Stats 101C Final

[cyc00]
Victor Shih

Our Best Model

- Algorithm: Generalized Boosted Regression Model (package: gbm)
- Training data used: 2315060 obs with 9 predictors
  - 5 variables from the original dataset
  - 4 newly created variables
- Parameters:
  - Maximum depth of variable interactions (interaction.depth) = 3
  - Total number of trees to fit (n.trees) = 300
  - Shrinkage = 0.15
  - Distribution = “Gaussian”

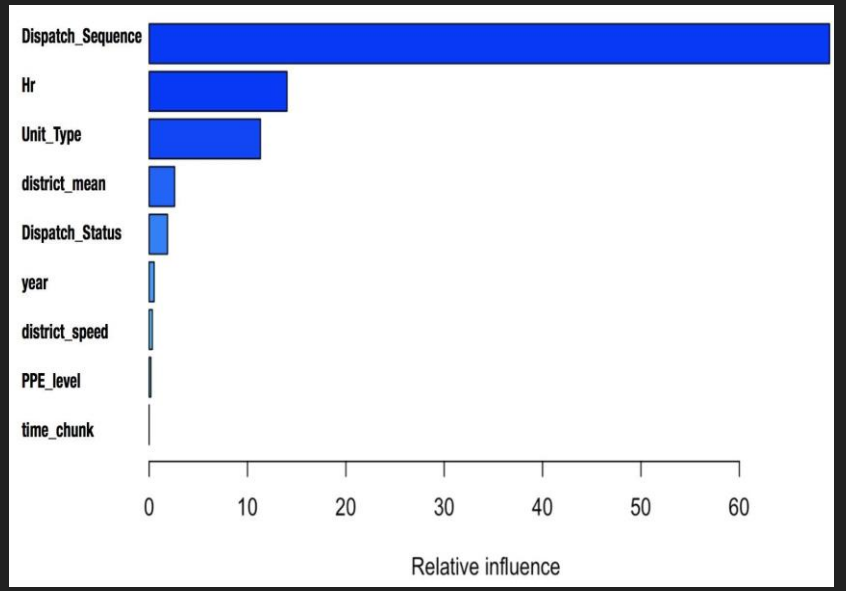
Our Best Model (Variables Used)

- **5 Predictors from the Original Data Set**
  - Dispatch_Status, Unit_Type, PPE_Level, year, Dispatch_Sequence
- **3 newly created variables**
  - Hour, time_chunk , district_speed, district_mean
- **9 Variables Used for Training and Testing Data Set**
  - Dispatch_Status_Time, Unit_Type, PPE_level, hour, year, time_chunk, district_speed, Dispatch_Sequence, district_mean

Summary of Model

```
bst3<-gbm(elapsed_time~.-row.id, interaction.depth = 3,
data=l[-idna,],distribution="gaussian",n.trees =300,verbose =F, shrinkage =0.15)
```

	var <fctr>	rel.inf <dbl>
Dispatch_Sequence	Dispatch_Sequence	69.1774313
hr	hr	14.0257283
Unit_Type	Unit_Type	11.3173671
district_mean	district_mean	2.5945573
Dispatch_Status	Dispatch_Status	1.8727827
year	year	0.5073719
district_speed	district_speed	0.3274878
PPE_Level	PPE_Level	0.1772736
time_chunk	time_chunk	0.0000000



Cleaning the Data

- Deleted Features:
  - **Emergency Dispatch Code** → Only contained one level (Emergency)
  - **Incident.ID** → Just random ID numbers. Did not provide useful for predicting.
  - **First_in_District** → Replaced with the
  - mean elapsed time for each station
- Added Features: **Hour** → extracted the hour from our Incident Creation Time (GMT)
  - **District_mean** → mean response time for each station
  - **Time_Chunk** → classified mean response time by the hour → rank which hour as faster or slower
  - **District_Speed** → rank median response time by fire stations from 1 to 9

Cleaning the Data

- Dealing with NAs:
  - There were missing values in both the training and testing dataset.
    - Most NAs in the training dataset were seen as response variables.
    - Best to delete these observations.
  - Only ~0.1% of the data which contained missing values for predictors after removing the NAs from the response variable in training dataset.
    - Removed these observations due to the low percentage.

Transforming the Data

- Training:
  - Changed Data Type to factor:
    - Dispatch_Status, hr, year, Unit_Type, time_chunk, and PPE_Level
  - Log transformation on predictor, “Dispatch_sequence”
- Testing:
  - Dealing with NA’s:
    - Linear model to predict dispatch_sequence from other predictors in the training dataset

- Gradient boosting model:
  - Has the highest predictive power, producing the minimum MSE.
  - Relatively computational inexpensive compared to other models (Using random forest, PCA, PLS, tree, Lasso, and ridge).
    - Making it possible to use ALL of the training data to build and tune our model.
  - Avoids overfitting by tuning the parameter shrinkage, creating a penalty for less useful predictors
  - Ability to use categorical predictors with large number of levels.
    - Avoid losing information by reducing the number of predictors.
- Other Methods Tried:
  - PCA, PLS, Random Forest, Tree, Lasso, and Ridge

Why Our Model May Not Have Worked

- Gradient boosting model:
  - Inability to find possible external variables which could have been helpful in decreasing our MSE.
  - Tuning the parameters manually, and not using package ‘caret’.
    - Limited number of combinations when tuning the parameters
    - The Range of Values We Tried for gbm Parameters:
      - Interaction_depth: 1-3 (Final Choice: 3)
      - N_trees: 80-400 (Final Choice: 300)
      - Shrinkage: 0.05-0.20 (Final Choice: 0.15)
      - n.minobsinnode: 10-20 (Final Choice: 10)