Figure7. Plot of total visa applications for different education level

Final plot represents the most popular countries that file for most visa applications. We can infer from the below plot that most citizens that file for visa applications are from India, China and South Korea.

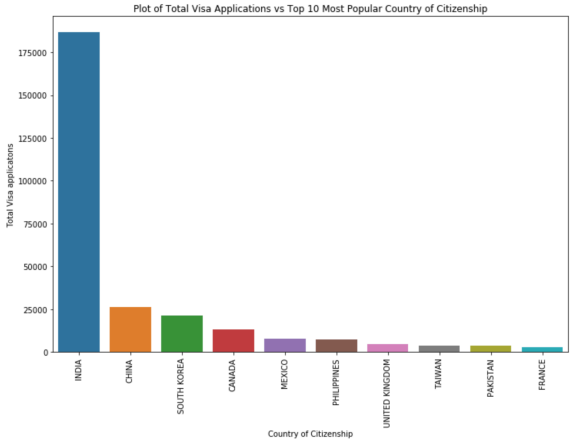

Figure8. Total visa applications filed against the country of citizenship

3.3 Machine Learning Models

We have made use of five different ML algorithms that work well with binary classification problems in general.

3.3.1 Logistic Regression

Logistic regression is a popular model whose objective is quite similar to linear regression in the way that it models the mean of the binary output after reading the set of independent features. Its name is derived from the logit function that is used by the model to make decisions on whether the output will be 1 or 0. This logistic function is also sometimes known as the sigmoid function because the sigmoid function, an s-shaped curve is used in splitting the binary outputs.

3.3.2 Random Forest

Random forest is an ensemble learning method which aims to improve its accuracy by combining the results of multiple models. The predictions of multiple classifiers may be aggregated to arrive at one final prediction. Ensemble learning also entails a set of base classifiers and the final classification is done by the weighted counts of outputs of each of these base models. Random Forest works best with Decision Tree classifiers wherein each base model is a Decision Tree built on an independent set of random vectors. There is a generalization error associated with ensemble methods with correlation between base decision trees, but it is reduced due to the randomness of these vectors.

3.3.3 Gradient Boosting Machine

Gradient boosting is another kind of ensemble learning method and is a sequential algorithm that keeps on adding predictors with each subsequent one being better than its previous one. The better model tries to fit the new predictors to the residual errors made by its predecessor. Gradient boosting models are becoming popular because of their effectiveness at classifying complex datasets and have recently been used to win many Kaggle data science competitions.

3.3.4 Linear SVM

Linear SVM is an extremely fast ML algorithm for solving multiclass classification problems from ultra large data sets that implements an original proprietary version of a cutting plane algorithm for designing a linear support vector machine. Linear SVM is a linearly scalable routine meaning that it creates an SVM model in a CPU time which scales linearly with the size of the training data set.

3.3.5 ANN

Artificial neural networks (ANNs), usually simply called neural networks (NNs), are computing systems vaguely inspired by the biological neural networks that constitute animal brains. An ANN is based on a collection of connected units or nodes called artificial neurons, which loosely model the neurons in a biological brain. Each connection, like the synapses in a biological brain, can transmit a signal to other neurons. An artificial neuron that receives a signal then processes it and can signal neurons connected to it. The "signal" at a connection is a real number, and the output of each neuron is computed by some non-linear function of the sum of its inputs. The connections are called edges. Neurons and edges typically have a weight that adjusts as learning proceeds. The weight increases or decreases the strength of the signal at a connection. Neurons may have a threshold such that a signal is sent only if the aggregate signal crosses that threshold. Typically, neurons are aggregated into layers. Different layers may perform different transformations on their inputs. Signals travel from the first layer (the input layer), to the last layer (the output layer), possibly after traversing the layers multiple times.

3.4 Model Evaluation

The F-score, also called the F1-score, is a measure of a model's accuracy on a dataset. It is used to evaluate binary classification systems, which classify examples into 'positive' or 'negative'. The F-score is a way of combining the precision and recall of the model, and it is defined as the harmonic mean of the model's precision and recall. The F-score is commonly used for evaluating information retrieval systems such as search engines, and also for many kinds of machine learning models, in particular in the NLP area. In our visa prediction problem, both false negatives and false positives may have dangerous consequences and hence we have decided to go ahead with f1-score as the most important metric to evaluate our models.

3.5 Deployment

We have used Dash and Heroku for deploying our web app. Dash is a Python framework for building web applications. It helps you to build dashboards using purely Python code. Heroku is one of the easiest platforms for deploying and managing public Dash applications. For this deployment, we have connected to our Github repository which was a straightforward process and we have done a manual deployment of the main branch of this repository. This repository contains the 'app.py' file with python code, the csv files with preprocessed data that we use for running the ML models, requirements.txt with the library dependencies used in 'app.py' and 'procfile' which has information about the python file

name and server that we have used for deployment. Our web app contains four tabs. The first talks briefly about our motivation for this project and the approach we went for. The other three are implementations of three of the ML models we have used for binary classification i.e., Gradient Boosting Classifier, Logistic Regression and Linear SVM. All the three tabs contain a table of the model performance metrics and a ROC curve with the AUC score. There is the ability for the user to experiment with hyperparameters for each model to see the best model according to the evaluation metric of their choice.

4 Results

We have implemented five machine learning algorithms using 3-fold cross validations for our research work to predict the outcomes of visa applications. We have executed the artificial neural network using 1 hidden layer with 20 nodes and epochs=10 and batch_size=20. We have evaluated our models using the evaluation metric as F1-score. Below, we have summarized the results obtained by all the machine learning models along with their accuracy, precision, recall, f1 score and area under the curve scores.

Model	**Accuracy**	**Precision**	**Recall**	**F1**	**AUC**
LR	66	70	64	67	71.3
RF	80	84	78	**81**	88.8
GBM	80	85	77	81	87.8
LSV	66	70	64	67	-
M					
ANN	68	68	70	69	75.2

Table1. Summary of models implemented along with their evaluation scores

We have plotted the feature importance plot using the best model i.e., Random Forest. The below plot shows that the key factors that are helpful in predicting whether a visa application will be approved or denied are casenumber, occupation code and income associated with the job, year of application, employer postal code, name, city and state.

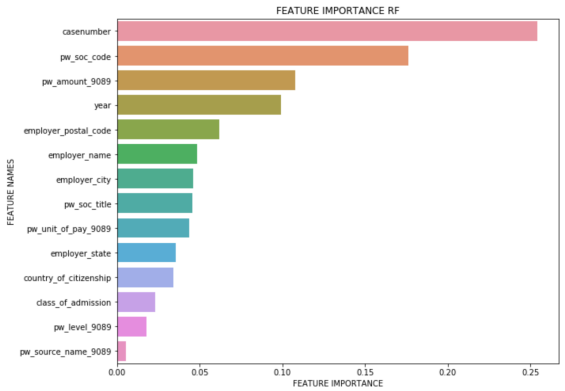

Figure9. Feature importance plot for Random Forest

We have also plotted the Area under the receiver operating characteristic curve for the Random Forest model. The below AUC-ROC curve represents that the Random Forest model is able to distinguish between the certified and denied visa applications 88.8% of the times.

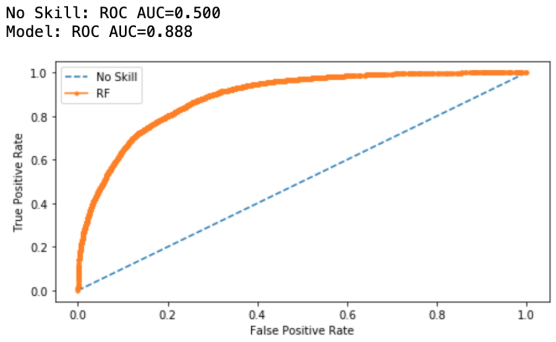

Figure10. AUC-ROC curve for Random Forest model

5 Discussion

Random Forest (RF) and Gradient Boosting Classifier (GBC) have significantly outperformed the other three models with respect to their F1 scores. As output classes are balanced, accuracy also is quite an important metric and even here, RF and GBC are way more functionally better than the others. We can attribute RF's impressive performance to higher accuracy through cross validation. Random forest classifier also handles the missing values and maintains the accuracy of a large proportion of data. Moreover, RF has the power to handle a large data set with higher dimensionality.

GBC performs well because it uses gradient descent to iterate over the prediction for each data point, towards a minimal loss function. In each iteration, the desired change to a prediction is determined by the gradient of the loss function with respect to the prediction of the previous iteration. Logistic Regression and Linear SVM uses a linear decision boundary that is not helpful for our problem, which has a non-linear decision boundary. The outcome prediction always depends on the sum of the inputs and parameters. In other words, the output cannot depend on the product (or quotient, etc.) of its parameters which greatly reduces its performance ability.

6 Conclusion & Future Scope

To conclude, we can say that through this research work, we have developed a system that can effectively predict binary outcomes of permanent visa applications. Random Forest and Gradient Boosting Classifiers work really well with the visa dataset and should be the focus of further improvements as more data is gathered. These models, with due modifications and refinement will be immensely helpful for employers that actively file for applications every year for their international employers. The features which can be interpreted to have maximum impact in predicting whether a visa will be rejected or accepted are pw_soc_code (occupational code associated with the job), pw_amount_9089 (income associated with the job), employer_postal_code (major cities may have higher acceptance rate)

In the future, we plan to involve a hybrid approach using oversampling technique and cost-sensitive learning while tackling the imbalanced dataset problem. An oversampling module with an optimal balancing ratio will give the best overall performance for the validation set. Then, using a cost-sensitive learning model, namely, CBoost algorithm, we can improve performance of models for visa prediction as compared to the existing approach.

References

[1]Kaggle dataset Link: https://www.kaggle.com/jboysen/us-perm-visas

[2]https://towardsdatascience.com/imbalanced-data-in- classification-general-solution-case-study-169f2e18b017

[3]https://www.neuraldesigner.com/blog/methods-binary-classification

[4]https://towardsdatascience.com/support-vector-classifiers-and-logistic-regression-similarity-97ff06aa6ec3

[5]https://towardsdatascience.com/an-implementation-and-explanation-of-the-random-forest-in-python-77bf308a9b76

[6]https://machinelearningmastery.com/binary-classification-tutorial-with-the-keras-deep-learning-library/

[7]https://machinelearningmastery.com/gradient-boosting-machine-ensemble-in-python/

[8]https://towardsdatascience.com/data-science-in-production-quickly-build-interactive-uis-for-your-data-science-projects-with-dash-6568e7df5528