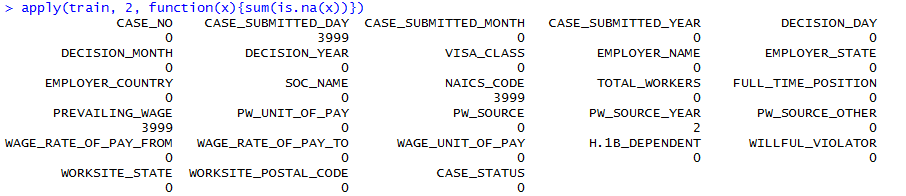
**Background Knowledge/Assumptions:**

* The normal response time for an application for temporary employment visa (3 years) is 1 week.
* The employee should be paid more than or equal to the prevailing wages in the country.
* The size of company determines the no of applications that can filed/approved by an Employer. Thus, the Employer is classified as H-1B\_DEPENDENT and WILLFUL\_VIOLATOR.
* All the applications must be filed online. However, there’s a procedure to file offline which increases the response time by some duration.
* Common reasons for rejection- salary standards, job profile mismatch, etc
* New requests are generally taken up in the month of April
* Change in H1B process after President Trump’s governance

**Data Anomalies:**

* Data description had a feature “*INSPECTION\_DATE*” which is not present in the training dataset.
* Difference between Submission Date and Decision Date is too large initially (in years). It drops down to months and days in 2017
* The date, month and year are separated in different columns. Combined these three to form a timestamp.
* Number of NA (missing values) in the dataset. Majority in Submission Date, NAICS code and prevailing wages



**Data Pre-processing:**

* Separate columns for day, month and year are combined to obtain a single Date object column.
* Wages are mentioned with unit as Year, Month, Bi-Weekly, Week and Hour.



These wages are all transformed to per year basis. The transformation is done using the formula below.

*Year = Wage\*1*

*Month = Wage\*12*

*Bi-Weekly = Wage\*(52/2)*

*Week = Wage\*52*

*Hour = Wage\*9\*(365-104)*

* Since data is sorted according to Case submission date. There are two ways to fill missing values in Case Submission Date

1. The closest data to the missing date can be said as the Submission date. Eg: Date value one row above or below the missing column can be used.
2. Take the difference between Decision date and submission date. This difference on quantile range has value 6 till 3rd quantile (after October’2017).



Hence, if there are many consecutive missing values after Octeber’2017, 2nd approach can be taken to interpolate. If single values are missing, 1st approach can be taken. In case, many consecutive values are missing before October’2017, those records can be ignored.

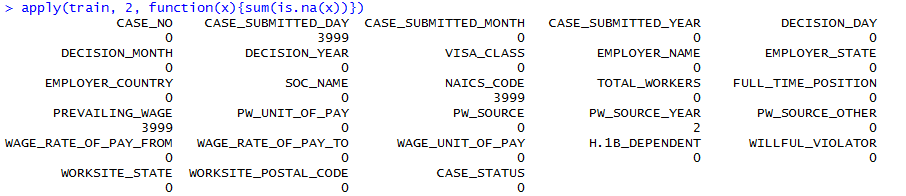
* Other missing data can be simply removed. 7998 rows in 4 lakh + rows won’t make any difference.
* Derive a new column as “DECISION\_TIME”- difference between Decision date and submission date.

**Some pre-analysis:**

Distribution of target variable



Number of NA values:



Difference in type of wage:



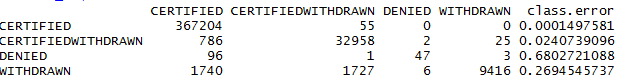
Aggregated wage to year basis for all. Thus, removed unit of wage column

**Model Choice Explanation:**

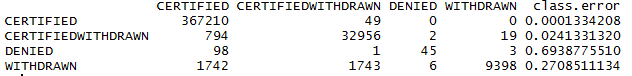
The model choice depends on various parameters:

* Type of data and outcome to predict
* The given data is a case of supervised classification problem. Hence, **NO** regression technique can be used.
* The outcome variable has 4 classes, hence algorithms with binary outcomes like naïve bayes can be eliminated.
* The dataset contains both numeric and categorical values as features. Therefore, for the given dataset, the best results can be obtained using – Decision Trees, Random Forests, ANNs.
* Moreover, the distribution of classes in target variable is not uniform. 2 out of 4 classes are very dominant whereas the other 2 have very few occurrences.
* **Decision Trees:** Due to large size of datasets and non-uniform distribution of target variable, the decision trees are prone to overfitting. Even after pruning the tress, the model was biased to predict the majority class.
* **Random Forest:** Random forest is a good choice for this kind of dataset. It produced better results when compared to decision trees.

With Factors:



W/O factors:



Knn

