Predicting H1B LCA Petition Success

My problem will focus on trying to predict whether or not a Labor Condition Application (LCA), which is a pre-requisite for H1B visa applications, will be accepted or denied. There are two prospective clients for this work. One would be a foreign citizen who would like to obtain employment in the United States. This person might want to know which types of jobs, what employment levels, and in what parts of the country they should be looking for employment to best maximize the chances that their H1B visa is ultimately approved. The other prospective client is an American company who may be looking to hire a specific foreign citizen. They may want to ensure that the combination of position, salary, employment status, etc. that they are offering their prospective employee is likely to succeed the application process.

The data that I will use are LCA petitions filed with the US Department of Labor between 2011 and 2016. These data will be obtained from the Kaggle H-1B Visa Petitions 2011-2016 dataset (<https://www.kaggle.com/nsharan/h-1b-visa>). This represents roughly 3 million LCA petitions. It should be noted that the data made available through Kaggle do not include all fields that are required on the LCA petition. However, they do include the final case status, employer name, standard occupational classification (SOC) system name, job title, whether the position will be full time or part time, the prevailing wage for that position, the city and state of the position, as well as the latitude and longitude of the employer.

My initial thoughts for solving this problem start with first exploring the data to determine if there are any readily-apparent trends. I would begin by visually exploring the relationships between year, position, location, and case status. If any interesting trends present themselves, I would follow up with examinations of relationships with the other variables. This will help determine which, if any, variables are corelated. The next step would be to build a model that best utilizes the available data in order to predict the outcome of the LCA application.

The deliverables for this project will be, primarily, my code. This will contain all the necessary tools to help predict application success. I will also prepare a presentation with the findings from my initial data exploration, the assumptions that went into my model, and finally example results and output from my model.