Project Proposal

Team Name: STATS Whispers

Team Members:

Masuti, Shraddha
Patel, Ketul
Sawant, Rashmi
Siriah, Deepika

Data Description and Problem Statement

H-1B Dataset comprehends data from employer's Labor Condition Application (LCA) and the case certification determinations processed by the Office of Foreign Labor Certification (OFLC). The LCA is a document that contains perspective H-1B employer files with U.S. Department of Labor Employment and Training Administration (DOLETA) when it seek to employ non-immigrant workers at a specific job occupation in an area of intended employment for not more than three years. The data is of 1 year decision dates i.e from October 2016 to September 2017. The goal of this project is:

1. To predict the case status of an application submitted by the companies to hire non-immigrant employees under the H-1B visa program

2. To find the features which affects the most to the decision about the visa status and to find the relationship between independent variables and dependent variable.

Data Set Details

Rows and Columns

The H1B dataset comprises of **52** predictors and **541,102** observations.

Prediction Column, Classification or Regression Problem

The main goal is to classify the status of the H1B applicant based on the predictive classification models. The application status i.e "CASE_STATUS" is a response variable and has two level classes which are "Certified' and 'Denied'.

Datatypes of the predictors

The dataset includes categorical, numerical, text, dates, and Geo-spacial which provides enough detail to assess if the dataset will work well with the analysis we are learning in class.
A few of predictors datatype are shown below:

Predictors	**Datatype**
Wage Level, PW Sources, Units of Pay	Categorial
Employer Name, Address, Job Title	Text
Wage, Total Workers	Numeric
Full Time Position, H1B Dependent	Binary
Application date, Employment Date	Date
City, Country, ZIP Code	Geo-spacial