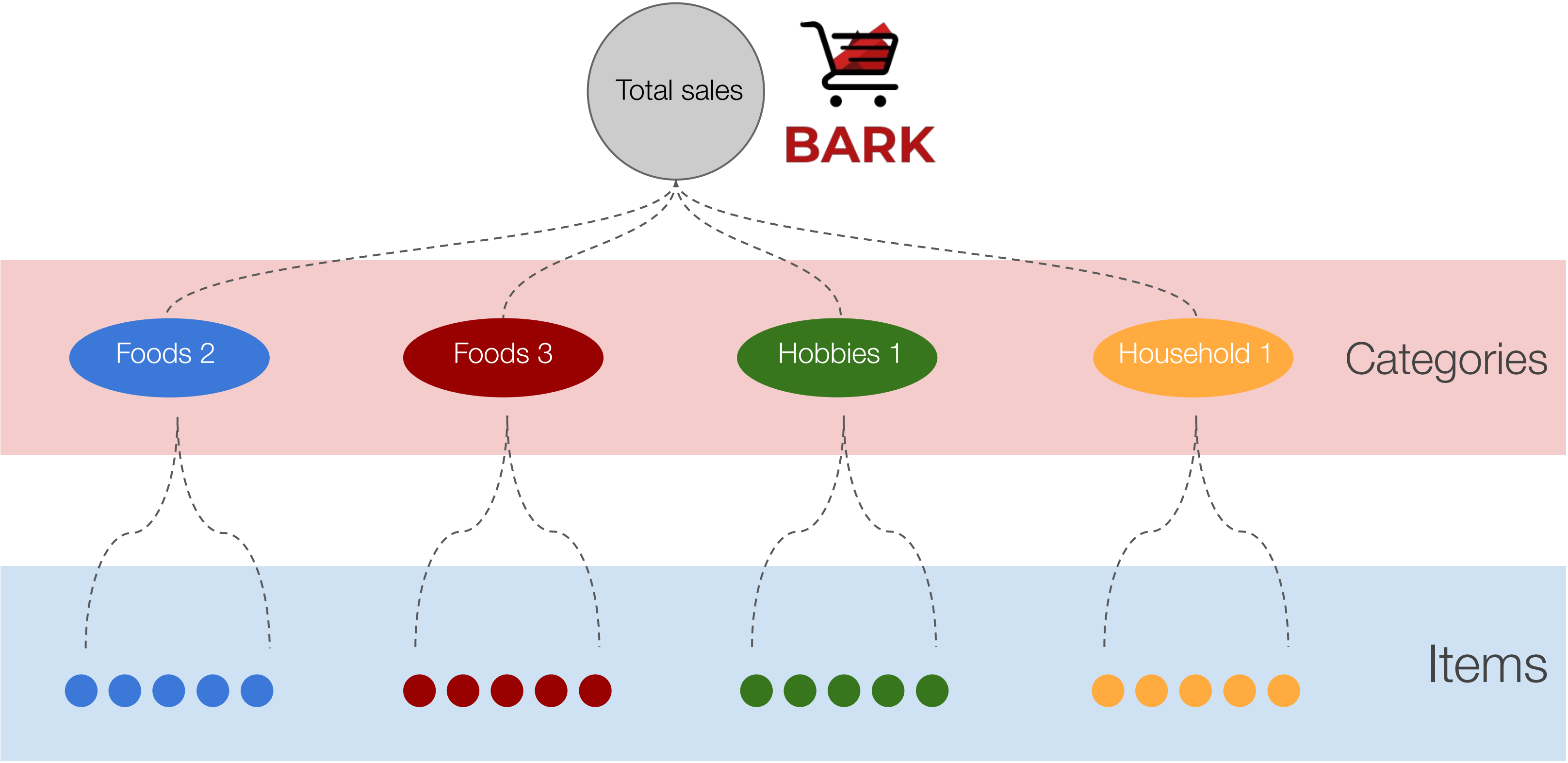
time series

all models are wrong

Exploring forecasting retail sales at scale

Benedict Au, Brea Beals, Mark Roberts, Yannik Kumar

BARK - our microcosm for exploration



BARK

> Introduction

> Models

> Selection criteria

> Feature extraction

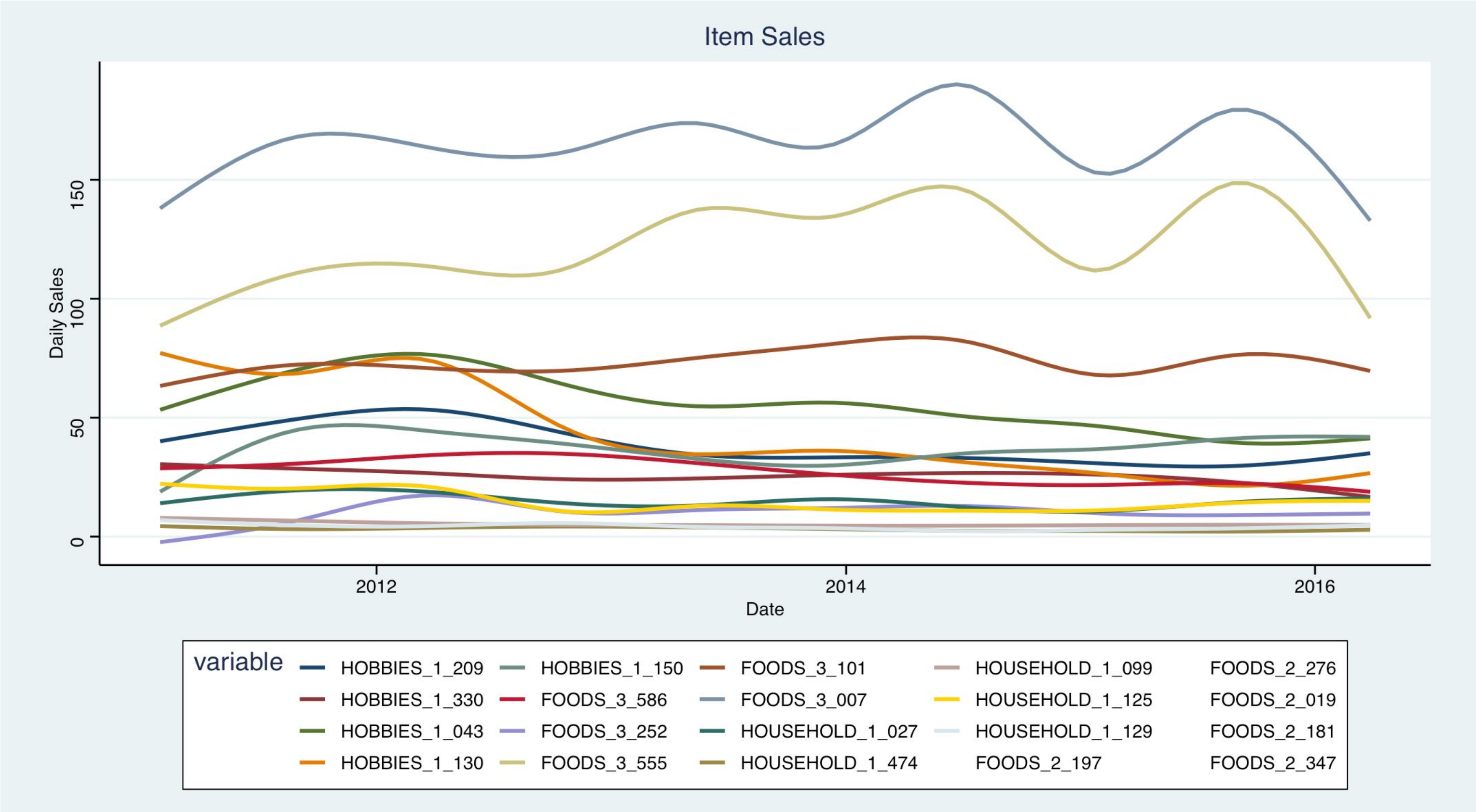
> Results

> Application

> Cross-validation

> Takeaways and future work

BARK - our microcosm for exploration



Objectives

1	what classes of models perform best across categories/items?
2	do we lose accuracy adopting a hierarchical approach? Are we ok with this tradeoff?
3	does cross-validation give us the same ranking of models?
4	how much data do we need to make a decent forecast?

BARK

4 categories

20 items

BARK

Models

> Ran all models below for all 20 items for both daily and weekly seasonality where appropriate

> All models were written as functions so this project can easily be expanded to additional product items

> **Simple forecasting methods**: Average, Naive, Seasonal Naive

> **ARIMA models**: ARIMA, Seasonal ARIMA, and ARFIMA

> **OLS**: using trend, year, month, day of week, snap benefits, prices, and cultural/religious events

> **Holt-Winters**: Additive Method

> **STL**

> **Prophet**

> **TBATS**

> **ARCH + GARCH** combination

> **Neural Networks**

> **Hierarchical Modeling**: Top-down (ARIMA and ETS) and Middle-out (ARIMA and ETS)

BARK

Selection criteria

- > Separated data into train and test sets, where test set was the last 28 days
- > **Mean absolute scaled error (MASE)** as measure of forecast accuracy
    - > Scale free error metric
    - > Well suited for intermittent-demand series because it never gives infinite or undefined values
- > **Ljung-Box Test** on residuals as a test for autocorrelation
    - > Used p-value to compare models to see which had residuals resembling white noise

Feature extraction

> specific to regression model:

dummy-coded days, months and added trend feature

> for autoregressive models:

dummy-coded event-types, and SNAP promotions

BARK

> Introduction

> Models

> Selection criteria

> Feature extraction

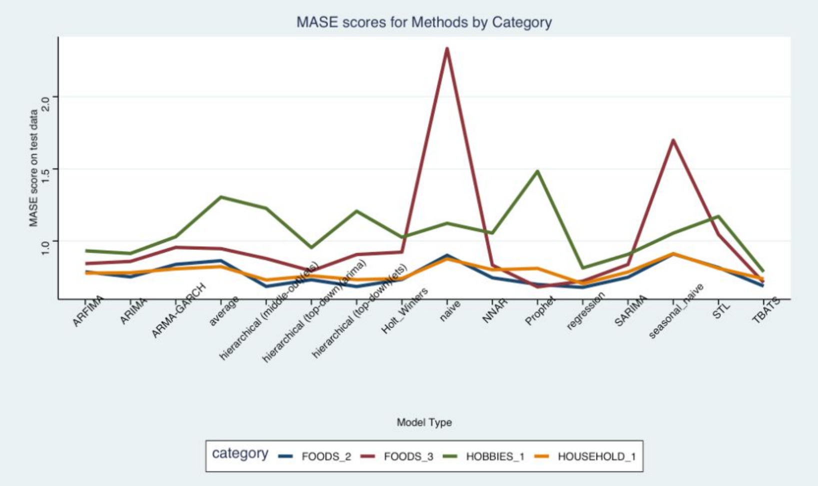
> Results

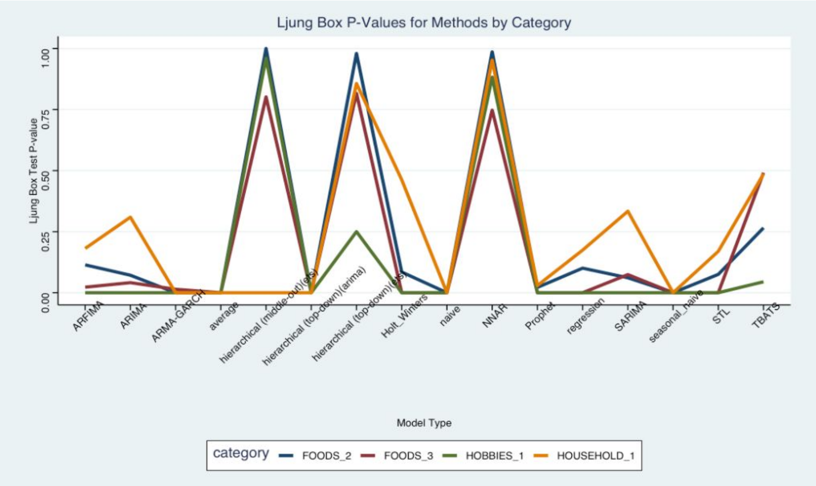
> Application

> Cross-validation

> Takeaways and future work

Results





> Grouped each item by category and calculated average scores

- MASE
- Ljung Box test P-value

BARK

> Introduction

> Models

> Selection criteria

> Feature extraction

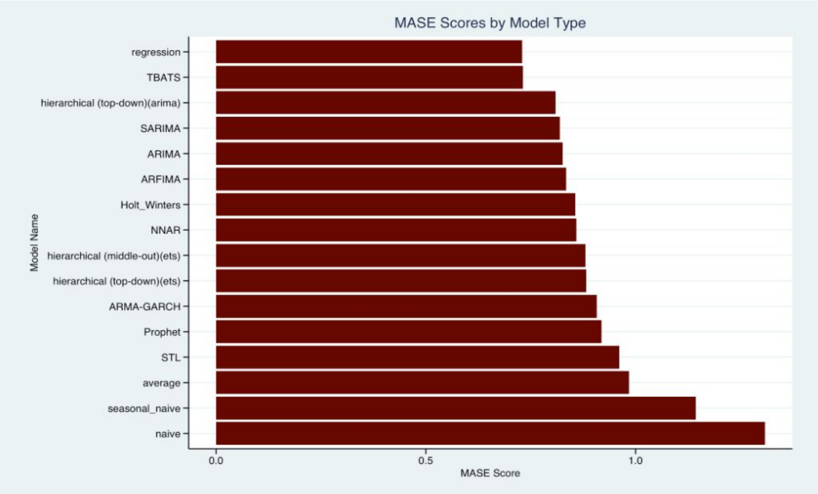
> Results

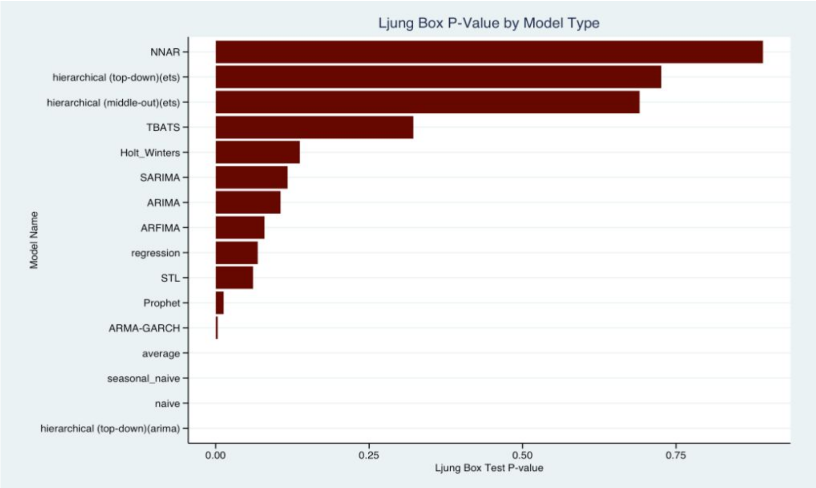
> Application

> Cross-validation

> Takeaways and future work

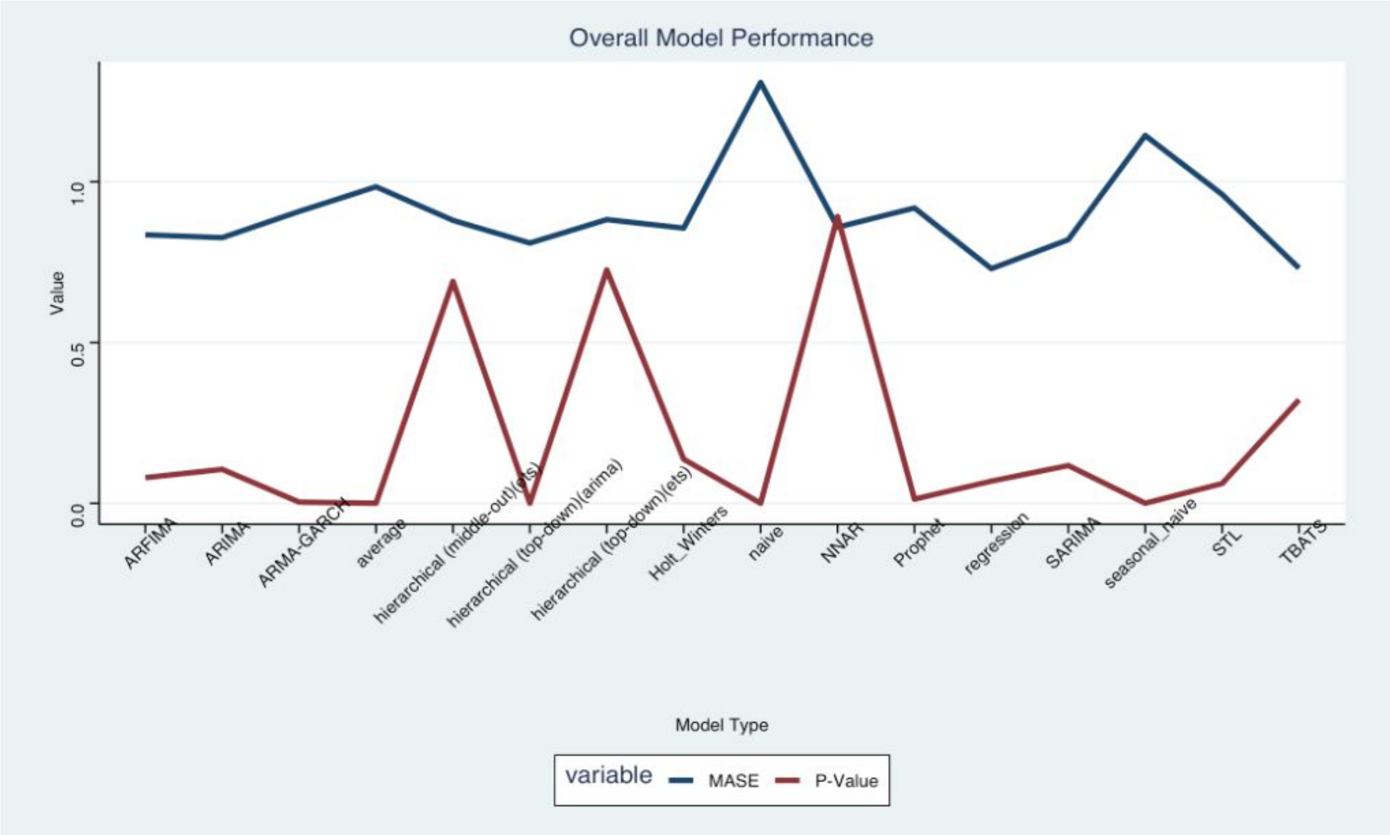
Results





> Best performing models: low MASE score and high P-value

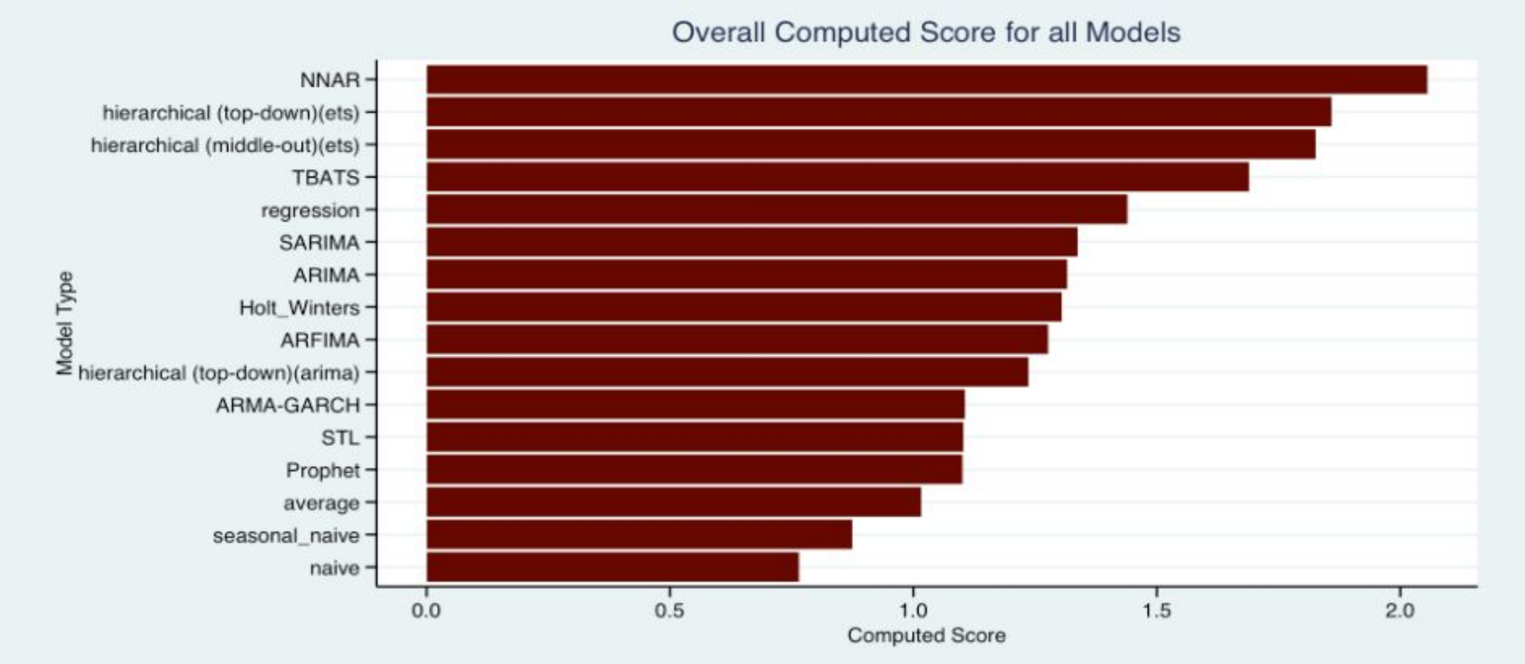
Results


Overall Model Performance
Value
0.0
0.5
1.0
ARFIMA
ARIMA
ARMA-GARCH
average
hierarchical (middle-out)(ets)
hierarchical (top-down)(arima)
hierarchical (top-down)(ets)
Holt_Winters
naive
NNAR
Prophet
regression
SARIMA
seasonal_naive
STL
TBATS
Model Type
variable
MASE
P-Value

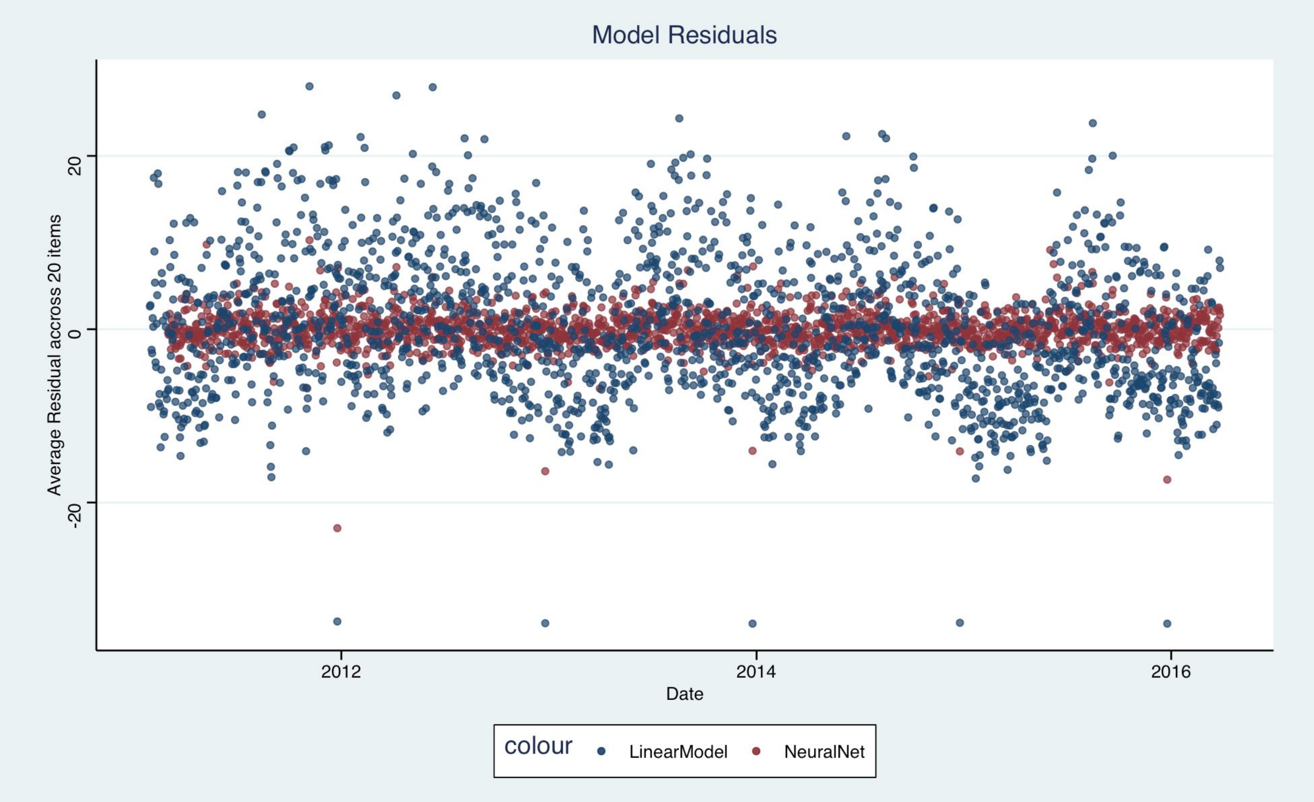
BARK

Results

$$\text{Computed Score} = \frac{1}{MASE\ Score} + LjungBox\ pvalue$$



Results- Whitening of Residuals



BARK

> Introduction

> Models

> Selection criteria

> Feature extraction

> Results

> Application

> Cross-validation

> Takeaways and future work

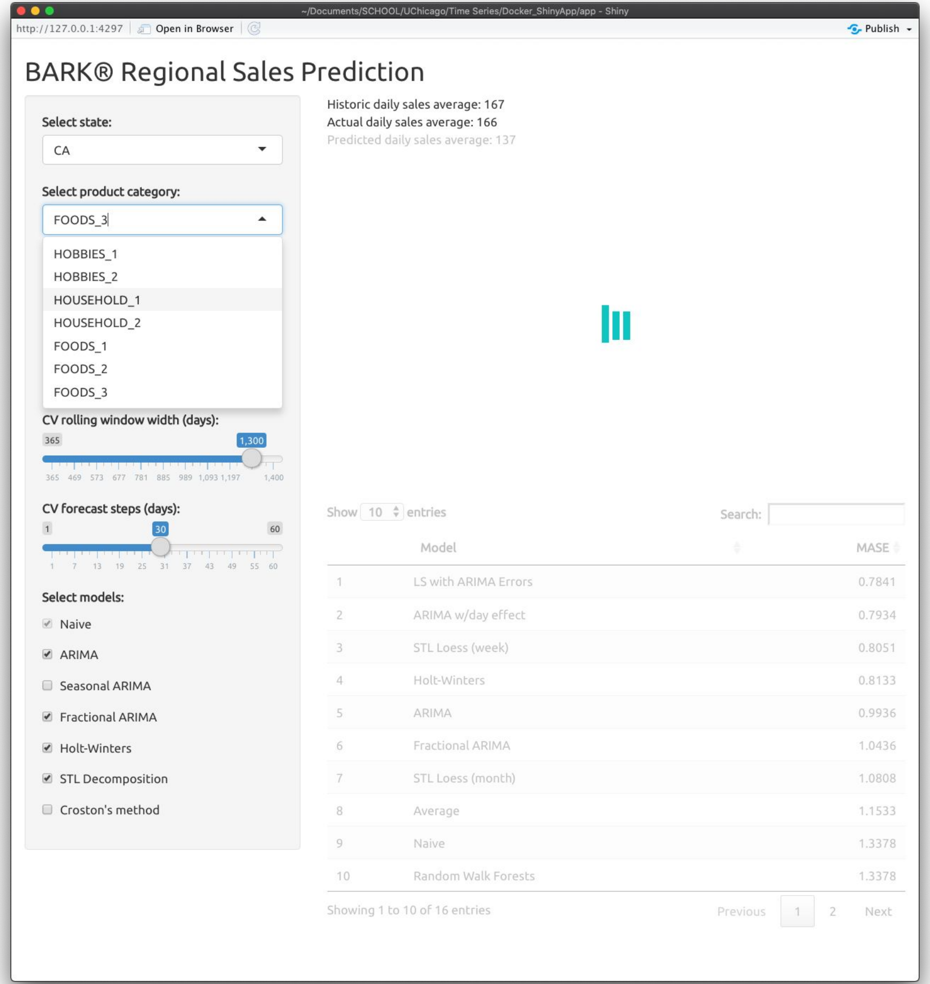
Application

> Shiny from R Studio

> Automated model selection with sliding-window CV

> Metric: avg. historical MASE

> Dockerfile for deployment on AWS/Google Cloud



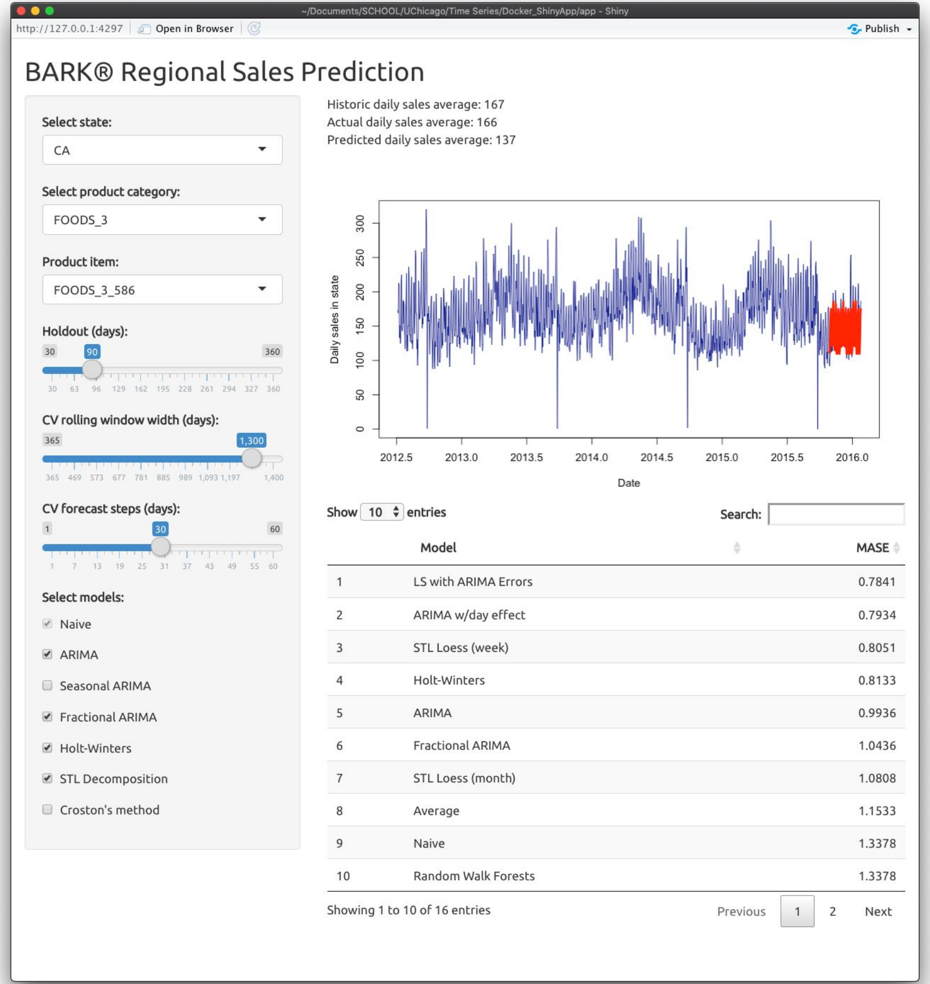
Application

> Shiny from RStudio

> Automated model selection with sliding-window CV

> Metric: avg. historical MASE

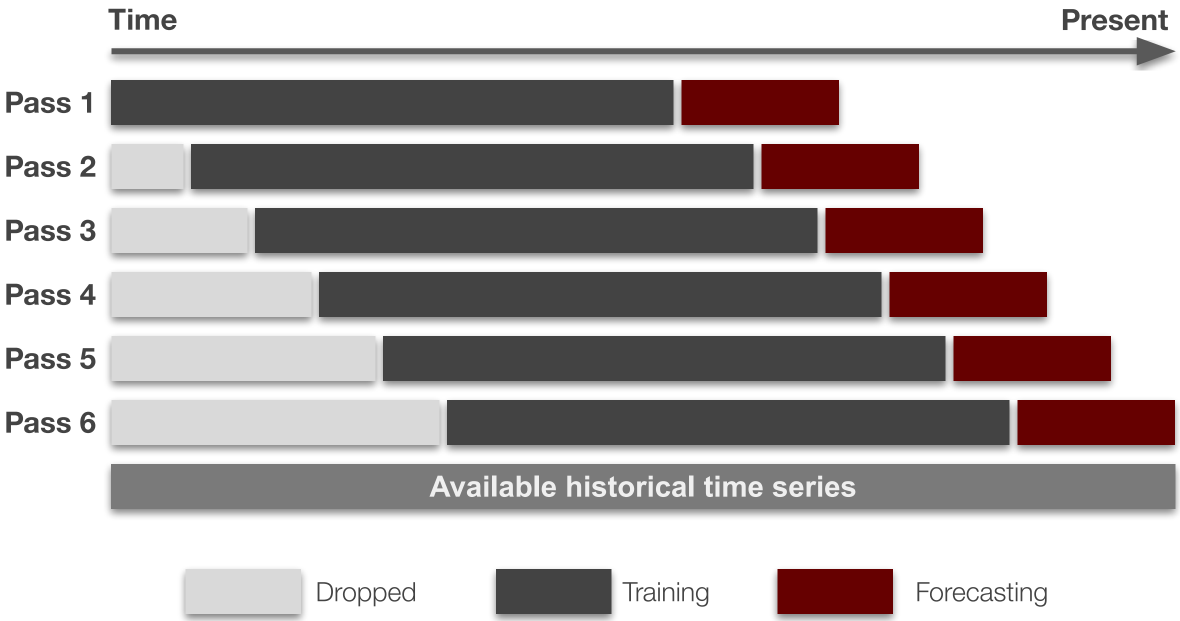
> Dockerfile for deployment on AWS/Google Cloud



BARK

Sliding-window cross-validation

> How does window width affect model selection prediction accuracy?



Sliding-window cross-validation

> How does window width affect model selection prediction accuracy?

Model performance	1-year window	2-year window	3-year window	4-year window
#1	LS w/ ARIMA errors **0.7326**	LS w/ ARIMA errors **0.7749**	LS w/ ARIMA errors **0.7841**	LS w/ ARIMA errors **0.7909**
#2	Croston's Method **0.7508**	STL weekly **0.7930**	ARIMA w/ day-of-week **0.7934**	ARIMA w/ day-of-week **0.7941**
#3	ARIMA w/ day effect **0.7520**	ARIMA w/ day-of-week **0.7945**	STL weekly **0.8051**	STL weekly **0.8093**

Number beneath models indicate average cross-validation MASE
Series: CA:FOODS_3_586

Train_test_split LS w/ ARIMA errors MASE: **0.8964**

BARK

> Introduction

> Models

> Selection criteria

> Feature extraction

> Results

> Application

> Cross-validation

> Takeaways and future work

1	what classes of models perform best across categories/items?

> Autoregressive neural networks! All you need is one hidden layer and <20 neurons and you can whiten your stubbornest residuals

2	do we lose accuracy adopting a hierarchical approach? Are we ok with this tradeoff?

> Surprisingly no. Hierarchical approaches using ETS (both top-down and middle-out) often worked better than models fitted to individual items. A regularizing influence?

3	does cross-validation give us the same ranking of models?

> Yes.

4	how much data do we need to make a decent forecast?

> Two years of data seems to be sufficient.

Future work

> Model inter-dependencies (complements vs substitutes) among products via estimating cross-price elasticity

> Extend our methodology and incorporate more items (do the same results hold when the number of items is 1000+?)

> Try to incorporate more external regressors (outside those provided in the dataset)

Appendix: Individual Contribution

Benedict Au	Brea Beals	Mark Roberts	Yannik Kumar
● R Shiny application for sales predictions ● Automated model selection with rolling-window cross val. ● Simple, ARIMA, Holt Winters, STL, and Croston’s method	● Simple Forecasting Methods: ○ Average ○ Naive ○ Seasonal Naive ● ARIMA Models: ○ ARIMA ○ Seasonal ARIMA ○ ARFIMA ● Regression Models	● Holt-Winters models ● STL models ● Prophet models ● TBATS models ● Created all data visualizations (aka: ggplot expert)	● ARCH + GARCH combo models ● Neural networks ● All hierarchical models ○ Top-down (ARIMA) ○ Top-down (ETS) ○ Middle-out (ARIMA) ○ Middle-Out (ETS)