Man vs Machine: A Comparison of Multivariate Machine Learning Techniques for Rooting Out Data Falsification

Scott Glendye, *U.S. Census Bureau*

Sheldon Waugh, *U.S. Census Bureau*

Rafael Puello, *U.S. Census Bureau*

Disclaimer: Any views expressed are those of the authors and not those of the U.S. Census Bureau

Background

Data Collection

- During the 2021 American Housing Survey (AHS) operation, several thousand interviewers took to the field
- Each day throughout the operation, the Field Quality Monitoring (FQM) team of the Census' Office of Survey and Census Analytics (OSCA) collected metadata on each incoming case across a variety of metrics
  - Metrics were then aggregated up to a series of rates at the interviewer level
- The FQM team then launched investigations into interviewer work in response to data anomalies

The Anomalies

- The FQM team identified anomalies in a variety of ad-hoc ways
- Automated flagging with Interquartile Range (IQR)
  - Across each metric
  - Outside of Q_1 - 1.5*IQR or Q_3 + 1.5*IQR in any metric
  - Flagged and sent to a human for investigation and manual verification
- This provided us with a natural experiment, where coded anomalies became the positive class for our experiment

United States® Census Bureau

Motivation for Improvement

- IQR is not the most accurate tool
  - Was easy to create and start with, but we believe that we can improve upon it
- Limiting False Negatives
  - Not detecting false data can have widespread impact on survey estimates
- Limiting False Positives
  - Investigating false positives takes resources away from working true positives

United States® Census Bureau

Experimental Design

Feature Selection

- During AHS, FQM monitored the data of *many* different metrics. For this experiment, we pared down our data set to just 4 metrics
- The metrics were normalized as percentages of all cases completed by the individual
- Combined with two one hot encodings (OHE) as control variables
  - Interviewer geography
  - Date in the operation

United States® Census Bureau

The Dataset

- Data was divided, by interviewer, into training and validation sets
- About 100 days worth of data was used in the training set
  - Used to tune hyperparameters of each model
- Each tuned model then made predictions on the holdout interviewers to identify outliers in each of the 100 days

Model Selection

- Three models were ultimately selected to compare against a benchmark of IQR, and against one another for accuracy
  - Multiple (cubic) linear regression with a cook's distance calculation
  - Isolation Forest
  - Extreme Gradient Boosting Outlier Detection (XGBOD)

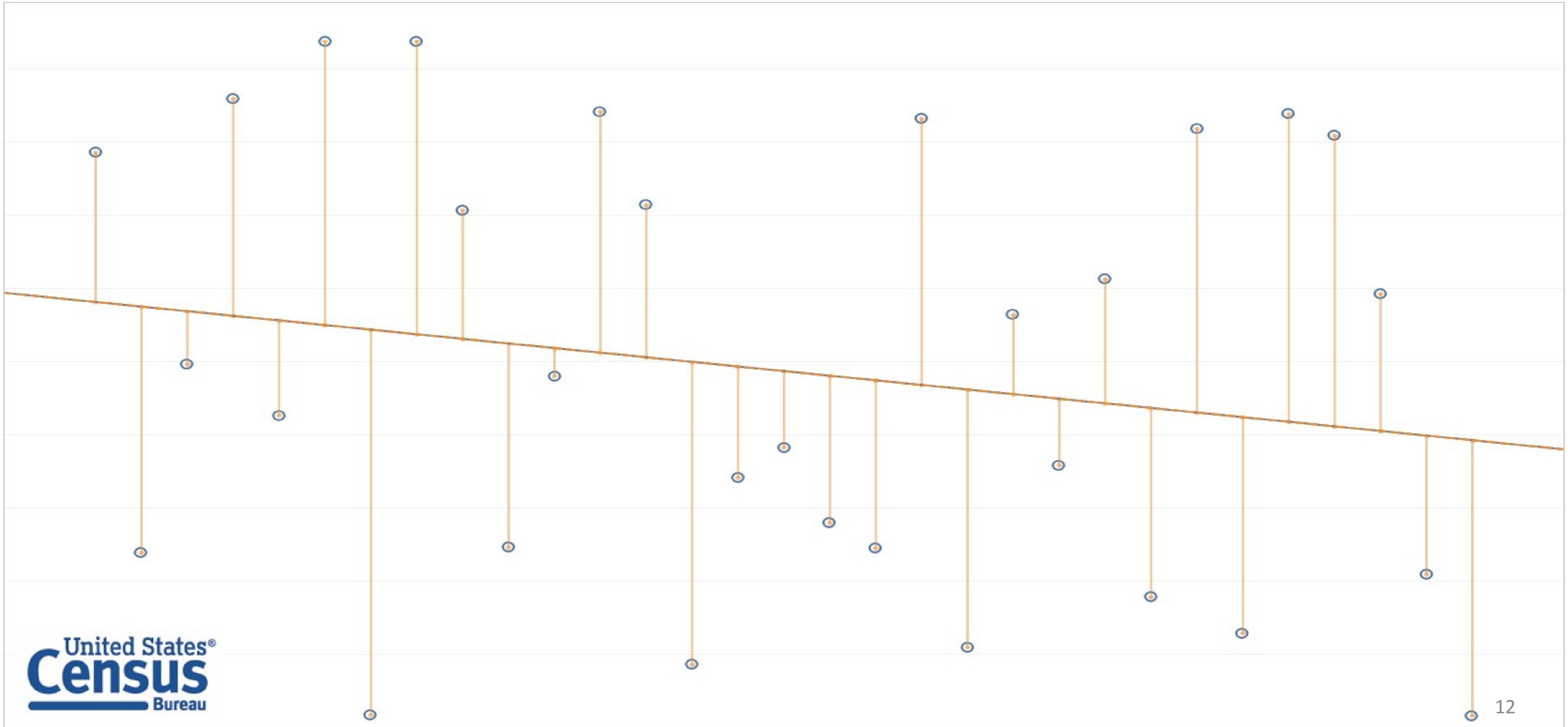
Model Overview

Multiple Regression With Cook’s Distance

- A cubic regression was selected to model the shape of interviewer case rates
- Cooks distance, or “delete one analysis”, was used to identify anomalous points
- Linear regression, being slightly different than the other pair of models required a regressand to fit against the regressors
    - Rate of completed cases of the interviewer

United States® Census Bureau

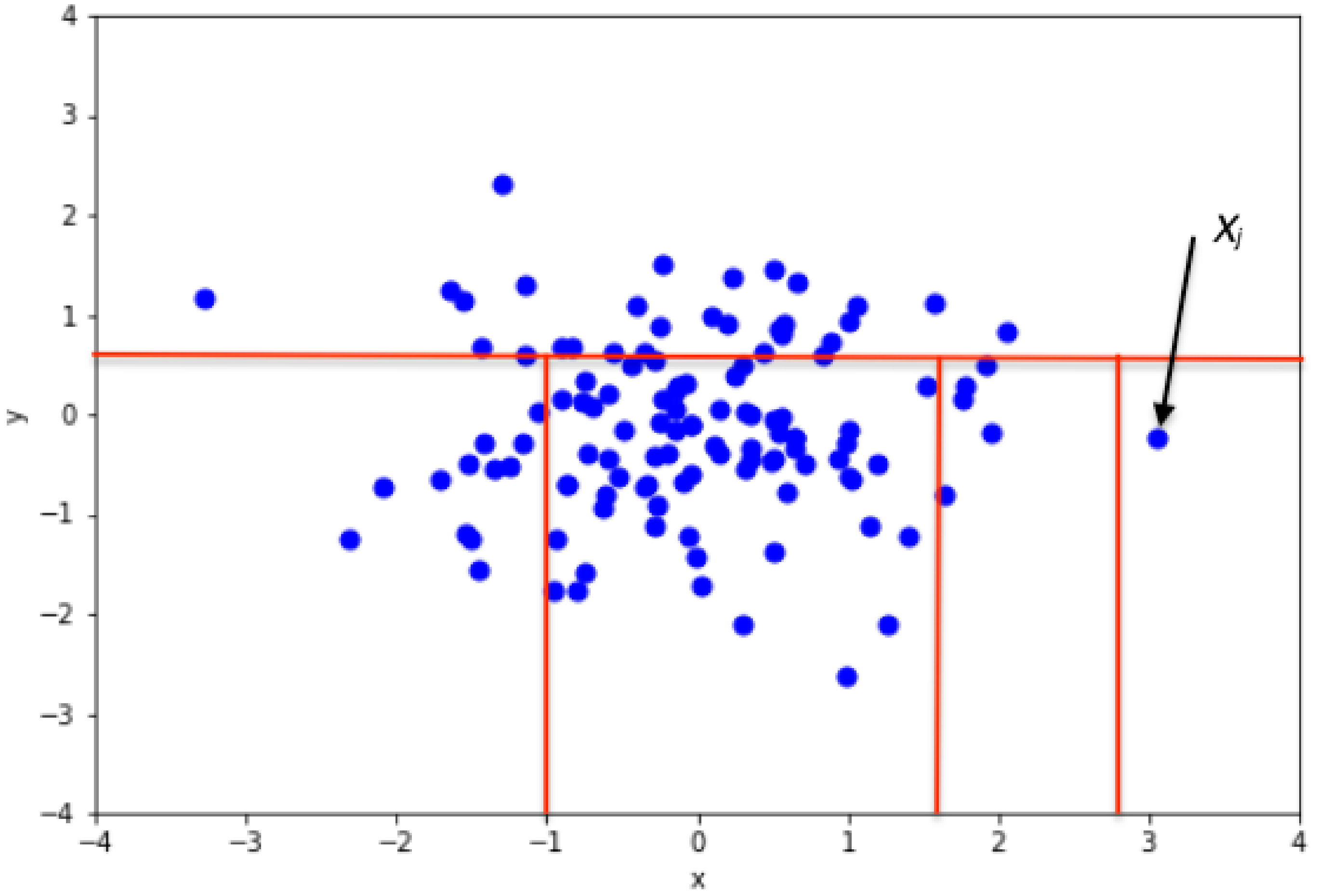
Cook’s Distance



Isolation Forest

- Tree based algorithm that achieves outlier “isolation” via random recursive partitioning of data to emphasize anomalous points
- Traditionally an unsupervised method
- We trained ours using a grid search with k-fold cross validation to identify the strongest hyperparameters
  - 4 folds were used for cross validation for each set of hyperparameters
  - Estimated grid search of hyperparameters

Isolation Forest



Source: Sal Borrelli - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=82709491

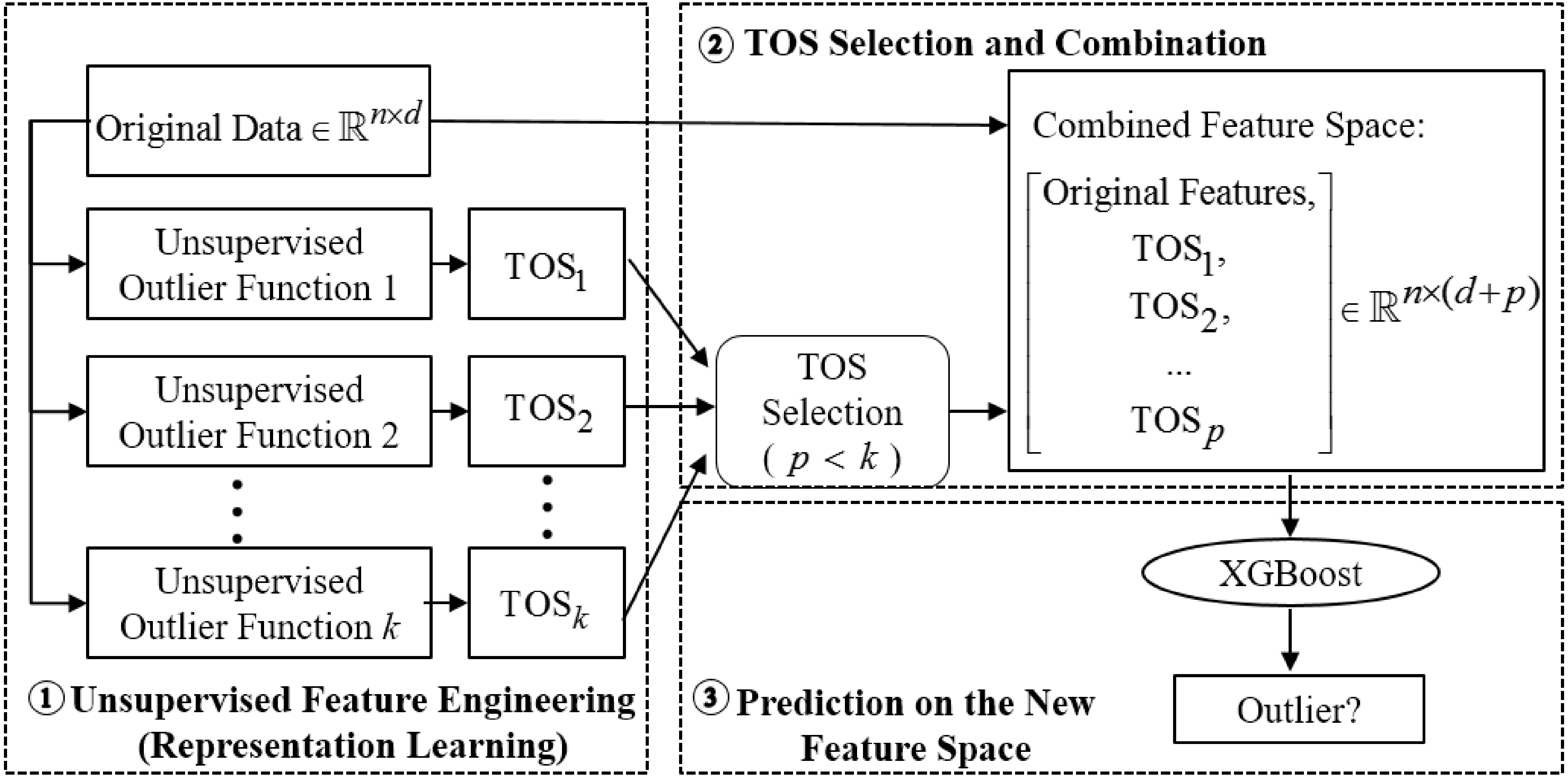
What is XGBOD?

- XGBOD is a framework established to improve the performance of xgboost classifiers[1]
- It is a three step semi-supervised learning algorithm
  - Generate “Transformed Outlier Scores” (TOS)
  - Pare off resultant scores
  - Perform a gradient boosted forest classifier on the newly modified feature space

United States® Census Bureau

[1]Zhao, Y. and Hryniewicki, M.K., "XGBOD: Improving Supervised Outlier Detection with Unsupervised Representation Learning," *International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2018.

XGBOD Framework



Source: Zhao, Y. and Hryniewicki, M.K., "XGBOD: Improving Supervised Outlier Detection with Unsupervised Representation Learning," *International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2018.

TOS Unsupervised Models

- Our XGBOD TOS models included only those in the original XGBOD code demonstration
  - k-nearest neighbors
  - One Class Support Vector Machine
  - Isolation Forest
- A range of values for k (KNN), mu (SVM), and number of trees (isolation forest) tested
  - Results randomly looped over and a portion were added to the model as features

United States® Census Bureau

Why These Models?

- Raising the bar
  - Cook’s D has been a highly effective tool for decades
  - Isolation forest is a leading SOTA model
  - XGBOD promises superior performance
- Supervised vs unsupervised
  - All 3 models can easily be trained for either scenario

Results

Overall Performance

Validation Set F1 Scores			
Model	Precision	Recall	F1
IQR	0.30	0.50	0.38
Cooks	0.24	0.35	0.29
Isolation	0.31	0.51	0.39
XGBOD	0.57	0.33	0.42

Interviewer Level Performance

Validation Set Interviewer Level Identification				
Model	Anomalies Correctly Identified	Anomalies Missed	Non-Anomalies Incorrectly Flagged	% Predictions Correct
IQR	83.8%	16.2%	49.9%	19.6%
Cooks	94.9%	5.1%	61.9%	18.8%
Isolation	74.7%	25.3%	34.8%	22.8%
XGBOD	70.7%	29.3%	12.2%	39.3%

United States® Census Bureau

Conclusion

Limitations

- Inaccurate coding of data irregularity start date
  - Many interviewers likely changed their patterns
- The dilemma of unknown unknowns
  - Some anomalies were likely missed and miscoded as non-anomalous
- XGBOD training resources
  - Training took many hours even with a slimmed down data set. Accuracy was left on the table

United States® Census Bureau

Closing Remarks

- We’ve known for some time that ensemble approaches used by some models (i.e. decision trees) are effective tools for improving model robustness
  - XGBOD uses a similar concept with an ensemble of models, albeit a more heterogenous selection than tree based approaches
- Based on this experiment, XGBOD did appear significantly more effective than a single model, as claimed by its authors
  - Future research should combine other algorithms to create additional features, as is currently being done with the SUOD project

United States® Census Bureau