A Paradigm for Assessing the Scope and Performance of Predictive Analytics

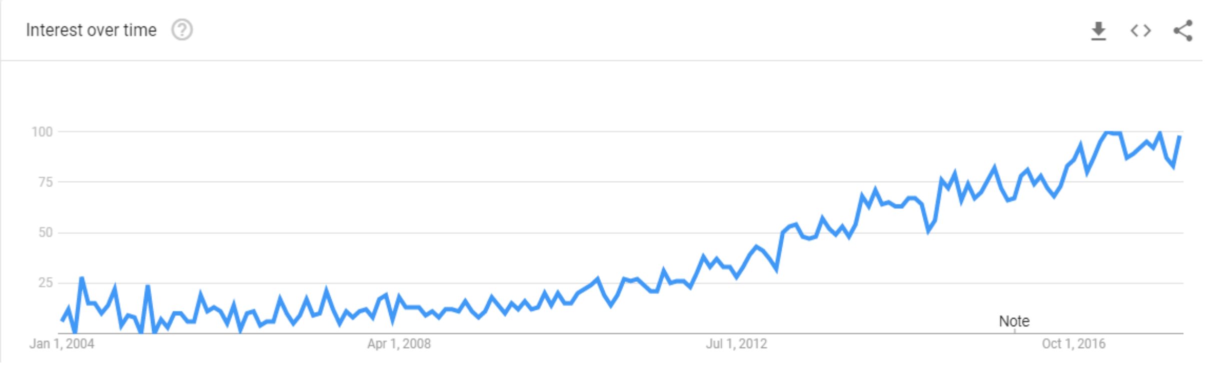
Jeff Prince

Kelley School of Business

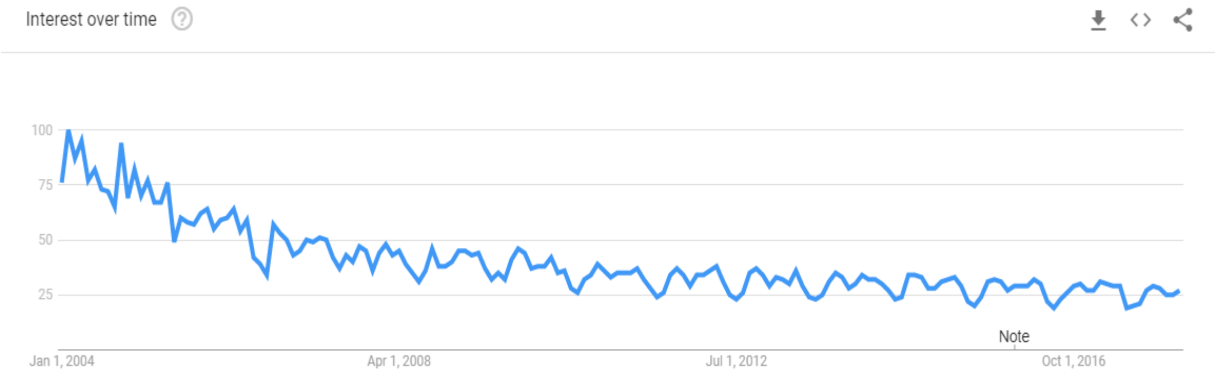
Indiana University

Terms and Definitions Matter

- Econometrics is as important as ever
  - Many of the questions we want answered require econometric analysis



- However, econometrics is losing interest, largely due to semantics
  - Econometrics vs. Predictive Analytics…



What is Prediction?

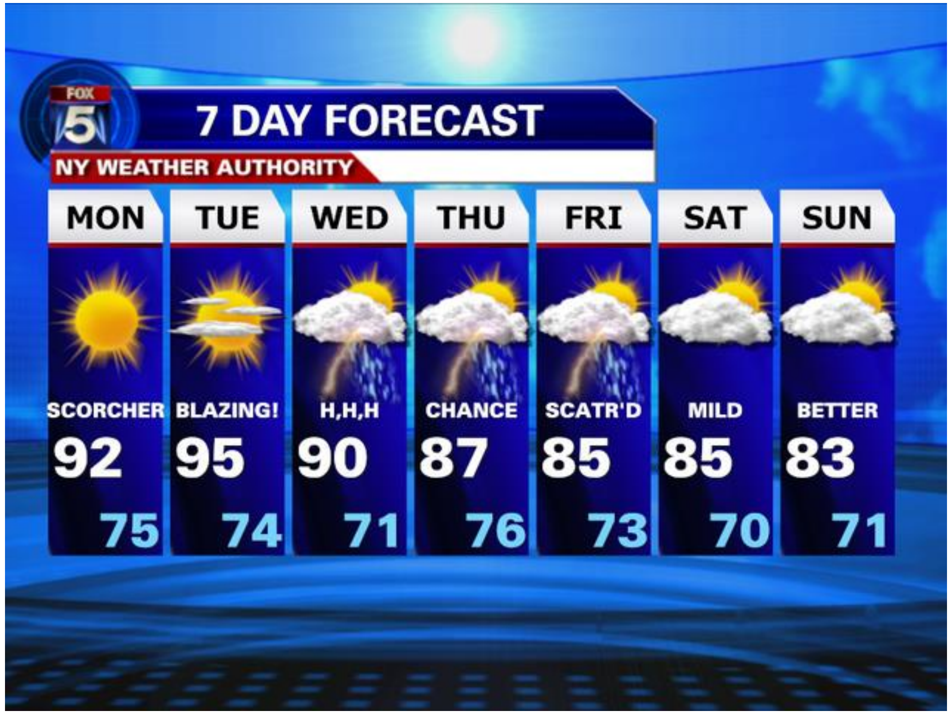
- Focus here is predictive analytics

- Predictive analytics is any use of data analysis designed to form predictions about future, or unknown, events or outcomes

- Aligns with supervised learning
  - Analyst picks the event or outcome to be predicted

Weather Prediction

- Short-run forecasts

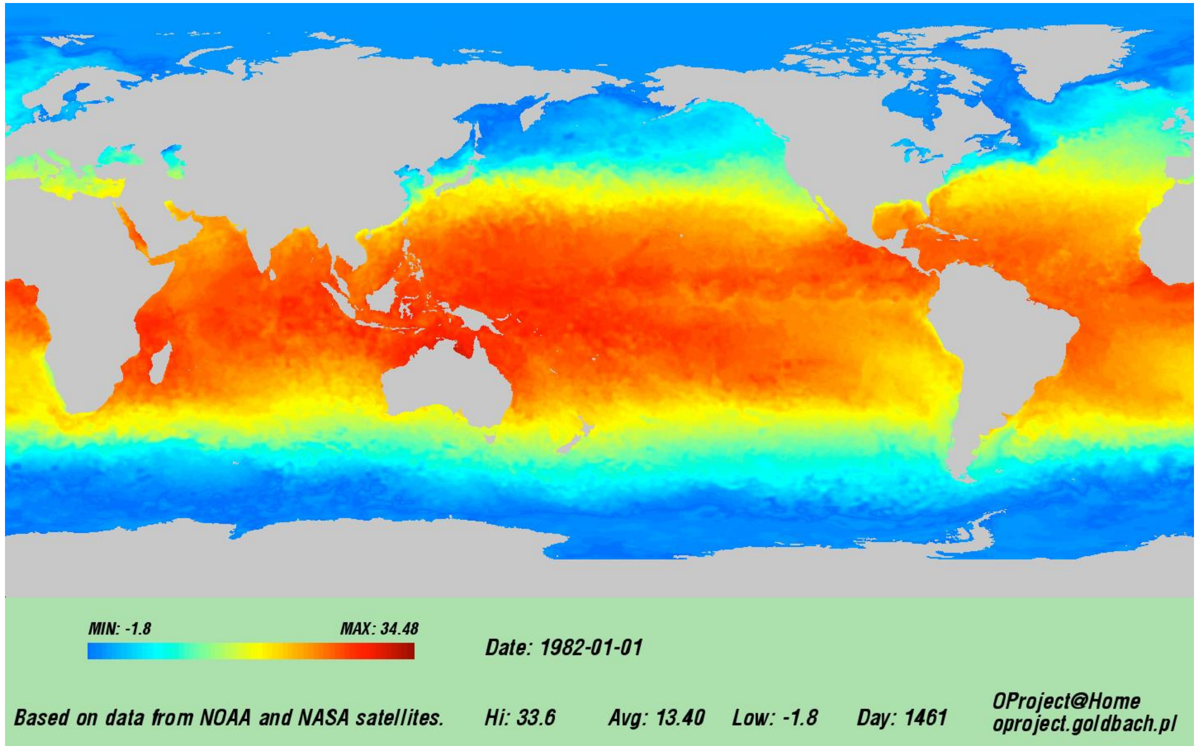


7 DAY FORECAST

NY WEATHER AUTHORITY

MON	TUE	WED	THU	FRI	SAT	SUN
SCORCHER	BLAZING!	H,H,H	CHANCE	SCATR'D	MILD	BETTER
92	95	90	87	85	85	83
75	74	71	76	73	70	71

- Climate prediction



Date: 1982-01-01

Based on data from NOAA and NASA satellites.

Hi: 33.6 Avg: 13.40 Low: -1.8 Day: 1461

OProject@Home
oproject.goldbach.pl

Churn Prediction

- Predicting who will churn

- Predicting changes in churn following a new strategy

Prediction, Bifurcated

- Passive Prediction
  - Make predictions based on data where no variables are exogenously altered
  - That is, **prediction without intervention**
    - Weather forecasts
    - Churn prediction

- Active Prediction
  - Make predictions based on data where at least one variable is exogenously altered
  - That is, **prediction with intervention**
    - Climate forecasts
    - Changes following churn strategies

Alternative Characterizations

- Descriptive-Predictive-Prescriptive
  - What happened?
  - What will happen?
  - What to do?
- Needs clarity on the link between what will happen and what to do
  - Often prediction largely treated as passive only

- Prediction vs. Inference/Explain
  - Prediction is all Passive
  - Inference/Explanatory Models look at causality
    - Essentially thinking about the data-generating process
- Again, a restrictive notion of prediction

Manager vs. Investor

- Manager intervenes via, say, a price change
- Can use prediction to help decide which price change to deploy
- Active predictions are key

- Investor does not intervene in business activities
- However, investor can use prediction to help decide whether or how much to invest
- Passive predictions are key

Why a Clearer Prediction Paradigm Matters

- Limited definition of prediction invites misuse of predictive models
    - Passive and Active predictions often require different modeling approaches
    - If the distinction isn’t made, it can become common for models suitable for passive prediction to be used for active applications
- Can help in assessing the scope and performance of predictions

Passive Prediction – Scope

- Lots of applications, but main concern is misapplication
  - Typically takes the form of finding notable variable co-movement, and then using it to make active predictions concerning, e.g., policy, strategy shift
  - Examples abound:
    - "Study: Marijuana Use Increases Risk of Academic Problems"
    - "A New Study Says Living Near a Pub Makes you Happier"
    - "Drinking More Coffee May Reverse Liver Damage from Booze"
    - "A Study Links Soda Consumption to Heart Failure"

Passive Prediction – Performance

- Fit is king
  - How close are model predictions to realized outcomes on “new” data?

- Examples:
  - Rate of correct predictions
  - R-squared

Passive Prediction – How Good Can It Get?

- Data-generating process (DGP) or “black box” to get best performance?
  - DGP may be best even for passive prediction (weather)
  - If DGP:
    - Deterministic or stochastic?
    - If stochastic, can randomness be reduced (weather factors, free will)?
    - Chaos theory apply (future vs. unknown)?
  - If “black box”
    - Hard to say how good these can get – must ask how much we are willing to bet on them despite little knowledge of the “why” behind the predictions
- “Shelf life” of a predictive model
  - Physical processes likely last a long time
  - For human behavior, are predictive relationships evolving in a way we aren’t capturing?
    - We can continue testing, but hard to assess how long into the future they will do well

Active Prediction – Scope

- Again lots of applications (policy actions, business strategy, etc.)

- Key limitation is our ability to attain exogenous variation for the treatment in question

- Lots of ways to try:
  - Controlled experiments
  - Natural experiments
  - Instrumental variables
  - Diff-in-diff
  - Matching estimators

Active Prediction – Performance

- Key idea is “identification”
  - Whether or not we are able to get measures for the parameters of the data-generating process

- Identification and fit are not at all the same
  - We can have great fit but poor identification
  - We can have poor fit but great identification

Active Prediction – How Good Can It Get?

- Depends on:
  - Sample/population relationship
  - How much exogenous variation we can find/generate

- How stable are these measured treatment effects over time?
  - Is there a data-generating process for the treatment effect?
  - Weather won’t alter its behavior due to us studying it, but people may
  - Even if we have a well-defined DGP for the treatment effect, does chaos theory come back into play?

A Note on Diagnostics

- A notable portion of “prediction” concerns the unknown, rather than the future

- These predictions are diagnostic:
  - Fraud detection
  - Image recognition
  - Presence of disease

- Most of these predictions fall in the passive category
  - However, suppose we are interested in image perception. We may then be interested in how alterations (interventions) to the image impact perception

Final Points

- Application and Performance measurement for prediction largely hinges on intervention

- If not intervening = Passive
  - Think of the Investor
  - Fit is key (model can be DGP or “black box,” whichever does better)

- If intervening = Active
  - Think of the Manager
  - Identification is key

- Misuse invites mistrust in analytics
  - Raises likelihood of contradictions
  - Invites practice of starting with conclusions