IDA

INSTITUTE FOR DEFENSE ANALYSES

Forecasting with Machine Learning

DATAWorks 2022

Akshay A. Jain
John W. Dennis

March 2022
Approved for public release;
distribution is unlimited.
IDA Document NS D-33017
Log: H 22-000103

INSTITUTE FOR DEFENSE ANALYSES
730 East Glebe Road
Alexandria, Virginia 22301

IDA

The Institute for Defense Analyses is a nonprofit corporation that operates three Federally Funded Research and Development Centers. Its mission is to answer the most challenging U.S. security and science policy questions with objective analysis, leveraging extraordinary scientific, technical, and analytic expertise.

About This Publication
The work was conducted by the Institute for Defense Analyses (IDA) under CRP C6608.

For More Information:
Dr. John W. Dennis III, Project Leader
jdennis@ida.org, 703-845-2166
ADM John C. Harvey, Jr., USN (ret) Director, SFRD
jharvey@ida.org, 703-575-4530

Rigorous Analysis | Trusted Expertise | Service to the Nation

Executive Summary

The U.S. Department of Defense has a considerable interest in forecasting key quantities of interest, including demand signals, personnel flows, and equipment failure. Many forecasting tools exist to aid in predicting future outcomes, and there are many methods to evaluate the quality and uncertainty in those forecasts. When used appropriately, these methods can facilitate planning and lead to dramatic reductions in costs.

This Institute for Defense Analyses presentation explores the application of machine learning (ML) algorithms to forecasting. ML tools and methods can be used to address increasingly common scenarios that trouble traditional forecasting methods, such as model selection and high dimensional covariate spaces. However, ML is not a panacea, as its methods are not appropriate in many situations, and understanding which methods apply to which questions remains an important issue. We conclude with an example that illustrates the use of gradient-boosted trees to forecast Air National Guard personnel retention.

This page is intentionally blank.

Forecasting with Machine Learning

Akshay Jain
John Dennis

April 28, 2022

Institute for Defense Analyses
730 E. Glebe Rd. Alexandria, VA 22305

Why is forecasting important?

Anticipate Demand

Reduce Costs

Prepare Safeguards

Assess Risk

Save Lives

IDA 3

Presentation Outline

Background on Forecasting:
What is forecasting?
How do we do it?

Application of Machine Learning (ML) to Forecasting:
What are some challenges of forecasting?
How can ML address challenges of forecasting?
How can we apply ML concepts to evaluate forecasts?

Case Study:
Applications of gradient-boosted trees to forecast the number of annual school slots required each year for the Air National Guard.

IDA 4

What is forecasting?

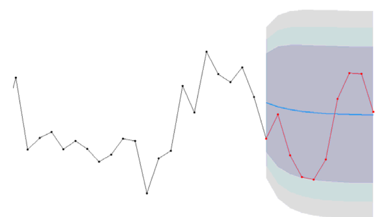
<u>Predicting future values based on past and current data</u>

Often implies the existence of a <u>time series</u> component to the data and an interest in estimating the unknown future value(s) in that data

Forecasting is not always appropriate for policy analysis and usually is not appropriate for causal inference

Possible forecasting problems:

- How many soldiers will leave service next year?
- When will a critical aircraft component part fail?
- When should I buy plane tickets?
- When will a country's president step down?
- What will gas prices be next month?



IDA 5

Forecasting Tools and Methods

Parametric

Uses a structure from a model with data

Nonparametric

Only uses data

ARIMA
GARCH

Parametric Kernels

Structural models

Neural Networks

Non-parametric Kernels

Trees and Forests

IDA 6

What are some common challenges of forecasting?

IDA | 7

How can ML address challenges of forecasting?

Model Selection and Overfitting	Regularization, Cross Validation, Ensemble Models, Sample Splitting
Many Covariates	Dimensionality Reduction, Interaction Detection
Forecast Uncertainty	Interval Forecasts and Quantiles, Density Forecasts
Distribution Drift	Depends on the Situation

IDA | 8

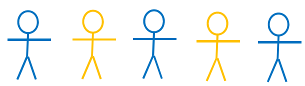
How good is my forecast?

Sample Splitting - Evaluate performance on unseen data

Split the sample over identifiers

ID	T 1	2	3	4	5
1	data	data	data	data	data
2	data	data	data	data	data
3	data	data	data	data	data
4	data	data	data	data	data
5	data	data	data	data	data
6	data	data	data	data	data
7	data	data	data	data	data

training sub-sample   evaluation sub-sample



Split the sample over time periods

ID	T 1	2	3	4	5
1	data	data	data	data	data
2	data	data	data	data	data
3	data	data	data	data	data
4	data	data	data	data	data
5	data	data	data	data	data
6	data	data	data	data	data
7	data	data	data	data	data

training sub-sample   evaluation sub-sample

IDA 9

IDA's Finite-Interval Forecasting Engine (FIFE) uses ML to estimate individual survival probabilities

FIFE forecasts the probability of individuals in period *t* remaining in the dataset h periods into future, conditional on prior survival

Classification problem (survival vs. non-survival) for interval of length h

No need to specify interactions ahead of time

Open source https://github.com/IDA-HumanCapital/fife https://pypi.org/project/fife/

LightGBM TensorFlow

IDA 10

Case Study: Retention Prediction for the Air National Guard (ANG)

ANG asked IDA to develop and apply a person-level Retention Prediction Model (RPM) to reduce inefficiencies in forecasting training school slot demand

Approach: We use a gradient-boosted tree model (from FIFE) to forecast individual retention probabilities

Performance: We test our method by forecasting exits from 2016 to 2020, and comparing predictions to observed exits

Correctly predicted **76.7%** of all exit-survival observations,

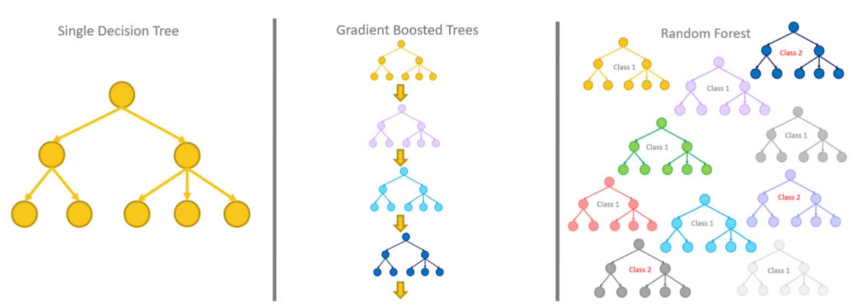
Achieve **51.8%** precision and **70.6%** recall

A closer look at gradient boosted trees

Gradient boosted tree (GBT) models are a collection of many decision trees

As opposed to random forest models, GBT models build trees successively to correct errors/residuals from preceding trees

Can set hyperparameters to improve model performance



We use 638 total features from DOD and public source data

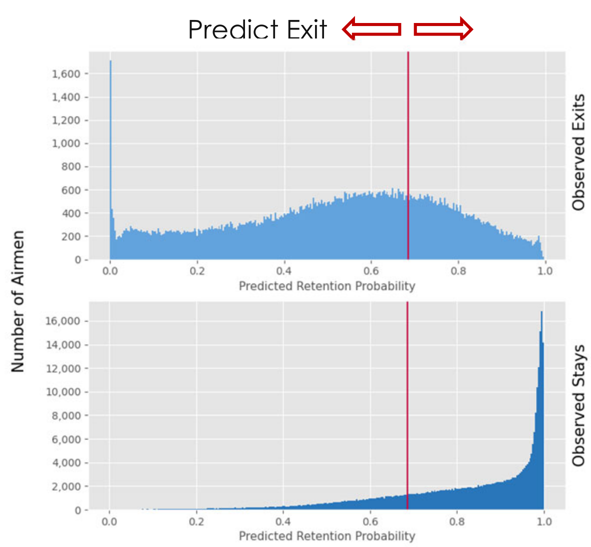
Activations	Career History	Family Life
Demographics	Drilling and Training	Pay and Bonuses
Deployments	External Labor Market	Trauma

	Evaluation Model	Forecasting Model
Training Set Years	2005-2014	2005-2019
Validation Set Years	2005-2014	2005-2019
Test Set Years	2015	2020
Forecasting Period	2016-2020	2021-2025

Hyper-optimized model (200 trials, 25% validation share)

IDA 13

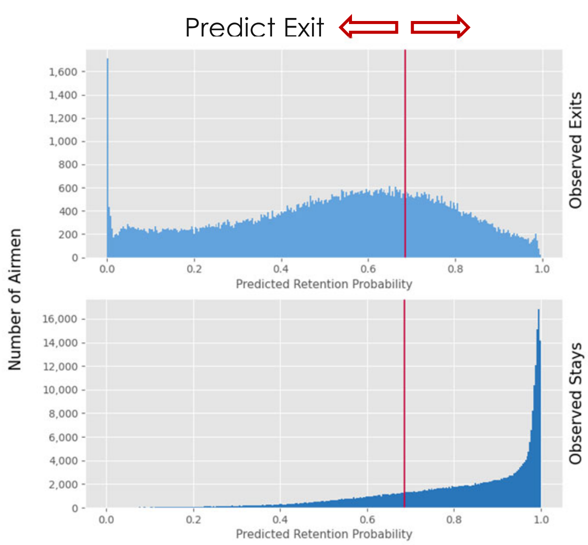
Visual Confusion Matrix: All Years Pooled



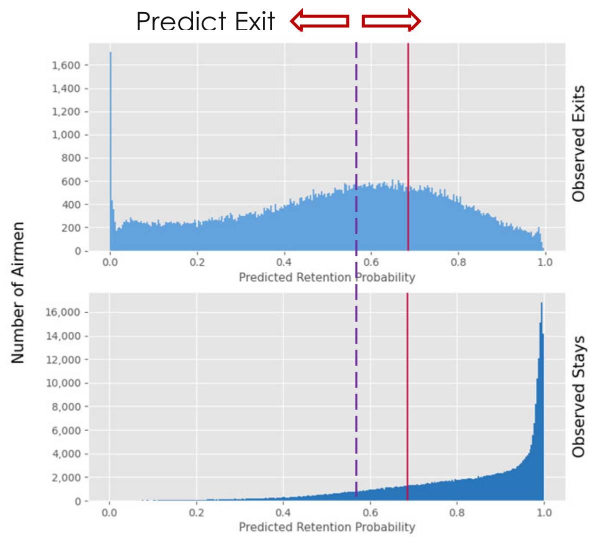
IDA 14

Visual Confusion Matrix: All Years Pooled

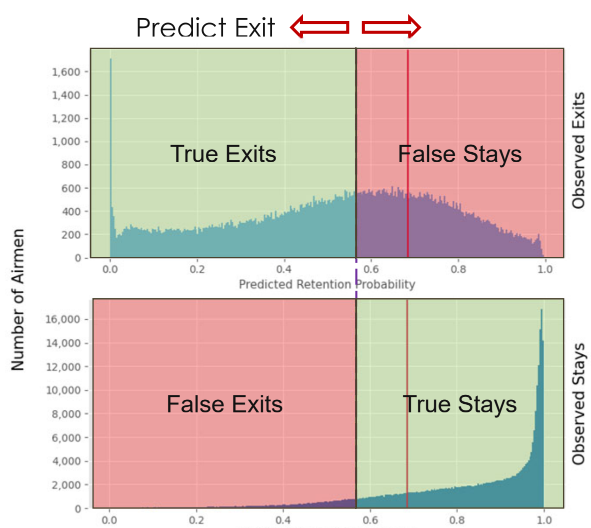
What if we move the line?



What if we move the line?



We generate more "True Stays" and more "False Stays"



Forecasting offers tremendous advantages!

Anticipate Demand | Reduce Costs | Prepare Safeguards | Assess Risk | Save Lives

IDA

Akshay Jain
ajain@ida.org

John Dennis
jdennis@ida.org

Image Sources

https://www.google.com/url?sa=i&url=https%3A%2F%2Ftowardsdatascience.com%2Fachieving-stationarity-with-time-series-data-abd59fd8d5a0&psig=AOvVaw1gRYxi3gIwGCLEWqioSXz5&ust=1646166057519000&source=images&cd=vfe&ved=0CAkQjhxqFwoTCJibw8qco_YCFQAAAAAdAAAAABAZ

https://cdn.freelogovectors.net/wp-content/uploads/2018/07/tensorflow-logo.png

https://community.rstudio.com/uploads/default/original/3X/b/3/b33220c9ecb0627a0185cb7f8feaba008a4f5ed8.png

https://medium.com/analytics-vidhya/ensemble-models-bagging-boosting-c33706db0b0b

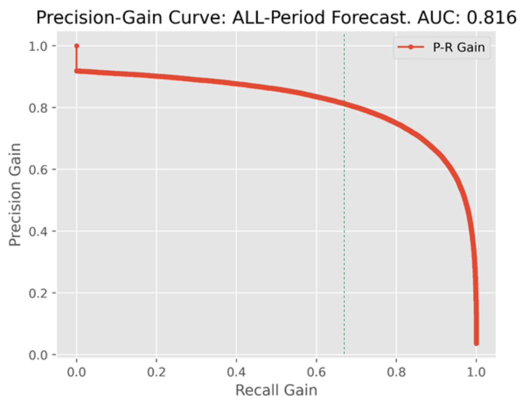
Appendix

Determining decision threshold with precision-recall calibration

Precision-recall calibration is especially useful in cases of class imbalance

We set “exits” to be the positive class

Optimal Decision Threshold is 0.685



$$\text{Precision} = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

$$\text{Recall} = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

$$F1\ Score = 2 * \frac{Recall\ * Precision}{Recall + Precision}$$

IDA 23