CUNY Data Challenge 2019

Team: Brooklyn College

Eugene Dorokhin[1,2] Paul Magrini[1]

[1]Brooklyn College, CUNY
[2]Institute for Neural and Intelligent Systems

August 14, 2019

Problem Overview

Objective: To predict the probability that a restaurant will get an A from the DOH.

Data Given: inspection result, partial location data, cuisine type, **violations cited**, actions taken, inspection date, inspection type

Minimization criteria: $-(y \log(p) + (1-y)\log(1-p))$

Final Private Dataset Error: 0.06780

Do the violations tell us everything?

Do the violations tell us everything?

- First, what do the violations tell us?

Do the violations tell us everything?

- First, what do the violations tell us?

- No, accuracy of linear separability via perceptron ~95%, accuracy with other features: ~97.5%

Do the violations tell us everything?

- First, what do the violations tell us?

- No, accuracy of linear separability via perceptron ~95%, accuracy with other features: ~97.5%

- Accuracy does not directly translate into optimal probability predictions.

Do the violations tell us everything?

- First, what do the violations tell us?

- No, accuracy of linear separability via perceptron ~95%, accuracy with other features: ~97.5%

- Accuracy does not directly translate into optimal probability predictions.

- How can we use the information contained within the violations?

Assign Weights to Violations

- Idea: Can we guess what score an inspection yielded?

Assign Weights to Violations

- Idea: Can we guess what score an inspection yielded?

$$
\begin{array}{cc}
 & \texttt{Violation Type} \\
\texttt{Inspection \#} & \begin{bmatrix} 0 & 0 & 0 & 1 \cdots & 0 \\ 1 & 1 & 0 & 0 \cdots & 0 \\ 0 & 0 & 0 & 0 \cdots & 0 \\ 0 & 0 & 1 & 0 \cdots & 0 \\ 1 & 0 & 0 & 1 \cdots & 1 \\ 1 & 0 & 1 & 0 \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & 0 & \ldots & 1 \end{bmatrix}
\end{array}
\begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \\ v_6 \\ \vdots \\ v_{92} \end{bmatrix} = \begin{bmatrix} 15 \\ 2 \\ 11 \\ 9 \\ 23 \\ 3 \\ \vdots \\ 4 \end{bmatrix} \ \texttt{Score Guess}
$$

Assign Weights to Violations

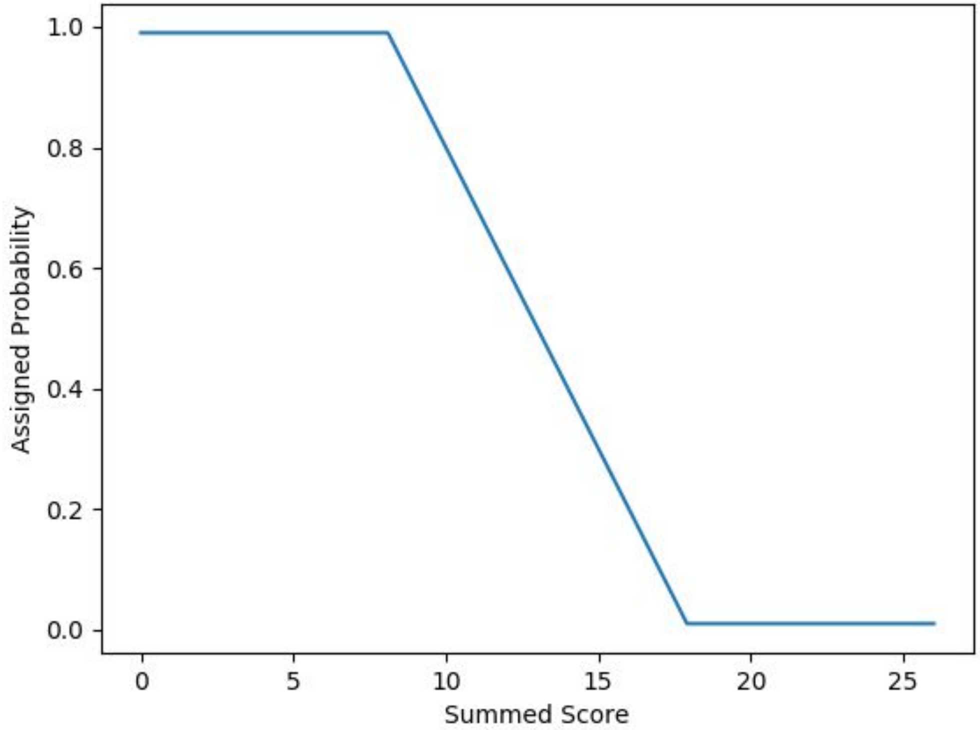
- Idea: Can we guess what score an inspection yielded?

$$
\begin{array}{cc}
 & \text{Violation Type} \\
\text{Inspection \#} & \begin{bmatrix} 0 & 0 & 0 & 1 & \cdots & 0 \\ 1 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 0 & 1 & \cdots & 1 \\ 1 & 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \ddots & \vdots \\ 0 & 1 & 0 & & \ldots & 1 \end{bmatrix}
\end{array}
\begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \\ v_6 \\ \vdots \\ v_{92} \end{bmatrix}
=
\begin{array}{cc}
\begin{bmatrix} 15 \\ 2 \\ 11 \\ 9 \\ 23 \\ 3 \\ \vdots \\ 4 \end{bmatrix} & \text{Score Guess}
\end{array}
$$

- But how should we pick the violation score vector?

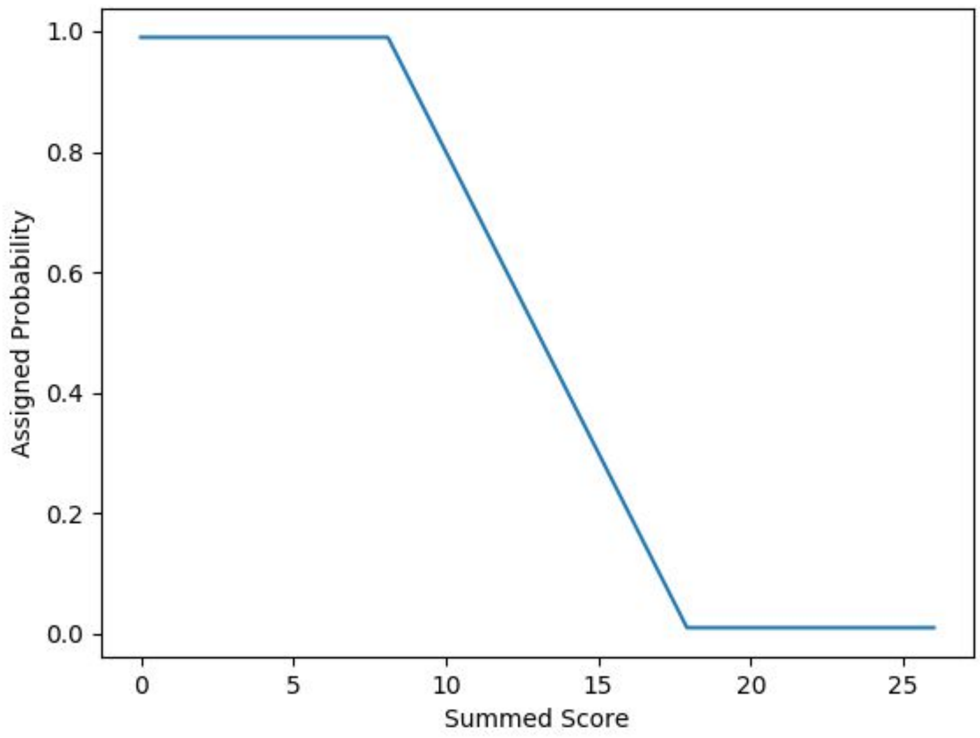
Picking scores with simple rules

- Guess every pass scored a 7, and every fail scored a 21!



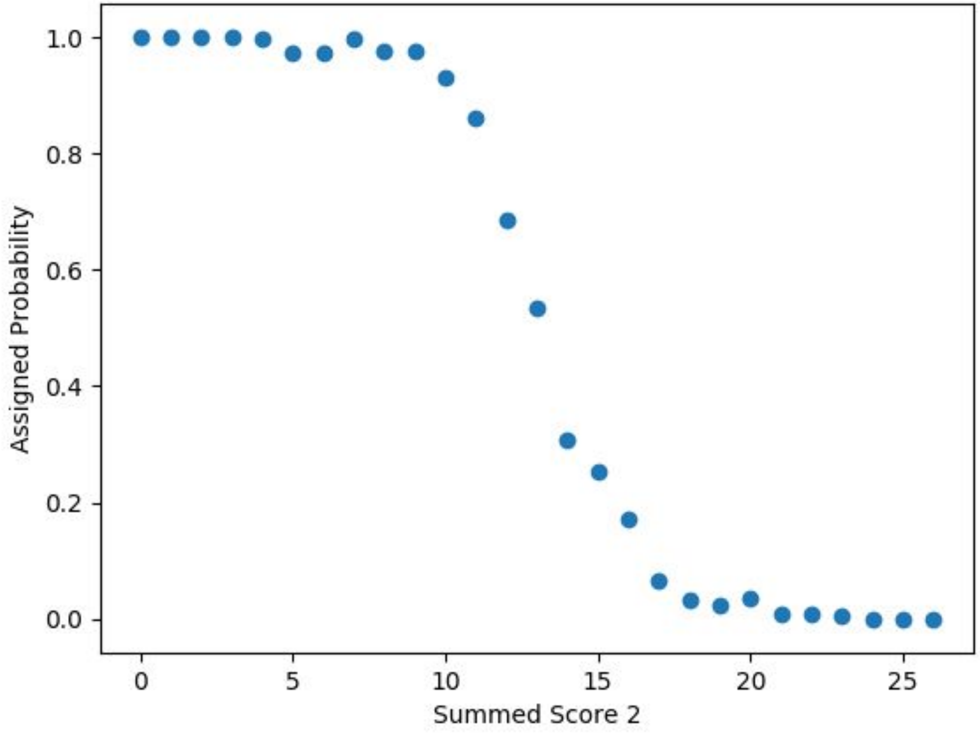
Picking scores with simple rules

- Guess every pass scored a 7, and every fail scored a 21!
- Let's make this a little more sophisticated.



Picking scores with simple rules

- Guess pass scored a 7, and fail score of 21.
- Exclude low information cases.
- Round violation to nearest integer.
- Include high pass (11) and high fail (45) scores.



Final Feature Engineering

- Sum Feature obtained by optimizing:
  $$\min P\hat{v} - 1 < 0 \quad \text{and} \quad F\hat{v} - 1 > 0$$

Final Feature Engineering

- Sum Feature obtained by optimizing:
  $\min P\hat{v} - 1 < 0 \quad \text{and} \quad F\hat{v} - 1 > 0$

- Use prior pass information

Final Feature Engineering

- Sum Feature obtained by optimizing:
  $\min P\hat{v} - 1 < 0 \quad \text{and} \quad F\hat{v} - 1 > 0$

- Use prior pass information

- Use frequency of restaurant name as a feature

Final Feature Engineering

- Sum Feature obtained by optimizing:
  $$\min P\hat{v} - 1 < 0 \quad \text{and} \quad F\hat{v} - 1 > 0$$
- Use prior pass information
- Use frequency of restaurant name as a feature
- Convert all actions, inspection types, and violations to binary columns

Data Quality / Observations

- Inspection Type column was slightly misleading.

Data Quality / Observations

- Inspection Type column was slightly misleading.
- Kaggle test data all took place on a later date than in training data.

Data Quality / Observations

- Inspection Type column was slightly misleading.
- Kaggle test data all took place on a later date than in training data.
- Weather data was accurate and complete, but not localized.

Data Quality / Observations

- Inspection Type column was slightly misleading.
- Kaggle test data all took place on a later date than in training data.
- Weather data was accurate and complete, but not localized.
- Location data had potential for higher level analysis.

Data Quality / Observations

- Inspection Type column was slightly misleading.
- Kaggle test data all took place on a later date than in training data.
- Weather data was accurate and complete, but not localized.
- Location data had potential for higher level analysis.
- Some possible mislabels / strange outliers occured in the actions column.

Classifier Selection

- Histogram Gradient Boosting Classifier worked best for this problem.

Classifier Selection

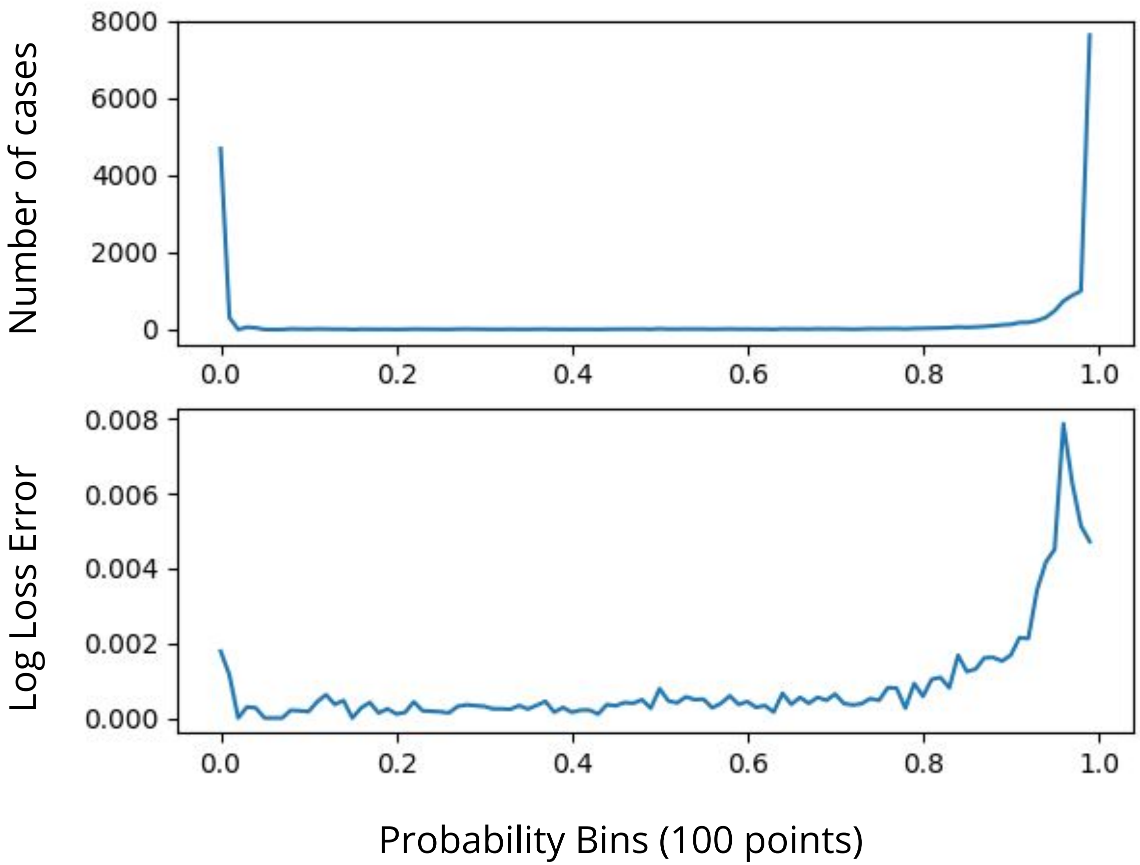
- Histogram Gradient Boosting Classifier worked best for this problem.
- Tested for robustness by frequently working with different subsets of our training data.

Classifier Selection

- Histogram Gradient Boosting Classifier worked best for this problem.
- Tested for robustness by frequently working with different subsets of our training data.
- Overfitting somewhat reduced by introducing L2 regularization.

Error Visualization

- Where does our model do well?
- Where does it do poorly?
- Integrate over the bottom graph to get total score! (0.084 on training data)



Closing Discussion

- How can we improve?

Closing Discussion

- How can we improve?
- Questions?

Appendix A

Many methods of altering the optimized sum method were attempted, none however offered any statistically significant improvement.

- These were: Constraint: $0 \leq v_i \leq 1$
- Separate data into groups for prior Failed, Passed, Unknown (assigned the regular sum)
- Separate data by cuisine type
- Separate data into weather groups
- Separate data by number of past passes / fails
- Let the prior discussed split groups be multipliers on an inspections score, rather than splitting the data
- Median of K distributions
- Retrain distribution excluding those violations which when multiplied by 13 were less than 1
- Include other binary information from action and inspection columns in the training

Appendix B

Here we exhaustively list all features that went into the final classifier that earned the best private score.

- The Second listed Sum Method
- The Third listed Sum Method
- 92 column vectors corresponding in binary to violations received during the inspection.
- Number of passes from prior inspections, Number of fails from prior inspections, result of previous inspection (-1 if failed, 0 if None, 1 if passed)
- Borough the inspection took place
- Cuisine type (as an integer label)
- Number of times the restaurant appeared in the data (larger values indicate a chain restaurant)
- Binary values for inspection type, and action type
- Precipitation, minimum temperature, maximal temperature, and average temperature that took place that day in Central Park.