WORLD BANK GROUP
Health, Nutrition & Population

2018 SKILLS BUILDING PROGRAM

BIG DATA, ARTIFICIAL INTELLIGENCE AND DECISION SCIENCE IN HEALTH AND NUTRITION

USING SUPERVISED LEARNING TO SELECT AUDIT TARGETS IN PERFORMANCE-BASED FINANCING IN HEALTH: AN EXAMPLE FROM ZAMBIA

Jed Friedman

In partnership with

datasciencedojo — data science for everyone

Burnet Institute — Medical Research. Practical Action.

UCL

techemergence symphony

University of Copenhagen

iDSI — Better decisions. Better health.

Machine Intelligence Institute of Africa — *Transform Africa through Machine Intelligence*

GHCC — Global Health Cost Consortium

UNDP

Bill & Melinda Gates foundation

What is Performance-based Financing (PBF)?

- Contracting mechanism that aims to increase the performance and quality of service providers.
- Offer financial incentives to health care facilities for provision of services
- Bonus payment based on a broad measure of quality

Monitoring of Performance-based Financing

- 3 layers of verification:
  - District or provincial supervisors visit all facilities on monthly or quarterly basis to confirm the accuracy of the reported quantity data.
  - District teams visit all facilities on a quarterly basis to complete a quality assessment.
  - Independent third-party conducts quarterly counter-verification visits to a sample of facilities.
- Aids in detection and determent of misreporting through random sampling of providers.
- Targeting of facilities varies from program to program, and has varied associated costs.

Zambia's performance-based financing pilot

2012-2014

WORLD BANK GROUP
Health, Nutrition & Population

Zambia's Performance-based Financing Pilot

- To realign health financing towards outputs rather than inputs
- To address various health system concerns such as relatively low coverage of key maternal and child health services
- Pilot operated in public health centers in 10 rural districts, covering a population of 1.5 million (11% of Zambia's population)
- 2 core features: financial rewards and equipment upgrades.

Zambia's Performance-based Financing Pilot

- Varying fee-for-service bonus payments for indicators measuring the quantity of 9 maternal and child health, and 10 structural and process quality domains.
- Health centers also received emergency obstetric care equipment.
- Participating health centers were subject to enhanced monitoring.
- Substantial financial rewards.

Zambia's Performance-based Financing Pilot

- Independent population surveys found gains in selected targeted indicators, such as rate of facility deliveries.
- Targeted indicators at already high levels of coverage saw little change (e.g. ante-natal care).
- Extensive auditing of reported data through continuous internal verification and a one-off external process.

Evaluating the performance of different classification methods

Using data from the Zambia PBF pilot

Data from Zambia PBF pilot

- Combined from facility reports and a dedicated facility survey (reproduction of external verification activities)
- 140 facilities: 105 primary health care centers in the 10 PBF pilot districts and 35 centers in another 8 non-pilot districts.
- Verification data were collected on the complete set of 9 incentivized indicators, for every calendar month of 2013.

Measurement of over-reporting

- Binary measure equal to 1 if bonus payment based on the reported vs. verified data is ≥ 10% of the reported value.

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

Classification Methods

Sampling-based approaches (overall sample size: 28):

- 50% clinics chosen at random
- Stratify by district, then select 50% clinics at random
- Random 50% of clinics that over-reported in prior quarter, plus random 50% from the remaining clinics
- Select up to 28 clinics that are prior offenders, randomly sample from remaining facilities to achieve target number.

Accuracy of sampling-based approaches reported as averages of 1000 independent sampling iterations without replacement

Classification Methods

Alternate approaches (including supervised machine learning):

- Naïve Bayes
- Logistic Regression
- Support Vector Machines
- Random Forest

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

Supervised machine learning for PBF data

- For verification in PBF:
  - Input: subset of facility-specific data points
    - 9 quantity measures that were rewarded in the RBF program
    - District identifier
    - Categorical variable indicating treatment arm from related audit experiment
  - Output: binary indicator for whether or not a facility over-reported
- Algorithm learns which facilities are at risk of over-reporting.
- Applies this learning to predict this risk for other facilities not included in the training data.

Naïve Bayes

- Calculates the probability of an input (or specific set of predictive features) belonging to each class (over-reported, or not), and then chooses the class with the highest score.
- Assumes strong independence between these predictive features. Correlations between features, if any, are disregarded.

Logistic Regression

- Uses a logistic function at its core to estimate a relation between the binary classification (over-reported or not) and its possible predictors.
- Assumes that the input space can be partitioned by a linear boundary, separating the data into two classes

Support Vector Machine

- Defined by a hyperplane that maximizes the separation between the two classes.
- 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 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

- Averages 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

Choosing the appropriate machine learning technique

- Size of training data set
- If there is a need to learn interactions between the various features or whether can they be treated as independent variables
- Whether additional training data may become available in the future and would need to be easily incorporated into the model.
- Whether the data is non-parametric and not linearly separable.
- Whether overfitting of the model to the training data is expected to be a problem.
- Requirements in terms of speed, performance and memory usage.

Choosing the appropriate machine learning technique

- Small training sets: use Naïve Bayes. Logistic Regression has tendency to overfit.
- Larger training sets:
  - Roughly linear data features: Logistic Regression. Robust to noise, can avoid overfitting, allows updates. Also can give probability output (instead of classification).
  - Non linearly separable: Support Vector Machines (SVMs). High accuracy, works with high dimensional spaces, avoids overfitting. Cons: Memory intensive, hard to interpret, challenging to tune for optimal performance.

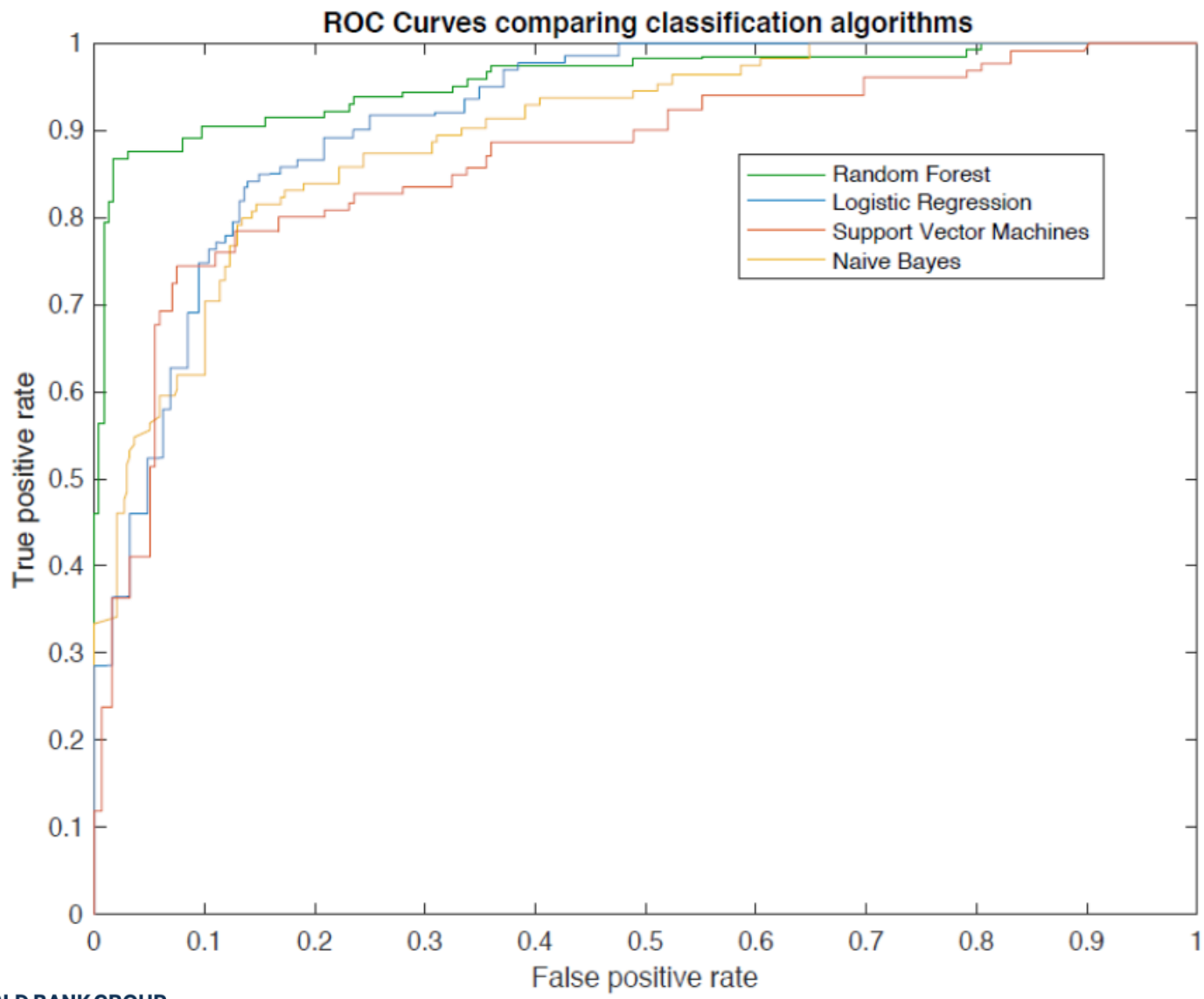
Advantages of tree ensemble-based learning methods

- Do not expect linear features or even features that interact linearly (unlike with Logistic Regression)
- Handle high dimensional spaces as well as large number of training examples (advantage over SVMs)
- Random Forest methods:
  - Are fast and scalable (unlike SVMs)
  - Avoid overfitting
  - Don’t require tuning of parameters

Analysis of methods: Performance metrics

- Prediction accuracy
- F-score
- Area under the ROC
- Average precision rate
- Root mean squared error (RMSE)

Results



WORLD BANK GROUP
Health, Nutrition & Population

Results

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).

Results

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.

Conclusions

- Over-reporting is a highly non-linear function of covariates
- Predictions from traditional regression analysis will not be particularly accurate
- Supervised learning approaches, such as Random Forest, could substantially improve the prediction accuracy of counter-verification in PBF
- Hence also increase the cost-effectiveness of verification.
- These methods are operationally feasible, especially in settings with electronic routine reporting systems