Using Supervised Learning to Select Audit Targets in Performance-Based Financing in Health: An Example from Zambia

Dhruv Grover, Sebastian Bauhoff, and Jed Friedman

Abstract

Independent verification is a critical component of performance-based financing (PBF) in health care, in which facilities are offered incentives to increase the volume of specific services but the same incentives may lead them to over-report. We examine alternative strategies for targeted sampling of health clinics for independent verification. Specifically, we empirically compare several methods of random sampling and predictive modeling on data from a Zambian PBF pilot that contains reported and verified performance for quantity indicators of 140 clinics. Our results indicate that machine learning methods, particularly Random Forest, outperform other approaches and can increase the cost-effectiveness of verification activities.

Keywords: performance-based financing, performance verification, audits, machine learning, health care finance, health care providers

JEL: C20, C52, C55, I15, I18

Center for Global Development
www.cgdev.org

Working Paper 481
April 2018

Using Supervised Learning to Select Audit Targets in Performance-Based Financing in Health: An Example from Zambia

Dhruv Grover
University of California, San Diego

Sebastian Bauhoff*
Center for Global Development

Jed Friedman
World Bank

*Corresponding author, sbauhoff@cgdev.org.

Resources for the data collection were provided by the Health Results Innovation Trust Fund administered by the World Bank. The authors are grateful to Kalipso Chalkidou, Collins Chansa, Sean Dalby, Ashis Das, Erik Josephson, Harrison Mkandawire, Jumana Qamruddin, Petra Vergeer, and Monique Vledder for support and comments.

The Center for Global Development is grateful for contributions from Australian Department of Foreign Affairs and Trade and Johnson & Johnson in support of this work. These funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Dhruv Grover, Sebastian Bauhoff, and Jed Friedman, 2018. "Using Supervised Learning to Select Audit Targets in Performance-Based Financing in Health: An Example from Zambia." CGD Working Paper 481. Washington, DC: Center for Global Development. https://www.cgdev.org/publication/using-supervised-learning-select-audit-targets-performance-based-financing-health

Center for Global Development
2055 L Street NW
Washington, DC 20036

202.416.4000
(f) 202.416.4050

www.cgdev.org

The views expressed in CGD Working Papers are those of the authors and should not be attributed to the board of directors, funders of the Center for Global Development, or the authors' respective organizations.

Contents

Introduction

Performance-based financing (PBF) is a contracting mechanism that aims to increase the performance and quality of service providers. PBF programs for health care services in low and middle-income countries typically offer financial incentives to health care facilities for the provision of a defined set of services, with an adjustment to the bonus payment based on a broad measure of quality [1]. For example, a PBF program may offer a bonus payment of $5 for each delivery in a clinic, and scale down the bonus if the clinics' quality is found to be low. In recent years, PBF has generated substantial interest among policy-makers in low- and middle-income countries. Donors and international organizations are actively engaged in supporting countries in developing, implementing and evaluating PBF programs; for instance, as of 2015 a dedicated trust fund at the World Bank alone supported 36 PBF programs [2].

Regular verification by an independent third-party is a central component of PBF, as of all contracts that condition compensation on performance and must contend with asymmetric information between the provider and payer. Providers may react to the incentives not only by increasing performance but also by gaming, e.g., by deliberately over-reporting relative to their actual performance. Verification serves to establish the veracity of performance data (the basis of payment) and to mitigate incentives for gaming by introducing a threat of detection with consequent penalties for identified over-reporting.[1]

Effectively targeting verification activities is an important concern for PBF programs. In order to balance the costs and benefits of verification activities, these programs need approaches that detect and deter misreporting by focusing on a sample of providers [1]. However, sampling schemes vary in their performance. For instance, simple random sampling is only effective in identifying misreporting if either the sample or the proportion misreporting are large, as otherwise the odds of capturing those facilities are very small. Sampling performance can be increased by incorporating background and contextual knowledge and using that information to produce models that encode the relationships in the data. This latter sampling problem is a natural application of machine learning, as these techniques use automated learning to identify data attributes that are empirically relevant based on past observations and can achieve highly accurate predictions. This, in turn, can tailor verification activities to high-risk facilities and thus lower cost and/or increase the precision of verification.

Here, we compare strategies for targeted sampling of health clinics for third-party verification of PBF quantity indicators, with a view toward increasing the cost-effectiveness of these essential program activities. Specifically, we compare the performance of a random sampling-based approach (with and without stratification) with four common supervised-learning based classification methods in correctly identifying health clinics that over-reported service volumes relative to verified data: Naïve Bayes, Logistic Regression, Support Vector

[1] Verification also serves to help under-reporting of providers (e.g., through misunderstandings or data entry errors) and promote a focus on results.

Machines and Random Forest. We apply these methods to reported and verified data from a PBF program in Zambia and evaluate each method on a range of performance measures. Our results indicate that machine learning methods, particularly Random Forest, outperform other approaches and, as result, can substantially increase the cost-effectiveness of verification activities.

Background

Verification in PBF programs

PBF offers financial incentives for increasing the quantity and/or quality of a specific set of health care services. Rewards for increasing quantities are generally "fee for service" payments, so that providers receive a bonus for every additional service they report. Many programs inflate or deflate the quantity bonus based on a broad index of quality.

Verification serves to counteract the incentive to over-report and, in some cases, to fulfill the fiduciary responsibilities of the implementing agency. A typical PBF program has three layers of verification [1], [3], [4]. First, district or provincial supervisors visit all facilities on a monthly or quarterly basis to confirm the accuracy of the reported quantity data. Second, district teams visit all facilities on a quarterly basis to complete a quality assessment. Third, an independent third-party such as a community or non-governmental organization conducts quarterly counter-verification visits to a sample of facilities. Core counter-verification activities include reconciling health service records at different points in the reporting chain, from individual services recorded in patient attendance books up to service aggregation sheets at the district level. Additional activities can include client tracer surveys that re-contact patients to verify the receipt of recorded service and inquire about patient satisfaction. Discrepancies between the reported and verified data result in penalties, such as deductions from the PBF bonus or exclusion from future rounds of the program.

PBF programs vary in their approach to targeting facilities for counter-verification and the associated costs.[2] In Burundi, the third-party agency randomly selects one district in each of four provinces every quarter; the sample of provinces is randomized but rotating so that all provinces are visited within a year [6]. The district hospital and a random sample of 25 percent of health centers are audited in the selected districts using quantity verification, technical quality assessment and household surveys. Facilities with positive or negative discrepancies of more than 5 percent should be subject to fines. The penalty increases in the size of the discrepancy so that, e.g., 10 percent of the PBF bonus is withheld if the discrepancy is 10-20 percent. In Benin, community-based organizations are contracted to conduct quarterly counter-verification through unannounced visits to facilities as well as tracing a random sample of patients to confirm the receipt and experience of service [4].

[2] For a series of detailed case studies of verification in PBF, see [5].

Counter-verification can represent a substantial share of overall program financial costs as well as staff time and effort, so that improved targeting could yield important savings. Although there is no systematic overview of verification costs in PBF, data from individual programs suggests that the costs vary substantially with the program design. In Burundi in 2011, the financial costs of these activities represent 1 percent of overall spending, not accounting for the time costs of district and facility staff [6]. In Benin in 2013-2014, the financial costs represent about 30 cents for every dollar paid to providers in bonuses, and there are additional time costs of district and facility staff [4]. In Argentina's Plan Nacer, which incentivizes provinces rather than facilities, verification costs may be equivalent to about 10 percent of the maximum bonus payments.[3]

Zambia's performance-based pilot

Zambia operated a PBF pilot project from 2012 to 2014 in an attempt to realign health financing towards outputs rather than inputs, and to address various health system concerns such as relatively low coverage of key maternal and child health services. The pilot operated in public health centers in 10 rural districts, covering a population of 1.5 million, or about 11 percent of Zambia's population [8]. It comprised two core features, financial rewards and equipment upgrades. Specifically, the program offered varying fee-for-service bonus payments for indicators measuring the quantity of nine maternal and child health (see Appendix Table A.1) and 10 quality domains covering aspects of both structural and process quality. Health centers also received emergency obstetric care equipment. In addition, participating health centers were subject to enhanced monitoring.

The financial rewards from the PBF were substantial, with an individual staff's bonus representing on average 10 percent of government salary [8]. An evaluation of the pilot based on independent population surveys found gains in selected targeted indicators, such as the rate of facility deliveries [9]. Other targeted indicators, especially those at already high levels of coverage such as ante-natal care, saw little change.

The program extensively audited reported data through continuous internal verification and a one-off external process.[4] Dedicated district steering committees served as internal verifiers, reconciling the facility reported information which served as the basis for incentive payments with the paper based evidence of services provided at the facility. An independent third party verified reports by primary health clinics on a sample basis in a one-off exercise after two years of program operation. The total cost of this one-off external verification

[3] A calculation for Argentina's Plan Nacer suggests that third party verification contracts cost the equivalent of 6.7 percent of the program's capitation payments that, in turn, represent 60 percent of allowable transfers under Plan Nacer (the remaining 40 percent consist of the PBF component) [7]. This suggests that verification costs may reach 10 percent of overall allowed disbursements.

[4] This external verification was only conducted once over the life of the two-year project while the design plans originally called for this exercise to be conducted every two quarters.

came to 1.5 percent of the total US$ 15 million spent on the pilot project. The cost of internal verification activities was assuredly greater given the scale of activities involved.

Conceptual framework

Conceptually, the primary fiduciary objective of counter-verification is to reduce or eliminate over-reporting while minimizing verification costs. The costs of this scheme can be modeled as a simple cost function where the total cost, *tc*, is a linear function of several parameters:

$$tc = n[p(mc - s) + (1 - p)mc] + fc$$

n is the total number of facilities selected into the verification sample. *p* is the proportion of sampled facilities found to have misreported in a given time period (e.g. a quarter). *mc* is the marginal cost of verification at a facility, which is assumed to be the same regardless of whether the facility is a mis-reporter or not. *s* is the financial sanction that the over-reporter must pay to the health system, and *fc* is the fixed cost of verification activities that are constant across time periods.

The financial efficiency gains from improving the accuracy of predicting what facilities are over-reporting (i.e. increasing *p*) while keeping *n* constant arises through two channels: (1) an increase in *p* directly leads to a lower cost of the verification by increasing the amount earned back through sanction payments, (2) an increase in *p* may have a dynamic effect in so far as facilities identified as over-reporters may be less likely to over-report in the future. If accurate prediction increases this deterrence effect, then future verification efforts can be reduced in size as the incidence of future over-reporting declines.

Data and methods

Data from Zambia's PBF pilot

We use operational data from the Zambia pilot project to evaluate the performance of different classification methods. Although a simulated dataset would have been sufficient for this assessment, the Zambia data are realistic with regards to data attributes and parameters.

Specifically, we combine data from facility reports and a dedicated facility survey that was designed to reproduce the stipulated external verification activities in a sample of facilities. The data cover 105 primary health care centers in the 10 PBF pilot districts and 35 centers in another 8 non-pilot districts, for a total of 140 facilities. The population of facilities were stratified by district and then selected on a proportional-to-size basis with respect to the facility catchment area. Reported performance stems from the Health Information Aggregation (HIA) 2 forms in which health facilities summarized services provided for each indicator. Verification data were collected on the complete set of nine incentivized indicators and cover every calendar month of 2013. These data are derived from tally sheets, activity sheets and registers, as these records indicate the individual services delivered to a specific client. These data sources contain the date of the service, client register number, and other information and were used to check errors relating to summing, recording and data entry.

Our focus is on whether or not a facility over-reported. Our primary measure of interest is a binary measure of over-reporting relative to the verified data. The measure is equal to one when the difference in the bonus payment based on the reported vs. verified data represents 10 percent or more of the reported value. We calculate this measure in two steps.[5] First, for each facility in each quarter, we calculate the product of the quantity and price (reward) for indicator and then sum these products to obtain the total bonus payment based on (separately) the reported and verified data. Second, we calculate the difference in the bonus payments and evaluate whether it is substantively large. We classify a facilities' quarterly report as over-reported if the bonus calculated on the reported data exceeds the bonus calculated on the verified data by 10 percent or more. Our label of "over-reporting" emphasizes that a regulator would primarily be concerned about cases when a facility's report exceed the actual volume of services delivered, because of the associated over-payment. The cutoff of 10 percent allows for possible leniency for smaller mismatches and generates sufficient variation in the data; it also corresponds to thresholds used in at least some operational PBF programs.

Tables 1 and 2 illustrate the structure of the Zambia data over the four quarters of 2013. 15-23 percent of facilities over-report in a given quarter. There is a strong correlation of over-reporting over time: of the facilities that over-report in quarter 1, 58 percent also over-report in quarter 2, and 42 percent also over-report in quarters 3 and 4. Table 2 shows that about 58 percent of facilities never over-report and only 4 percent over-report in all four quarters.

Classification methods

We evaluate the performance of sampling-based approaches in classifying a facilities' quarterly report as over-reporting, and compare their performance to four alternative approaches including supervised machine learning: Naïve Bayes, Logistic Regression, Support Vector Machines and Random Forest. In this section we briefly describe these approaches and considerations for their use.

Sampling-based approaches

Random sampling is the current default selection method for counter-verification in many PBF programs [1].

In this study, we examine how well four approaches to random sampling perform with regards to identifying clinics that over-report. First, we use simple random sampling to

[5] The calculations for the quarterly value are first on the month-level and then summed. Both data sources have some missing values that we fill where possible, e.g., we calculate a single month's missing value as the difference between the quantities for the quarter and the other two months. When calculating the bonus payment, we only consider indicator-months that are available for both the reported and verified data, i.e., we ensure that the bases for the calculations are identical.

determine the probability of an over-reported event, wherein 50 percent of the clinics are chosen repeatedly at random. Second, we stratify the sample by district and then use simple random sampling to select 50 percent of clinics within these strata. This approach ensures that counter-verification takes place in all districts. Third, we leverage knowledge of what clinics previously over-reported. Specifically, we use simple-random sampling to draw half the audit sample from those clinics that over-reported in the immediate prior quarter, and simple random sampling to draw from the remaining clinics. The overall sample size is 20 percent of all clinics, i.e., 28 clinics. Fourth, we select up to 28 clinics that were prior offenders. If the number of prior offenders is greater than 28, we use simple-random sampling to select the target number. If the number is less than 28, we randomly sampled from the remaining facilities to achieve the target number.

We report the accuracy of the sampling-based approaches as averages of 1000 independent sampling iterations without replacement.

Supervised learning

Supervised learning are a class of machine learning algorithms that use labeled examples to infer a relationship between input and output variables, and then use that inferred relationship to classify new examples. The underlying basis of these algorithms is to generalize from training data to prediction of class labels of unseen instances from a test set. In practice, it is important that the test set of unseen instances be distinct from the training corpus. If training examples were reused during testing, then a model that simply memorized the training data, without learning to generalize to new examples, would receive misleadingly high scores.

In the context of verification in PBF, these training data are a subset of facility-specific data points (input) that contain a binary indicator for whether or not a facility over-reported (output). The algorithm learns from these data which facilities are at risk of over-reporting, and applies this learning to predict this risk for other facilities not included in the training data.

For the below analyses, we used as input features the reported and verified values for the nine quantity measures that were rewarded in the RBF program, along with the district identifier and a categorical variable indicating the treatment arm from a related audit experiment.[6] In a supplemental analysis, we expanded this set of 22 covariates to include an additional 6 facility-level covariates (see below).

Next, we describe briefly each of the four supervised learning approaches used here, their strengths and weaknesses, and suggest ways in which a practitioner might ascertain the method best suited for the classification problem at hand.

[6] The audit experiment randomly varied the probability of audit (10, 30 or 100 percent) in the RBF areas while clinics in non-RBF areas had a zero percent probability of audit. Facilities were told what their specific audit probability was.

Naïve Bayes is a simple and efficient classification technique that involves the application of Bayes theorem, wherein the probability of a feature is determined using prior knowledge of conditions that might be related to that feature [10], [11]. This algorithm calculates the probability of an input (or specific set of predictive features) belonging to each class (labeled output), and then chooses the class with the highest score. It assumes strong independence between these predictive features, i.e. correlations between features, if any, are disregarded, and all features contribute independently to the probability of the class variable.

Logistic Regression uses a logistic function at its core to estimate a relation between the binary classification and its possible predictors [12], [13]. Unlike standard regression, logistic regression does not try to predict the value of a numeric variable given a set of inputs, rather, the output is a probability of a given input belonging to a certain class. The central premise of logistic regression is the assumption that the input space can be partitioned by a linear boundary, separating the data into two classes.

The difference in learning mechanisms between Logistic Regression and Naïve Bayes can be subtle. Naïve Bayes attempts to model both inputs and outputs, by estimating the joint probability of a set of features and output label from the training data, and is termed as a generative classifier. Logistic Regression, described as a discriminative classifier, attempts to infer the output from input data: estimating output probabilities from the training data by minimizing error [14].

A support vector machine (SVM) is a discriminative non-probabilistic classifier formally defined by a hyperplane that maximizes the separation between the two classes [15]. Given labeled training data, this non-parametric algorithm outputs an optimal hyperplane that best separates the data belonging to the different categories, for instance, in two-dimensions the hyperplane separating data from different categories would be a line, in three-dimensions, a plane, and so on. To determine the optimal hyperplane or decision boundary between the categories, the SVM algorithm maximizes the margins from both categories, such that the distance from the boundary to the nearest data point on either side is the largest. Once an optimal hyperplane is found using labeled training data, features from the test set can then be classified into their respective categories by determining whether they fall on one side of the boundary or the other.

Random forests are ensemble-based classification algorithms [16], [17]. The main principle behind ensemble based methods is that a group of weak learners can be integrated during training time to form a strong learner.

Random forests are based on decision trees, that is, a graph that uses a branching method to illustrate every possible outcome of a decision. Each internal node represents questioning an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of that question, and the leaf nodes of the tree signify the class labels. The random forest starts with a decision tree, wherein the input entered at the top traverses down a tree structure to the leaves to get binned into smaller sets.

Random forests are a way of averaging multiple decision trees, trained on different parts or features of the same training set, with the goal of reducing variance. Individually, predictions made by decision trees may not be accurate, but combined together on different features, they achieve higher predictive power. In ensemble terms, the decision tree corresponds to the weak learner, a multitude of which form a strong learner in the shape of a random forest.

There are six factors to consider when choosing an appropriate machine learning technique for an application like counter-verification in PBF. First and foremost being the size of the data set used for training the classifier. Second, whether there is a need to learn interactions between the various features or whether can they be treated as independent variables. Third, whether additional training data may become available in the future and would need to be easily incorporated into the model. Fourth, whether the data is non-parametric and not linearly separable. Fifth, whether overfitting of the model to the training data is expected to be a problem. Finally, whether there are any requirements in terms of speed, performance and memory usage.

For small training sets, high bias/low variance classifiers (e.g. Naïve Bayes) have an advantage over low bias/high variance classifiers (e.g. Logistic Regression), since the latter can have a tendency to overfit [18]. However, low bias/high variance classifiers perform increasingly better as the training set grows (they have lower asymptotic error), since high bias classifiers are not powerful enough to provide accurate models. Logistic Regression can work quite well as long as the data features are roughly linear and the problem linearly separable. It is also relatively robust to noise, can avoid overfitting and allows updates to the model with new data. A final advantage of Logistic Regression is that the output can be interpreted as a probability and can be used for ranking instead of classification [19], [20].

SVMs are superior to Logistic Regression for problems that are not linearly separable, in which case a SVM with a non-linear kernel would perform well. SVMs provide high accuracy and useful theoretical guarantees against overfitting, and are popular algorithms where high-dimensional spaces are the norm. However, they can be memory intensive, hard to interpret and challenging to tune for optimal performance in most industry-scale applications [21]. While the size of the dataset used in this study is not of immediate concern for the application of SVM, scaling of this approach to larger datasets spanning an entire country (or several countries) that might include additional possible correlated features, and increased training size, can negatively affect performance by increasing variance of the prediction, thus diluting the true signal.

Finally, tree ensemble-based learning methods such as Random Forest have distinct advantages over both Logistic Regression and SVM. A major advantage over Logistic Regression is that they do not expect linear features or even features that interact linearly [22]. The advantage over SVM is that, because of how they are constructed (using bagging or boosting), these algorithms handle high dimensional spaces as well as large number of training examples. It captures the variance of several input variables at the same time and enables high number of observations to participate in the prediction. Random Forest methods are fast and scalable unlike SVMs, do not suffer from overfitting and require no tuning of parameters [23], [24], making them a—prima facie—ideal choice of a classification algorithm for our application.

Performance metrics

We assess the performance of the classification methods with five performance metrics: prediction accuracy, F-score, area under the ROC [25], average precision rate, and root mean squared error (RMSE).

Prediction accuracy is the ratio of the number of correct predictions made in relation to the total number of predictions. Classification accuracy alone is typically not sufficient to evaluate the robustness of the model's predictive capabilities. We therefore also utilize the other performance measures to evaluate the algorithms.

Metrics such as accuracy or RMSE have a range of [0, 1]. In the case of accuracy, higher values are better. However, low RMSE values indicate better performance. Metrics such as ROC area have baseline rates that are independent of the data, while others such as accuracy have baseline rates that depend on the data. Therefore, to allow comparisons across metrics for the various models tested, we scaled the performance for each metric from [0, 1], where 0 is baseline performance and 1 is best observed performance as a proxy for optimality. We also report on the prediction accuracy of simple random sampling for different sample sizes and with or without stratification.

Analysis

We compare four supervised learning algorithms (Naïve Bayes, Logistic Regression, Support Vector Machines and Random Forest) on five performance metrics (prediction accuracy, F-score, area under the ROC, average precision rate, and root mean squared error).

We used 10-fold cross-validation on the Q1 dataset (140 cases), in which the original sample is randomly partitioned into 10 equal size subsamples. Of the 10 subsamples, a single subsample is retained as the validation data for testing the model, and the remaining 9 subsamples are used as training data. The cross-validation process is then repeated 10 times, with each of the 10 subsamples used exactly once as the validation data. We then averaged the 10 results to produce a single estimation. Given that the size of our training dataset is relatively small, the advantage of this method is that all observations are used for both training and validation, and each observation is used for validation exactly once. Similarly, we report the accuracy of the sampling-based approaches as the averages of 10 independent sampling iterations.

For each metric we found the best parameter settings for each algorithm using the validation data set aside by cross-validation, then report that model's normalized score on the test set. Following training, we then used the estimated models to further classify data from Q2-Q4, i.e., we assess the models' prediction accuracy in subsequent periods.

Results

Figure 1 shows the ROC curves for random forest, logistic regression, support vector machine and Naïve Bayes classifiers using cross-validation with training data.

Table 3 shows the normalized scores for each supervised learning algorithm across five performance metrics. Each entry in the table averages these scores across the ten trials. Random Forest outperforms all other approaches on these five performance metrics. SVM also performs relatively better than Naïve Bayes and LR on most measures except RMSE, while LR performs marginally better than Naïve Bayes on all metrics.

Table 4 shows the prediction accuracy of the four supervised learning algorithms and the four sampling approaches. As in Table 3, Random Forest performs best with almost 87 percent prediction accuracy in a single cross-section and 77-89 percent accuracy in the time series. SVM also performs well relative to other methods but has substantively lower performance than Random Forests at 64 percent in the cross-section and 49-58 percent in the time-series. The sampling methods perform worse than most of the supervised learning approaches (except logistic regression): Simple random sampling (SRS) and SRS with district stratification have a low predictive accuracy. Sampling using historical information about over-reporting performs better, reflecting the correlation of over-reporting over time shown in Table 1. Revisiting the top-offenders further boosts the prediction accuracy. Note that Table 1 shows the accuracy of revisiting all offenders of a previous quarter, without sampling. Of those facilities that over-reported in the first quarter, 57.7 percent, 42.3 percent and 42.3 percent also over-report in quarters 2-4, respectively.

For each of the supervised learning algorithms, we thoroughly explored the parameter space related to the classification methods. For instance, we performed Naïve Bayes modeling as single normal, with kernel estimation, or discretizing them with supervised discretization. In Logistic Regression, we tested both un-regularized and regularized models, varying the ridge parameter by a factor of 10 from 10^-5 to 10^-2. For SVMs, we used the following kernels: linear, polynomial degree 2 and 3, radial with width ranging from 10^-3 to 1 by a factor of 10. For Random Forest, we used the Breiman-Cutler version with btrees of sizes (50, 100, 200, 500) and size of feature set considered at each split from 1-9. Appendix Figure A1 shows the tuning of the Random Forest *mtry* parameter (number of variables randomly sampled at each split) for a range of *ntree* (number of trees to grow). The most accurate values for *ntree* and *mtry* were 200 and 3, respectively, with an accuracy of 84.4 percent.

We also examined whether additional covariates would improve the predictive accuracy of the supervised learning algorithms. Appendix Table A2 shows the predictive accuracy when also accounting for the facility type, managing authority, location, size of the catchment population, and number of established and filled staff posts. Incorporating these covariates slightly reduced the performance for all algorithms. Appendix Figures A2 and A3 show the importance of covariates in the Random Forest algorithm and indicate that these additional covariates contribute little to the predictive accuracy.

Discussion

Verification of reported performance is a crucial issue in PBF in health, given that the payment incentives may not only encourage increased performance but also over-reporting. Independent verification serves to deter over-reporting and to ensure that payments reflect actual performance.

We described several approaches to identify the set of health clinics that should be audited, and tested performance of these approaches on self-reported and verified data for 140 clinics in Zambia. Our results indicate that sampling-based approaches do not perform well even with large sample sizes. Algorithm-based supervised learning methods perform substantially better, especially Random Forest which—in our data—has a prediction accuracy that remains high even over time. The finding that Random Forest outperforms a regression approach such as Logistic Regression indicates that over-reporting is a highly non-linear function of covariates—information commonly observed in administrative data or facility surveys—and consequently predictions from traditional regression analysis will not be particularly accurate.

These high-performing methods are feasible in operational PBF settings: they use existing data, can be updated as new information becomes available. Indeed, unlike sampling-based approaches, the supervised learning methods are likely to further improve as new and additional data becomes available. These methods can also be made user friendly and automated, by drawing data from existing data systems, such as the electronic District Health Information System (DHIS-2) or dedicated PBF data portals, and outputting the list of facilities that to be visited by verification teams.

Improving the prediction accuracy yields several benefits to PBF programs. It leads to clear cost savings as, on the margin, each detected case will result in a penalty that helps defray the costs of the verification activities.[7] Methods with high accuracy may also be more effective at deterring over-reporting over time. They will also reduce the time that staff of correctly reporting clinics spend (unnecessarily) to support verification activities. Finally, better detection may also be perceived as more fair and improve the acceptance of PBF among the clinics and policy-makers.

There are several directions for future research. First, the various approaches could be tested on different data, either from other real-life PBF program or simulated data, and with different definitions of what constitutes "over-reporting." Our findings are applicable to a particular setting and program design, and the performance of the methods may vary across contexts. Second, there may be other approaches that could perform well but that we did not test in this study. For instance, sampling-based approaches could use more productive strata than districts, e.g., strata constructed using machine learning or principal-component analysis of covariates that are predictive of over-reporting. Fourth, future testing or

[7] The average cost of verification was about $1,600 per clinic. The marginal cost may have been around $800 per clinic.

implementations could use additional covariates to improve the accuracy of the supervised learning methods, e.g., district and facility characteristics such as size, staffing and remoteness. Some of these data are readily available in settings with electronic health management information systems. As our example shows, adding covariates is not always an improvement and, indeed, can worsen predictive accuracy (see Appendix). Finally, additional research may be required on the behavioral response of clinics to the performance of the verification scheme and how the scheme could be adapted over time to address these responses as well as general improvements in reporting accuracy. We relied on quarter-1 data to predict over-reporting in subsequent quarters, so that any response after quarter 1 would not impact the predictions of the algorithm-based models. However, behavioral responses could affect the data basis for models built on later data; the sampling-based models are immune to this concern. For this reason, a hybrid approach may involve periodic retraining of the learning algorithm on a new random sample drawn from the population of participating clinics.

Overall our findings suggest that supervised learning approaches, such as Random Forest, could substantially improve the prediction accuracy of counter-verification in PBF and thus increase the cost-effectiveness of verification. These methods are operationally feasible, especially in settings with electronic routine reporting systems.

Recommendations for practitioners

Training of the algorithm:

- Create an initial training dataset for the supervised learning algorithm that contains facility-reported and verified data for a subset of the health facilities, e.g., for a 10 percent random sample of all participating facilities.
- Define a threshold for over-reporting that is considered unacceptable. For example, we used a discrepancy of 10 percent or more of the self-reported and verified data: facilities with a larger discrepancy are classified as over-reporting.
- Select an algorithm that is appropriate for the specific data and setting (see above for a brief list of criteria to that can inform this choice)
- The algorithm will learn patterns that are associated with over-reporting in the training data.

Applying the algorithm:

- Apply the trained algorithm to the reported data from all facilities in the program.
- The algorithm predicts which facilities are likely to over-report.
- Use the predictions to inform audit activities, e.g., send audit teams to verify data at those facilities that have the highest risk of over-reporting.

Continued use of the algorithm:

- Occasionally refresh the training sample, i.e., collect a new training dataset from all participating facilities, for instance, using random sampling. Then re-apply the algorithm.

References

[1] G. B. Fritsche, R. Soeters, and B. Meessen, *Performance-Based Financing Toolkit*. The World Bank, 2014.

[2] HRITF, "Achieving Results for Women's and Children's Health. Progress Report 2015.," Health Results Innovation Trust Fund. World Bank, 2015.

[3] J. Naimoli and P. Vergeer, "Verification at a Glance," World Bank, Washington DC, 2010.

[4] M. Antony, M. P. Bertone, and O. Barthes, "Exploring implementation practices in results-based financing: the case of the verification in Benin," *BMC Health Serv. Res.*, vol. 17, p. 204, 2017.

[5] P. Vergeer, A. Heard, E. Josephson, and L. Fleisher, "Verification in results-based financing for health," 2016.

[6] A. Renaud, "Verification of Performance in Result-Based Financing : The Case of Burundi," Washington D.C., 86190, 2013.

[7] A. Perazzo and E. Josephson, "Verification of performance in results based financing programs : the case of Plan Nacer in Argentina," World Bank, 95083, 2014.

[8] G. C. Shen *et al.*, "Incentives to change: effects of performance-based financing on health workers in Zambia," *Hum. Resour. Health*, vol. 15, p. 20, 2017.

[9] J. Friedman, J. Qamruddin, C. Chansa, and A. K. Das, "Impact Evaluation of Zambia's Health Results-Based Financing Pilot Project," World Bank, 2017.

[10] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 2 edition. Upper Saddle River, N.J: Prentice Hall, 2002.

[11] D. J. Hand and K. Yu, "Idiot's Bayes – not so stupid after all," *Int. Stat. Rev.*, vol. 69, pp. 385–399, 2001.

[12] D. R. Cox, "The regression analysis of binary sequences (with discussion," *J Roy Stat Soc B*, vol. 20, pp. 215–242, 1958.

[13] S. H. Walker and D. B. Duncan, "Estimation of the probability of an event as a function of several independent variables," *Biometrika*, vol. 54, pp. 167–178, 1967.

[14] A. Y. Ng and M. I. Jordan, "On Discriminative vs. Generative Classifiers: A Comparison of Logistic Regression and Naive Bayes," in *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, Cambridge, MA, USA, 2001, pp. 841–848.

[15] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, pp. 273–297, 1995.

[16] T. K. Ho, "Random decision forests," in *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, 1995, vol. 1, pp. 278–282.

[17] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[18] J. D. Rennie, L. Shih, J. Teevan, and D. Karger, "Tackling the poor assumptions of naive bayes text classifiers," in *International Conference on Machine Learning*, 2003, vol. 20, p. 616.

[19] M. Pohar, M. Blas, and S. Turk, "Comparison of Logistic Regression and Linear Discriminant Analysis: A Simulation Study," *Metodološki Zv.*, vol. 1, no. 1, pp. 143–161, 2004.

[20] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R*, 1st ed. 2013, Corr. 7th printing 2017 edition. New York: Springer, 2013.

[21] G. C. Cawley and N. L. C. Talbot, "Over-fitting in model selection and subsequent selection bias in performance evaluation," *J. Mach. Learn. Res.*, vol. 11, pp. 2079–2107, 2010.

[22] T. K. Ho, "The Random Subspace Method for Constructing Decision Forests," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, pp. 832–844, 1998.

[23] E. Kleinberg, "On the Algorithmic Implementation of Stochastic Discrimination," *IEEE Trans. PAMI*, vol. 22, 2000.

[24] E. Kleinberg, "An Overtraining-Resistant Stochastic Modeling Method for Pattern Recognition," *Ann. Stat.*, vol. 24, pp. 2319–2349, 1996.

[25] F. J. Provost and T. Fawcett, "Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions," *Knowl. Discov. Data Min.*, pp. 43–48, 1997.

Figures and tables

Figure 1: ROC curves for random forest, logistic regression, support vector machine and naïve bayes classifiers using cross-validation with training data.

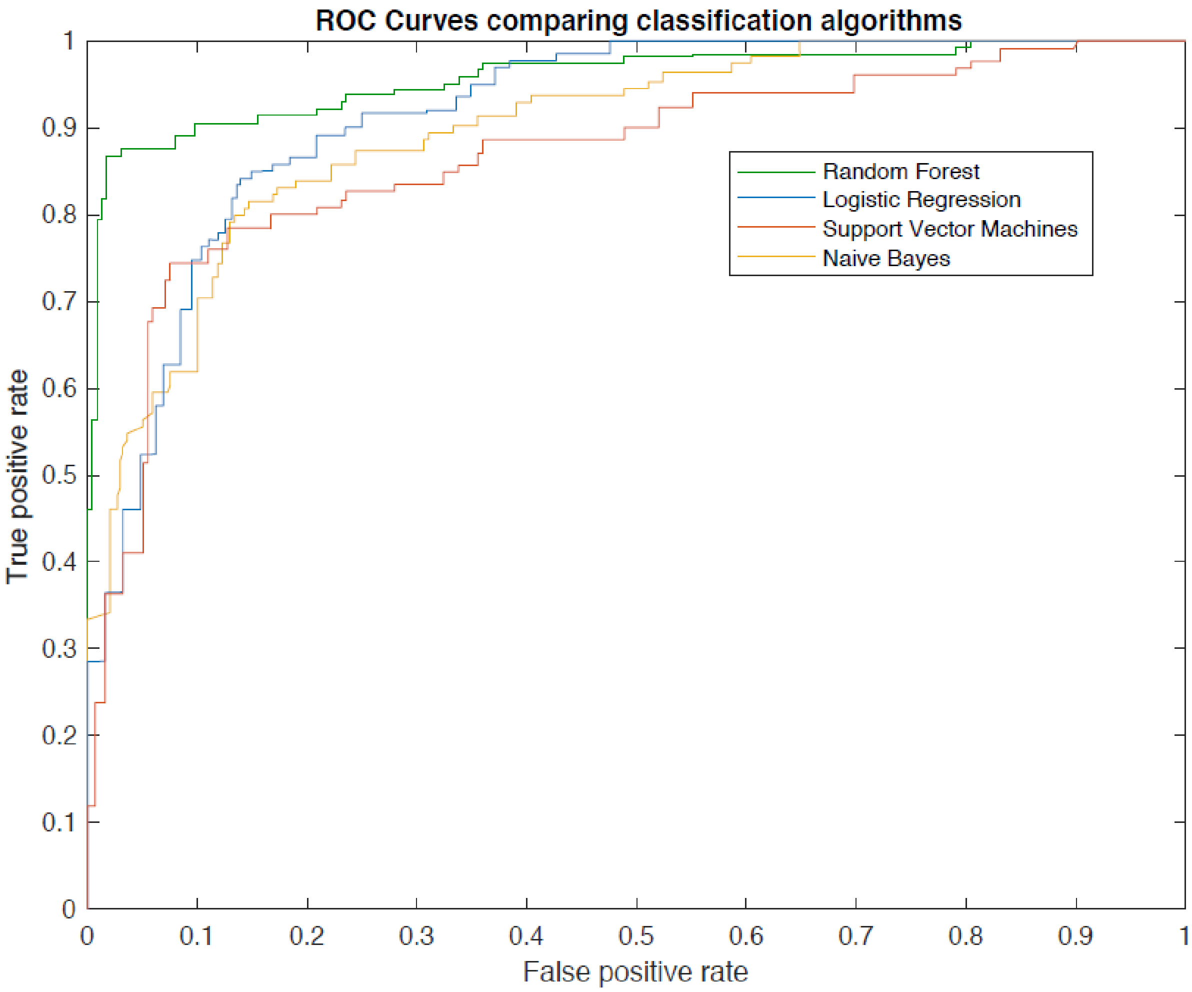


Table 1: Overview of data from Zambia pilot

	Quarter			
	1	2	3	4
Percent over-reporting	18.6	15	22.9	20
Count	140	140	140	140

Percent of facilities over-reporting if also over-reporting in…				
Quarter 1	100	57.7	42.3	42.3
Quarter 2	71.4	100	66.7	47.6
Quarter 3	34.4	43.8	100	43.8
Quarter 4	39.3	35.7	50	100

Table 2: Distribution of facilities that over-report

	N	Percent
Never	81	57.9
One quarter	32	22.9
Two quarters	12	8.6
Three quarters	9	6.4
All four quarters	6	4.3

Table 3: Normalized scores of learning algorithms across five performance metrics

Model	Accuracy	F-score	ROC area	Avg precision	RMSE
Logistic Regression	0.584	0.509	0.728	0.627	0.603
Naïve Bayes	0.552	0.425	0.583	0.523	0.488
SVM	0.647	0.651	0.783	0.691	0.501
Random Forest	0.866	0.821	0.901	0.896	0.817

Note: scores normalized to range from 0 (worst) to 1 (best).

Table 4: Prediction accuracy performance of different approaches

Approach	Prediction of over-reported event			
	Q1	Q2	Q3	Q4
Sampling approaches				
SRS	18.77%	14.98%	22.56%	20.04%
SRS with district stratification	18.83%	15.21%	23.22%	19.9%
SRS of offenders & non-offenders	-	34.5%	36.5%	27.87%
SRS of only offenders	-	44.5%	42.19%	38.81%
Supervised learning				
Logistic Regression	58.42%	32.84%	31.28%	34.76%
Naïve Bayes	55.24%	46.15%	32.05%	41.3%
SVM	64.75%	58.02%	49%	52.26%
Random Forest	86.6%	89.18%	84.92%	77.31%
Random Forest with district	87.84%	86.19%	81.99%	76.96%
Random Forest with intervention	85.08%	82.29%	77.83%	73.08%

Note: Accuracy is calculated as average of 1000 independent sampling without replacement iterations for SRS, and 10-fold cross-validation for supervised learning.

Appendix

Figure A1: Tuning of random forest *mtry* parameter (number of variables randomly sampled at each split) for a range of *ntree* (number of trees to grow).

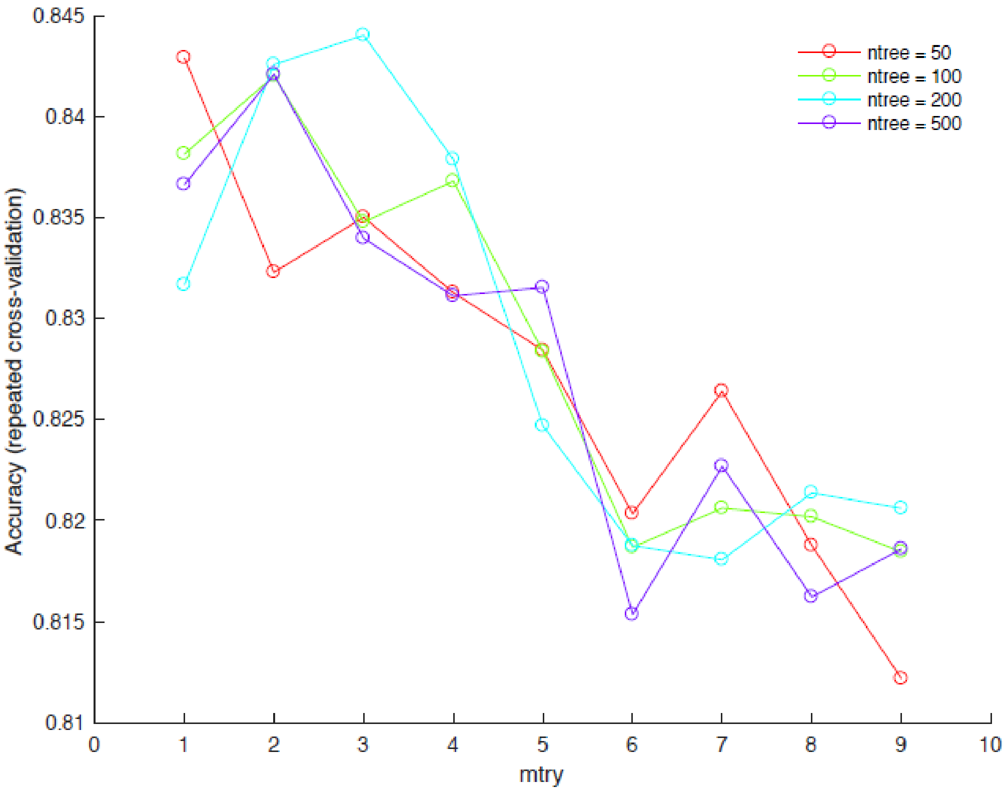
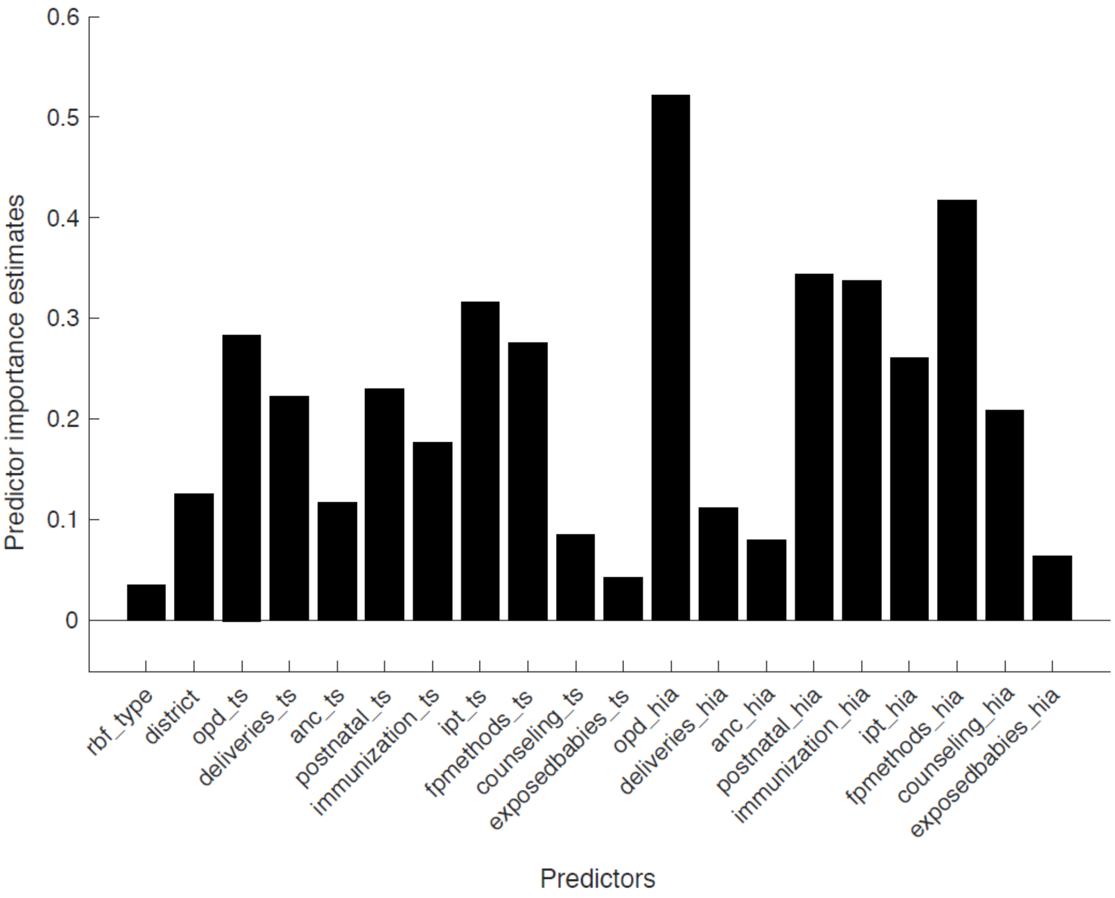
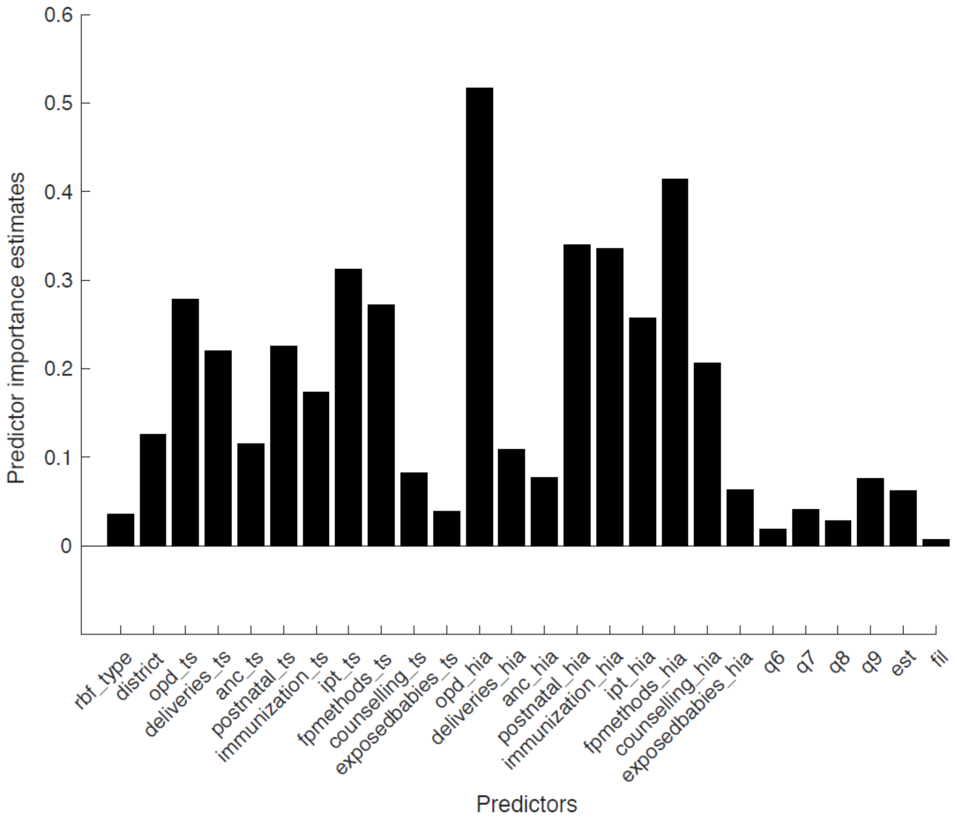


Figure A2. Feature importance criteria for random forest classification with basic set of variables



Note: "ts" refers to tally sheet; hia refers to Health Information Aggregation 2 forms.

Figure A3. Feature importance criteria for random forest classification with expanded set of variables



Note: “ts” refers to tally sheet; hia refers to Health Information Aggregation 2 forms.

Q6=type of facility (7 levels plus “other,” from central hospital to rural health post)

Q7=managing authority (government, mission/FBO, private, military, other)

Q8=location (rural, peri-urban, urban)

Q9=size of catchment population

est=total number of established staff posts

fill=total number of filled staff posts

Table A1. Rewarded indicators in Zambia's 2012-2014 pilot PBF

#	Indicator	Fee (Kwacha)	Fee (USD*)
1	Curative Consultation	1,000	0.20
2	Institutional Deliveries by Skilled Birth Attendant	32,000	6.40
3	ANC prenatal and follow up visits	8,000	1.60
4	Postnatal visit	16,500	3.30
5	Full immunization of children under 1	11,500	2.30
6	Pregnant women receiving 3 doses of malaria IPT	8,000	1.60
7	FP users of modern methods at the end of the month	3,000	0.60
8	Pregnant women counselled and tested for HIV	9,000	1.80
9	HIV Exposed Babies administered with Niverapine and AZT	10,000	2.00

* Approximate value. Source: [8].

Table A2. Prediction accuracy performance of different approaches with expanded set of variables

Approach	Prediction of over-reported event			
	Q1	Q2	Q3	Q4
Logistic Regression	49.31%	28.64%	26.49%	29.14%
Naïve Bayes	49.97%	40.58%	33.16%	38.68%
SVM	58.38%	52.97%	45.74%	49.82%
Random Forest	82.72%	79.38%	74.26%	71.16%