Methods of Anomaly Detection

A Mathematical Framework and Application in Health Care Policy Optimization

Nathaniel Islip[1]

[1]Department of Mathematics
Eastern Washington University

Masters Thesis Defense, June 2022

Table of Contents

Table of Contents

What is an Anomaly?

Definition (Anomaly or Outlier)

An instance or data point demonstrating behavior substantially different from what is considered normal

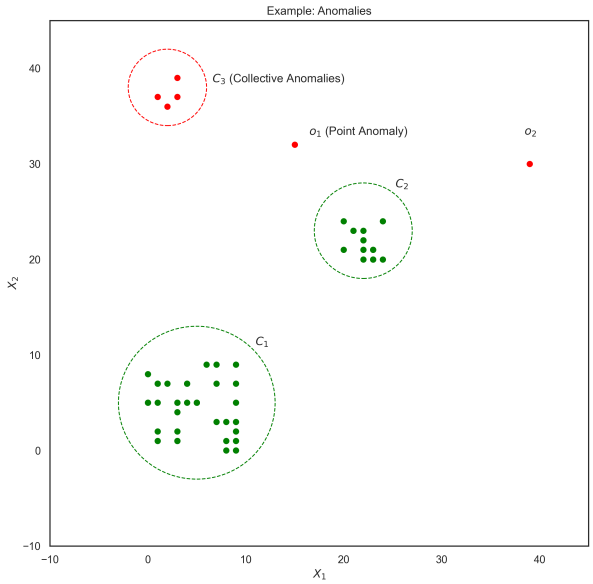


Figure: Example of anomalies

Anomaly Detection

Definition (Anomaly Detection)

Anomaly Detection (AKA novelty or outlier detection) is a sub-field of Machine Learning (ML) used in identifying multi-dimensional anomalies

Learning Problems

- Supervised
- Unsupervised
- Semi-supervised

Types of AD

- Distance-Based (LOF) [1]
- Rule-Based
- Class Based (OC-SVM) [8]

Applications of Anomaly Detection

- Astronomical Science [5]
- Cyber-security [9]
- Genomic Sequencing [7]
- Medical Imaging [11]
- Fraud Detection [12]
- Nuclear Science [6]

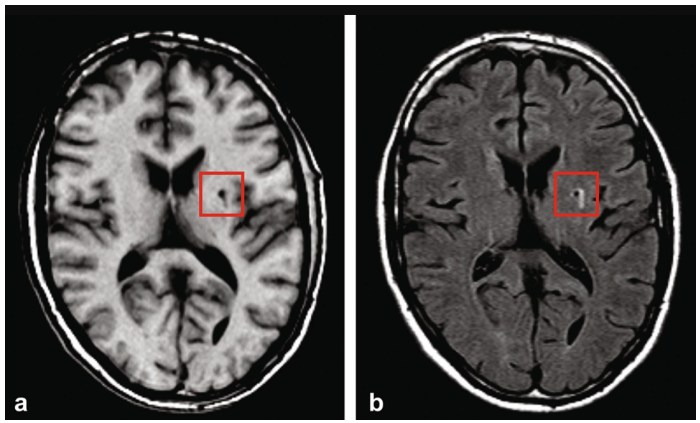


Figure: Identification of chronic brain infracts from an MRI

Table of Contents

Motivation

Motivation

Optimizing the structure of cost sharing is a long standing issue among policy makers to reduce Medicaid spending while maintaining the quality of care.

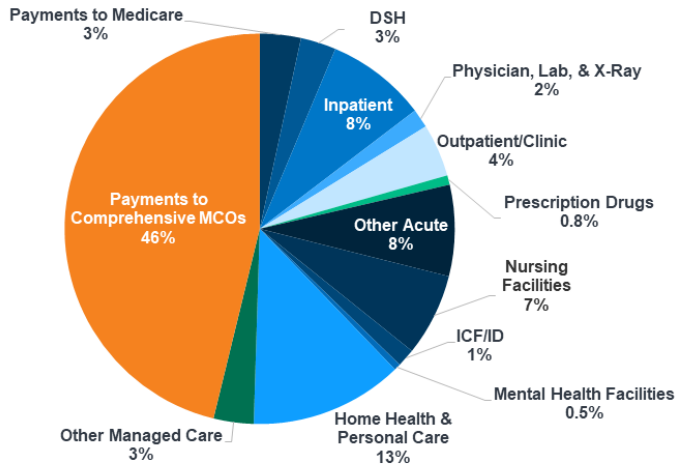
Medicaid

Definition (Medicaid)

Medicaid is a public insurance health program for individuals of low income jointly funded by the Federal and State government.

Figure 2

Payments to comprehensive MCOs account for almost half of total national Medicaid spending.



FY 2019 Total Medicaid Spending: $603.7 Billion

NOTE: "Other Managed Care" includes prepaid health plans (PHPs), primary care case management (PCCMs), programs of all-inclusive care for the elderly (PACEs) as well as premiums and coinsurance paid toward employer group insurance plans and premiums paid for other insurance or remedial care. Prescription Drug spending is for fee-for-service only, managed care prescription drug spending is included in "Payments to Comprehensive MCOs". Excludes administrative spending, adjustments, and payments to the territories.
SOURCE: Urban Institute estimates based on FY 2019 data from the CMS-64, prepared for the Kaiser Program on Medicaid and the Uninsured.

KFF

Figure: Medicaid spending across services

Medicaid Cost Sharing

Definition (Cost Sharing)

States charge groups of individuals for medical services (see Figure 3) contingent upon limitations. Costs include deductibles, premiums, co-payments, and coinsurance

Limitations as they pertain to cost sharing...

- Are these limitations the same across all **States**? **No**
- Are these limitations the same across all **Medicaid Services**? **No**
- Are these limitations the same across all **age groups**? **No**

Previous Work

Motivation

Optimizing the structure of cost sharing is a long standing issue among policy makers to reduce Medicaid spending while maintaining the quality of care.

Potential impacts of cost sharing on Medicaid spending...

- Higher premiums induce dis-enrollment
- Interruptions in Medicaid coverage
- Increased premiums induce lower revenue
- Incentive programs and preventative care

Research Statement

Research Statement

Using AD, identify abnormal spending, utilization, and access to care at the family-level. Moreover, characterize these clusters of abnormal and normal spending

Potential implications

1. Define an **average population**
2. Define an **abnormal population**

Data Source

Data provided by **Agency for Healthcare Research and Quality** (AHRQ) [4]

- Healthcare Cost and Utilization Project (HCUP)
- **Medical Expenditure Panel Survey (MEPS)**
- U.S. Health Information knowledge base (USHIK)

Components of **MEPS** data

- MEPS Household Component (HC) 2018 ($n = 30461$, $q = 1501$)
- Expenditures, insurance coverage, health status, employment...
- Individual and family level

Data Cleaning

Goal: Obtain a family-level data set constrained to the Medicaid population

---

1. Filter: **FAMWT18F** > 0
2. Concatenate **DUID** and **FAMIDYR** to **DUIDFAMY**
3. Aggregate features using **DUIDFAMY**
4. **MCDHMO31** $=$ **MCDHMO42** $=$ **MCDHMO53** $= 1$
5. Aggregate numeric features by **SUM**
6. Aggregate categorical features by **MAX**

Output: Family-level data set containing features related to Expenditures, Utilization, and Access to Health Care constrained to Medicaid population ($n = 1541$, $q = 31$)

Data Cleaning (Continued)

Index	DUID	PID	DUIDFAMY	FAMRFPYR	POVLEV18
2	2290002	101	2290002A	1	163.92
3	2290002	102	2290002A	0	163.92
4	2290002	103	2290002A	0	163.92
5	2290002	104	2290002A	0	163.92
6	2290002	105	2290002A	0	163.92
7	2290002	106	2290002A	0	163.92

Table: Records from a single family unit or unique DUID

Index	DUIDFAMY	FAMRFPYR	POVLEV18	RXMCD18
0	229001A	1	190.31	1103
2	2290002A	1	163.92	0
8	2290003A	1	1013.48	0
13	2290005A	1	203.18	0

Table: Records from multiple family units after aggregating data

Model Pipeline

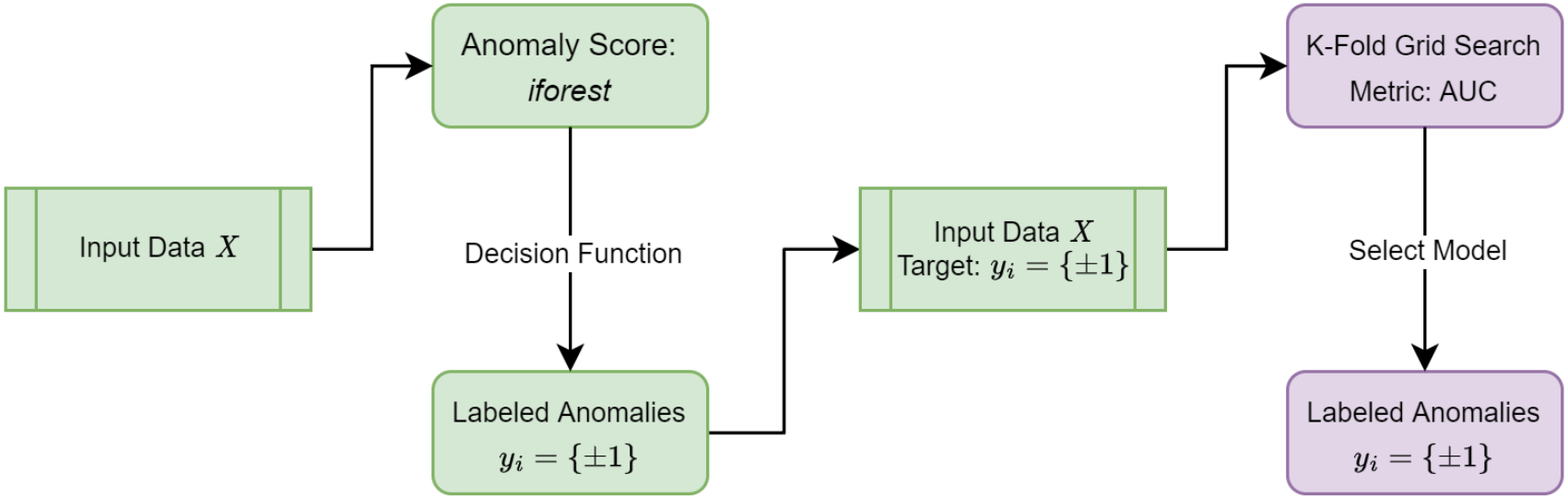


Figure: Model pipeline

Concept: Random Partitioning

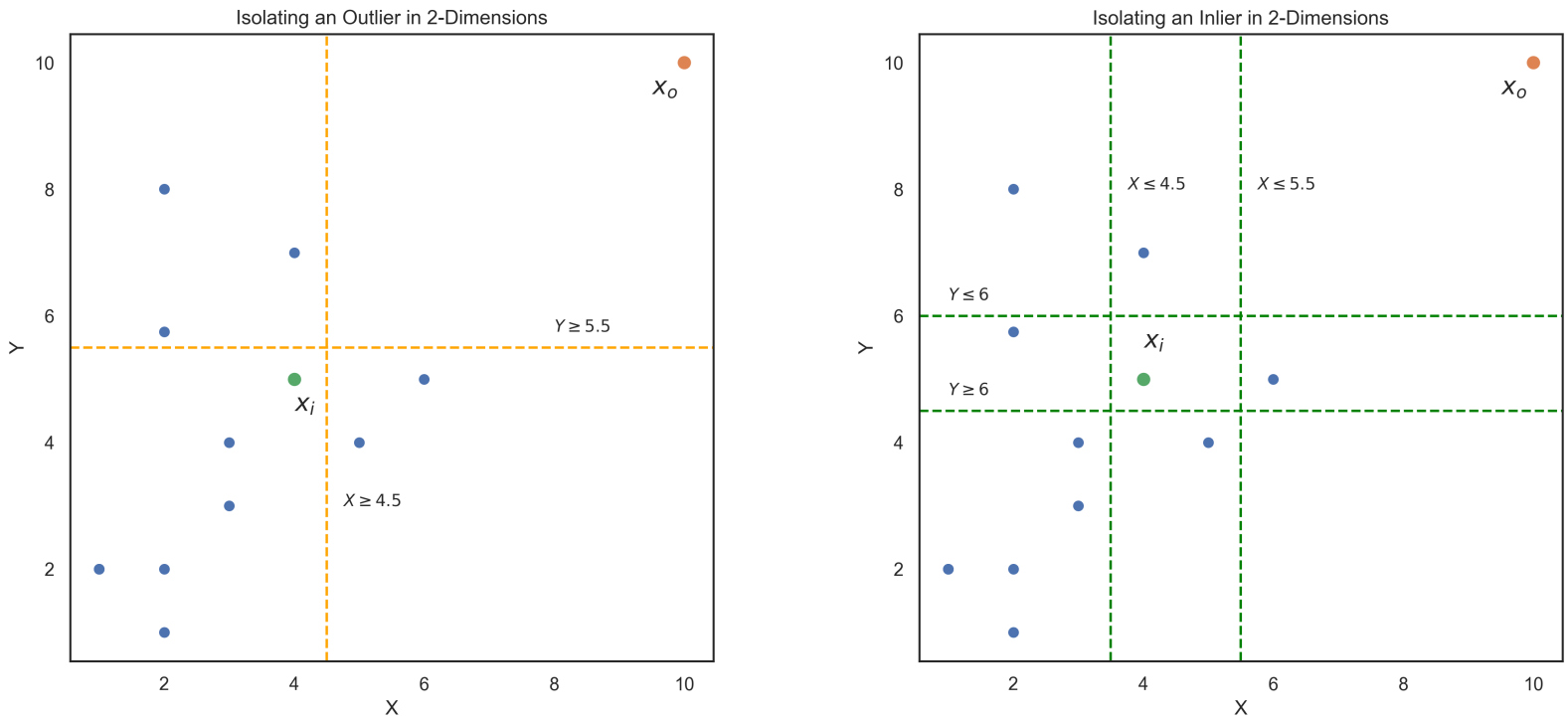


Figure: Isolating the data points x_o and x_i using random partitioning

Properties of Binary Search Trees (BST) and *itree*

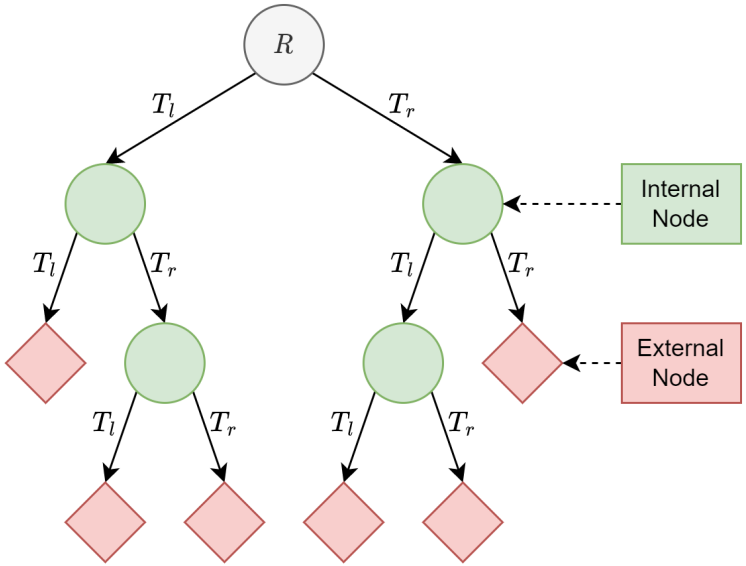


Figure: Binary Search Tree (BST)

Binary Search Trees (BST)

- Every node (T_l, T_r) has at most 2 children [10]
- *itree* is a proper binary tree
- Number of external nodes n
- Number of internal nodes $n-1$
- Total nodes in *itree* $2n-1$

Isolation Tree (*itree*)

Definition (Isolation Tree)

Let T be a node of an isolation tree. T is either an external-node with no child, or an internal-node with one test and exactly two daughter nodes (T_l, T_r). A test consists of an attribute q of an attribute q and a split value p such that the test $q < p$ divides the data points into T_l and T_r.

Given $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$ recursively divide $\mathcal{X}$ until ...

1. Height limit ℓ
2. $|\mathcal{X}| = 1$
3. All data in $\mathcal{X}$ have the same values

itree Algorithm

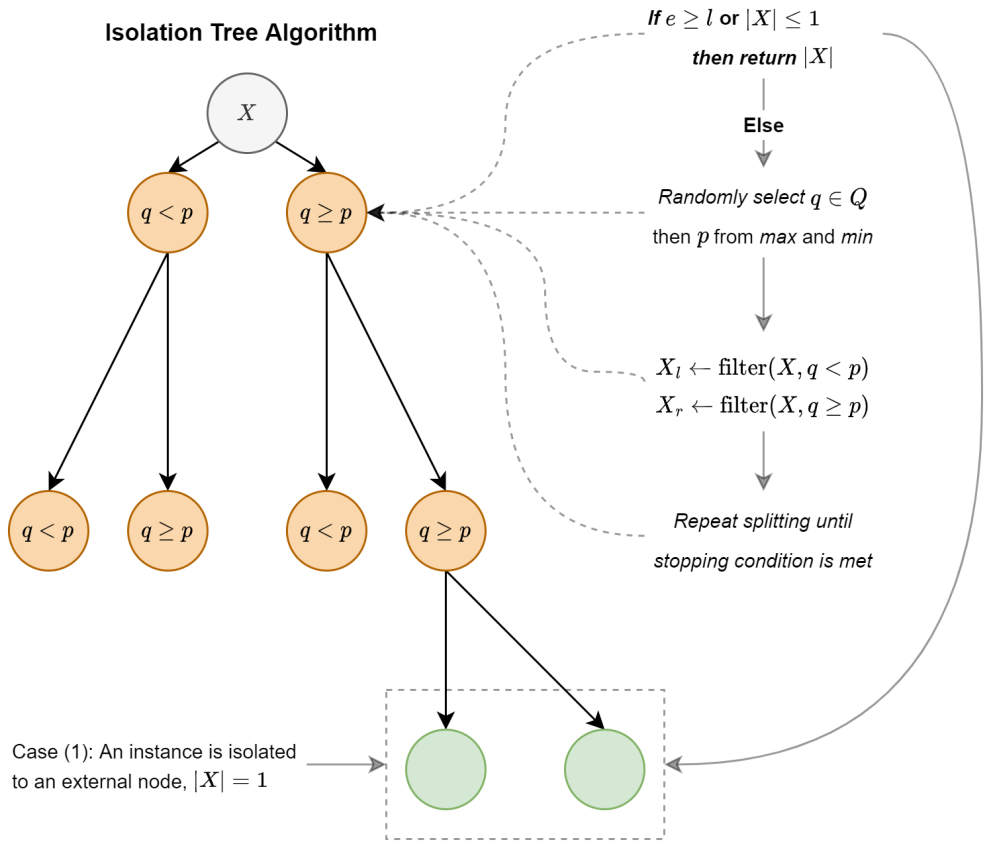


Figure: *itree* algorithm schematic

iforest Anomaly Score Method

Definition (Anomaly Score)

let $s(x, n)$ be the anomaly score of an instance x given n instances.

$$s(x, n) = 2^{-\frac{E(h(x))}{c(n)}} \tag{1}$$

- $h(x)$ is the path length of a data point x traversing an *itree*
- $c(n) = 2H(n-1) - (2(n-1)/n)$ is the average path length of an unsuccessful BST
- $E(h(x))$ is the average of $h(x)$ from a collection of isolation trees

1. Instances close to $s = 1$ are anomalies
2. Instances smaller than $s = 0.5$ are considered normal
3. All instances with $s \approx 0.5$ indicate no definitive anomalies

iforest Anomaly Score: Results

Figure: Scatter plot of anomaly scores

Modeling Anomalies

Interpreting a Black-Box Model

What combinations of features characterize the anomalies?

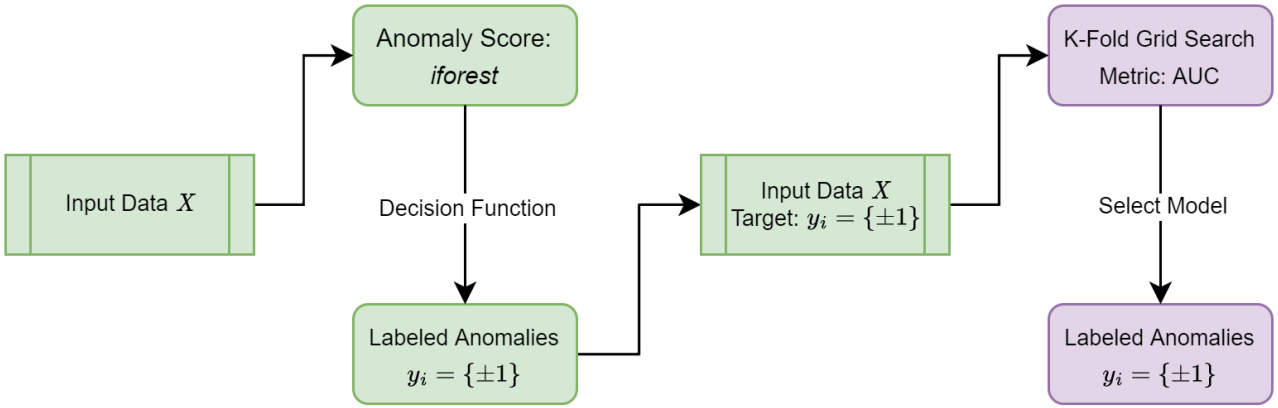


Figure: Model pipeline

Feature Importance (FI)

Goal: Add up weighted impurity decreases for each split in a tree and average over all trees in the forest [13]

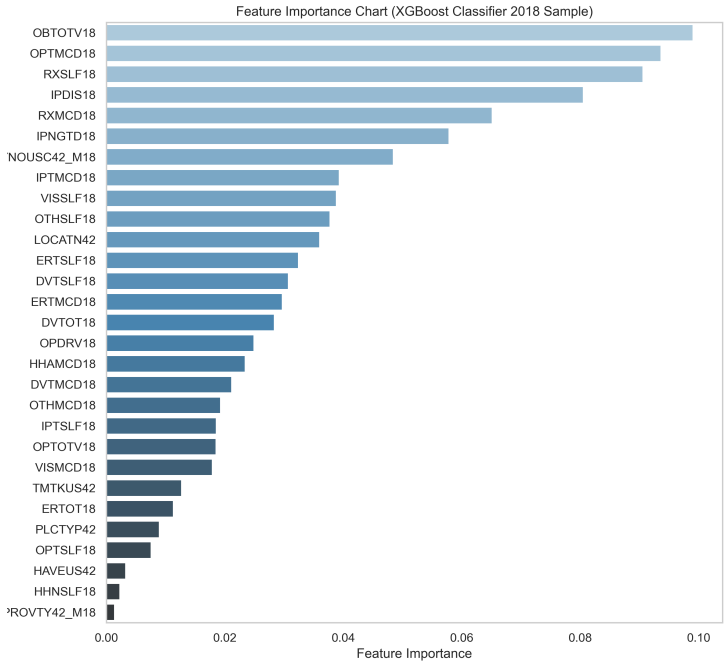


Figure: Feature importance from a gradient boosted random forest classifier

Partial Dependence Plots (OBTOTV18)

Goal: The Partial Dependence Plot (PDP) measures the average effect of the jth feature on the prediction. For classification, the PDP plot outputs the probability for a specific class given different features.

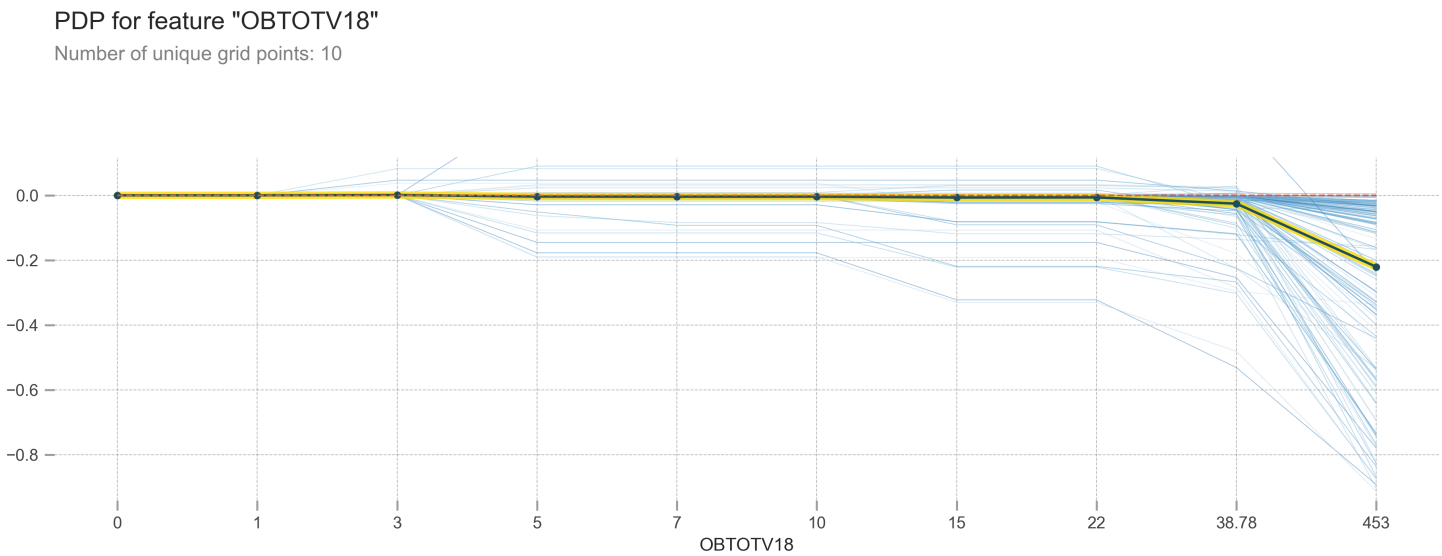


Figure: PDP OBTOTV18

PDP (OPTMCD18)

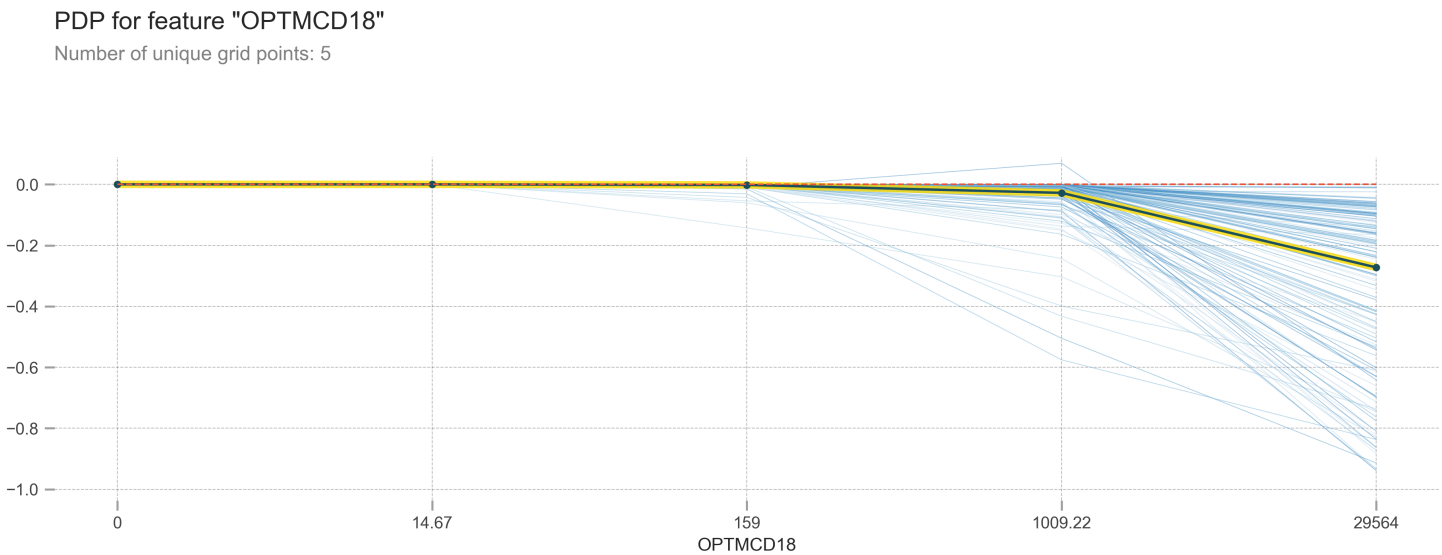


Figure: PDP OPTMCD18

PDP (RXMCD18)

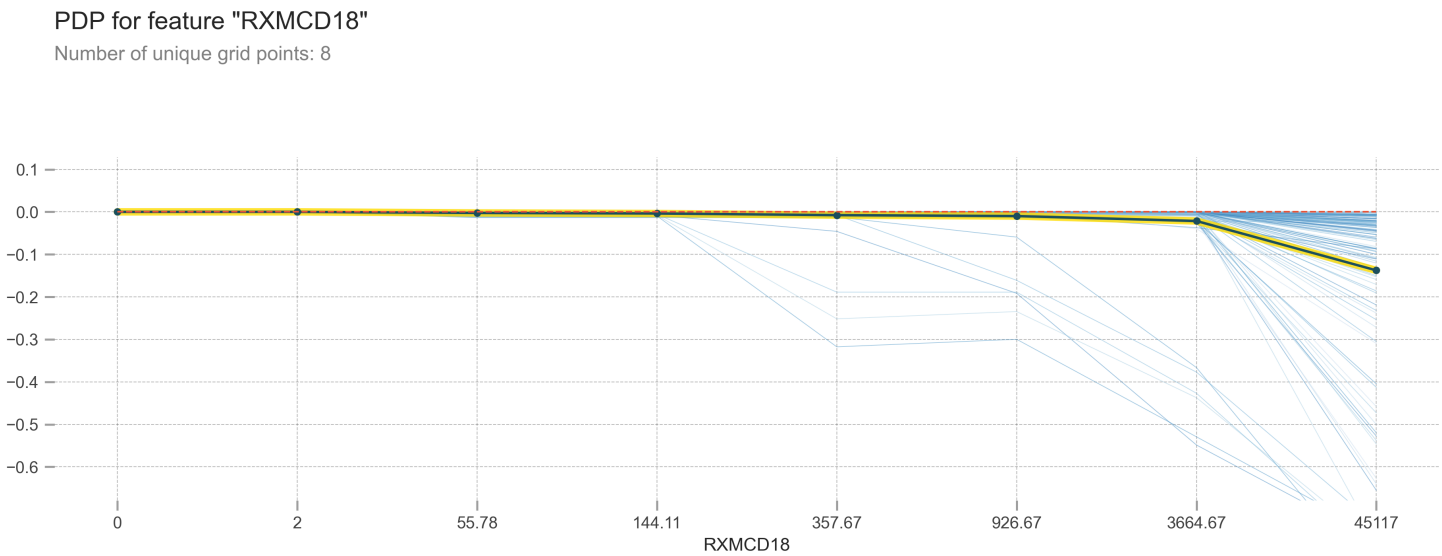


Figure: PDP RXMCD18

PDP (RXSLF18)

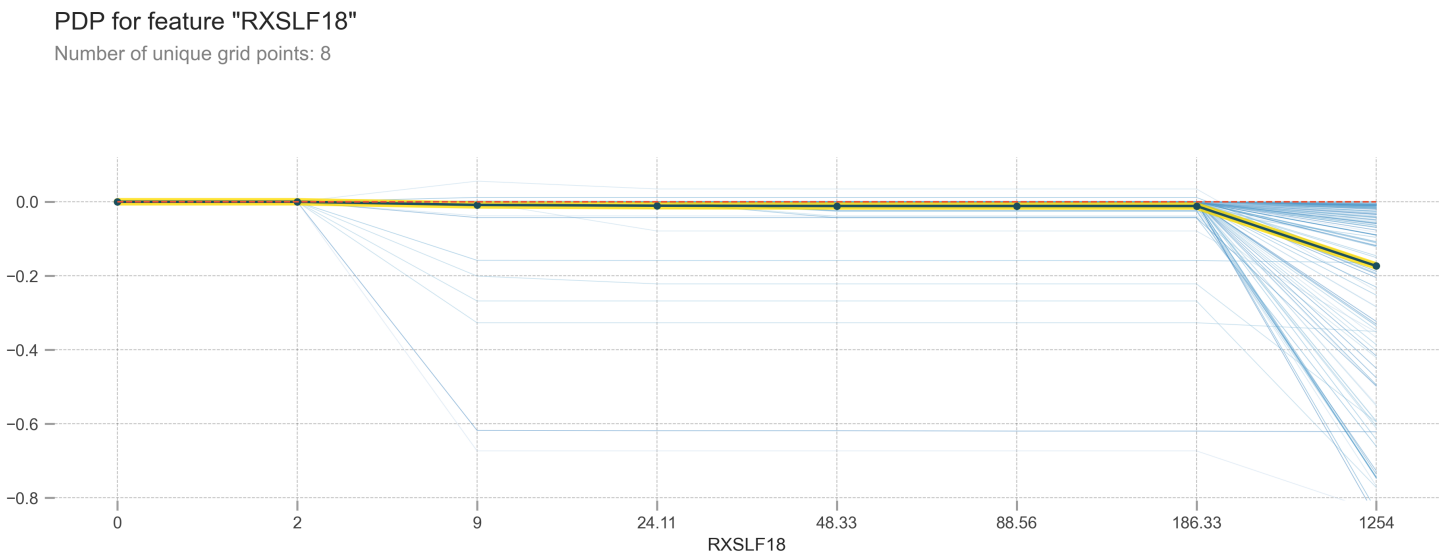


Figure: PDP RXSLF18

Discussion and Next Steps

Takeaway

Using AD, instances of abnormal expenditures or utilization, or access to health care are identified using iforest's. Moreover, the features driving these anomalous and normal populations are identifiable by modeling the anomaly scores.

Limitations

- Small N
- Control (i.e. State Policy)
- Interpretation (i.e. PDP)

Future Research

- Large N (e.g. ResDAC)
- Medicaid eligibility
- Cluster Analysis

Questions?

Table of Contents

References I

[1] Markus M Breunig et al. "LOF: identifying density-based local outliers". In: *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*. 2000, pp. 93–104.

[2] Lars Buitinck et al. "API design for machine learning software: experiences from the scikit-learn project". In: *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*. 2013, pp. 108–122.

[3] Tom Fawcett. "An introduction to ROC analysis". In: *Pattern recognition letters* 27.8 (2006), pp. 861–874.

[4] Agency for Healthcare Research and Quality. *MEPS HC-209: 2018 Full Year Consolidated Data File*. URL: https://meps.ahrq.gov/mepsweb/data_stats/download_data_files_detail.jsp?cboPufNumber=HC-209.

[5] Emille EO Ishida et al. "Active anomaly detection for time-domain discoveries". In: *arXiv preprint arXiv:1909.13260* (2019).

[6] Xin Jin et al. "Anomaly detection in nuclear power plants via symbolic dynamic filtering". In: *IEEE Transactions on Nuclear Science* 58.1 (2010), pp. 277–288.

References II

[7] Jonathan Kaufmann et al. "One-class ensembles for rare genomic sequences identification". In: *International Conference on Discovery Science*. Springer. 2020, pp. 340–354.

[8] Minh-Nghia Nguyen and Ngo Anh Vien. "Scalable and interpretable one-class svms with deep learning and random fourier features". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2018, pp. 157–172.

[9] Animesh Patcha and Jung-Min Park. "An overview of anomaly detection techniques: Existing solutions and latest technological trends". In: *Computer networks* 51.12 (2007), pp. 3448–3470.

[10] Bruno R.. Preiss. *Data Structure and Algorithms: With Object-oriented Design Patterns in Java*. John Wiley & Sons, 1999.

[11] Maximilian E Tschuchnig and Michael Gadermayr. "Anomaly Detection in Medical Imaging–A Mini Review". In: *arXiv preprint arXiv:2108.11986* (2021).

References III

[12] Weijia Zhang and Xiaofeng He. "An anomaly detection method for medicare fraud detection". In: *2017 IEEE International Conference on Big Knowledge (ICBK)*. IEEE. 2017, pp. 309–314.

[13] Zhengze Zhou and Giles Hooker. "Unbiased measurement of feature importance in tree-based methods". In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 15.2 (2021), pp. 1–21.

Table of Contents

Appendix A: Grid Search K-Fold Cross Validation

Definition (K-Fold Cross Validation)

k-fold cross validation (K-CV) is a method of re-sampling data for k partitions of the sample. For each partition, the data is split into a holdout (test) and training data set.

- Grid Search with CV was ran with Logistic Regression, SVM, Random Forest, and Gradient Boosted Random Forests [2]
- `gridsearchCV(model_paramters = dict, X_train = X_train_18, Y_train = Y_train_18, scoring = "roc_auc", k = 5)`
- `GradientBoostingClassifier(Learning_rate = 0.01, n_estimators = 300, tol = 0.01, validation_fraction = 0.3)`

Appendix B: Receiver Operator Curve (ROC)

Definition (Receiver Operator Characteristic Curve)

A Receiver Operator Characteristic Curve is a method for evaluating the performance of a classification model using the False Positive Rate (FPR) and True Positive Rate (TPR). The Area under the Curve is a scalar value between 0 and 1.0 where a value of 0.5 is equivalent to random guessing, and a value of 1.0 is a perfect classifier [3].

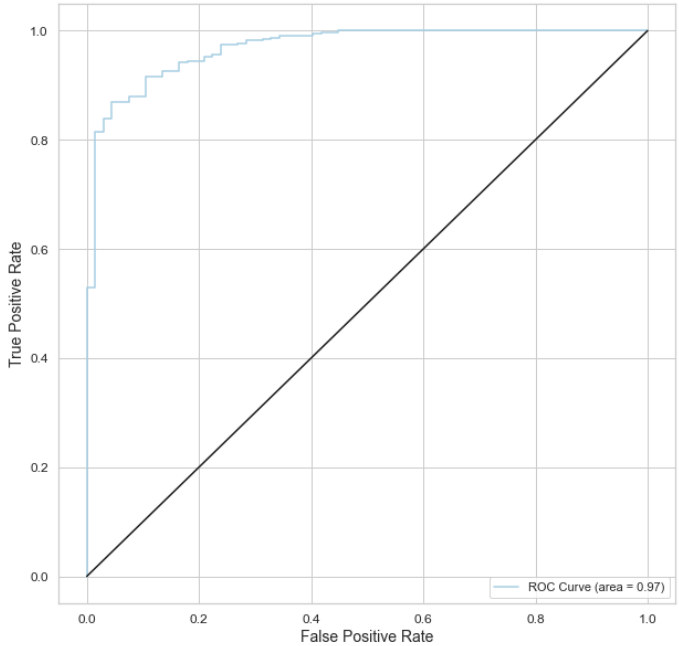


Figure: ROC for GBRF with 0.97 AUC

Appendix C: FI Derivation

Definition (Split-Improvement Feature Importance)

Let $H(m)$ be the impurity function (Gini Index) for a node m

$$H(m) = \sum_{k \neq k'} p_{mk} p_{mk} = 1 - \sum_{k=1}^{K} p_{mk}^2 \tag{2}$$

where p_{mk} is the proportion of class K in node m and n_m observations at node m.

$$p_{mk} = \frac{1}{n_{mk}} \sum_{x_i \in m} \mathbb{1}(y_i = K)$$

Let the best split at m be θ_m^* by splitting the jth variable resulting in two child nodes l and r. Decrease in impurity for θ^*

$$\Delta(\theta_m^*) = \omega_m H(m) - (\omega_l H(l) + \omega_r H(r))$$

where ω is the proportion of observations falling into each node. The split feature importance for an ensemble of Random Forests is given by

$$VI_j^{\text{RF}} = \frac{1}{B} \sum_{b=1} \sum_{m, j \in \theta_m^*} \Delta_b(\theta_m^*) \tag{3}$$

where B is base learner [13].

Appendix D: Harmonic Numbers

Deriving H_{n-1}, or $H(n-1)$

Using the definition of harmonic series H_n

$$H_n = \sum_{i=1}^{n} \frac{1}{i} \tag{4}$$

we can compute the area under the stair descending staircase as

$$\int_1^n \frac{1}{x} dx = \ln(n)$$

therefore approximating H_{n-1} is approximately $\ln n$ for $n > 1$. This derivation yields Euler's constant γ

$$\begin{aligned}
\gamma &= \sum_{i=1}^{\infty} \left(\int_i^{i+1} \left(\frac{1}{i} - \frac{1}{x} \right) dx \right) \\
&= \sum_{i=1}^{\infty} \left(\frac{1}{i} - \ln \left(\frac{i+1}{i} \right) \right) \\
&\approx 0.5772156649
\end{aligned}$$

Therefore, the usual approximation for the $(n-1)$th harmonic number is $H_{n-1} = \ln n + \gamma$

Appendix E: BST Unsuccessful Search

The average external path length of a binary search tree is given by

$$\begin{aligned} E(n) &= I(n) + 2n \\ &= 2(n+1)H_n - 2n \\ &\approx 2(n+1)(\ln(n) + \gamma) - 2n \\ c(n) &= \frac{E(n)}{(n+1)} \\ &= 2H_n - \frac{2n}{(n+1)} \\ &\approx 2(\ln(n) + \gamma) - \frac{2n}{(n+1)} \end{aligned}$$

Appendix F: Isolation Forest Example

Example of *iforest* on synthetic data.

- $N = 115$, $q = 2$
- Hyper-parameters: n-estimators $= 150$

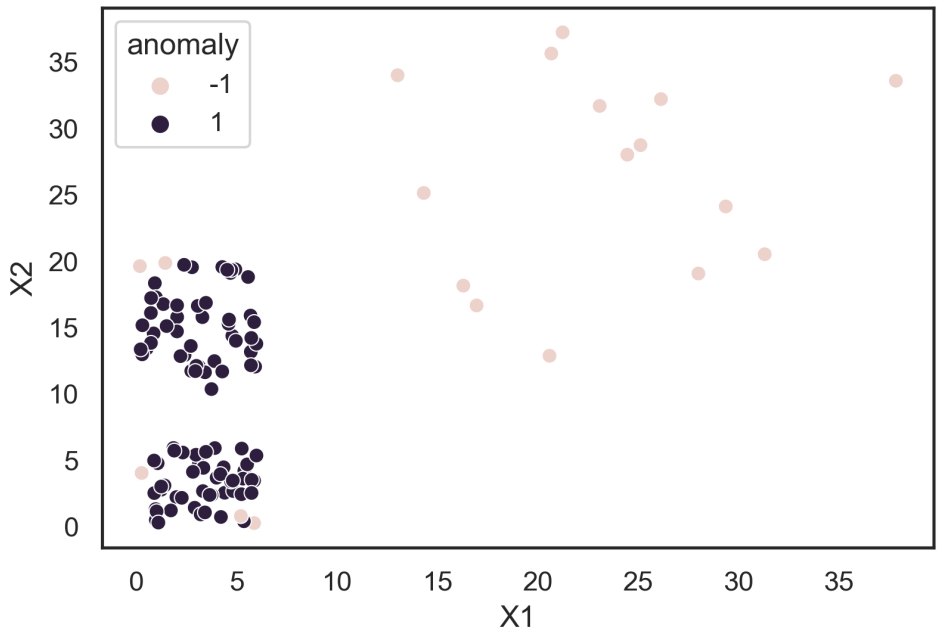


Figure: Synthetic example of *iforest*