Forecasting and Election Forecasting

Marco Morales
mam2519@columbia.edu

GR5069
Topics in Applied Data Science
for Social Scientists

Spring 2017
Columbia University

Housekeeping

- number of students keeps changing?
  - does everyone have a team?
- does every team have a topic?
- next week, guest speaker
  - guest speaker
  - data challenge
  - do not forget: **teams start reporting progress**

Mini-Lab: GitHub

Version control (and Git)

though this be madness...

- **version control** allows you to keep track of changes/progress in your code
  - keeps “snapshots” of your code over time
  - helpful to debug, and to enhance reproducibility
  - also great for team collaboration (everyone can see who changed what!)
- **Git** is a version control software
- **GitHub** is an online Git repository (on steroids)
  - widely used by data scientists (and in academia)
  - not (strictly) a “software development” tool

Version control (and Git)

...yet there is method in't!

- some Git concepts to keep in mind
  - **clone;** a local copy of a repository that can be updated as changes happen
  - **fork**; a fork is a thread a repository.
  - **pull;** brings changes into master repository
  - **branch**; a local mirror copy of a repository at a given point in time

Version control (and Git)

...yet there is method in't!

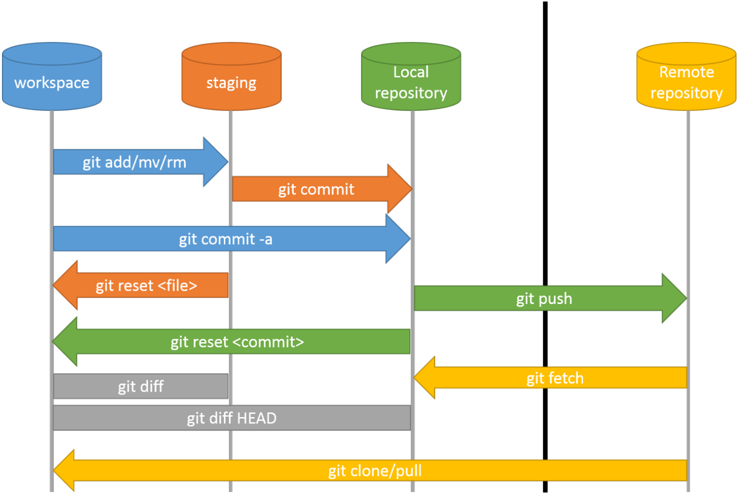


Figure: http://www.moxie.io/images/git-operations.png

Version control (and Git)

...yet there is method in't!

- some useful actions in GitHub
  - `git init`: initializes Git, and indicates that the folder should be tracked
  - `git add`: brings new files to the attention of Git to be tracked as well
  - `git commit`: takes a snapshot of alerted files
  - `git push`: sends changes in your local file to the GitHub repository

Forecasting

Forecasting

"Unfortunately, we were a little off-target again this quarter."

Forecasting

What is a forecast / forecastable?

- what do we mean by forecasts?

  forecast**:** *fore-* before + *casten* to prepare

  prognosticate**:** *pro-* before + *gnoscere* to know

- in essence, we use past information to estimate the future

$$\hat{Y}_{t+1} = f(Y_t, X_t, \epsilon_t)$$

Forecasting

What is a forecast / forecastable?

- **Predictability** depends on (Hyndman & Athanasopoulos 2013):
  - how well we know factors that influence the forecast
  - how much data (and of what quality!)
  - recursive influence of the forecasts
- Key question: **what to forecast?**
  - every item?
  - at what level of aggregation?
  - at what frequency? daily? weekly? quarterly? yearly?
- **Remember:** explain $\neq$ predict

Forecasting

Models to explain vs models to forecast

- social scientists typically trained to fit models to **explain**, and derive "predictions" from them
  - we are taught to approximate the data-generating mechanism
- the type of uncertainty associated with explanation is different from that of prediction (Shmueli 2010)
  - both rely on the relationship of $\mathcal{Y} = \mathcal{F}(\mathcal{X})$ with $E(Y) = f(\mathbf{X})$
  - **explanation** tries to match $\mathcal{F}$ with f, using $\mathbf{X}$ and Y as tools
  - **prediction** uses f as a tool to generate future values of Y given $\mathbf{X}$

Forecasting

Ways, Means and Tools to Forecast...

- **Cross-Sectional models**
  - regression-based
  - ML-based
- **Time-Series models**
  - Naïve

$$\hat{Y}_{t+1} = Y_t$$

  - Exponential Smoothing

$$\hat{Y}_{t+1|t} = \sum_{j=0}^{t-1} \alpha(1-\alpha)^j Y_{t-j} + (1-\alpha)^t \ell_0$$

  - ARIMA models

$$\hat{Y}_{t+1} = c + \phi_1 Y_t + ... + \phi_p Y_{t-p} + \theta_1 \epsilon_t + ...\theta_q \epsilon_{t-q} + \epsilon_{t+1}$$

Forecasting

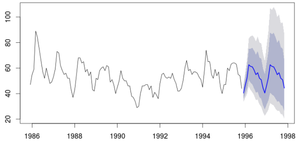
Fitting models vs forecasting: some (empirically validated) rules of thumb

1. **keep it simple:**
   - start parsimonious and add complexity (*iff* called for)
   - increased complexity typically reduces forecast accuracy
2. **rely on domain expertise to select inputs**
   - statistical significance a faulty guide for inclusion
   - domain expertise should drive variables to include
3. **include more (useful) information**
   - high correlation in predictors (and multicollinearity) not an issue
4. **fit $\neq$ accuracy**
   - well-fitting models may impose unwarranted “structure” and “certainty” to the forecast
5. **update models constantly**
   - update parameters as new information arrives

Forecasting

Forecast uncertainty

- by definition, forecasts are uncertain
  - we should be interested in the **point estimate** of the forecasts and its **prediction interval**
- it is possible to estimate the **range of values** where the forecast may lie **with a given probability**



- the **prediction interval** ($\hat{y}_{t+i} \pm k\hat{\sigma}$) is a function of an estimate of the standard deviation of the forecast ($\hat{\sigma}$) and a multiplier k

Forecasting

Time-Series cross-validation

- usual k-fold validation **inadequate** for time-series because of lagged values in these models
- an appropriate (rolling) **time-series cross-validation algorithm** (Hyndman):
  1. fit your time-series model and compute the error (ϵ^*_{t+h}) for the forecasted observation ($\hat{Y}_{t+h}$) h steps into the future per

  $$\epsilon^*_{t+h} = Y_{t+h} - \hat{Y}_{t+h}$$

  2. repeat step 1 for $t = m + h, ..., n - 1$ where m is the minimal number of obs to estimate model
  3. compute appropriate error measure (i.e. MAPE, RMSE..) with estimated errors

Forecasting

Time-Series cross-validation

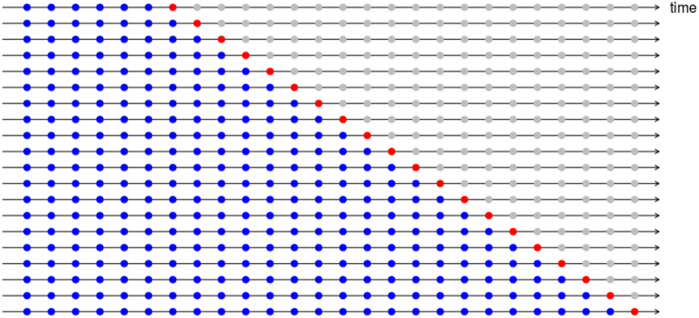


Figure: Rob Hyndman

Forecasting
Time-Series cross-validation

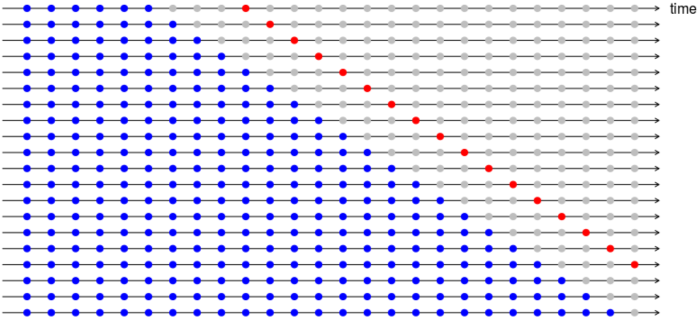


Figure: Rob Hyndman

Error Measure	Definition
Root Mean Squared Error (RMSE)	$\sqrt{\frac{\sum_{i=1}^{n}(Y_{t+i}-\hat{Y}_{t+i})^2}{n}}$
Mean Absolute Percent Error (MAPE)	$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\lvert Y_{t+i}-\hat{Y}_{t+i}\rvert}{Y_{t+i}} * 100\right)$

- when evaluating forecasts remember:
  - is the measure **valid** (makes sense to experts)?
  - is the measure **sensitive to outliers**?
  - is the measure **affected by scale**?
  - **do not use R^2 to assess models**

Forecasting

Forecast Ensembles

Forecasting
Forecast Ensembles

- we typically think of a single model to produce forecasts
  - what if we have various “informative” models?
- simple averaging of forecasts has proven in many cases superior to single forecasts
  - complex methods have been devised to optimize forecast weights, not always best
- particularly useful when models/methods are sufficiently different

Forecasting Elections

What do we know about elections?

- most important resource is **contextual / expert knowledge**
  - do not re-invent the wheel!
  - listen to:
    - people with intuitive knowledge
    - people with empirical knowledge
    - prior research
- Political Scientists have been studying elections for over half a century...
  - we must know something...

Forecasting Elections

Dense knowledge: economic conditions

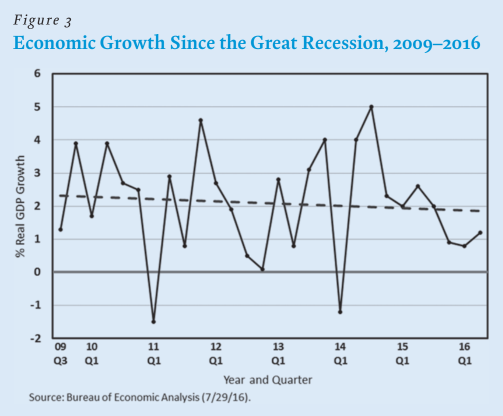


Figure: Campbell 2016

- good economic performance typically helps the candidate of the incumbent party

Forecasting Elections

Dense knowledge: presidential approval

Table 1

Presidential Approval in Mid-July of Open Seat Election Years, 1952–2012

Rank	Departing President (Year)	Approval %	Election Outcome
1.	Bill Clinton (2000)	59	Won (Lost EV)
2.	Ronald Reagan (1988)	54	Won
3.	Barack Obama (2016)	51	?
4.	Dwight Eisenhower (1960)	49	Lost? (Lost EV)
5.	Lyndon Johnson (1968)	40	Lost (Close)
6.	George W. Bush (2008)	31	Lost
7.	Harry Truman (1952)	29	Lost

Source: Gallup.

Figure: Campbell 2016

- good presidential approval of the incumbent typically helps the in-party candidate

Forecasting Elections

Dense knowledge: ideological polarization

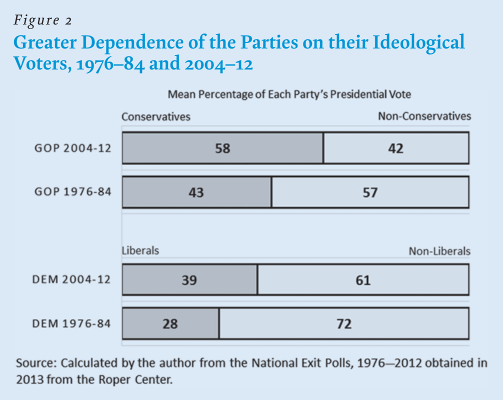
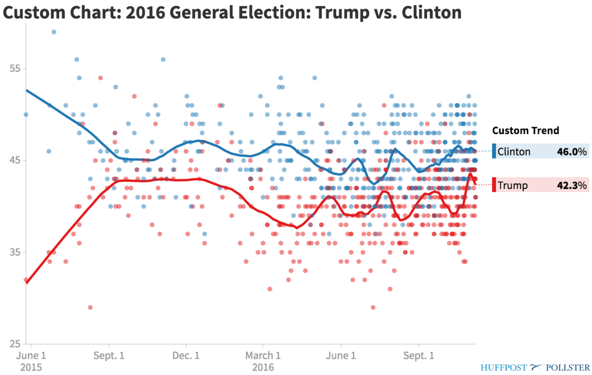


Figure: Campbell 2016

- the more polarized the electorate, the harder it becomes to move sway it

Forecasting Elections

Dense knowledge: survey data



- surveys start relying information useful for forecasting right after the Conventions

Team Planning

Forecasting and Election Forecasting

Marco Morales
mam2519@columbia.edu

GR5069
Topics in Applied Data Science
for Social Scientists

Spring 2017
Columbia University