Exploring Machine Learning in Production Processes: Experiences from Statistics Canada

Delivering insight through data for a better Canada

Overview

- With advances in computing power and availability of big data, more complex machine learning algorithms have gained prominence
- Statistical agencies have been incorporating some of these techniques into production and analysis
    - Classifier algorithms (for categorizing products)
    - Optical character recognition (for digitizing grocery receipts)
    - Natural language processing (for identifying economic events from news articles)
    - **Predictive modelling** (for predicting price movements)

Overview

- From a statistical agency point of view, two questions arise:
  - What advantages, if any, do more modern predictive methods present over more traditional ones?
  - How easy is it to incorporate these methods into a production process that faces tight deadlines?
- Today's talk: answering these questions in the context of Statistics Canada's Wholesale Services and Retail Services prices programs

Table of Contents

Motivation

- At Statistics Canada, monthly GDP section requires wholesale services and retail services price deflators approximately 6 weeks after reference period
- Wholesale Services and Retail Services price programs are quarterly: not timely enough to meet this requirement
- Thus, wholesale and retail price deflators need to be modelled until actual data are received, at which time monthly GDP is revised

Context

- Timeliness
  - Not much time between receipt of data (from retail/wholesale programs and auxiliary sources) and deadline for results
- Computing resources
  - Do not have access to unlimited computing power
- Limited input data
  - Microdata often unavailable for “nowcasting”
  - Must rely on contemporaneously published series
  - Proxies must be published at same or greater frequency than series being predicted
    - Much of what statistical agencies publish is produced on an annual basis
- Business continuity
  - Process should be understood by anyone who uses it (both math and code)
  - Interpretability is valued

Description of Indices and Data Sources

- Wholesale and Retail Services Price Indices (WSPI/RSPI)
- Monthly indices produced quarterly (with a three-month revision)
- Actually three indices in one: margin; selling price; purchase price
- Coverage
  - WSPI – wholesale trade services under NAICS 41, excluding 419 (B2B brokers)
  - RSPI – retail trade services under NAICS 44 & 45, excluding 44112 (used cars) and 454 (non-store retailers)
- Most price data from quarterly Wholesale/Retail Price Report
  - Randomly selected sample of wholesalers and retailers
  - RSPI also uses some scanner data from major retailers and auto data from J.D. Power

Production and Dissemination of Indices

- Produced in an R-based pipeline
- Margin indices disseminated publicly; selling and purchase price indices available internally
  - Selling prices needed for deflators
- Disseminated 2.5 months after end of reference quarter
  - Not timely enough for monthly GDP deflators!

Methods Reviewed

- Wholesale
  - Basic linear model
  - ARIMA with stepwise selection
  - Simplified ARIMA
- Retail
  - Basic linear model
  - Neural network model
  - Linear time-trend model

Wholesale: Basic Linear Model (1/2)

- Description
  - Old model; used pre-2020
  - Predictions are convex combinations of Consumer Price Index (CPI), Industrial Product Price Index (IPPI), and Raw Materials Price Index (RMPI) series
  - Weights for convex combination come from NAPCS commodity shares in Annual Wholesale Trade Survey
  - Implemented in SAS (computation) and PowerBI (reports)

Wholesale: Basic Linear Model (2/2)

- Upsides
  - Very simple model: prediction just a linear combination of contemporaneous values
- Downsides
  - Did not incorporate trends, just contemporaneous values of other series
  - Annual weights
    - Do not vary by reference month
    - Not available contemporaneously with price data
  - Model parameters not obtained by training on our data set, but by estimation from a completely different data set
  - SAS knowledge not widespread throughout the organization, dwindling

Wholesale: ARIMA Model with Stepwise Selection

- Description
  - Used from early 2020 to early 2022
  - Retrained every month on an expanding window
  - ARIMA model with at most 5 autoregressive and moving average lags and up to 70 covariates
  - Lags, covariates selected by BIC
  - Instead of estimating all possible models for each series, stepwise selection used
  - Implemented in R
- Upsides
  - Flexible; could incorporate both contemporaneous information and trends
  - Let the data “do the talking”: data will tell us which covariates matter and to what extent
- Downsides
  - Large number of potential covariates could lead to spurious correlation, high variance in predictions, numerical instability
  - Different predictors could be used every month: compromises interpretability
  - Long runtime

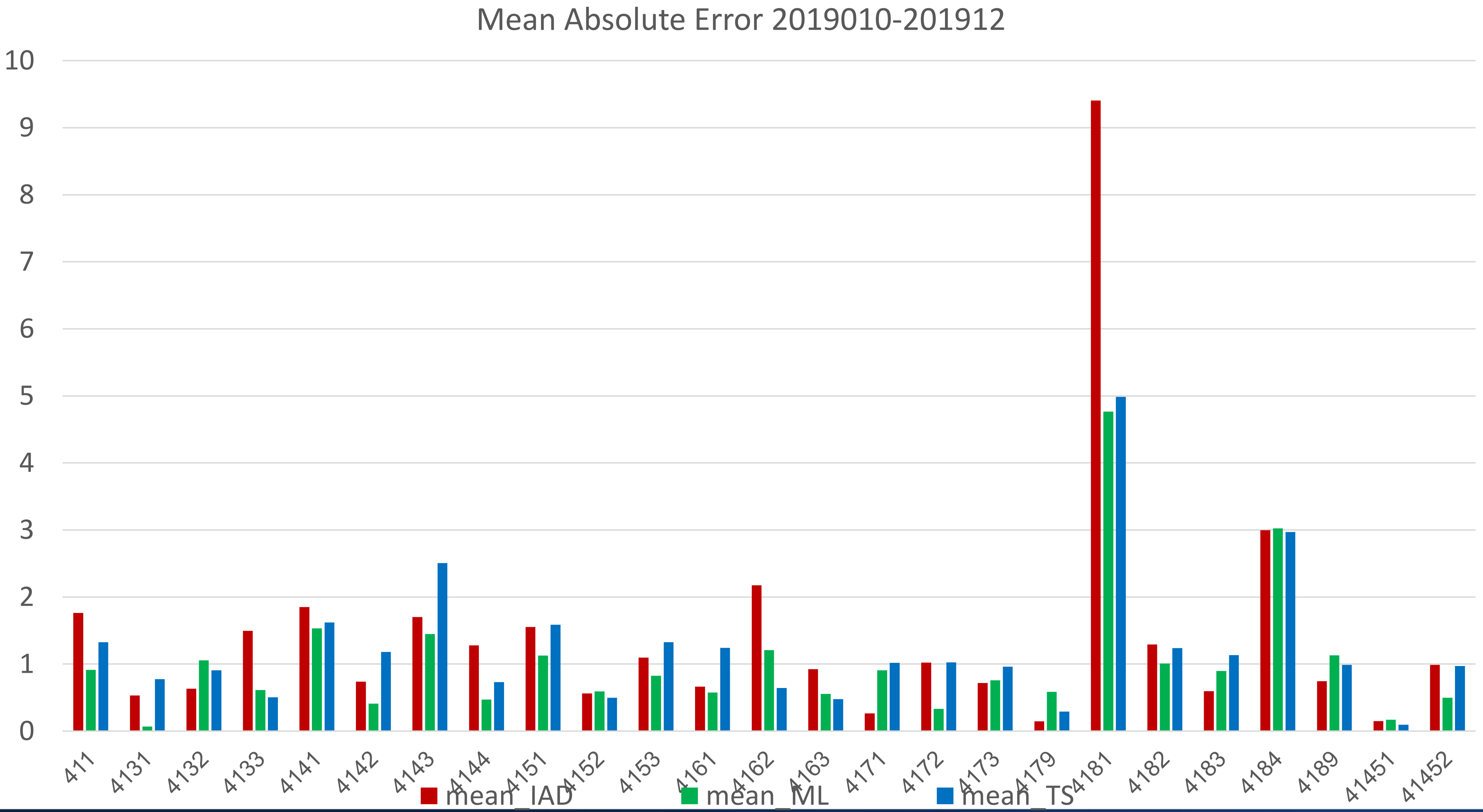
Wholesale: Simplified ARIMA Model (1/2)

- Description
  - Currently in use
  - Retrained every month on a rolling window of 5 years
  - ARIMA model with at most 5 autoregressive and moving average lags
  - Each WSPI-SP series uses pre-defined set of covariates (typically around 5 but up to 8), plus possible seasonality adjustment in both AR and MA
    - Covariates selected from subject-matter knowledge
    - Include CPI, IPPI, and RMPI series
  - Scanner data used for NAICS 413 (food, beverage, & tobacco); J.D. Power sales data for 4151 (motor vehicles); Kalibrate data for 4121 (petroleum)
  - Models selected by AICc
    - Model selection only on lags and seasonality dummies; stepwise selection used
    - Set of possible models much smaller than earlier ARIMA model (2^18 vs 2^80)
  - Implemented in R

Wholesale: Simplified ARIMA Model (2/2)

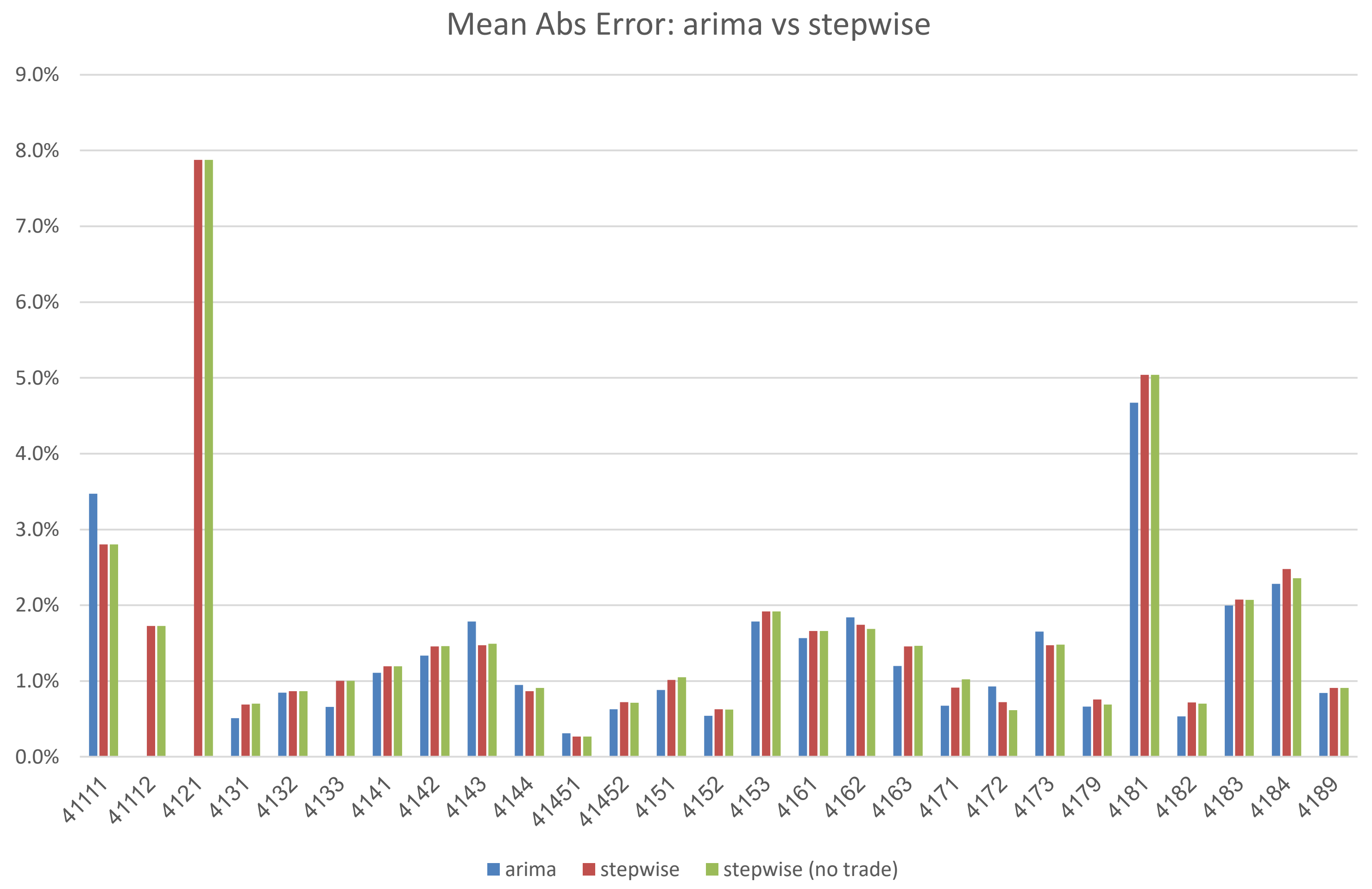
- Upsides
  - Flexible; could incorporate both contemporaneous information and trends
  - Manageable number of covariates
  - Stable models; same covariates used every month – enhances interpretability
  - Simple, single-environment implementation
  - Runs in about two minutes
- Downsides
  - Does not fully let the data “do the talking”
  - Using a rolling window results in only T = 60 data points per series

Wholesale: Basic Model vs ARIMA



- ARIMA performs similarly to simple linear model in terms of mean absolute forecast error
- ARIMA outperforms linear model for some series but not others
- Averaging across series, ARIMA reduces mean absolute forecast error 0.17% relative to basic linear model

Wholesale: ARIMA vs Simplified ARIMA



- Simpler ARIMA model outperforms more complex one for some 4-digit NAICS; vice versa for others
- Averaging across series, using the simpler ARIMA model leads to an improvement in MAE of 0.04%
- MAE improvement modest, but runtime improvement large (minutes vs. hours)

Wholesale: Key Takeaways

- Good predictions incorporate both trends and contemporaneous data
- Letting the data “do the talking” is a good idea… to a point
  - Cannot fully substitute for specialist knowledge
- Having stable models makes interpretation and diagnostics easier
- Implementation trade-off: increased complexity and flexibility come at the expense of runtime, convenience, business continuity
  - Worth switching to a faster method even when improvements to accuracy of prediction are modest

Retail: Basic Linear Model (1/2)

- Description
  - Used up to December 2020
  - Predictions are convex combinations of CPI and other series
  - Weights for convex combination come from NAPCS commodity shares in Retail Commodity Survey
    - Projected to reference month based on historical data
  - Later revised to take weights from Annual Retail Trade Survey to match wholesale methodology
  - Implemented in SAS and Microsoft Excel

Retail: Basic Linear Model (2/2)

- Upsides
  - Very simple model: prediction just a linear combination of contemporaneous values
- Downsides
  - Did not incorporate trends in price movements, just contemporaneous values of other series
  - Monthly weights not available contemporaneously, while annual weights do not vary by reference month
  - Model parameters not obtained by training on our data set, but by estimation from a completely different data set
  - SAS knowledge not widespread throughout the organization, dwindling

Retail: Neural Network Model (1/2)

- Description
  - Used from January 2021 to April 2022
  - Neural network with two hidden layers
    - Rectified linear unit activation function (avoids vanishing gradient problem)
    - Loss function is asymmetric squared loss (penalizes predicting incorrect movement direction 50% more)
    - Uses L_1 and L_2 regularization (both parameters set to 0.001) to guard against overfitting
    - Learning rate is adaptive (uses adaptive moment estimation to accommodate sparsity)
    - Input variables for training pre-selected by QR decomposition
  - Incorporated contemporaneous scanner data
  - Multi-environment implementation
    - R: data cleaning; preparation of inputs; prediction
    - Python: model training (using TensorFlow and Keras)
    - SAS: preparation of outputs

Retail: Neural Network Model (2/2)

- Upsides
  - Flexible model that could potentially capture nonlinearities
- Downsides
  - Large number of potential covariates could lead to spurious correlation, high variance in predictions, numerical instability
  - Model needed to be retrained quarterly, but retraining could take 2-3 weeks
  - Estimation not rapid either
  - Laborious to update model when data sources changed or basket updated
  - Multi-environment implementation complicates production process
  - Opaque for end users

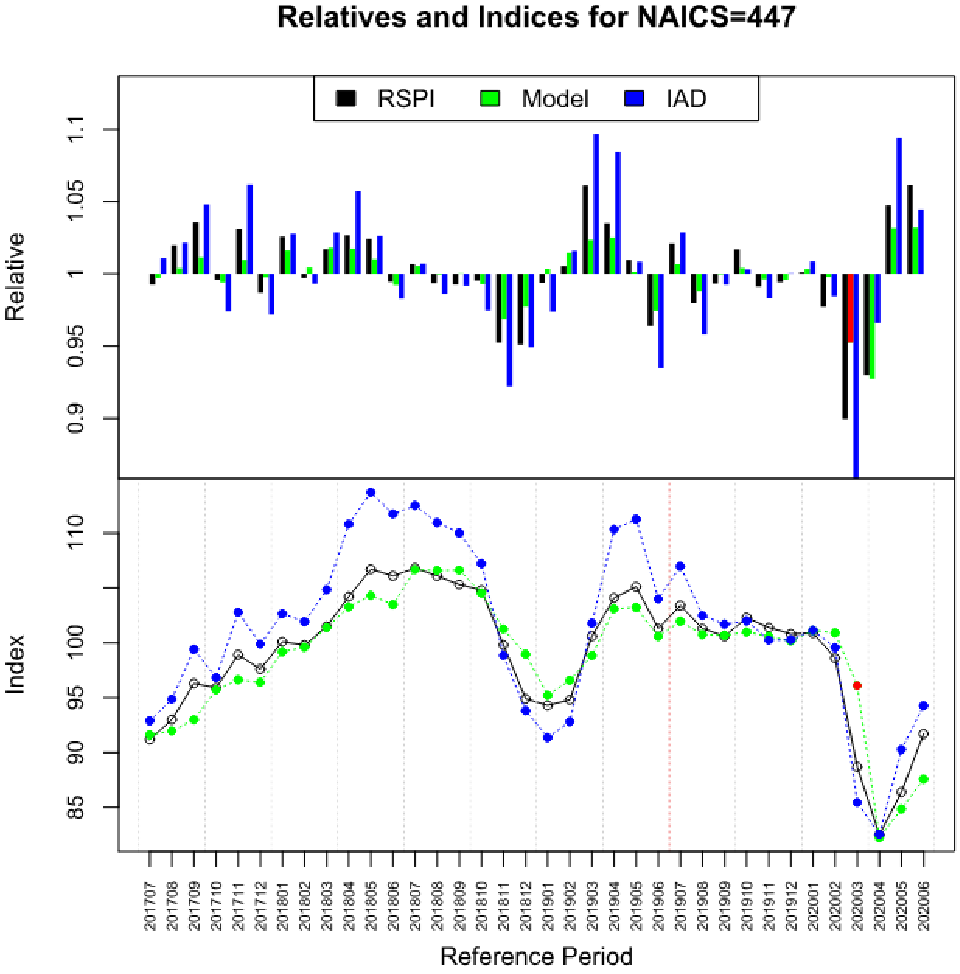
Retail: Linear Time-Trend Model (1/2)

- Description
  - Currently in use
  - Retrained every month on a three-month rolling window
  - Regress each series on a time trend and a single, series-specific controlling parameter
    - Parameter is a convex combination of relevant CPI, IPPI, and Kalibrate series, with weights corresponding to North American Product Classification System (NAPCS) shares from the Quarterly Retail Commodity Survey (QRCS)
    - Use of a single controlling parameter cuts down on degrees of freedom
  - Essentially extends the old basic linear model by adding trend data
  - Scanner and administrative data from some retailers used instead of model prediction for some series, because those data are available monthly
  - Implemented in R

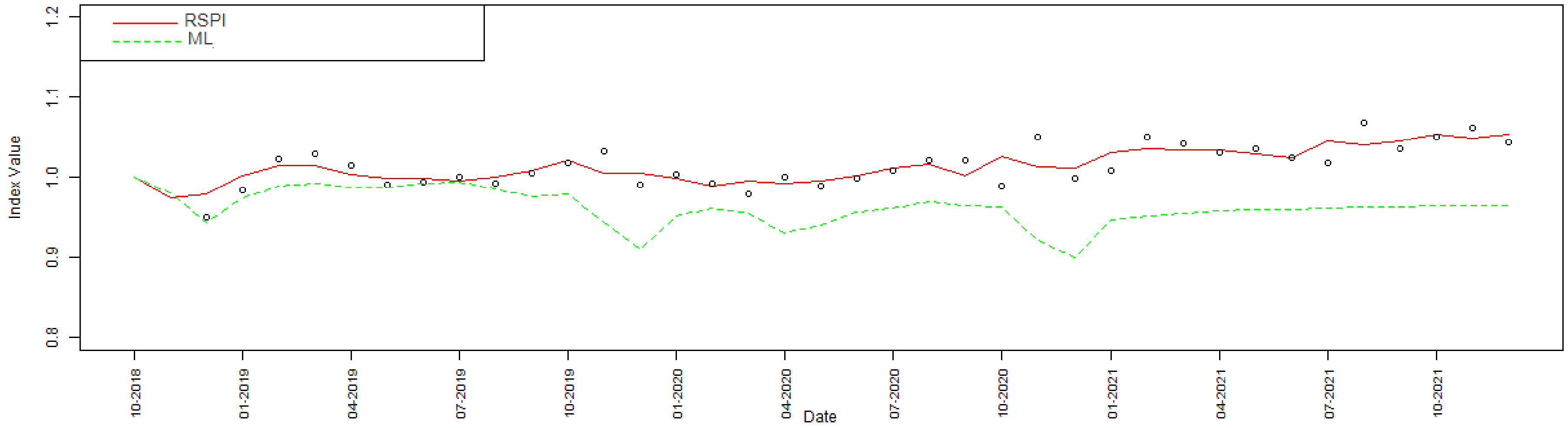
Retail: Linear Time-Trend Model (2/2)

- Upsides
  - Extremely simple
  - Stable models; same covariates used every month
  - Runs in about an hour, most of which is processing scanner data
  - Model is easily interpretable
- Downsides
  - Somewhat non-standard model
  - Rigid functional form: does not fully let the data “do the talking,” and assumes a linear time trend

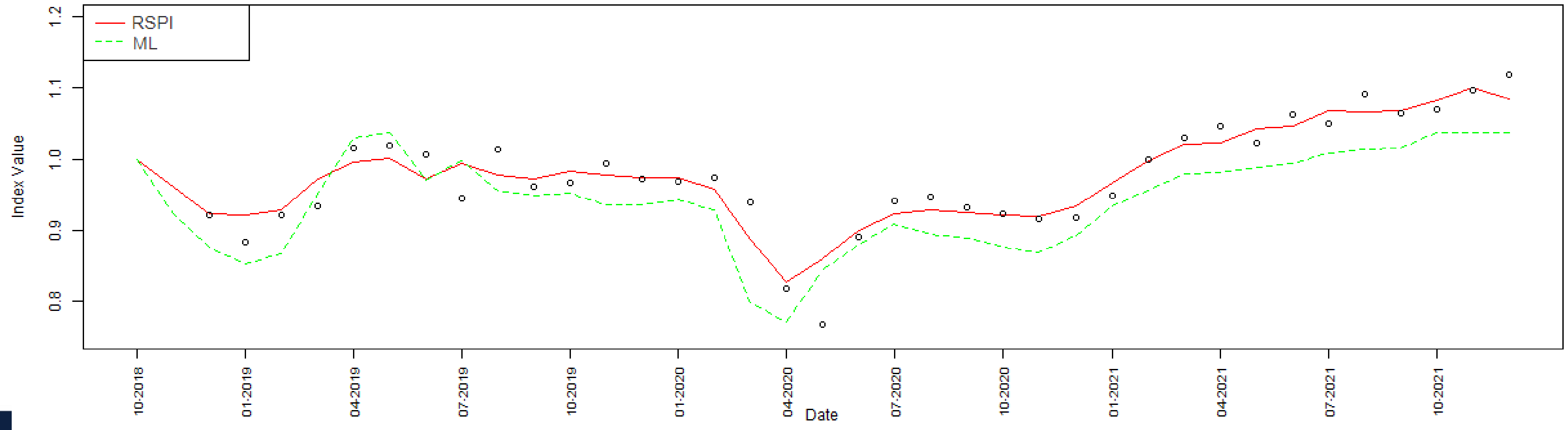
Retail: Basic Model vs Neural Network



Retail: Neural Network vs Time-Trend Model (1/2)

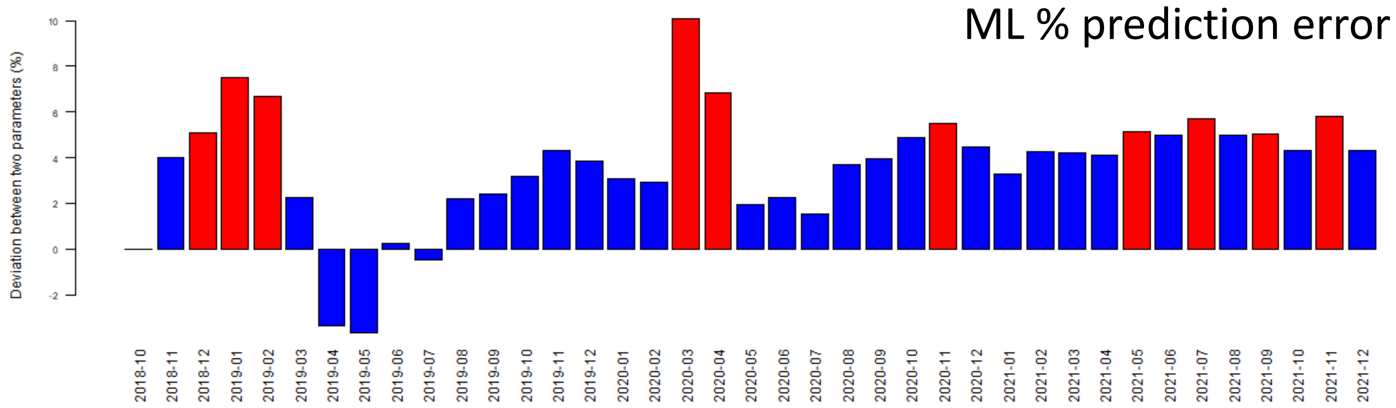


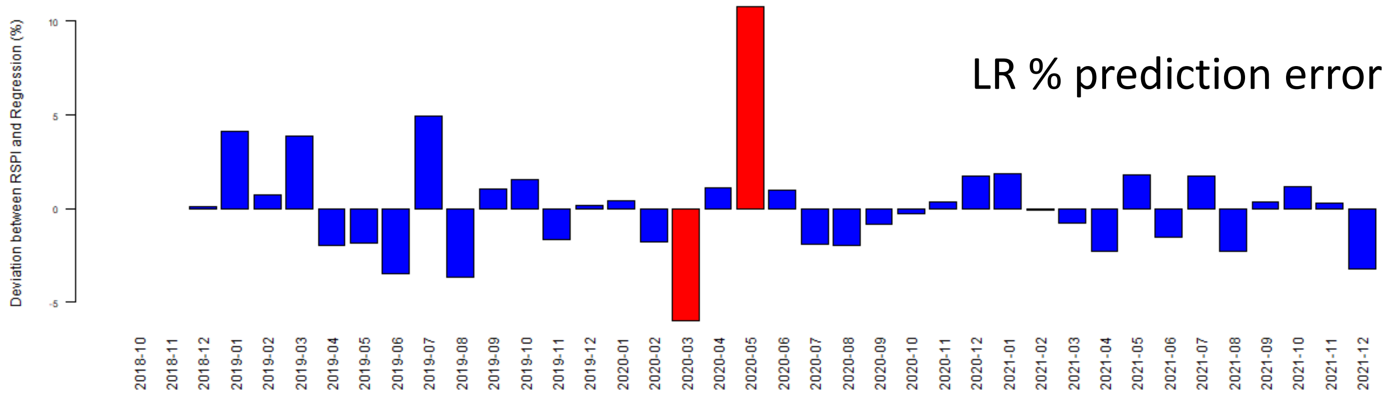
NAICS 443



NAICS 447

Retail: Neural Network vs Time-Trend Model (2/2)





- NAICS 447 (gasoline stations)
- ML has higher average % prediction error, more months where % error exceeds 5%
- Direction of prediction error in ML model upward biased
  - Indicates overfitting on training data, lack of retraining
- Direction of prediction error in LR model not biased
  - Rolling window allows model to adapt

Retail: Key Takeaways

- Neural network models often not suitable for limited input data
- Overly complex models on limited data can overfit in sample and perform poorly out of sample, even with regularization
- Subject matter knowledge should guide model construction
- Simplicity of implementation can be just as important as simplicity of computation
  - For one-off analysis, may be better to train in one environment (e.g. Python) and estimate in another (e.g. R)
  - But this is messy for production purposes; mixing environments complicates production

Implementation at a Statistical Agency

- Model performance (i.e. goodness of fit) not the sole criterion by which we judge models
- Operational concerns (e.g. business continuity, compatibility with other processes, software/package management) matter
- Personnel and computing power are scarce resources
- Statistical agencies must be able to explain what they do to a broad audience
- Continuity in methods is valued
- A complicated model that takes a week to run, is understood by few people, and differs immensely from previous models is unlikely to be used, regardless of out-of-sample performance

Conclusion

- Vast array of powerful and innovative machine learning techniques available for prediction
- With limited training and input data, gains from using more complex methods are modest or even negative
  - Little advantage to allowing for nonlinearity and complexity on small data sets
- Regular statistical production subject to constraints on data, time, computation, and personnel
  - Complex methods sometimes unable to meet deadlines
- Success can often be found where the machine learning toolkit and the standard econometric toolkit intersect

Acknowledgements

- Ivan Carrillo-Garcia
- Lydia Couture
- Benoit Germain
- Xin Ha
- Quentin Hillis
- Steve Martin
- Thomas McDowell
- Roozbeh Mollaabbasi
- Zekai Shao
- Shaoxiong Wang
- Zhaoxin Ye
- Sisi Zhang