Using data science to predict key indicators in HIV programs: Is it possible, and if so, how can we apply our learnings to better understand program performance?

vantage
health technologies

A part of the
BROADREACH™
group

Vantage Health Technologies creates solutions to the world's most complex health challenges. We provide decision support, operational tools and step-by-step workflows to empower healthcare workers across the spectrum to achieve predictable, cost-effective and improved health outcomes - at scale.

Client
Large PEPFAR funded HIV program in Africa

Technology Partner
Microsoft

The Challenge and Our Approach

Understanding performance on large-scale, multi-year, multi-faceted and geographically dispersed programs is complex as comparing partners, districts, and facilities based on pure achievement to target can be fraught: Their circumstances vastly differ from each other and there are external factors that impact on their ability to perform.

This is a topic that we have heard many of our clients debate, and we therefore set out to explore ways in which we can use data and technology to better understand performance of selected key indicators. We wanted to see if we could accurately predict the performance of these indicators; if we could determine what impacts on their performance; and ultimately, if the findings could be used to improve program performance.

This case study serves to document our process and findings. We hope it will trigger a conversation around how we use data and technology to better understand and improve performance in large and complex HIV treatment programs.

Our Research Questions

We applied data science methodologies, specifically predictive analytics, to look at performance through a different lens. The idea was to assess whether these kinds of predictions could serve to measure performance without bias and to highlight areas for attention.

The following questions helped guide our exploration

1) Can we estimate or predict the values of selected key performance indicators?
2) If yes, how well can we do it?
3) What are the essential ingredients that allow us to make such a prediction?
4) What can we learn from this?

Our Dataset and Indicators

We used a publicly available PEPFAR data set[1] containing the values of 40 indicators for every quarter starting from 2019-Q1 and ending 2022-Q1. This data set covers 10 quarters in total.

We are interested in forecasting or predicting the following indicators:

- HTS_TST: Number of individuals who received HIV Testing Services (HTS) and received their test results
- HTS_TST_POS: Number of individuals who received HIV Testing Services (HTS) and received their test results, and found positive
- TX_NEW: Number of adults and children newly enrolled on antiretroviral therapy (ART)
- TX_NET_NEW: The quarterly net increase or decrease in ART patients
- TX_CURR: Number of adults and children currently receiving antiretroviral therapy (ART)

[1] PEPFAR Monitoring, Evaluation, and Reporting Database https://mer.amfar.org/

Can we estimate or predict the values of selected key performance indicators?

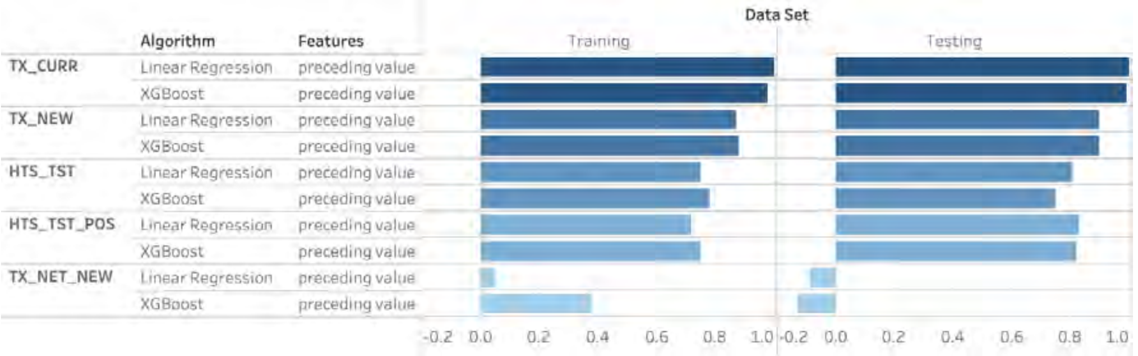
**Refer to the Glossary overleaf for an explanation of the technical terms used.*

We used forecasting as our experiment design. We divided the data set into training and testing sets.

- Training set: everything up to and including 2021-Q3.
- Testing set: 2021-Q4 & 2022-Q1.

We then applied two methods: Linear Regression model and Gradient Boosting (XGBoost) algorithm. We trained on the training data set and then made the predictions on the testing data set.

Method 1: Forecasting using only historical values of the considered indicator.
The table below reports the confidence values on the training and testing data.



In Summary

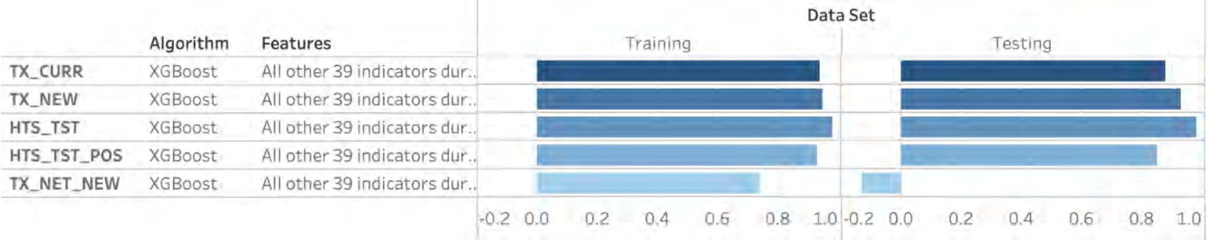
There are some indicators whose value may be well predicted from their previous values, and there are some indicators for which the previous value cannot serve as a predictor using this approach (TX_NET_NEW).

There are some signs of overfitting for the TX_NET_NEW indicator using Gradient Boosting. Some further tuning may allow us to reduce overfitting and make better predictions.

Can we improve the predictions?

We again split our data into training and testing sets, but here we ask a different question: How well can we predict the indicator's value during the considered quarter using the values of the other 39 indicators during the same quarter?

Method 2: Forecasting using the values of the other 39 indicators during the same quarter
The table below reports the confidence values on the training and testing data. It also shows the most important indicator for predicting the considered indicator according to the various methods applied.



In Summary

Predictions improved by forecasting using other indicators during the same quarter.

Our TX_NET_NEW indicator remains less predictable than other indicators. This prompted further investigation into predicting this indicator, as it is one of the keyways of measuring performance.

A part of the

Is there another way to predict TX_NET_NEW?

With the key indicators identified, we applied the same methodologies - as well as Random Forest - with the aim to predict the TX_NET_NEW indicator. The findings are shown below:

Gradient Boosting
R(training data) = 0.88
R(testing data) = 0.63

Linear Regression
R(training data) = 0.42
R(testing data) = 0.44

Random Forest
R(training data) = 0.97
R(testing data) = 0.66

Polynomial regression, including the 2nd-order terms, arrives at a correlation of an order R=0.50 and includes the 3rd-order terms at R=0.58.

In Summary

Linear regression proved to be a viable method in predicting TX_NET_NEW, however, the results could be strengthened by considering other factors, such as concentrating our data sets.

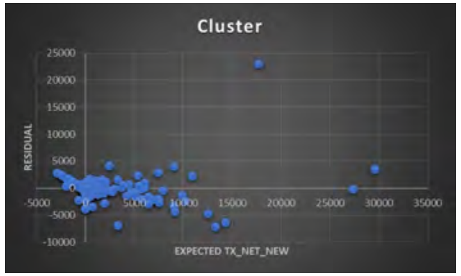
Conclusion

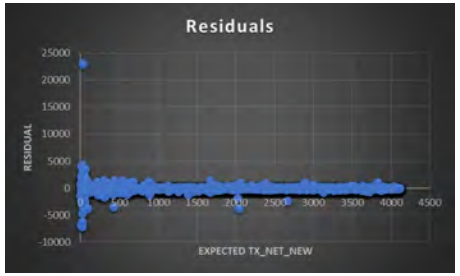
This exercise showed that some indicators are ‘better’ at being predicted than others. It was clear that there is a sweet spot in terms of which level the analysis is applied: This is in relation to the time-period as well as the geographic level selected. Although it is not a perfect science, it is felt that the results can be used by HIV treatment programs to inform programmatic discussions and decision-making. Using predictions to highlight areas that need attention can support proactive- and focused intervention. Comparing predicted values to actual values can also help identify outliers and be used to interrogate other elements that are important for driving improved performance.

What do you think? What other suggestions do you have for using data science and technology to better understand - and ultimately improve - HIV program performance?

How is this prediction affected by the geo-level?

With the predictive formula generated by the linear regression, we set out to apply it at various geo-levels. From all levels, it was evident that the best results were seen at the Cluster level.





Technical Glossary

- **Training data set:** An initial dataset that is used to teach a machine learning model
- **Testing data set:** A secondary (or tertiary) data set that is used to test a machine learning model after it has been trained.
- **Regression analysis:** A set of statistical methods used for the estimation of relationships between a dependent variable and one or more independent variables.
- **Linear Regression:** A model that estimates the relationship between one independent variable and one dependent variable using a straight line
- **Gradient Boosting:** A machine learning technique used for regression analysis that relies on the intuition that the best possible next model, when combined with previous models, minimizes the overall prediction error.
- **Random Forest:** A machine learning technique that combines the output of multiple decision trees to reach a single result
- **Polynomial regression:** A machine learning model that can capture non-linear relationships between variables
- **Confidence value:** Refers to the statistical R2 value in the applied methods.
- **Residuals:** The residual for each observation is the difference between predicted values of the dependent variable and observed values of the variable.