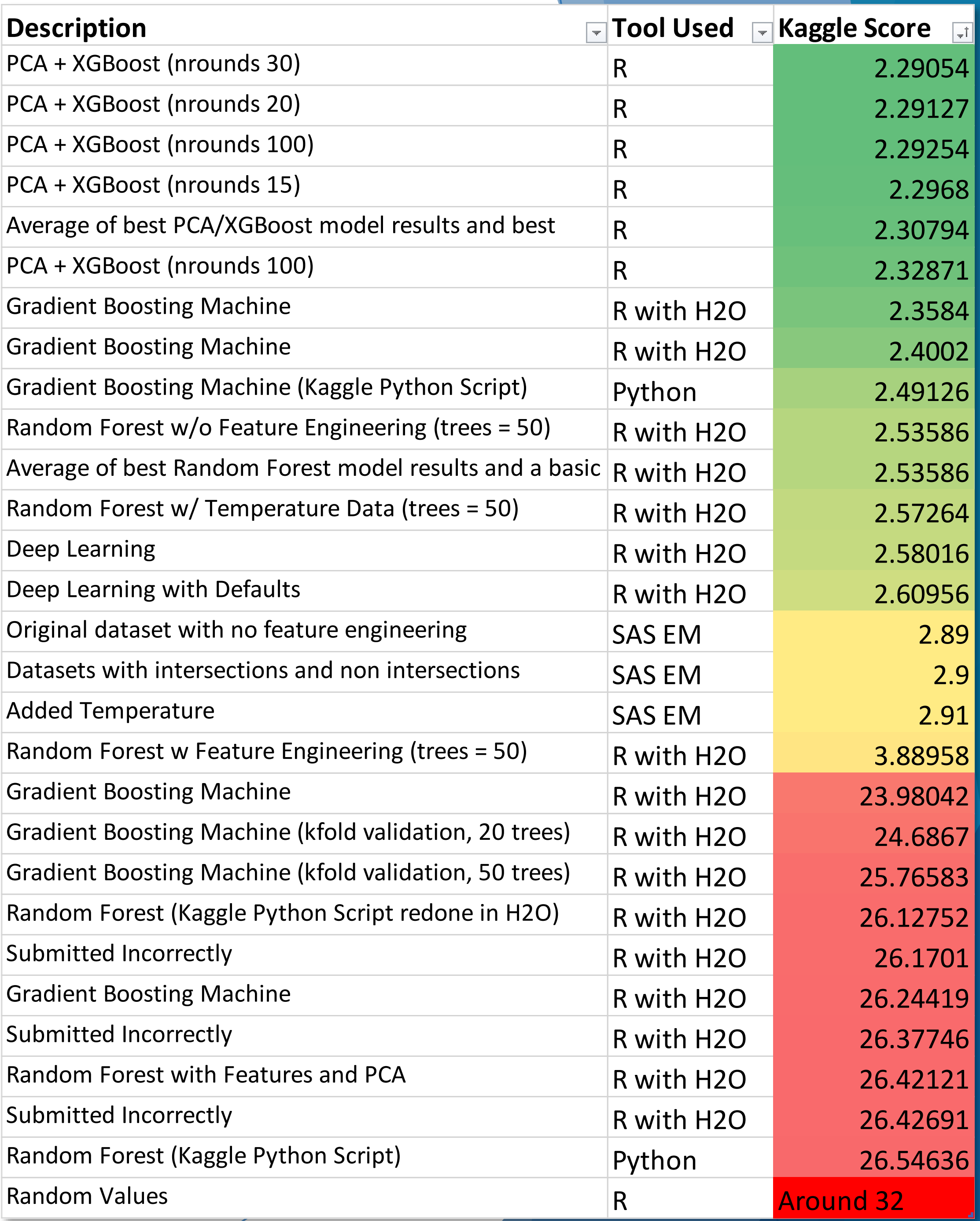
Kaggle:
San Francisco Crime Classification

Amit Talapatra, Wendy Zhang, Wei Zheng

Models and Methods Used

- Through SAS EM, Python Packages, R Packages, and H2O, we tried:
  - PCA (for latitude/longitude)
  - Feature engineering from existing data
  - Adding temperature data
  - Random Forest Models
  - Deep Learning Models
  - Gradient Boosting
  - Extreme Gradient Boosting



Description	Tool Used	Kaggle Score
PCA + XGBoost (nrounds 30)	R	2.29054
PCA + XGBoost (nrounds 20)	R	2.29127
PCA + XGBoost (nrounds 100)	R	2.29254
PCA + XGBoost (nrounds 15)	R	2.2968
Average of best PCA/XGBoost model results and best	R	2.30794
PCA + XGBoost (nrounds 100)	R	2.32871
Gradient Boosting Machine	R with H2O	2.3584
Gradient Boosting Machine	R with H2O	2.4002
Gradient Boosting Machine (Kaggle Python Script)	Python	2.49126
Random Forest w/o Feature Engineering (trees = 50)	R with H2O	2.53586
Average of best Random Forest model results and a basic	R with H2O	2.53586
Random Forest w/ Temperature Data (trees = 50)	R with H2O	2.57264
Deep Learning	R with H2O	2.58016
Deep Learning with Defaults	R with H2O	2.60956
Original dataset with no feature engineering	SAS EM	2.89
Datasets with intersections and non intersections	SAS EM	2.9
Added Temperature	SAS EM	2.91
Random Forest w Feature Engineering (trees = 50)	R with H2O	3.88958
Gradient Boosting Machine	R with H2O	23.98042
Gradient Boosting Machine (kfold validation, 20 trees)	R with H2O	24.6867
Gradient Boosting Machine (kfold validation, 50 trees)	R with H2O	25.76583
Random Forest (Kaggle Python Script redone in H2O)	R with H2O	26.12752
Submitted Incorrectly	R with H2O	26.1701
Gradient Boosting Machine	R with H2O	26.24419
Submitted Incorrectly	R with H2O	26.37746
Random Forest with Features and PCA	R with H2O	26.42121
Submitted Incorrectly	R with H2O	26.42691
Random Forest (Kaggle Python Script)	Python	26.54636
Random Values	R	Around 32

Methods and Key Findings

1. Feature Engineering and Dimension Reduction
2. Parameter Tuning and Cross Validation
3. Model Selection

1. Feature Engineering and Dimension Reduction

What Helped:

- Breaking date and time into year, month, day, and hour.
- Principal Components Analysis of latitude and longitude coordinates.

What Did Not Help:

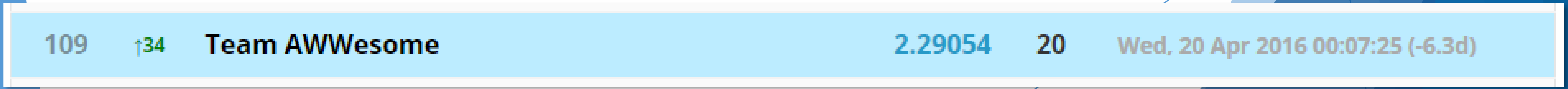
- Adding temperature data (Highs, Lows, Average).
- Identifying addresses as ‘intersection’ or ‘non-intersection’.

2. Parameter Tuning and Cross Validation

- 3 folds for Cross Validation
- Grid Search for Hyperparameter Tuning
  - GBM Key Parameters:
    1. Number of Trees
    2. Learning rates
    3. Maximum depth
  - Random Forest Key Parameters
    1. Number of Trees
    2. Maximum depth
    3. Number of Features
  - Extreme Gradient Boosting
    1. Round

3. Model Selection

- Best Approach: PCA and Extreme Gradient Boosting (Using R packages, based on a script found on Kaggle) **(Score: 2.2905)**
  - Adjusting nrounds (iterations) improved our score beyond other Kaggle submissions using the same method.
- Second Best Approach: GBM (Using R with H2O) **(Score: 2.3584)**
  - The best combination of parameters is (ntrees : 200, learn_rate:0.1, max_depth: 10)
- Third Best Approach: Random Forest (Using R with H2O) **(Score: 2.5726)**
  - Used ntrees = 50, could be improved by adjusting this parameter



109	↑34	Team AWWesome	2.29054	20	Wed, 20 Apr 2016 00:07:25 (-6.3d)

Questions?