IDENTIFYING ANOMALOUS ER CLAIMS USING HIERARCHICAL CLUSTERING AND RANDOM FORESTS

Jesse B. Crawford[12] **and Nicholas A. Petela**[12]
[1]Department of Mathematics
[2]Tarleton Analytics Institute
Tarleton State University
Stephenville, TX 76402
jcrawford@tarleton.edu

ABSTRACT. Improper health insurance payments resulting from fraud and upcoding result in tens of billions of dollars in excess health care costs annually in the United States, motivating machine learning researchers to build anomaly detection models for health insurance claims. This article describes two such strategies specifically for ER claims. The first is an upcoding model based on severity code distributions, stratified by hierarchical diagnosis code clusters. A statistically significant difference in mean upcoding anomaly scores is observed between free-standing ERs and acute care hospitals, with free-standing ERs being more anomalous. The second model is a random forest that minimizes improper payments by optimally sorting ER claims within review queues. Depending on the percentage of claims reviewed, the random forest saved 12% to 40% above a baseline approach that prioritized claims by billed amount.

1. INTRODUCTION

The U.S. Government Accountability Office [1] estimates that over $50 billion dollars each year are spent on improper health insurance payments, with fraud possibly accounting for $18B to $61B. According to Che and Janusz [2], the loss due to upcoding alone is approximately $2.4 billion, prompting data analytics research in upcoding detection [3]. He et. al [4] have investigated applications of Multi-Layer Perceptron (MLP) algorithms to human labeled data from the Australian Health Insurance Commission. While they originally used four degrees of anomaly classification, they found that using only two presented a higher rate of agreement between the MLP and consultants. Rosenburg et al. [5] built a hierarchical logistic Bayesian model based on the principal reason for admission that predicted whether audits on Diagnosis Related Group (DRG) codes were performed. This method relied on a high number of local models (one for each principal reason for admission) and achieved useful approximations of the underlying distribution of audits and related covariates, such as length of stay.

Yamanishi et al. [6] developed the SmartSifter algorithm, an on-line process for the detection of anomalous claims, by applying unsupervised machine learning to the underlying claims distribution. This method was inspired by research in network detection intrusion algorithms and was successfully applied to certain outlier detection problems in a dataset from the Australian Health Insurance Commission. Luo and Gallagher [7] also implemented an unsupervised algorithm in an Australian dataset using DRG codes, comparing individual hospitals under the assumptions that upcoding was likely adjacent (upcoded a single level) and unintentional. Their model uses decision trees to create homogenous subgroups and then tests the null hypothesis that each group originated from the same underlying DRG distribution. Che and Janusz [2] investigated another on-line algorithm by using an adjustable time-window model to create semi-supervised rulesets. Other research includes Thornton et al. [8], who outlined a systematic model for claims review that includes upcoding but relies entirely on human reviewers, and Johnson and Nagarur's multi-step method based on distance and relative density of resource expenditures by individual providers [9].

In Sections 2 and 3, we present findings for two ER claims analytics, developed in partnership with Blue Cross and Blue Shield of Texas. The first is an upcoding anomaly score based on ICD-10 diagnosis codes and HCPCS/CPT severity codes. Hierarchical agglomerative clustering is used to group diagnosis codes into risk categories, and the severity code for each claim is compared to the severity code distribution of claims within the same diagnosis code cluster. A statistically significant difference in average upcoding anomaly scores ($P < 10^{-74}$) is observed between acute care hospitals and free-standing ERs.

The second analytic addresses an optimization problem posed by the claims review process: reviewing all claims is prohibitively expensive, but reviewing an insufficient number causes fraudulent claims and improper payments to be overlooked. The solution is to build a predictive model that estimates cost avoidance for claims before they are reviewed and to prioritize review queues accordingly. A random forest is applied to claims data at the header and line level and engineered features describing prescription drug abuse, doctor shopping, unnecessary unbundling, association analysis metrics, and excessive CT scans. These features and reimbursement method anomalies are among the most predictive claim attributes, according to variable importance metrics. The model's performance is then evaluated in Section 3.3, showing that its cost avoidance is 12% to 40% above a baseline approach that prioritizes claims by billed amount. Section 3.4 concludes with interpretation and visualization of the random forest model.

2. Upcoding Detection

Below, we present an algorithm for assigning anomaly scores to ER claims for the purpose of upcoding detection. The data under consideration consists of $n =$ 9,793 claims, indexed $i = 1, \ldots, n$. Each claim i has an ICD-10 diagnosis code d_i and a CPT severity code s_i taking integer values from 1 (least severe) to 5 (most severe) [10, 11].[1] Upcoding is the practice of systematically assigning high severity codes to medical diagnoses where they are unwarranted [12]. This suggests an upcoding detection strategy where each claim is compared to other claims with the same diagnosis code. If the claim being evaluated has an unusually high severity code compared to the others, it is flagged as anomalous.

2.1. **Upcoding Anomaly Scores.**

To formalize this approach, let (D, S) be a randomly selected pair of diagnosis/severity codes from the entire data set $\{(d_i, s_i) \mid i = 1, \ldots, n\}$. Given a claim (d_i, s_i), define the *background data* to be the set of all other claims $\{(d_j, s_j) \mid j \neq i\}$. Probability and expectation operators applied to the entire data set are denoted by P and E respectively, while the same operators applied to the background data are denoted $P_{(i)}$ and $E_{(i)}$. The *upcoding anomaly score* (UAS) for the ith claim is given by the conditional probability

[1] These are HCPCS codes 99281 through 99285.

$$\text{UAS}_i = P_{(i)}(S \geq s_i \mid D = d_i).$$

In other words, the upcoding anomaly score for claim i is computed by first identifying all claims in the background data with the same diagnosis code d_i and then computing the proportion of those claims whose severity code is greater than or equal to s_i. If s_i is an unusually high severity code for diagnosis d_i, this will be reflected by a low value of UAS_i (low values are more anomalous). The problem with this approach is high granularity of the diagnosis codes[2], potentially resulting in high variance estimates of the anomaly scores.

2.2. **Hierarchical Clustering.** This problem is solved by optimally reducing granularity of the diagnosis codes with a clustering algorithm. Hierarchical clustering is well suited to dimensionality reduction of categorical variables due to its flexibility [13]. While other clustering algorithms, such as k-means [14] and DBSCAN [15], make strict assumptions about the geometry or density of data clusters, hierarchical clustering can be applied to any data set where a dissimilarity metric is defined.

To define a dissimilarity metric on the set of diagnosis codes, let

$$\mu_{s|d} = E(S \mid D = d),$$

the average severity code for all claims with diagnosis code d. The dissimilarity metric between two diagnosis codes d_1 and d_2 is the distance between their corresponding average severity codes,

$$\Delta(d_1, d_2) = |\mu_{s|d_1} - \mu_{s|d_2}|.$$

Applying hierarchical clustering to this dissimilarity metric results in the dendrogram in Figure 1. Each of the 1,951 diagnosis codes are represented by points at the bottom of the dendrogram. Moving upwards, diagnosis codes that are close to each other with respect to the metric Δ are joined together into clusters, and these clusters are recursively joined to form even larger clusters, resulting in a hierarchy. Cutting the dendrogram at various heights determines the resulting number of clusters. For instance, cutting at height 15 would place diagnosis codes into only two clusters, causing underfitting and poor model performance [16]. Cutting at height zero places each diagnosis code in a separate cluster, causing overfitting, high variance estimates, and poor model performance [17]. Instead of choosing one of these extremes, we would like to select the optimal number of diagnosis code clusters k by maximizing the model performance metric described below.

[2] There are 1,951 diagnosis codes present in the data.

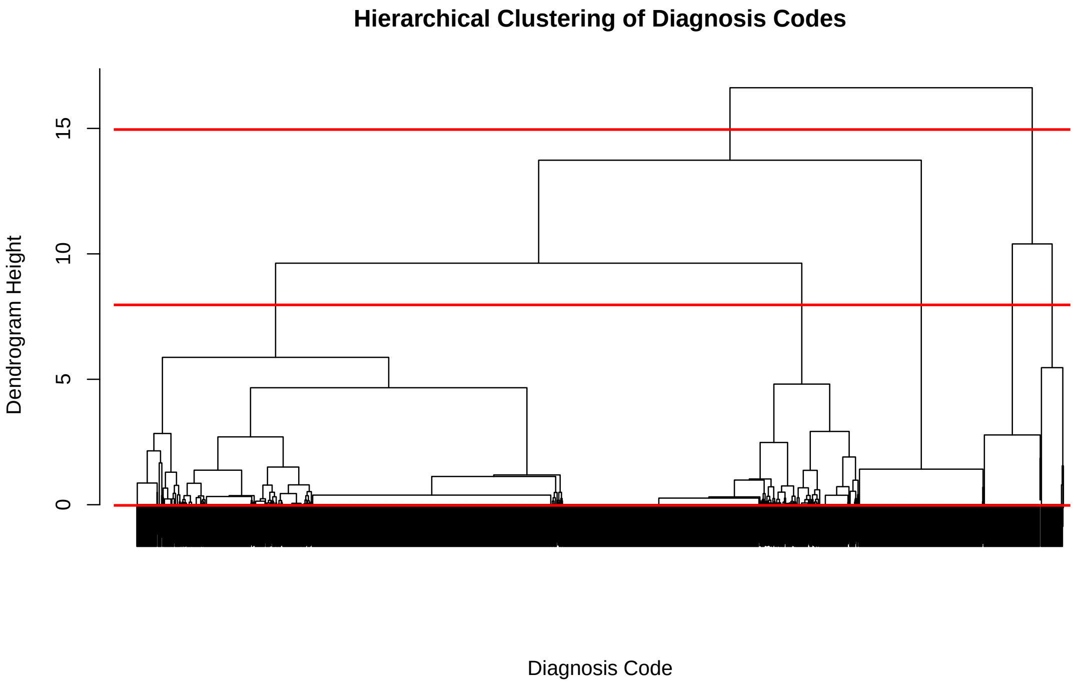

Figure 1. Relationship Between Cut Height and Number of Clusters. Cutting at height 15 produces two clusters, cutting at height 8 produces five clusters, and cutting at height 0 produces 1,951 clusters, with each diagnosis code in a separate cluster.

First, randomly divide the claims into two subsets A and B for the purpose of two-fold cross-validation.[3] Consider a dendrogram cut resulting in k clusters, let $d_i^{(k)}$ denote the ith claim's diagnosis code cluster, and randomly select $(D^{(k)}, S)$ from the training data.

If claim i is in the test data, then the probability that its severity code is equal to s is estimated by

(1) identifying all claims in the training data with the same diagnosis code cluster $d_i^{(k)}$ and
(2) computing the proportion of those claims with severity code s, that is,

$$\hat{p}_i(s) = P\left(S = s \mid D^{(k)} = d_i^{(k)}\right), \text{ for } s = 1, \ldots, 5.$$

In this manner, each claim i in the test data is assigned a vector of predicted probabilities $(\hat{p}_i(1), \ldots, \hat{p}_i(5))$, and these vectors are compared to the true severity codes s_i to compute

[3] Initially, A will serve as a training set to build a model and B will serve as a test set to compute model performance, and then the roles of A and B will be reversed. After both steps are complete, the model performance scores are averaged.

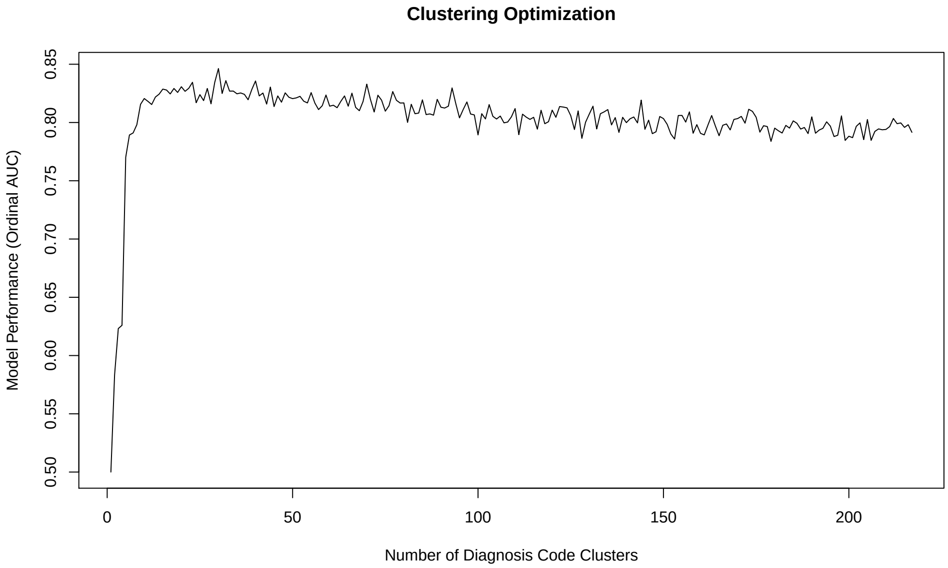

Figure 2. Ordinal AUC vs. Number of Diagnosis Code Clusters k.

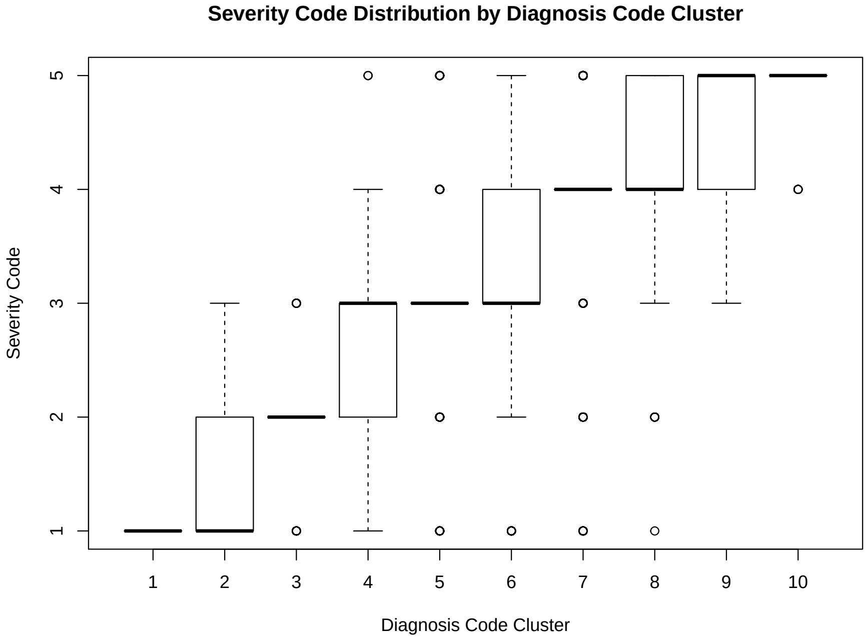

Figure 3. Box-plot of Severity Codes Stratified by Diagnosis Code Cluster.

model performance (ordinal AUC) [18]. Repeating this process for both folds A and B and averaging the results yields the ordinal AUC for a dendrogram cut resulting in k clusters.

Figure 2 shows that the maximum ordinal AUC (0.846) occurs at $k = 30$, however, this yields some clusters with samples sizes of six claims, which risks unstable model estimates. Decreasing the number of clusters to $k = 10$ achieves sample sizes greater than 30 for all

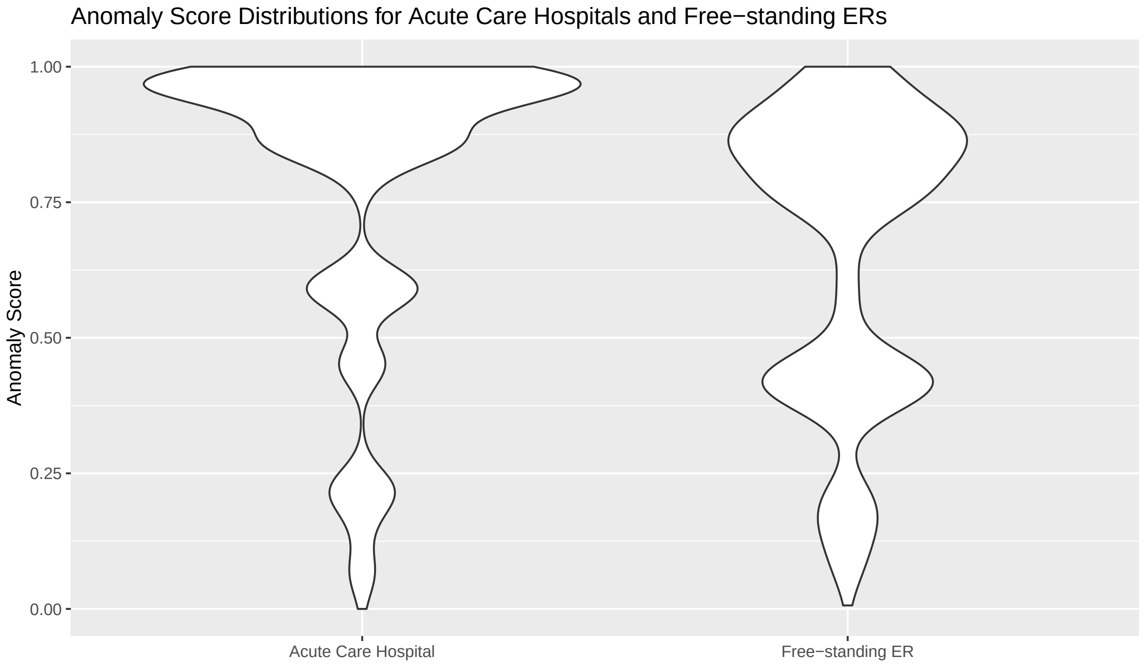

Figure 4. Upcoding Anomaly Score Distributions for Acute Care Hospitals and Free-standing ERs.

clusters while only decreasing ordinal AUC to 0.821. Figure 3 displays box plots of severity codes stratified by diagnosis code cluster, showing that this clustering effectively partitions diagnosis codes into different severity code profiles.

2.3. **Upcoding Anomaly Scores Revisited.** After optimally reducing granularity of the diagnosis codes, we are prepared to improve the upcoding anomaly scores. Previously, the anomaly score for claim i was based on claims in the background data with the same diagnosis code d_i. The key modification is to consider claims in the background data with the same *diagnosis code cluster* $d_i^{(k)}$:

$$\text{UAS}_i = P_{(i)}(S \geq s_i \mid D^{(k)} = d_i^{(k)}). \tag{2.1}$$

An additional modification is useful for stratifying the anomaly scores based on categorical risk factors. For a given risk factor X, the background data for claim i is $\{j \mid X_j \neq X_i\}$. For instance, to compare upcoding anomaly scores for acute care hospitals and free-standing ERs, the background data would be defined as follows:

- if claim i is from an acute care hospital, its background data consists of all claims from free-standing ERs
- if claim i is from a free-standing ER, its background data consists of all claims from acute care hospitals.

With this modification in place, Equation (2.1) is used to calculate upcoding anomaly scores, with the understanding that $P_{(i)}$ is applied to the modified background data. Keeping in mind that smaller values of UAS are more anomalous, Figure 4 shows a greater concentration of moderately anomalous claims (in the 0.25 to 0.5 range) at free-standing ERs. In fact, the mean UAS at acute care hospitals and free-standing ERs are 0.79 and 0.66, respectively, and this difference is highly statistically significant ($P < 10^{-74}$).

3. Cost Avoidance Optimization

The previous section described a method for detecting one specific type of health insurance anomaly, upcoding, on the basis of two predictor variables, diagnosis and severity codes. This section presents a more extensive model for detecting several types of health insurance anomalies within the unified framework of cost avoidance optimization. The data consist of $n =$ 20,086 ER claims that have been inspected by claims reviewers. The cost avoidance for a claim is the difference between its claim amounts before and after review:

$$\text{Cost Avoidance} = (\text{Prereview Claim Amount}) - (\text{Postreview Claim Amount})$$

Our objective is to predict cost avoidance *before* a claim is reviewed using the following explanatory variables, which are defined for the claim (header level) or for individual transaction lines on a claim (line level).

- **Billed Amount**
- **Number of Claim Lines**
- **ICD-10 Diagnosis Code**
- **ICD-10 Procedure Code**
- **HCPCS/CPT Code**
- **HCPCS/CPT Modifier Codes**
- **Facility Observational Units**
- **Product Type**
- **Network Status**. In/out of network.
- **Revenue Code**. Used on institutional claims to identify the type of service or procedure.
- **Acute/Ancillary Code**, e.g., acute care hospital, or free-standing ER.
- **Servicing Provider Type**, e.g., medical doctor, dentist, or optician.
- **Servicing Provider Specialty**, e.g., obstetrics, pediatrics, or anesthesiology.
- **Servicing Provider Key**. Identifies each provider for the purpose of estimating fixed effects at the provider level.
- **Place of Treatment Code.** Identifies the place of treatment where health care services were rendered.
- **Institutional Pricing Code**. Indicates if the service is contained within the facility allowed amount or is subject to financial carve-outs.
- **Reimbursement Method**, e.g., percent of charge or flat rate.

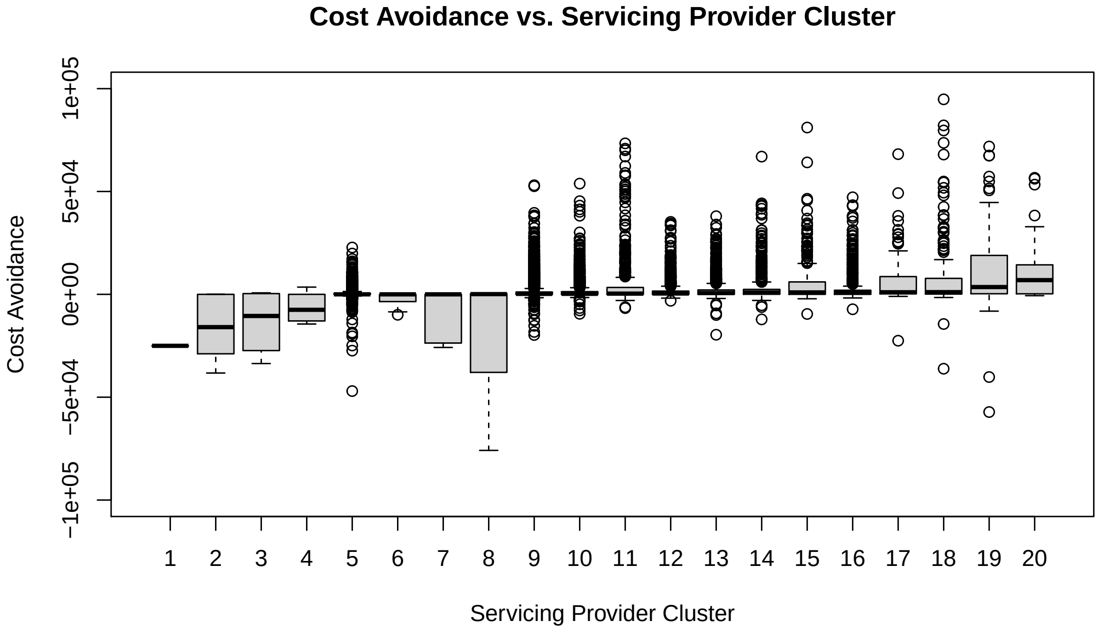

Figure 5. Box-plot of Cost Avoidance Stratified by Servicing Provider Cluster. Servicing providers in clusters 1 through 8 tend to understate ER claims, while clusters 9 through 20 tend to overstate them.

The model also bases its predictions on features that have been engineered to detect the following types of anomalies:

- Prescription Drug Abuse
- Doctor Shopping
- Unnecessary Unbundling
- Association Analysis Metrics
- Excessive CT Scans.

3.1. **Hierarchical Clustering for the Cost Avoidance Model.** As discussed in Section 2.2, hierarchical clustering is used to reduce dimensionality of the categorical predictor variables at the header and line levels. The main difference for the cost avoidance model is the measurement scale of the dependent variables. Cost avoidance is quantitative, while severity codes are ordinal.

Given a categorical variable D, e.g. diagnosis code, we first fit an ANOVA model for predicting cost avoidance in terms of D. For any two possible values d_1 and d_2 of D, let $T(d_1, d_2)$ denote the t-statistic for testing equality of their coefficients in the ANOVA model. The dissimilarity metric is then simply the absolute value of this t-statistic:

$$\Delta(d_1, d_2) = |T(d_1, d_2)|.$$

With this dissimilarity metric in place, hierarchical clustering proceeds as previously described. First, a dendrogram is constructed, and then the number of clusters is determined by optimizing model performance, the coefficient of determination R^2, with two-fold cross-validation. For example, the 361 servicing provider keys represented in the data are binned into only 20 servicing provider clusters, as displayed in Figure 5. The servicing provider clusters can be viewed as an analytic that groups providers into different risk categories. At this point, it should be mentioned that cost avoidance can be negative in cases where a claim review resulted in an increased claim payment, and machine learning can identify these claims to ensure that patients receive the claim they are entitled to. Figure 5 reveals that servicing providers in clusters 1 through 8 tend to understate ER claims, while clusters 9 through 20 tend to overstate them.

3.2. **Building the Cost Avoidance Model with Random Forests.** Once hierarchical clustering is complete, we are prepared to build the cost avoidance model. The data is $\{(\mathrm{CA}_i, \mathbf{x}_i) \mid i = 1, \ldots, n\}$, where CA_i is the ith claim's cost avoidance, and $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ is the vector of header/line data and engineered features for the ith claim. The goal is to build a model for predicting CA_i in terms of $\mathbf{x}_i$, and this is achieved in three steps: cost avoidance downscaling, random forest predictions, and cost avoidance upscaling.

(1) Define the cost avoidance ratio (CAR) of the ith claim to be the ratio of its cost avoidance to the billed amount:

$$\mathrm{CAR}_i = \frac{\mathrm{CA}_i}{(\text{Billed Amount})_i}.$$

Cost avoidance is always less than or equal to billed amount, i.e., $\mathrm{CAR}_i \leq 1$, and the predictive model needs to preserve this feature.

(2) A random forest model $\widehat{\mathrm{CAR}}_i = \mathrm{RF}(\mathbf{x}_i)$ is then built for predicting CAR_i using the explanatory variables $\mathbf{x}_i$.

(3) The random forest predictions are then upscaled to produce cost avoidance predictions:

$$\widehat{\mathrm{CA}}_i = (\widehat{\mathrm{CAR}}_i)(\text{Billed Amount})_i.$$

The random forest model described in step (2) is built as an ensemble of many decision trees. Decision trees are essentially flow charts, built by recursively partitioning the training data $(\mathrm{CAR}_i, \mathbf{x}_i)$, $i = 1, \ldots, n$, based on the values of the independent variables. When the independent variables x_j are quantitative, the splits can be binary, as displayed in Figure 6, or multiway splits can be used by binning the values of x_j. Any given split partitions the data into several nodes, and the *impurity* of each node is defined to be the sum of square deviations from the mean for the values of CAR in that node. Decision tree algorithms, such as AID and CART [19, 20], generate splits which minimize the sum of the impurity measures for the resulting nodes until a stopping criterion is met. When presented with a new test case $\mathbf{x}$, the decision tree classifies it into one of the leaf nodes, and the estimate $\widehat{\mathrm{CAR}} = f(\mathbf{x})$ is determined by fitting a linear model to that node.

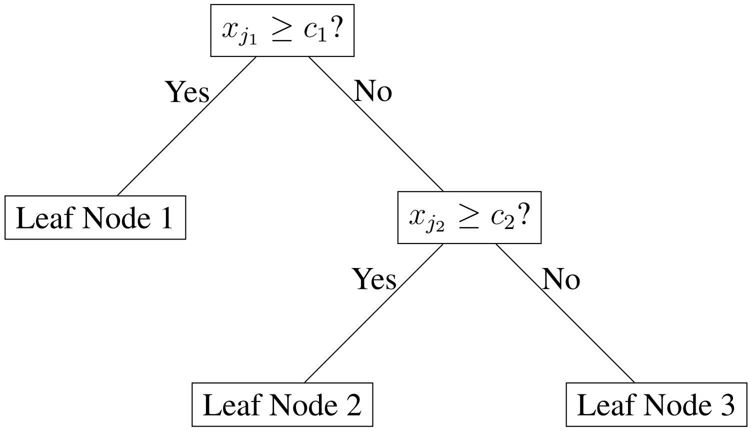

Figure 6. Example of a Decision Tree with Two Binary Splits.

Random forests [21] are constructed by first generating a sequence of independent and identically distributed random vectors $\Theta_1, \ldots, \Theta_K$, which are used to build K decision trees, $f_k = f(\cdot, \Theta_k)$. The random forest then produces estimates for the cost avoidance ratio by averaging the estimates of these decision trees,

$$\widehat{\text{CAR}} = \text{RF}(\mathbf{x}) = \frac{1}{K}\sum_{k=1}^{K} f(\mathbf{x}, \Theta_k).$$

The randomness from the vectors $\Theta_1, \ldots, \Theta_K$ can influence growth of the trees in several ways. In bootstrap aggregating (bagging), the training data for each tree is a random sample with replacement from the original training data. Other ways to introduce randomness include randomizing each tree's input features x_j, or randomly choosing each decision tree split from a set of top performing splits.

As the number of trees tends to infinity, the mean square error of a random forest converges to $E_{y,\mathbf{x}}(y - E_\Theta f(\mathbf{x}, \Theta))^2$, where $y = \text{CAR}$, and $E_{y,\mathbf{x}}$ and E_Θ are expectations with respect to the probability distributions of $(y, \mathbf{x})$ and Θ, respectively [21]. Therefore, there is no danger of overfitting when increasing the number of trees in a random forest, but one should ensure that a sufficient number of trees are being used to achieve high performance.

3.3. **Economic Performance of the Cost Avoidance Model.** As discussed in the introduction, reviewing all claims is not feasible, but reviewing too few risks overlooking a large number of fraudulent claims and improper payments. To solve this problem, claims in review queues should be sorted by predicted cost avoidance $\widehat{\text{CA}}$, so that high priority claims are reviewed first.

The red curve in Figure 7 displays the cumulative cost avoidance achieved by the predictive model vs. the percentage of claims reviewed. Initially, as highly anomalous claims are reviewed, cost avoidance rises rather steeply, and then it levels off, eventually decreasing once negative value claims are encountered.

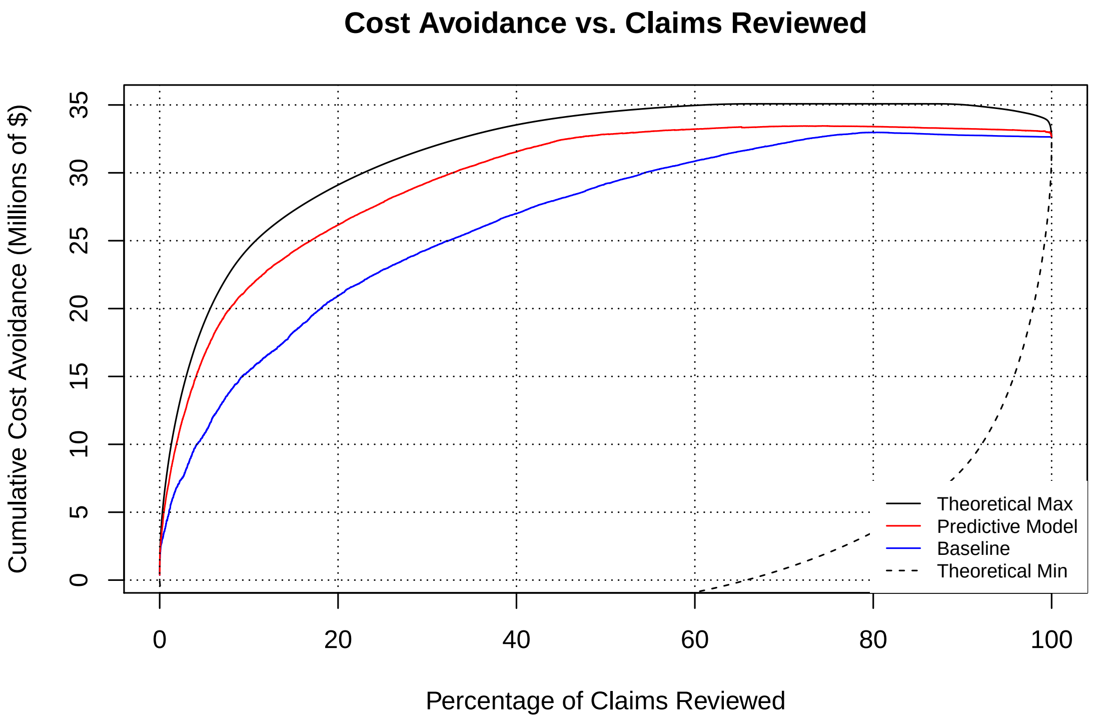

Figure 7. Cumulative Cost Avoidance as a Function of Percentage of Claims Reviewed.

Reviewed (%)	Baseline	Model	Improvement Over Baseline	
10	$15.4M	$21.6M	$6.2M	40%
20	21.0	26.2	5.2	25%
30	24.4	29.3	4.9	20%
40	27.0	31.5	4.5	17%
50	29.2	32.8	3.6	12%

Table 1. Comparison of the Cost Avoidance Model to Baseline Performance.

Three other scenarios are also presented to make comparisons.

(1) **Baseline.** Claims are sorted by billed amount.
(2) **Theoretical Maximum.** Claims are sorted by a hypothetical model with perfect accuracy.
(3) **Theoretical Minimum.** Claims are sorted by a hypothetical model that deliberately prioritizes low-value claims.

These curves reveal that the cost avoidance model approaches the theoretical maximum and significantly exceeds baseline performance. Optimizing savings requires accounting for the cost of claims reviews, and ideally most savings can be achieved by reviewing

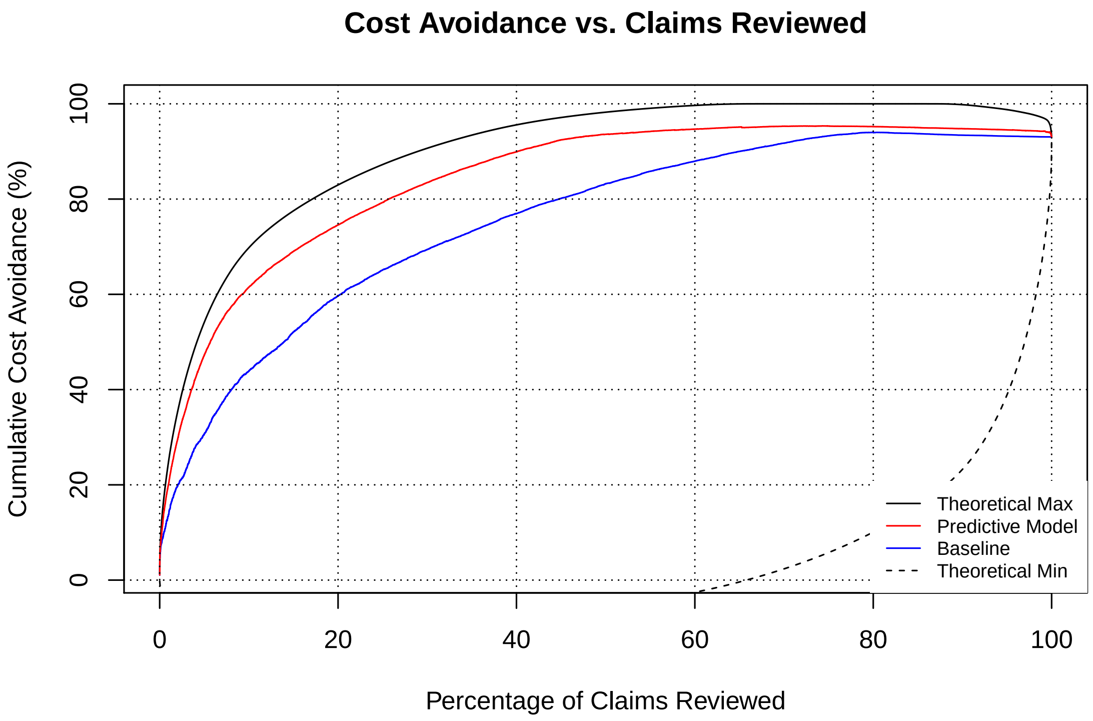

Figure 8. Cumulative Cost Avoidance, Expressed as a Percentage of Potential Savings.

50% of claims or less. In this range, the cost avoidance model saves 12% to 40% over baseline (see Table 1). Figure 8 demonstrates that the model achieves 94% of potential savings when 50% of claims are reviewed, bordering the theoretical maximum of 98% and substantially exceeding the baseline of 83%. These performance statistics are based on the previously mentioned sample of n = 20,086 ER claims. To put these savings into perspective, note that there were over 144 million ER visits in the United States in 2016, and approximately 132 million (92%) of these visits were by insured patients [22, 23]. If the statistical distribution of relevant variables for national claims are approximately equal to those in this sample, adopting a cost avoidance optimization model nationally could save $24B to $41B over baseline annually. This is a coarse approximation, but it suggests the potential savings are likely in the tens of billions of dollars per year.

3.4. **Model Interpretation and Visualization.** Sections 3.2 and 3.3 demonstrated how to build a cost avoidance model and evaluate its overall performance. We now turn our attention to ranking the relative importance of the predictor variables and visualizing them graphically. Random forests achieve this objective with two variable importance metrics, increased mean square error (MSE) and increased node purity [21]. The first of these measures the difference in MSE caused by permuting the data for each predictor x_j, normalized by the standard deviation of these differences. The second metric is the total decrease in

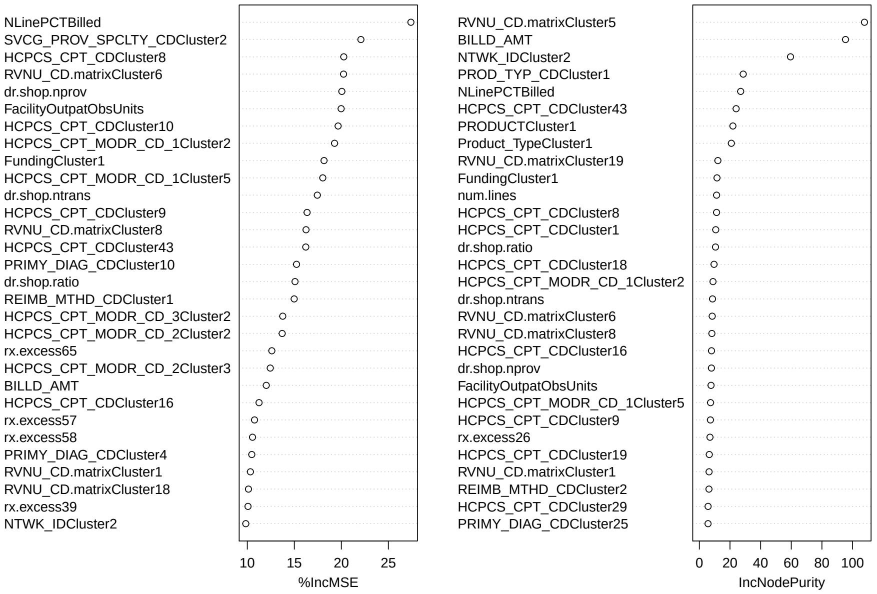

Figure 9. Variable Importance Plots for the Cost Avoidance Model.

node impurities from splitting on the variable, averaged over all trees in the random forest. The variable importance plots for the cost avoidance model are displayed in Figure 9.

The most important variable, as measured by increased MSE, is the percentage of N-lines on a claim, referring to an institutional pricing code for transactions that are subject to financial carve-outs in a contract. Unsurprisingly, claims with a disproportionately high percentage of N-lines are more likely to be anomalous, as confirmed by the heat-map in Figure 10.

The hierarchical clusters developed for categorical variables (e.g., HCPCS, Revenue, and Diagnosis Codes) appear prominently in the variable importance plots, as do churn metrics, such as facility observational units, unnecessary unbundling, and doctor shopping. Another notable predictor is prescription drug abuse, specifically for the following General Product Identifier (GPI) codes:

- Opioids (GPI 65)
- Antianxiety Agents (GPI 57)
- Antidepressants (GPI 58)
- Antihyperlipidemics (GPI 39)
- Progestins (GPI 26)

4. Acknowledgments

We would like to thank Blue Cross and Blue Shield of Texas for funding this research under the Affordability Cures^{SM} Initiative.

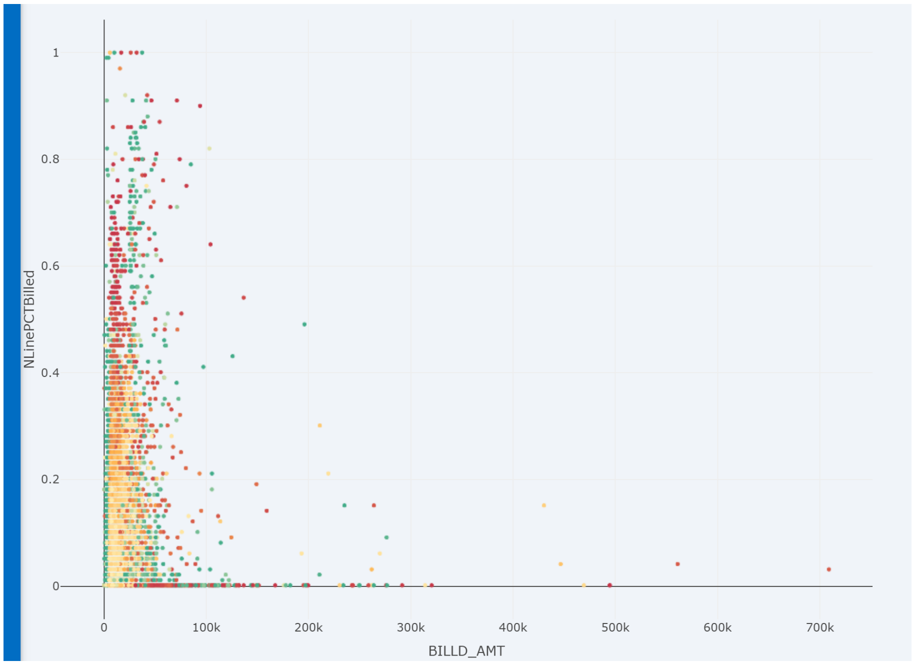

Figure 10. Anomaly Score Heat Map. Claims with a higher percentage of financial carve-outs (N-lines) are more anomalous on average.

References

[1] K.M. King. Medicare fraud: Progress made, but more action needed to address medicare fraud, waste, and abuse, 2014. https://www.gao.gov/products/gao-14-560t.

[2] Ngufor Che and Wojtusiak Janusz. Unsupervised labeling of data for supervised learning and its application to medical claims prediction. *Computer Science*, 14(3):191, 2013.

[3] Richard Bauder, Taghi M. Khoshgoftaar, and Naeem Seliya. A survey on the state of healthcare upcoding fraud analysis and detection. *Health Services and Outcomes Research Methodology*, 17(1):31–55, July 2016.

[4] Hongxing He, Jincheng Wang, Warwick Graco, and Simon Hawkins. Application of neural networks to detection of medical fraud. *Expert Systems with Applications*, 13(4):329–336, 1997.

[5] Marjorie A. Rosenberg, Dennis G. Fryback, and David A. Katz. A statistical model to detect DRG upcoding. *Health Services and Outcomes Research Methodology*, 1(3/4):233–252, 2000.

[6] Kenji Yamanishi, Jun-Ichi Takeuchi, Graham Williams, and Peter Milne. On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms. *Data Mining and Knowledge Discovery*, 8(3):275–300, 2004.

[7] Wei Luo and Marcus Gallagher. Unsupervised DRG upcoding detection in healthcare databases. *2010 IEEE International Conference on Data Mining Workshops*, 2010.

[8] Dallas Thornton, Roland M. Mueller, Paulus Schoutsen, and Jos van Hillegersberg. Predicting healthcare fraud in medicaid: A multidimensional data model and analysis techniques for fraud detection. *Procedia Technology*, 9:1252–1264, 2013. CENTERIS 2013 - Conference on ENTERprise Information Systems / ProjMAN 2013 - International Conference on Project MANagement/ HCIST 2013 - International Conference on Health and Social Care Information Systems and Technologies.

[9] Marina Evrim Johnson and Nagen Nagarur. Multi-stage methodology to detect health insurance claim fraud. *Health Care Management Science*, 19(3):249–260, 2015.

[10] Medicare coverage general information ICD-10. *Centers for Medicare and Medicaid Services*, 2021. https://www.cms.gov/Medicare/Coverage/CoverageGenInfo/ICD10.

[11] HCPCS - general information. *Centers for Medicare and Medicaid Services*, 2021. https://www.cms.gov/Medicare/Coding/MedHCPCSGenInfo.

[12] Richard Bauder, Taghi M. Khoshgoftaar, and Naeem Seliya. A survey on the state of healthcare upcoding fraud analysis and detection. *Health Services and Outcomes Research Methodology*, 17:31–55, 2017.

[13] Zdenek Sulc and Hana Řezanková. Dimensionality reduction of categorical data: Comparison of HCA and CATPCA approaches. 2015.

[14] Edward W. Forgy. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics*, 21(3):768–769, 1965.

[15] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. pages 226–231. AAAI Press, 1996.

[16] Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, and Vipin Kumar. *Introduction to Data Mining, 2nd ed.* Pearson, 2018.

[17] David A. Freedman. *Statistical Models, Theory and Practice, Revised ed.* Cambridge University Press, 2009.

[18] Pelin Yıldırım, Ulaş K. Birant, and Derya Birant. EBOC: Ensemble-based ordinal classification in transportation. *Journal of Advanced Transportation*, 2019.

[19] James Morgan and John Sonquist. Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association*, 58(302):415–434, 1963.

[20] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. *Classification And Regression Trees*. Routledge, 1984.

[21] Leo Breiman. Random forests. *Machine Learning*, 45(1), 2001.

[22] Trends in emergency department visits. *Healthcare Cost and Utilization Project. Agency for Healthcare Research and Quality*, 2021.

[23] Adam J. Singer, Jr Thode, Henry C., and Jesse M. Pines. US Emergency Department Visits and Hospital Discharges Among Uninsured Patients Before and After Implementation of the Affordable Care Act. *JAMA Network Open*, 2(4), 2019.