Areté Associates
Influenza Forecasting Discussion

CDC Influenza Forecasting Workshop

30-31 August, 2018

K. Apfeldorf

E. Luckwald

C. Neisinger

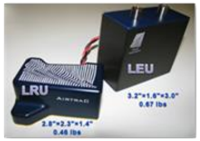
K.Heinzelmann

Areté Associates – Who We Are

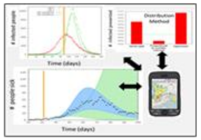
- Traditionally involved with ‘hard’ physics based detection and tracking problems
- Occasionally branch out into other hard, data driven problem spaces (health, genomics,…)

Modeling & Simulations

Low SWaP Sensors



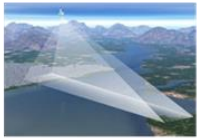
Networked Surveillance



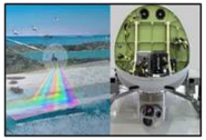
Weak Signals Intelligence Applications

Real-Time Processing

Remote Sensing

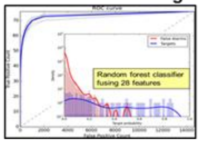


Systems Integration

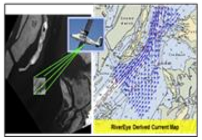


Production

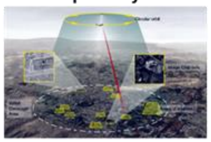
Machine Learning



Field Tests & Measurements



Deep Analytics



Rapid Prototyping

- Employee-Owned Small Business
- 260 Employees
- 75% Advanced Degrees
- Eight US Locations
- > 40 years of government experience

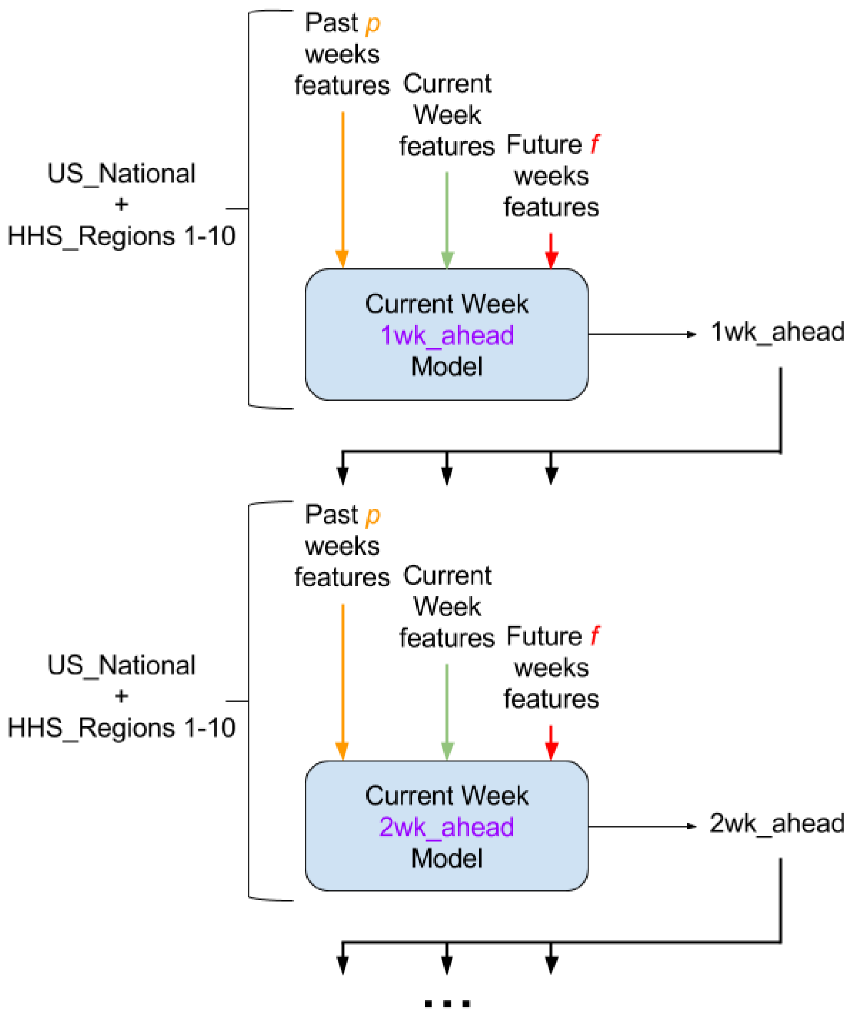
Physics → Phenomenology → Concepts → Algorithms → Solutions

Our Approach

- Adapted a scikit-learn ExtraTrees regressor model originally developed as a component for the Dengue Forecasting Challenge
- Due to limited resources:
  - Used mostly data available on FluView portal
  https://gis.cdc.gov/grasp/fluview/fluportaldashboard.html
  - Forced to focus on “transferable” aspects of problem
  - Did not tailor our predictive model to the current competition metrics
  - Got a late start

Resulted in quite a few instances of achieving the worst possible score

Model Architecture



- Chain of ExtraTrees Regressors
- Predictions of each regressor feed in as a feature to the next
- State model is almost identical, but with additional features to account for geographic proximity of populations (“neighbor-weighted ILI”)

Insights and Lessons Learned

- Machine Learning (ML) models can
  - Find correlations ignored by traditional models
  - Ingest more data and give insights into what is important
- ML models can not (typically)
  - Predict an event that's never been seen before
  - Tell you when it's taking a wild guess
- Data used in testing must be similar to data used in training
  - Training on "corrected" data can hurt you if you're testing on "uncorrected" data

Insights and Lessons Learned

- Machine Learning (ML) models can
  - –
  - – important

In Backups, we suggest a number of seemingly available datasets that may be useful for this problem. Suggest creating a way for collaborators to share data and methods for acquisition

- ML models can not (typically)
  - Predict an event that's never been seen before
  - Tell you when it's taking a wild guess
- Data used in testing must be similar to data used in training
  - Training on "corrected" data can hurt you if you're testing on "uncorrected" data

Insights and Lessons Learned

- Machine Learning (ML) models can
  - Find correlations ignored by traditional models
  - Ingest more data and give insights into what is important
- ML models can not (typically)
  - Predict an event that's never been seen before
  - Tell you when it's taking a wild guess
- Data used in testing must be similar to data used in training
  - Training on "corrected" data can hurt you if you're testing on "uncorrected" data

Insights and Lessons Learned

- Machine Learning (ML) models can
  - Find correlations ignored by traditional models
  - Ingest more data and give insights into what is important
- ML models can not (typically)

Suggest using a hybrid model (e.g., combine ML model with Seasonal ARIMA model), depending on ML model's "certainty" (see later slides)

- Data used in testing must be similar to data used in training
  - Training on "corrected" data can hurt you if you're testing on "uncorrected" data

Insights and Lessons Learned

- Machine Learning (ML) models can
  - Find correlations ignored by traditional models
  - Ingest more data and give insights into what is important
- ML models can not (typically)
  - Predict an event that’s never been seen before
  - Tell you when it’s taking a wild guess
- Data used in testing must be similar to data used in training
  - Training on “corrected” data can hurt you if you’re testing on “uncorrected” data

Insights and Lessons Learned

- Machine Learning (ML) models can
  - Find correlations ignored by traditional models
  - Ingest more data and give insights into what is important
- ML models can not (typically)
  - Predict an event that's never been seen before
  - Tell you when it's taking a wild guess
- Data used in testing must be similar to data used in trai
  - Trai ou're testing on "uncorrected" data

Suggest creating a version-controlled database so that teams can pull "uncorrected" data from an earlier date

Areté Associates

Useful Ideas and Tools

- NOCATS Library
  - Categorical splits for tree-based learners
  - Developed internally at Areté (J. Blackburne) as extension to scikit-learn
  - https://github.com/scikit-learn/scikit-learn/pull/4899

- Considerable thought on “self-evaluation” of ML algorithms
  - Where/how they can indicate when they are uncertain about their predictions.
  - Where/how they indicate when they are seeing something unfamiliar
  - Could be used to weight ML models vs. epidemiological models

Towards a Forecasting Solution

Ingredient	Comment
MODEL	
Hybrid model with fully data-driven & more structured components	Fully data-driven model highly predictive for novel season, can fail otherwise
Optimization using metrics defined by stakeholders (here, in challenge)	
Epidemiologically important features	Take inspiration from epi-models; Include spatial, temporal, categorical
DATA	
Availability of timely, relevant sources	Weather, traffic, …
Handle erroneous/incomplete data	
DECISION-SUPPORT	
Explain & interpret results	Feature importance, analysis
Explore impact of interventions	Supporting simulation capability

Backups

NOCATS Library

- Traditionally, branching methods like Random Forests have difficulty with categorical variables with modest to large number of categories (2^n possible splits)
  - Standard methods include brute force, artificial order, one-hot encoding
  - NOCATS extends the extra trees functionality to allow categoricals by splitting randomly on a possible split
  - Developed internally at Areté (J. Blackburne) as extension to scikit-learn
  - https://github.com/scikit-learn/scikit-learn/pull/4899

Recommendations – Data Sets

In which we recommend data sets we think could be useful for Machine Learning algorithms used in Flu Forecasting

These datasets come from consideration of what would affect terms in a compartmental model. For ML model, we skip the compartmental model and ingest the data directly

Potentially Useful Data Sets

- ILI values as they were believed to be *at the time of forecast*
  - Training a ML algorithm on corrected data sets biases the algorithm when using uncorrected data in current forecasts
  - Ideally, we want the data as they were believed at the time that an algorithm would be expected to make a forecast.
  - Suggest version controlling the FluView data and allowing teams to grab versions with different timestamps.

Potentially Useful Data Sets

- Some common external time series data could prove useful, provided it indicates a serious departure from the seasonal norm.
- For example, severe cold weather, large atypical demographic changes (see next few slides)
- The data sets mentioned on the following slides are available, but maybe not in a timely manner
- Suggest creating a way for collaborators to share data and methods of acquisition

Potentially Useful Data Sets

- Historical weather data and forecasts
  - Temperature: Flu virus is more infectious in cold weather
  - Humidity: Drier mucous membranes prevent the body from fighting off respiratory infections.
  - Cloud cover: Vitamin D production affects the immune system and UV radiation kills exposed viruses.

- Air Quality
  - Affects immune system and data should be available through EPA

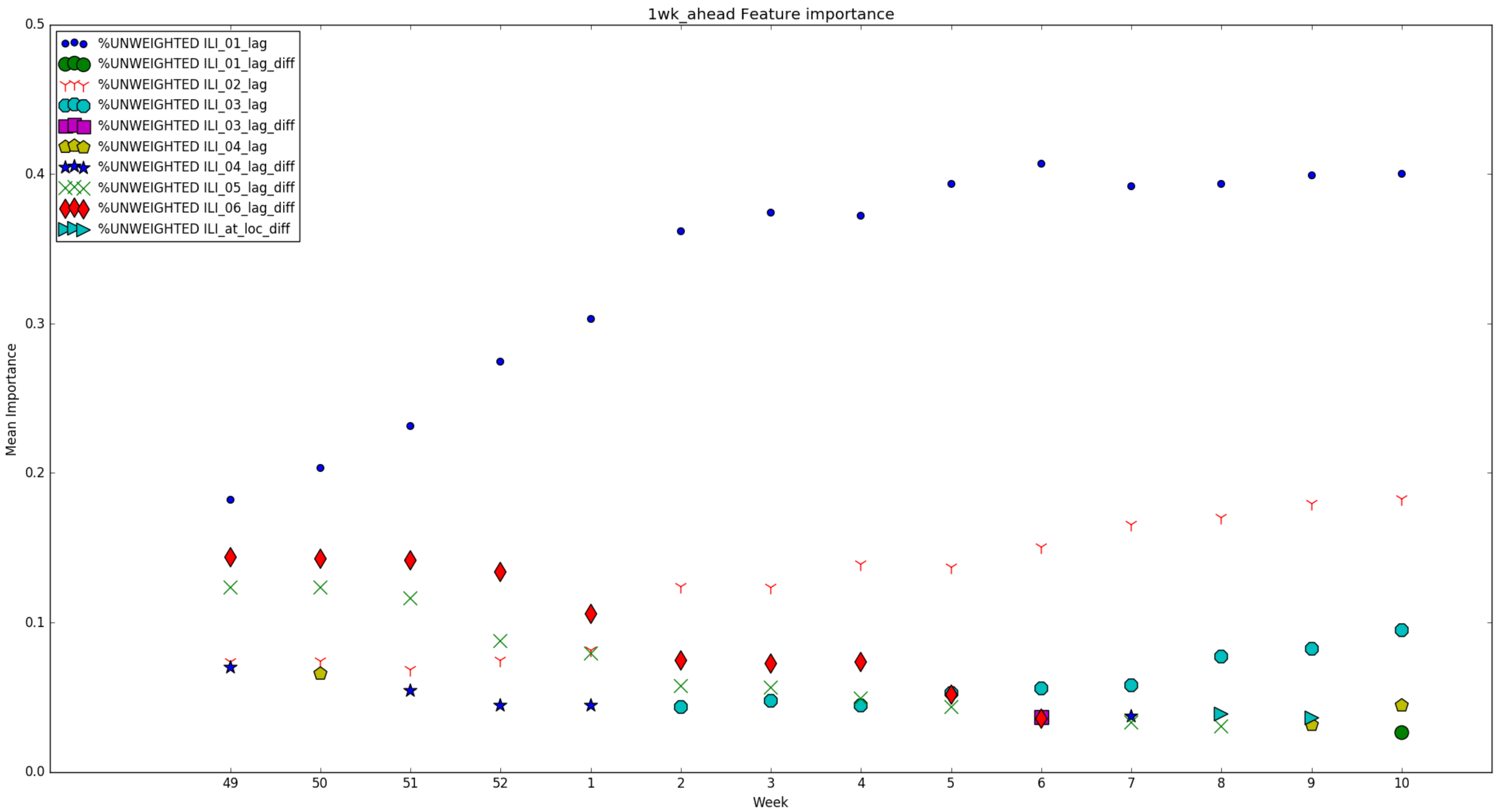
Potentially Useful Data Sets

- Age demographics (Census Bureau)
  - Very old and very young most susceptible
  - Demographics at State/County/City level
- Vaccination
  - Rate and Availability of vaccination at the State/County/City level (or finer)
- Travel Rates (Dept. of Transportation)
  - Flight statistics linking cities/states

Results: Feature Importances Over Time

Summary: Most important feature for determining ILI for a given week is the (realized or predicted) ILI of the preceding week.

Features vs. Time – State Predictions



Areté Associates

Features vs. Time – National/Regional Predictions