Big Data
Volume 6 Number 1, 2018

DOI: 10.1089/big.2017.0083

ORIGINAL ARTICLE

FraudBuster: Reducing Fraud in an Auto Insurance Market

Saurabh Nagrecha,* Reid A. Johnson, and Nitesh V. Chawla

Abstract

Nonstandard insurers suffer from a peculiar variant of fraud wherein an overwhelming majority of claims have the semblance of fraud. We show that state-of-the-art fraud detection performs poorly when deployed at underwriting. Our proposed framework "FraudBuster" represents a new paradigm in predicting segments of fraud at underwriting in an interpretable and regulation compliant manner. We show that the most actionable and generalizable profile of fraud is represented by market segments with high confidence of fraud and high loss ratio. We show how these segments can be reported in terms of their constituent policy traits, expected loss ratios, support, and confidence of fraud. Overall, our predictive models successfully identify fraud with an area under the precision–recall curve of 0.63 and an f-1 score of 0.769.

Keywords: insurance fraud; cost-sensitive techniques; market segmentation; loss ratio; regulation compliance

Introduction

Insurance fraud is a perennial problem in the United States and throughout the world. Its low-risk, high-reward nature attracts criminals and is a financial burden to insurers and policyholders alike, costing U.S. auto insurers $7.7 billion in excess payments for the year 2012.[1] The risk of fraud is particularly high for nonstandard insurance companies, which provide insurance to individuals who fall outside of the "low-risk" category. These are drivers who pose a high risk to the insurer due to multiple possible reasons, for example, drivers with multiple accidents, prior convictions, or those who prefer to carry the state minimum insurance coverage. Although an estimated 5% of personal injury claims in the United States are fraudulent,[2] nonstandard insurers in certain states see fraud rates as high as 84% of their personal injury claims.[3] Ultimately, these high rates of fraud and abuse manifest as increased insurance premiums that are passed on to all policyholders.[4] The problem of high fraud rates raises several important challenges for cost-sensitive and highly regulated insurance markets.

The high rate of fraud necessitates a solution that goes beyond classifying claims as fraudulent/nonfraudulent, instead proactively predicting bad risks at the underwriting stage. The lack of actionability at the claims stage forces us to look for more proactive signals of fraud, that is, before a claim even occurs.* This change in prediction timeline alters the very problem statement when fighting fraud. In fraud detection, we classify a small, fully labeled subset of the overall policy data, that is, the 7% who filed personal injury claims. Underwriting data consists of a superset of all policies, where only a small fraction of the true labels are known. Figure 1a pictorially depicts the difference between claims and underwriting data. Fraud detection only handles personal injury claimants (fraud and not fraud), underwriting data expand this concept to include all policies that did not result in personal injury claims.

We address the combination of these challenges in our proposed framework "FraudBuster" for insurers in high fraud rate markets. Specific contributions of this framework are as follows:

- Restructuring the problem definition of insurance fraud: Since conventional fraud detection techniques are inapplicable to a fraud-majority market, we pivot our focus toward identifying the worst affected segments of the market. We outline

*For a more detailed description of insurance stages, refer to Supplementary Data Section 1 (Supplementary Data are available online at www.liebertpub.com/big).

Department of Computer Science and Engineering, iCeNSA, University of Notre Dame, Notre Dame, Indiana.

*Address correspondence to: *Saurabh Nagrecha, Department of Computer Science and Engineering, iCeNSA, University of Notre Dame, Notre Dame, IN 46556,* E-mail: nchawla@nd.edu

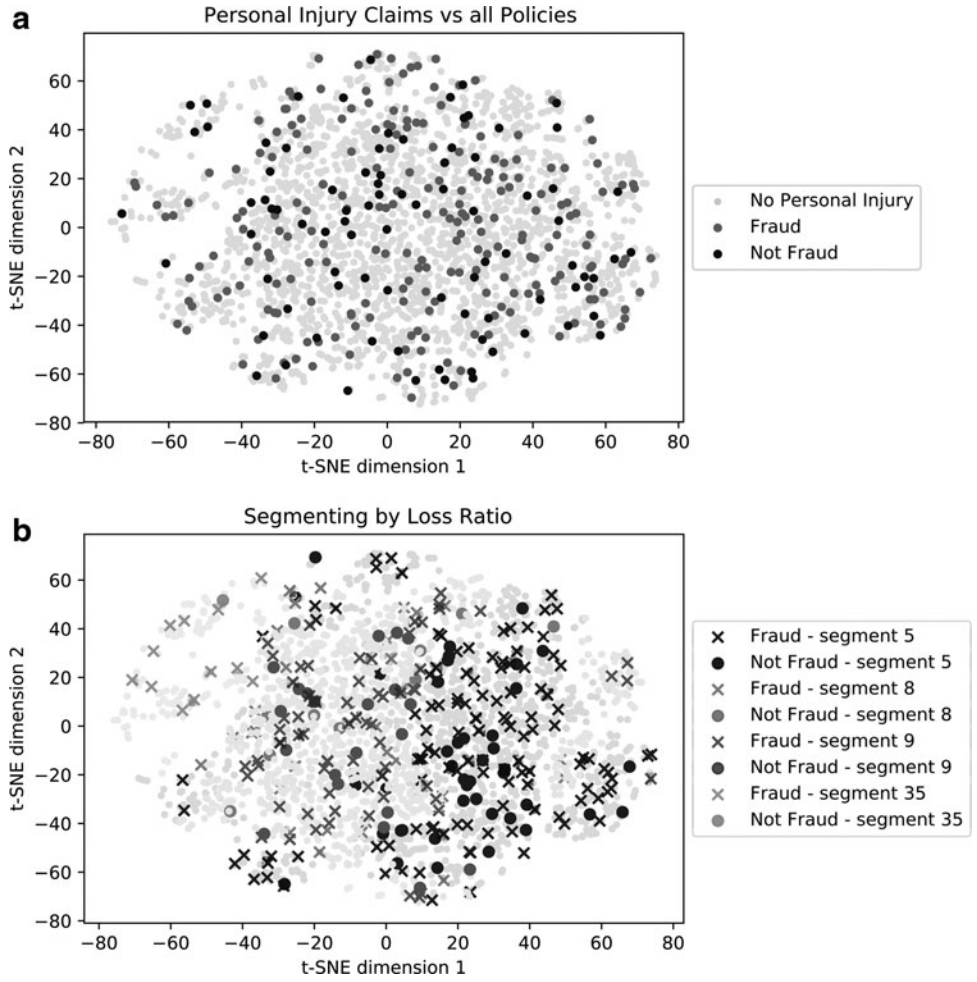

FIG. 1. Where naive fraud detection fails, only a small fraction of instances in the training and testing data set have observed labels **(**6.9% of policies as per **a)**—this makes it impossible to use fraud detection at a policy level. Instead, computing loss ratios for buckets of instances makes it possible to identify heterogeneity in confidence of fraud and profitability for *all policies* **(b)**.

the associated design decision tradeoffs and deployment in Proposed Technique: FraudBuster section.

- Identifying *actionable* fraud: We combine traditional machine learning metrics such as area under the precision–recall curve (AUPR) and f-1 scores with actuarial metrics of profitability like the loss ratio to create a hybrid definition of *actionable* fraud. In Experiments section, we experimentally show that this definition is more generalizable than fraud individually defined by (A) high likelihood of fraud and (B) high loss ratios.
- Ensuring end-to-end compliance with regulations: We factor in industry regulations at each stage of the predictive pipeline. We use prepruned decision trees (DTs) in our underlying classification model and show that this interpretability does not come at the cost of predictive performance (as shown in Table 1). Our choice of classifier also makes it possible to express individual market segments in terms of their support, confidence, and loss ratio profitability (as shown in Table 4).

FraudBuster addresses several key aspects of business viability of an applied machine learning solution—business value, interpretability, and compliance. By reconciling predicted cost-sensitive classification segments

Table 1. Predictive performance on test data using various classification pipelines

Classifier	Resampling	AUPR	f-1 score	Precision		Recall	
				Class 0	Class 1	Class 0	Class 1
DT	—	0.176	0.012	0.829	0.517	0.999	0.006
	RUS	0.182	0.287	0.848	0.188	0.452	0.609
	ROS	0.530	0.698	0.970	0.583	0.871	0.869
SGD	—	0.173	0.001	0.828	0.350	1	0.001
	RUS	0.172	0	0.828	0	1	0
	ROS	0.172	0.178	0.828	0.168	0.796	0.203
NB	—	0.210	0.226	0.849	0.354	0.912	0.209
	RUS	0.211	0.232	0.848	0.342	0.910	0.209
	ROS	0.209	0.248	0.853	0.306	0.878	0.257
RF	—	0.191	0.043	0.831	0.593	0.999	0.023
	RUS	0.313	0.485	0.928	0.377	0.736	0.715
	ROS	**0.631**	0.769	0.968	0.720	0.922	0.848
GBT	—	0.172	0	0.828	0	1	0
	RUS	0.170	0.222	0.823	0.165	0.636	0.345
	ROS	0.488	0.661	0.964	0.545	0.852	0.845

Bold and underline indicate best performing model in terms of AUPR and f-1 score, respectively.

Interpretable DTs are the closest second to the highest performing RF model. Classes 1 and 0 here refer to the respective FraudBuster labels of "actionably fraudulent" and "not actionably fraudulent," respectively.

AUPR, area under the precision–recall curve; DT, decision trees; GBT, Gradient Boosted Trees; NB, Naive Bayes; RF, Random Forests; ROS, Random Oversampling; RUS, Random Undersampling; SGD, Stochastic Gradient Descent.

to actuarial metrics, the proposed framework demonstrates its bottom-line impact in eliminating bad risks. This is given in Table 4, where the loss ratio for each predictive segment is used as a measure of bottom-line profitability from that segment. FraudBuster's reliance on interpretable predictive pipelines makes it open to scrutiny by business analysts and simultaneously fuels its regulation compliance. Through this article, we provide a template for insurers to tackle fraud as and when policies are underwritten as opposed to retroactively classifying incoming claims.

The rest of this article is structured as follows—in Related Work section, we discuss how fraud research in automobile insurance is situated at the convergence of actuarial, legal, and machine learning literature. We then introduce the reader to insurance data in Background section, outlining which features are collected and when true labels are identified. It is here that we identify our challenge—predicting unprofitable fraud at underwriting instead of claims. We present our solution framework "FraudBuster" to address this challenge in Proposed Technique: FraudBuster section. We then validate our experimental assumptions and demonstrate the generalizability of our hybrid definition of actionable fraud in Experiments section. In Discussion section, we analyze the market segments that FraudBuster identifies and comment on how insurers can act upon these segments to improve their bottom-line. Finally in Conclusion section, we revisit FraudBuster's contributions toward the major challenges already outlined.

Related Work

FraudBuster addresses what is traditionally framed as an actuarial problem, applying principles drawn from the field of data science to develop a feasible and implementable solution. Our related work thus includes literature drawn from the insurance, actuarial, machine learning, and data science domains.

We borrow our definitions of insurance fraud from Derrig,[5] which provides a general account of what actions constitute modern insurance fraud and describes the many ways in which the insurance industry is vulnerable to fraud. Another article by the same author also provides a thorough description of how fraud manifests specifically in the automobile insurance sector.[2] Hoyt[6] studied the economic impact of fraud, whereas Schiller[7] investigated the cost of measures used to combat fraud. Specific to the states with no-fault personal injury protection (PIP) laws, Delegal and Pittman[3] discuss its legal and economic ramifications, examining the effectiveness of these laws in preventing PIP fraud.

Insurance fraud has typically been approached as a supervised classification problem. Of the early attempts to detect fraud in automobile insurance, Viaene et al.[8] present detailed experimental results that compare several standard binary classification techniques used to detect insurance fraud claims. From these, we use Naive Bayes (NB), DTs, and linear models as classification techniques in this article. Belhadji et al.[9] propose a procedure to identify and select the most significant features by measuring their sensitivity and specificity in detecting fraudulent claims. Artís et al.[10] used a modified logit-based model to demonstrate the predictability and interpretability of binary discrete choice models for predicting auto insurance fraud. Caudill et al.[11] extend this model in the presence of missing information. Throughout these studies, a resonant emphasis on interpretable models has been driven by regulatory pressure and motivation to implement these models as business rules. Although "interpretability" of classifiers is a topic of major debate,[12] regulatory authorities in the United States recognize models that can be phrased as lists of rules. Prime examples of these include falling rule lists[13] DTs,[14] and linear models.[15] Examples of noninterpretable classifiers include ensembles,[16,17] which combine the predictions of multiple classifiers.

Custom cost-based metrics have been used throughout the years to imbue classifiers with the ability to quantify dollar values of predictions. Fawcett and Provost[18] and Phua et al.,[19] respectively, use explicit costs as a performance metric to develop models for fraud classification. Fraud is an imbalanced-class problem, for which naive metrics like accuracy do not capture the nuance between models. The area under the receiver operating characteristic curves has been used as a metric,[20] along with the more nuanced AUPR[21] and f-1 score,[22] which we use in this article. In addition to choice of metrics, the class imbalance in fraud problems also merits resampling of the data.[19] The article by Phua uses Random Undersampling (RUS), Random Oversampling (ROS), and Synthetic Minority Oversampling Technique (SMOTE) for resampling. RUS randomly undersamples the majority class and ROS randomly oversamples the majority class instances to alleviate the class imbalance in two separate ways. SMOTE creates artificial minority class instances in the neighborhood of the existing minority class instances; however, this is unsuitable in our case wherein we have small disjuncts in our data (Fig. 1). In this article, we restrict our resampling to RUS and ROS, while noting that SMOTE can be a valuable tool for other use cases.

In addition to these classification metrics, we also make use of the actuarial profitability metric "loss ratio",[23] which assesses the health of an insurer's market segments. In this article, we combine the loss ratio with AUPR to determine drivers who are both *unprofitable* and likely to be fraudulent.

Background

In this section, we cast the problem of conventional fraud detection in terms of machine learning terminology and demonstrate its limitations in a fraud-majority insurance market. Instead, we propose the use of an actuarial metric for the impact of fraud (loss ratio), to use in combination with likelihood of fraud. It is through the amalgamation of these characteristics that we define "actionable" fraud, that is, sections of the market that exhibit significantly high amounts of fraud and are insufficiently priced for the level of claims originating from them. These metrics and definitions serve as the building blocks for FraudBuster's design in Proposed Technique: FraudBuster section.

Data description

The data for this study come from a nonstandard insurer in a no-fault PIP law state in the United States.

Table 2. Overview of underwriting features used

Family	*Type*	*Description*
Driver	Mixed	Gender, age, marital status, credit score, points
Vehicle	Mixed	Make, model, year, lienholder status, and number of vehicles insured
Location	Categorical	County level granularity
Coverage	Categorical	Lines of insurance coverage purchased
Discounts	Categorical	Prior coverage, vehicle safety equipment, home ownership, etc

Each feature corresponds to potentially risk-intensifying/mitigating factors for the insurer.

Overall, the data represent information on 1,037,914 drivers, each characterized by 44 descriptors or features. Since auto insurance policies in the United States are typically underwritten for a period of 6 months, we collected data on all policies in two phases (I and II), each covering a 6-month cycle. Two distinct sets of features are collected during each phase: (1) underwriting data, collected at the beginning for all policies, and (2) claims data, collected as and when claims are filed throughout the 6-month duration of each phase. Out of all the policies that are underwritten, only 7% file for PIP claims, which is typical of most insurers in the United States. Although the rate of fraudulent policies for most insurers is 5%,[2] nonstandard auto insurers are inundated with a staggering 84% fraud—that is, >8 out of every 10 PIP claims are fraudulent.

In our data, each of the ≈ 1 M policies represents an instance. Each of these policies has associated underwriting feature data (X_{UW}), as per Table 2. For the 7% policies that file for PIP claims, we additionally have claims feature data (X_{Claim}) and true labels of fraud (0) and not fraud (1).[†] For the other 83%, we use the label "no PIP claim" because we do not know whether these policies would have committed fraud, had they developed a PIP claim.

Fraud detection: underwriting versus claims

The predominant approach to combating auto insurance fraud is the deployment of a fraud detection model at the claims stage. In terms of the mentioned formalism, fraud detection trains a classifier on $(X_{\mathrm{UW}}, X_{\mathrm{Claim}})$ to predict $Y_{\mathrm{PIP|Claim}}$. This model achieves an almost perfect classification performance, meriting further inspection. We find that

[†]It should be noted that fraud is the majority class here and as a result, the labeling conventions are the reverse of typical fraud detection problems.

certain Current Procedural Terminology (CPT)[‡] codes are almost always (ab)used in fraudulent claims and are highly indicative of fraud. Although this is not a case of data leakage, it identifies CPT codes as an extremely strong signal of fraud.

Although this classifier seems to work perfectly when claims are made, its utility is limited by practical concerns in a fraud-majority market. Each predicted fraudulent claim requires costly investigative follow-up. Instead of reactively denying claims, we would like to proactively identify market segments that are probably entrenched in such fraudulent behavior.

This is tantamount to predicting Y_{PIP} at the underwriting stage, wherein we do not have "prescient" knowledge of X_{Claim}. The unique challenge here is that the likelihood of a given policy developing a claim is entirely random in actuarial science. Staged accidents are the only notable exception to this randomness, and they constitute <0.5% of the total number of claims.[2] The introduction of this random event into the prediction pipeline dramatically reduces the probability of fraud assigned to any given PIP claim, $\mathbb{P}(Y_{\text{PIP}} = \text{Fraud} | Y_{\text{Claim}} = 1, X_{\text{UW}})$, by a factor of 15. This *a priori* scaling of $\mathbb{P}(Y_{\text{PIP}})$, combined with the small random subset of underwritten policies (6.9%) with actual observable labels for $Y_{\text{PIP}|\text{Claim}}$, makes it impossible to directly apply existing fraud detection techniques to our problem.

Proposed Technique: FraudBuster

Through FraudBuster, we have built a framework that can identify actionably fraudulent segments *across* insurance cycles. As a result, it operates in two distinct phases to mimic the cadence followed by insurance cycles. In Phase 1 (Fig. 2), we train an interpretable model to identify actionably fraudulent drivers, which we then apply to drivers in Phase 2 (Fig. 3). At the end of Phase 2, we evaluate our predictions against observed ground truth for all PIP claimants. When deployed, FraudBuster returns segments of the market that are actionably fraudulent and *remain* to be so across insurance cycles.

Loss ratio: a litmus test for profitability

Given the limitations of using likelihood of fraud alone for a policy (as used in fraud detection), our problem requires metrics that capture the bottom-line impact of fraud. Such a metric should encompass *all* of the underwritten policies and their costs (losses claimed) and benefits (premiums earned). In actuarial science, this profitability is measured using the loss ratio. The loss ratio is simply defined as the amount of total losses to claims divided by the total earned premiums.[§] For a given population of policies S_i, the loss ratio is defined as $LR_i = \frac{Loss_i}{EP_i}$.

Here, $Loss_i$ is the total amount of loss incurred by PIP claims in segment of policies S_i, and EP_i is the total number of premiums earned over all policies in segment S_i. The loss ratio, therefore, provides insight into the profitability of a given set of policies. In the insurance domain, this segment-wise loss ratio quantifies the risk presented by that segment and is thus heterogeneous. Each of these segments' loss ratios needs to be considered in the context of the overall baseline loss ratio ($\overline{LR}$), which is computed across all underwritten policies. This baseline loss ratio is a key business metric for insurers, and is often considered an indicator of the overall "health" of the pool of insured policies—a higher loss ratio is worse for the insurer. That is, if a given segment's loss ratio is significantly higher than the baseline, it means that the cost of claims filed by policyholders in this segment significantly outweighs the amount of premium that they are being charged.

Observed loss ratio, by definition, is an aggregate over multiple policies in the training set. However, the *expected loss ratio* for a policy given their underwriting data makes it possible to assign a profitability value at an individual level. This expected loss ratio places a given policy in a bucket with policies that exhibit similar levels of profitability to the insurer.

"Actionably" fraudulent

Although our predictive models segment insurance policies based on their propensity of fraud, loss ratios help segment them by profitability. Considered together, these two metrics (likelihood of fraud and loss ratio) divide the underwritten policies into the following segments. The most important segment to be noted here is the *high confidence of fraud and high loss ratio*, which we show to be actionable.

Insurers may have already priced segments with high likelihood of fraud proportionate to their risk of fraud, reflected by acceptable loss ratios. Conversely, nonfraudulent segments in need of legitimate nonfraudulent treatment often fall under the high-loss ratio category.

[‡]A medical code set maintained by the American Medical Association through the CPT Editorial Panel; https://web.archive.org/web/20160511115308; http://www.ama-assn.org/ama/pub/physician-resources/solutions-managing-your-practice/coding-billing-insurance/cpt/cpt-process-faq/code-becomes-cpt.page

[§]In this article, we do not consider the reserves and expenses of the insurer in addition to the definition of loss ratio mentioned. These may be interesting to consider in future work.

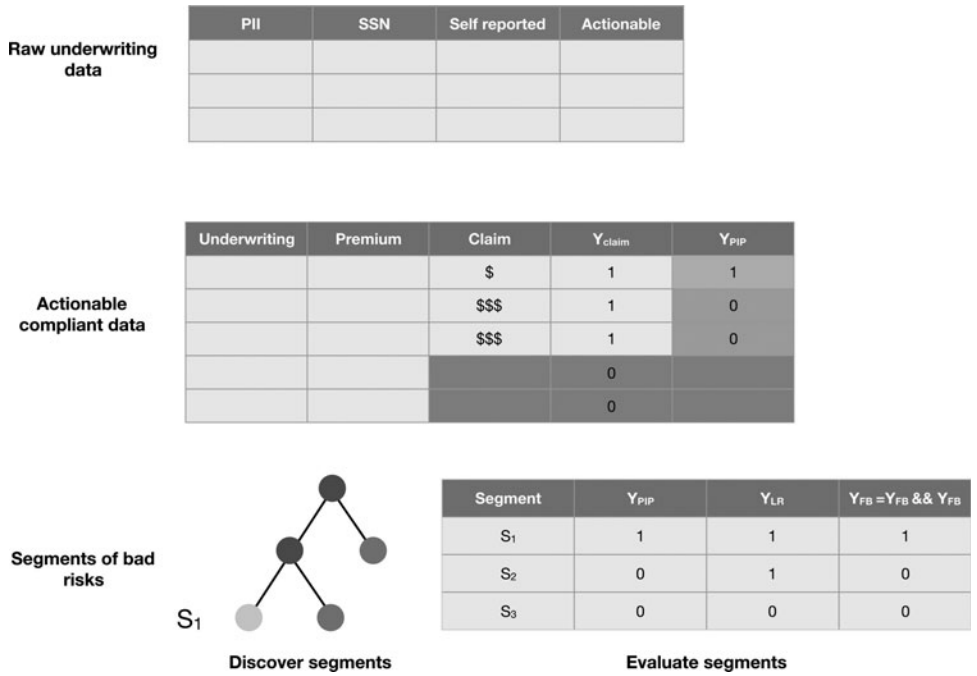

FIG. 2. Overview of FraudBuster's Phase 1. LR, loss ratio.

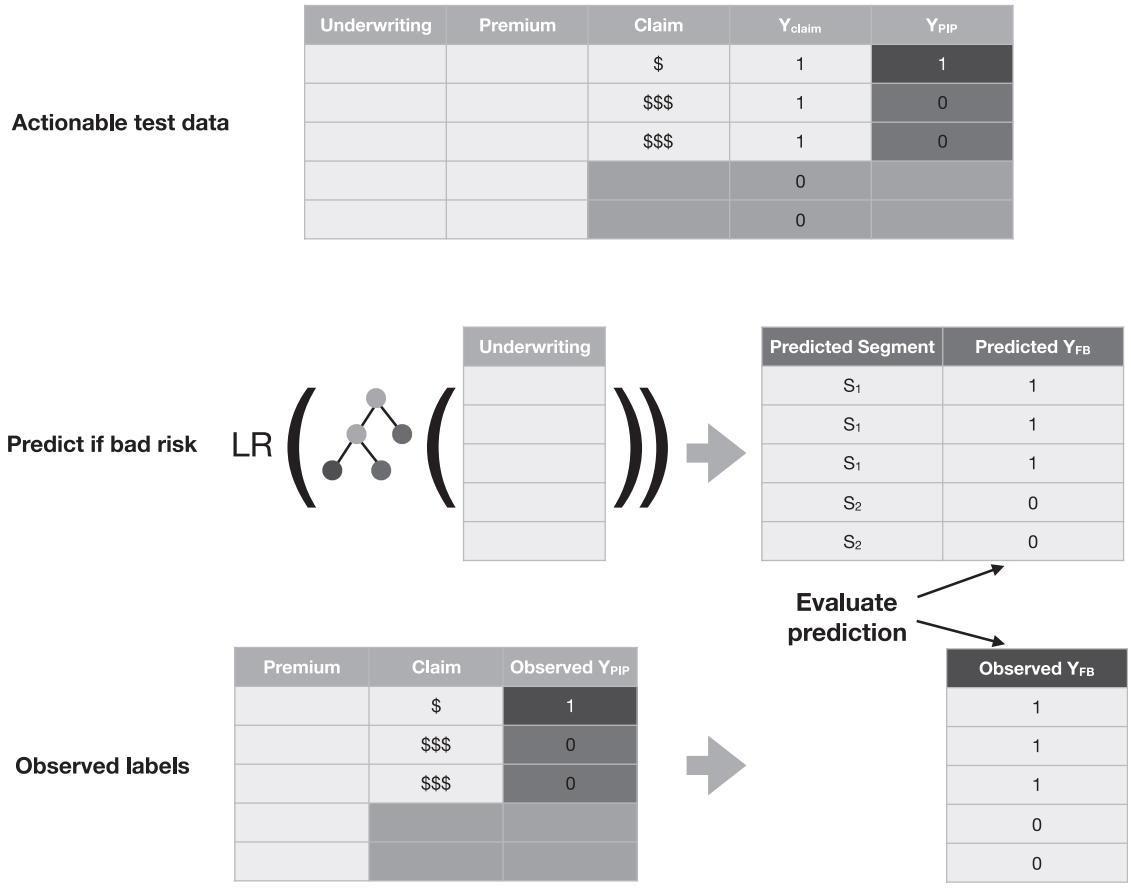

FIG. 3. Overview of FraudBuster's Phase 2.

Identifying high likelihood-of-fraud policies that exhibit a disproportionately high loss ratio sidesteps these false positives. Moreover, because these composite segments are statistically significant and compliant to insurance regulations, we deem them to be "actionable."

Phase 1: training

Data preparation. We begin by filtering raw underwriting data for any potentially noncompliant fields, such as race, religion, or country of origin. In addition, we filter out personally identifiable information, such as names, addresses, contact information, and social security numbers, as well as any other potential proxy identifiers. We also identify several fields in our data that are collected at the time of underwriting but that are considered prone to manipulation or misrepresentation by drivers to obtain a lower quote. An example of such a field in our data is the self-reported occupation field. Drivers who are currently unemployed often report a job on their underwriting data out of fear that they would otherwise face higher surcharges, making such data unreliable. The resulting data can be divided into features that characterize each driver ($X_{\mathrm{UW},1}$), the earned premiums for each driver (EP_1) and claimed losses on the policy ($Claim_1$).

Learning high likelihood of fraud and high loss ratio segments. In keeping with our definition of *actionable* fraud, we identify segments that have a high likelihood of fraud, $Y_{\mathrm{PIP},1}$, and significantly high loss ratio relative to the baseline, $Y_{\mathrm{LR},1}$. We first train a (cross-validated) DT on known PIP claimant data and observe outcomes of fraud/not fraud, $Y_{\mathrm{PIP},1}$. The DT's depth is deliberately regularized and a minimum support of 10 observed PIP cases in each leaf is set to abide by insurance regulations. The trained DT model produces mutually disjoint segments in the form of paths-to-leaves $\{S_i\}$. For each unique path, we can compute a loss ratio, giving us $Y_{\mathrm{LR},1}$ according to the following formalism:

$$Y_{\mathrm{LR}} = \begin{cases} 1, & \text{if } \mathbb{E}[LR] > \overline{LR} \\ 0, & \text{otherwise} \end{cases}.$$

Actionable fraud labels. We then derive a composite label, $Y_{\mathrm{FB}} = Y_{\mathrm{PIP}} \wedge Y_{\mathrm{LR}}$, based on the values predicted for each driver who filed a PIP claim. This composite label can be used to identify drivers who fit our profile of *actionable* fraud, representing high-propensity and unprofitable claimants. The resulting model from this phase can be readily interpreted using the paths-to-leaves of the trained DT. Drivers belonging to each segment share a propensity of fraud in case of PIP claims, an expected loss ratio, and Y_{FB}.

Phase 2: prediction and evaluation

Predicting actionable fraud in the new batch of policies. Incoming raw data in the new insurance cycle are filtered using the same process outlined in Phase 1. As a result, the data are similarly divided into feature data ($X_{\mathrm{UW},2}$), earned premiums (EP_2), and claimed losses ($Claim_2$). Using the model trained in Phase 1, we then partition the insurees into segments corresponding to $\{S_i\}$ and predict their respective labels for $\hat{Y}_{\mathrm{PIP},2}$, $\hat{Y}_{\mathrm{LR},2}$, and $\hat{Y}_{\mathrm{FB}}$. The ground truth values for the respective labels are observed at the end of Phase 2, and are used to evaluate how effectively our model from Phase 1 captures actionable fraud across these insurance cycles.

Evaluation. FraudBuster's prediction tasks take the form of a supervised binary classification problem. Although the prediction is made at the beginning of Phase 2 for all drivers, only a small fraction of the drivers eventually file PIP claims. This restricts the amount of data for which ground truth is available, but does not invalidate the presence of data on the drivers who did not file PIP claims. Earned premiums from these drivers are still relevant in computing observed loss ratios, $Y_{\mathrm{LR},2}$ and $Y_{\mathrm{FB},2}$.

Actionable fraud consists of 1/6th of our test instances, making Y_{FB} an imbalanced class problem. We measure the performance of these prediction tasks in using precision, recall, f-1 score, and AUPR on observed test data.

Typical deployment

Insurers can deploy FraudBuster to identify actionably fraudulent segments in their market as follows. First, past insurance can be used to train FraudBuster's Phase 1. If data on multiple insurance cycles are available, the loss ratio impact of discovered segments can be readily computed by creating tables such as Table 4 on these past data. Once the model is trained, insurers can apply it to their current set of underwritten policies to predict which segments are likely to have a high fraud propensity and high loss ratio. This pretraining eliminates the need to wait for 6 months of observation period otherwise required in Phase 1. At the end of the most current test phase, the insurer can evaluate their observed segments in terms of their predicted and observed Y_{FB} values. Using the interpretable

feature value bounds to define these segments, they can then file with their respective regulatory boards to reprice or even block said segments. This process can then be repeated for future insurance cycles.

Experiments

Given our outline of FraudBuster, several design decisions merit experimental validation. What is our target adverse driver population—is it those who exhibit high likelihood of fraud, high loss ratio, or both? Given the imbalance in the data, should we resample? Are we losing out by using a prepruned classifier?

Pipeline design

All of the mentioned questions can be directly stated as components of a grid search over various target labels, resampling techniques, classification techniques and hyperparameter spaces, while optimizing the AUPR as discussed in Related Work section. Here, we compare traditionally interpretable models such as DTs, Stochastic Gradient Descent (SGD), and NB against ensemble approaches such as Random Forests (RF) and Gradient Boosted Trees (GBT).

Resampling. Resampling is typically performed on data sets with severe class imbalance to mitigate classifier bias when training. In our experiments, we use RUS, ROS, and compare both of these approaches with the baseline unresampled performance. Across the board, we see that resampling the data has a positive impact on predictive performance. Even in the case of SGD technique wherein randomly undersampling the data drops the predictive performance to an f-1 score of 0, it is within statistically significant bounds. Focusing on the predictive performance of the minority class, we see that ROS performs better than RUS. This is consistent with the small disjuncts observed in Figure 1.

Classifier regularization. Classifier regularization penalizes complexity and favors training more compact models to reduce overfitting. This can be demonstrated by showing predictive performance across a wide range of maximum depths of a DT. As part of our grid search, we find that the optimum hyperparameter value for maximum depth is at 5. An added advantage of training smaller DTs (maximum depth of <6–7) is that they are now interpretable and, therefore, compliant with regulations.

Ensembles of classifiers. Ensembles of classifiers such as RF and GBT help reduce biases of individual DTs by combining the predictions from several predictors. Even though their constituent classifiers may be interpretable, the ensembles themselves are not. We compare their predictive performance against our DT to identify any potential tradeoffs involved in selecting DTs instead of sophisticated ensembles of classifiers. As expected, both ensemble models achieve high predictive performance, but at the cost of interpretability. This quantifies the predictive trade-offs when choosing an interpretable model (DT, SGD, etc) against a high performing ensemble model as summarized in Table 1.

Know *thy* enemy

In FraudBuster, we identify the following profiles of undesirable driver outcomes (which may not necessarily be related)—(A) high likelihood of fraud, (B) high loss ratios, or (A + B) high likelihood of fraud combined with high loss ratios.

Each of these profiles directly corresponds to a predictive target variable identified within FraudBuster. We would like to identify which one of these hypotheses leads to a predictor that is consistently performant across both insurance cycles. Formally, this means that we compare models trained to identify Y_{PIP} for hypothesis A, Y_{LR} for hypothesis B, and Y_{FB} for hypotheses A + B on Phase 2 data.

Which of these profiles of adverse driver behavior is most predictable over the two phases? Table 3 suggests that the combined profile of fraud (A+B) has the best f-1 score and AUPR, which means that it is the most generalizable out of the three. In comparison, the AUPR for hypothesis B (i.e., demographics with high loss ratio) comes within striking distance of hypotheses A+B, but fails to outperform a random predictor on f-1 score. This can be directly traced back to hypothesis B's majority class recall (0.4590) being significantly less than that of the baseline predictor (0.8301) and that of hypotheses A + B (0.9947).

Table 3. How well do our hypotheses generalize?

Hypothesis	*f-1 score*	*AUPR*
Baseline	0.724	0.240
Likelihood of fraud (A)	0.767	0.295
High loss ratio (B)	0.533	0.380
FraudBuster (A + B)	<u>0.769</u>	**0.631**

Individually, high likelihood of fraud and high loss ratio definitions of fraud do not generalize across multiple insurance cycles.

Instead, actionable fraud identified by FraudBuster can outperform these on both f-1 score and AUPR. Bold and underline indicate best performing model in terms of AUPR and f-1 score, respectively. The baseline considered here is a random predictor.

Table 4. Discovered segments: expressed in terms of loss ratio, support, and confidence

		Phase 1		*Phase 2*			
Segment	$\hat{Y}_{FB}$	*No. of drivers*	*E*[LR]	*No. of drivers*	LR_{obs}	*PIP supp.*	*PIP conf.*
1	0	169	2.640	157	4.401	14	1.000
2	1	6696	2.400	6956	6.303	274	0.500
3	1	741	1.665	805	1.338	26	0.769
4	1	328	1.107	224	1.320	31	1.000
5	0	2736	1.042	2737	0.748	256	0.715
6	0	2539	0.801	2792	0.810	88	0.545
7	0	2596	0.718	2575	0.430	251	0.721
8	0	6,75	0.709	683	0.161	53	0.925
9	0	463,522	0.557	463,121	0.571	29,974	0.849
10	0	12,911	0.555	12,813	0.582	1303	0.738
11	0	22,429	0.523	22,245	0.613	3424	0.799
12	0	1339	0.330	1367	0.637	84	0.833

Each row here corresponds to a discovered market segment in the data. For each segment, we compute the expected loss ratio and likelihood of fraud. If a segment scores high in both components, we then say it is "actionably fraudulent" ($Y_{FB} = 1$). We then check whether these segments remain highly fraudulent and exhibit high loss ratios in the subsequent insurance cycle. Segment definitions in terms of feature value bounds are hidden to preserve insurer privacy.

LR, loss ratio; supp., support; conf., confidence; PIP, personal injury protection.

Interpreting the segments

Table 4 shows each of the learned segments from Phase 1 of FraudBuster. We see how only a small fraction of these constitute actionable fraud ($Y_{FB} = 1$). Each of these segments has an associated number of policies and expected loss ratio. More importantly, using these predicted outcomes for Phase 2, we can see that the predictions for bags-of-instances generalize well over time. Note how segments 2, 3, and 4 exhibit both notoriously high confidence of fraud *and* loss ratio, indicating a high true positive rate. The segments that do not generalize well are also useful to insurers—segments 7, 8, and 12 represent such segments wherein the loss ratio does not hold up. Segments that are stable across cycles are repriced, whereas volatile segments are taken out of consideration for repricing.

Discussion

Can we apply fraud detection techniques directly to nonstandard markets?

In Background section, we show how fraud detection techniques that classify claims are not useful when fraud is the majority class. If we implement the same solutions at underwriting, we run into the problem of lack of ground truth and data quality concerns (Fig. 1). Fraud prediction does not work at a policy level, however, through FraudBuster, we show that it does work for bags of instances. Such predictions are robust to the noise in underwriting data, interpretable, and also generalizable across insurance cycles.

What is actionable fraud?

FraudBuster identifies policies that exhibit high confidence of fraud and high loss ratios (represented by Y_{FB}) as the most actionable type of fraud. We show that either of the singular measures of fraud fails to generalize as well across insurance cycles (Table 3).

How do we ensure end-to-end regulation compliance?

FraudBuster is deployed when policies are underwritten, as opposed to existing solutions that identify fraud when a claim is filed. In terms of data availability and data quality, we rely on extremely limited and potentially noisy data to make early predictions at the underwriting stage. We use regularized DTs with ROS and achieve performance comparable with ensembles of classifiers. In addition, we create interpretable profiles of fraud versus not-fraud segments, each with their expected loss ratios—this connects a segment's predicted outcome to its respective profitability.

Conclusion

FraudBuster represents a novel paradigm combating fraud when it is present as a majority class. We show that although we cannot predict *which drivers* are likely to get into an accident and commit fraud, we *can* identify drivers who are unprofitable and likely to be fraudulent risks. Through FraudBuster, we show how fraud can be predicted within bags of instances (policies). Merging the conventional likelihood of fraud with the profitability-driven loss ratio, we create an operationally viable framework to identify the segments that are demonstrably worst affected by fraud. Although our models can predict this fraud with an AUPR of nearly 0.63, the highlight of this approach is that it does so while staying compliant with industry regulations. FraudBuster is intended to be used by insurers to identify and reduce fraud within their markets. The underlying framework can be extended to markets where outcomes of *future* cost-sensitive risks need to be predicted using *current* data. This approach can be used to predict and assess the risk-based health of markets segments in credit, lending, healthcare, and marketing.

Author Disclosure Statement

No competing financial interests exist.

References

1. Corum D. 2015. Insurance research council finds that fraud and buildup add up to $7.7 billion in excess payments for auto injury claims. Available online at www.insurancefraud.org/downloads/InsuranceResearchCouncil02-15.pdf (last accessed February 17, 2017).
2. Derrig RA, Johnston DJ, Sprinkel EA. Auto insurance fraud: Measurements and efforts to combat it. Risk Manage Insur Rev. 2006;9:109–130.
3. Delegal MK, Pittman AP. Florida no-fault insurance reform: A step in the right direction. Fla St UL Rev. 2001;29:1031.
4. McChristian L. No fault auto insurance in Florida: Trends, challenges, and costs. Insurance Information Institute, 2011.
5. Derrig RA. Insurance fraud. J Risk Insur. 2002;69:271–287.
6. Hoyt RE. The effect of insurance fraud on the economic system. J Insur Regul. 1990;8:304.
7. Schiller J. The impact of insurance fraud detection systems. J Risk Insur. 2006;73:421–438.
8. Viaene S, Derrig RA, Baesens B, Dedene G. A comparison of state-of-the-art classification techniques for expert automobile insurance claim fraud detection. J Risk Insur. 2002;69:373–421.
9. Belhadji EB, Dionne G, Tarkhani F. A model for the detection of insurance fraud. Geneva Papers on Risk and Insurance. Issues Pract. 2000; 25:517–538.
10. Artís M, Ayuso M, Guillén M. Detection of automobile insurance fraud with discrete choice models and misclassified claims. J Risk Insur. 2002; 69:325–340.
11. Caudill SB, Ayuso M, Guillén M. Fraud detection using a multinomial logit model with missing information. J Risk Insur. 2005;72:539–550.
12. Lipton ZC. The mythos of model interpretability. arXiv. 2016; DOI: arXiv: 1606.03490.
13. Zhang H. The optimality of naive bayes. AA. 2004;1:3.
14. Quinlan JR. C4. 5: Programs for machine learning. Elsevier, 2014.
15. Zhang T. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In: Proceedings of the Twenty-First International Conference on Machine Learning. ACM, 2004, p. 116.
16. Breiman L. Random forests. Mach Learn. 2001;45:5–32.
17. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SigKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016, pp. 785–794.
18. Fawcett T, Provost F. Adaptive fraud detection. Data Min Knowl Discov. 1997;1:291–316.
19. Phua C, Alahakoon D, Lee V. Minority report in fraud detection: Classification of skewed data. ACM SIGKDD Explor Newsl. 2004;6:50–59.
20. Bradley AP. The use of the area under the roc curve in the evaluation of machine learning algorithms. Pattern Recognit. 1997;30:1145–1159.
21. Davis J, Goadrich M. The relationship between precision-recall and roc curves. In: Proceedings of the 23rd International Conference on Machine Learning. ACM, 2006, pp. 233–240.
22. Powers DM. Evaluation: From precision, recall and f-measure to roc, informedness, markedness and correlation. J Mach Learn Technol. 2011;2: 37–63.
23. Brown RL, Gottlieb LR. Introduction to ratemaking and loss reserving for property and casualty insurance. Actex Publications, 2007.

Cite this article as: Nagrecha S, Johnson RA, Chawla NV (2018) FraudBuster: reducing fraud in an auto insurance market. *Big Data* 6:1, 3–12, DOI: 10.1089/big.2017.0083.

Abbreviations Used

AUPR = area under the precision–recall curve
DTs = decision trees
GBT = Gradient Boosted Trees
NB = Naive Bayes
PIP = personal injury protection
RF = Random Forests
ROS = Random Oversampling
RUS = Random Undersampling
SGD = Stochastic Gradient Descent
SMOTE = Synthetic Minority Oversampling Technique