Capstone Project 1- H1B ML

Sheila Torrico - October 2018 Cohort

Phase 1

The first phase on my project consisted of several steps such as collecting the data, reviewing it and deciding how to transform it for the purpose of building a model that will be able to predict if a visa application will be Certified or not.

Dataset

From the Department of Labor, I collected the H1B-visa-applications file from the past fiscal year (Oct 1, 2017- Sept 30, 2018). The file was available in an .xlsx format.

Data Review

Before uploading the file to a Jupyter notebook, I did a preliminary analysis in Excel by using a sample of 2000 observations. The goals of this review were to understand the type of information that each of the 52 variables had, the quality of the observations, and if there were any mandatory vs optional variables.

The results gave me an insight around the data types I should expect to see once the file is available in a Jupyter notebook and a glimpse of the variables that might have missing observations or outliers. For the next step, I uploaded the sample file in Jupyter notebook and used available functions from pandas to confirm the data types and that all variables and corresponding observations were successfully uploaded.

Given the positive outcome, I uploaded the complete file and sure enough all my variables and corresponding observations were in the notebook. The variables were divided into the following data types: integer, float, datetime and object.

Data Wrangling

To continue with the transformation process of the data, I divided my variables into non-object and object data types so I could apply the appropriate techniques.

For the non-object variables, I identified the variables that had any missing observations and calculated the percentage of missing observations. Also, I used statistical functions to identify variables that might have outliers. For variables with outliers, I performed an additional analysis to find out potentials reasons/patterns. I.e., in the case of the wage variable, the large standard deviation was the result of having values with different pay rates such as bi-weekly, yearly.

I created a separated dataframe for variables with an object data type so I could assign them to the categorical data type. Before performing such step, I identified all variables that had any missing observations and calculated the corresponding percentage. Then, I extracted a subset from these variables to perform further analysis and identify potential patterns.

Results

From performing these steps, the dataset was enhanced by:

- Removing variables that had at least 90% of missing observations.
- Filling the NaN values of variables that had at least 95% of all observations.
- Not filling or removing variables that have 30% to 60% of missing observations. Instead, will look for a ML algorithm that could handle large number of NaN values.
- Ensuring numerical variables have the same units.
- Transforming categorical variables to new numerical variables.