USA Permanent Visa Application

Mengjie Li, Wilson Tran, Marjan Emadi, Kyle Nolan

Problem:

U.S is one of the best countries in science, technology and art around the world, that's why many people dream about coming to the U.S and make their dream come true.
The problem is, getting a permanent visa for staying in the U.S is very difficult.
In this project, we are observing different probabilities for different features and trying to analyze which features can help people more for acceptance in getting permanent visa.

Question:
What are the qualities that lead to a higher chance of being accepted?

Dataset:
https://www.kaggle.com/jboysen/us-perm-visas/data
Data covers 2012-2017 and includes information on employer, position, wage offered, job posting history, employee education and past visa history, associated lawyers, and final decision.
This dataset contains 600k+ rows, 100+ items and reasonably sparse. To process the data, certain cleaning methods and dimensionality reductions are necessary.

Hypothesis:

1. Class of admission code, current country of citizenship, and current employer have a great effect on being accepted for a permanent Visa.
2. Amount of wage contributes to the possibility whether a permanent Visa would be approved

Proposed Methods:
We will look at a US Permanent Visa Application dataset to examine the type of questions asked during the application process. From the results of each applicant, we can learn if certain types of answers will lead to a greater chance of being accepted. We can extract and visualize information to help us better understand how each question is related to the overall process and maybe even find out how much each question is weighted in importance.

We can clean the data by getting rid of attributes that are irrelevant to the question we want to answer and getting rid of junk entries. To analyze the data, we will use a pairwise correlation to see how each attribute affect each other. After we get a sense of which factors are important, we will see if there are any outliers affecting our data and remove if necessary. Additionally, we can use Principal Component Analysis to help reduce the dimensionality of our data further in order to determine the most important attributes and figure out the potential latent factors if there are any. By fitting a linear model using Multiple Linear Regression, it is possible to see if the results still apply today.

We plan to create a word cloud of the applicant's country of citizenship, class of admission code, employer city, and employer name to have a better understanding of the ideal applicant and employers that issued the permanent Visa. In addition, we will report pairwise correlations of the different attributes to see how each feature relate to one another. This can be further visualized using histograms, bar charts, and plots.

For the prediction model, we plan to fit different models such as Linear Regression, Logistic Regression and Support Vector Machine.

In order to test whether our models make sense, we would split the dataset into training set, validation set and test set to further train, tune any possible parameters we would face with.

Tools

To approach this problem, we would be using Python as our main programming language, well-formed machine learning package such as sklearn and TensorFlow/Keras if we find Neural Network is needed.

Tasks

Task	Person in Charge
Data Cleaning	Mengjie Li
Data Visualization	Wilson Tran
Data Analysis	Marjan Emadi + Kyle Nolan
Model + Prediction	Kyle Nolan + Mengjie Li

Timeline

Task	Estimated Duration
Data Cleaning	1 week
Data Visualization + Analysis	2 weeks
Model + Prediction	2-3 weeks