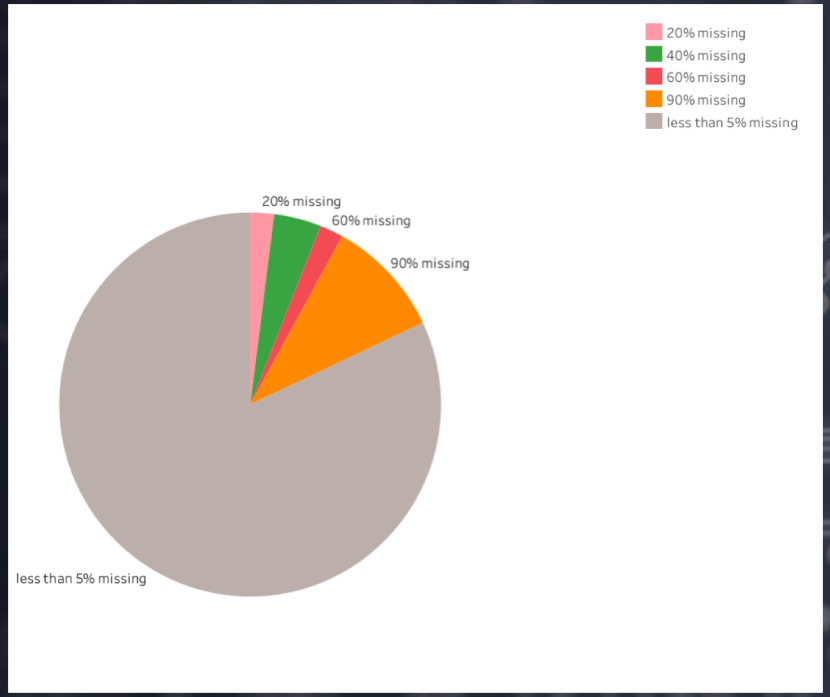
Denied or Certified: Predicting H1B VISA

TEAM Extreme Value Theorem

Citina Liang, Yan Kang, Zheng Wang

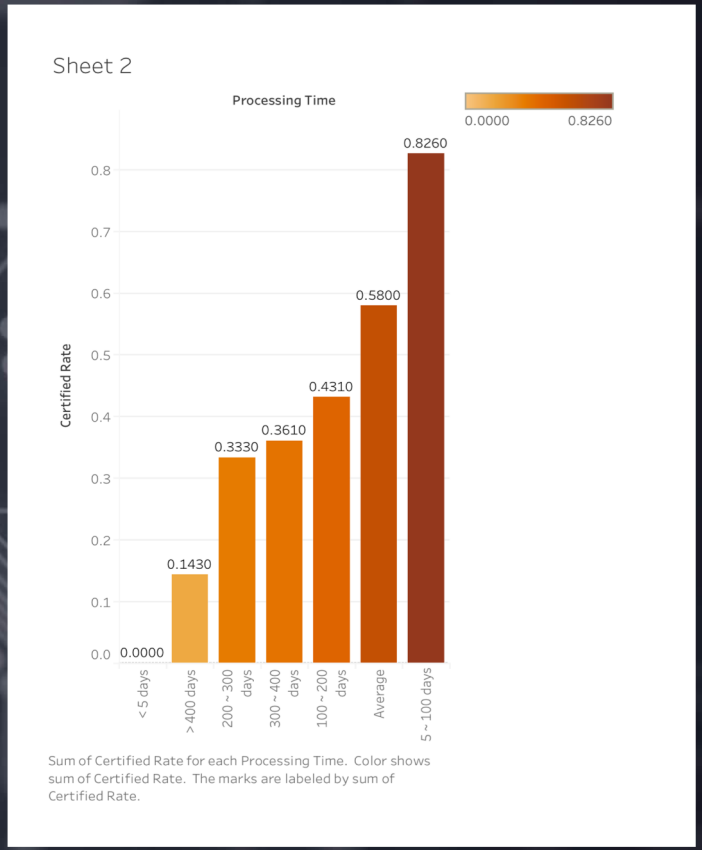
Data Cleaning



- Data Description
  - The h1b data set contains 4918 observations and 52 variables.
  - There are 4 date-time variables, 3 numerical variables and 45 categorical variable. However, some of the categorical variables contains over 50 levels.
  - Nearly every observation in this data set contains NA in at least one of the variables.
  - A few variables contains over 90% of NA values.

Data Cleaning

- Data Transformation
  - We unify the unit of Wages to year so that they can be compared mutually.
  - We used fill the missing values to be “N/A” and take it as a new level of factor to indicate that information are not provided.
  - For factor variables with too many levels, we combine some of the levels to simplify their structures.
  - For Date-time variables, we take the differences of the starting and ending date or use it to identify which weekday it is.
- Exam how the variables after cleaning contribute to the acceptance ratio.
- Delete highly correlated variables or uneffective variables.
- After cleaning data, we obtained a dataset with 4918 observations and 29 variables



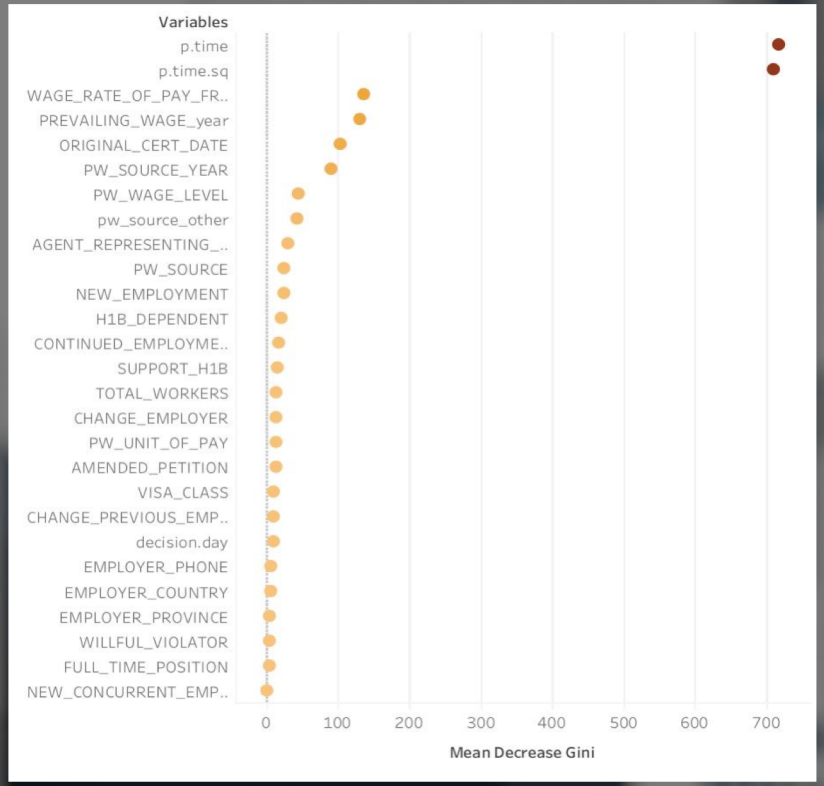
An Example of an effective predictor

Method Choosing

- What methods did we consider?
  - Logistic Regression
  - Random Forest
  - Generalized Boosted Regression

- Why Random Forest?
  - Compared to the other methods, it gives highest accuracy
  - It gives estimates of what variables are important
  - Resistant to outliers
  - Most of our predictors are categorical variables.

Model Adjustment



- What arguments can we change in the function?
  - number of variables to use
  - number of trees to grow
- How to do it?
  - Doing Cross Validation to reduce number of unimportant variables
  - Doing “for” loop to find the best size for tree
  - Knowing what variables contribute the most
- Does the change make an improvement in the accuracy?
  - We conclude the default values are the best

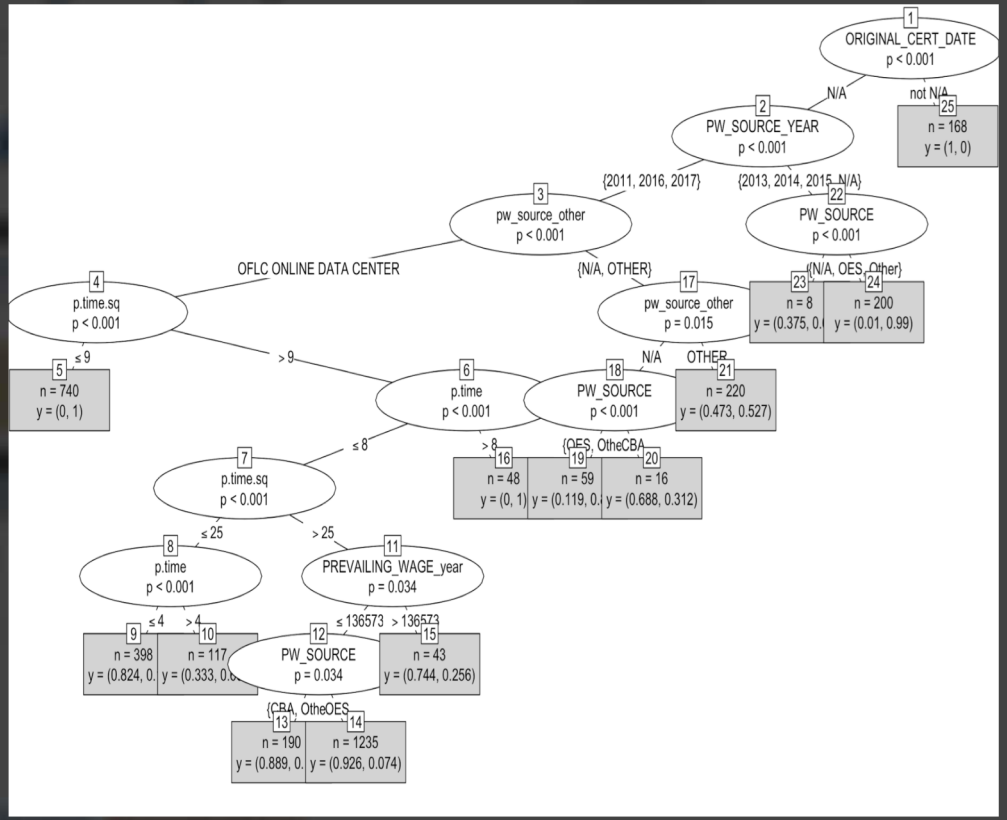
Assumptions and Disadvantages

- Assumptions
  - The only assumption that it relies is that sampling is representative. This is a common assumption. In our case, we have nearly 50% CERTIFIED cases and 50% DENIED cases where assumption holds.

- Disadvantages
  - For our dataset, it includes categorical variables with different number of levels. RF are biased in favor of those attributes with more levels.
  - For minimizing the bias, we have reduced number of levels for categorical variables while we were cleaning data.

Result and Summary

- After performing different models, random forest give the best performance.
- Our model produce an approximately 91.5% accuracy.
- Final model consists 29 predictors which includes squared processing time as one polynomial predictor.



Result and Summary

- The graph plots number of trees against error rate for each level of CASE_STATUS.
- **Red curve** - CERTIFIED cases
  Green curve - DENIED cases.
  Black curve - out-of-bag errors.
- The error rate of DENIED cases is always higher than the rate of CERTIFIED cases.