STATS 101C Final Project

Team **Stats Trek**
Jiahao Huo | Jia Qi Wang | Xiao Li

Our Best Model

- **Algorithm:** e**X**treme **G**radient **B**oosting (package: xgboost)
- **Target Variables:** elapsed_time (numeric)
- **Independent Variables:** 10 variables
  - 6 variables from the original dataset
  - 4 newly created variables
- **Parameters**
  - Eta: 0.3
  - Max_depth: 10
  - Nround : 100

Feature Engineering

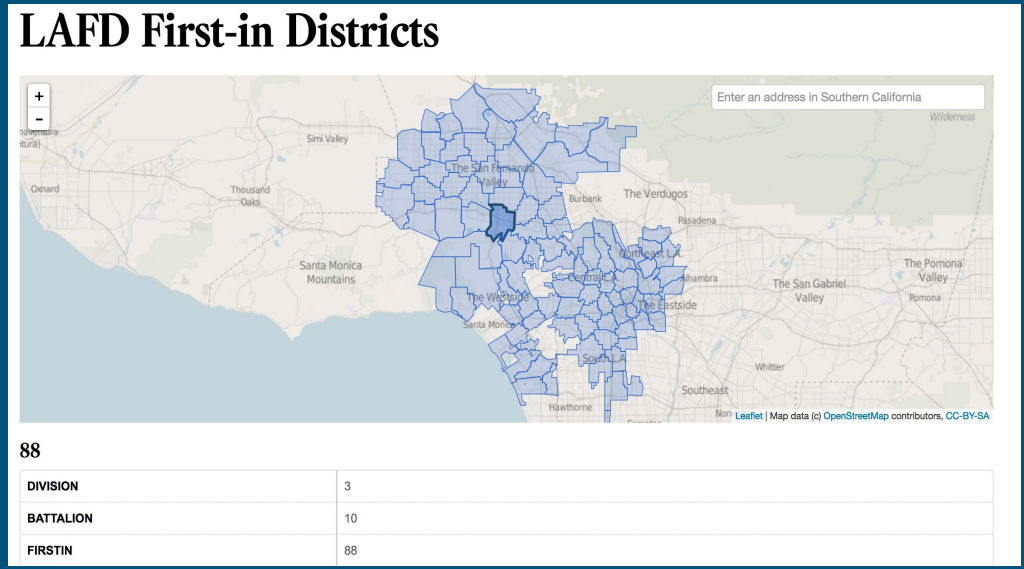
- 6 Variables from Original Dataset
  - year — 4 levels
  - First.in.District — 102 levels
  - Dispatch.Status — 12 levels
  - Dispatch.Sequence — int
  - Unit.Type — 41 levels
  - PPE.Level — 2 levels
- 4 Newly Created Variables
  - Fd (fire department) — 2 levels
  - Incident — num
  - Creation — 4 levels
  - Cnt (count) — int

Without Utilizing External Data

- **Incident.Creation.Time**
  - Convert Incident.Creation.Time into 4 levels
    - (00:00~06:00, 06:00~12:00, 12:00~18:00, 18:00~00:00)
- Separate incident ID into 2 new variables by using substring
  - **fd** (fire department; factor with 2 levels)
  - **incident** (incident number: numeric)
- Create a new variable, **cnt** (int), from incident.ID
  - the number of vehicles dispatched in the same incident

With Utilizing External Data

- ➔ Import Data from LA Times
  - ◆ Division
    - 3 levels
  - ◆ Battalion
    - 17 levels



Dealing with NAs

- ★ elapsed_time, PPE.Level and Dispatch.Sequence
  - Deleted all NAs in elapsed_time since it is target variable
  - For other NAs in PPE.Level and Dispatch.Sequence, our model has a built-in feature to deal with them.

Variables Selection

- After applying different techniques, we dropped insignificant variables
  - row.id
  - incident.id
  - Emergency.Dispatch.Code
  - Incident.Creation.Time
  - DIVISION
  - BATTALION

Parameter Tuning

- Range of values we have tried
  - Eta: 0.01-1.0 (Final choice: 0.3)
  - Max_depth: 1 - 10 (Final choice: 3)
  - Nround: 1 - 1000 (Final choice: 100)

WORST MODEL :)

1. Import dataset with negative values of elapsed_time
2. Do the same process of analyzing
3. Add 86400 to all negative prediction values

FINAL MSE: 1816388753.76668

Thank You!