Springboard – DCS

Capstone Project 3

Forecasting Sales of Multiple Products Across Multiple Favorita Stores

By Michael Bobal

April 2023

Introduction

- In order to stay in business, restaurants and commercial grocery stores must offer prices that are commensurate with competitors, offer deals to entice customers, and accurately predict which products, and the quantity of those products, to keep in stock.

- According to Retail Wire, overstocking costs the average retailer 3.2% in lost revenue, while understocking items can cost 4.1%. A review of the data has shown that overstocks are costing retailers $123.4 billion every year, and understocks remove another $129.5 billion from net inflows.

Introduction

Goals:

- Use machine learning time series analysis to forecast sales of different types of items across dozens of stores.
- Empower Favorita to become more efficient with its distribution of resources
- Inform the company of:
    - the best times to offer discounts
    - whether to stock up on certain items
    - knowledge of general market trends

Approach: Data Acquisition and Wrangling

Data is from Kaggle competition

Alexis Cook, DanB, inversion, Ryan Holbrook. (2021). Store Sales - Time Series Forecasting. Kaggle. https://kaggle.com/competitions/store-sales-time-series-forecasting

This project consisted of 1782 total time series to analyze and forecast (16-day horizon)

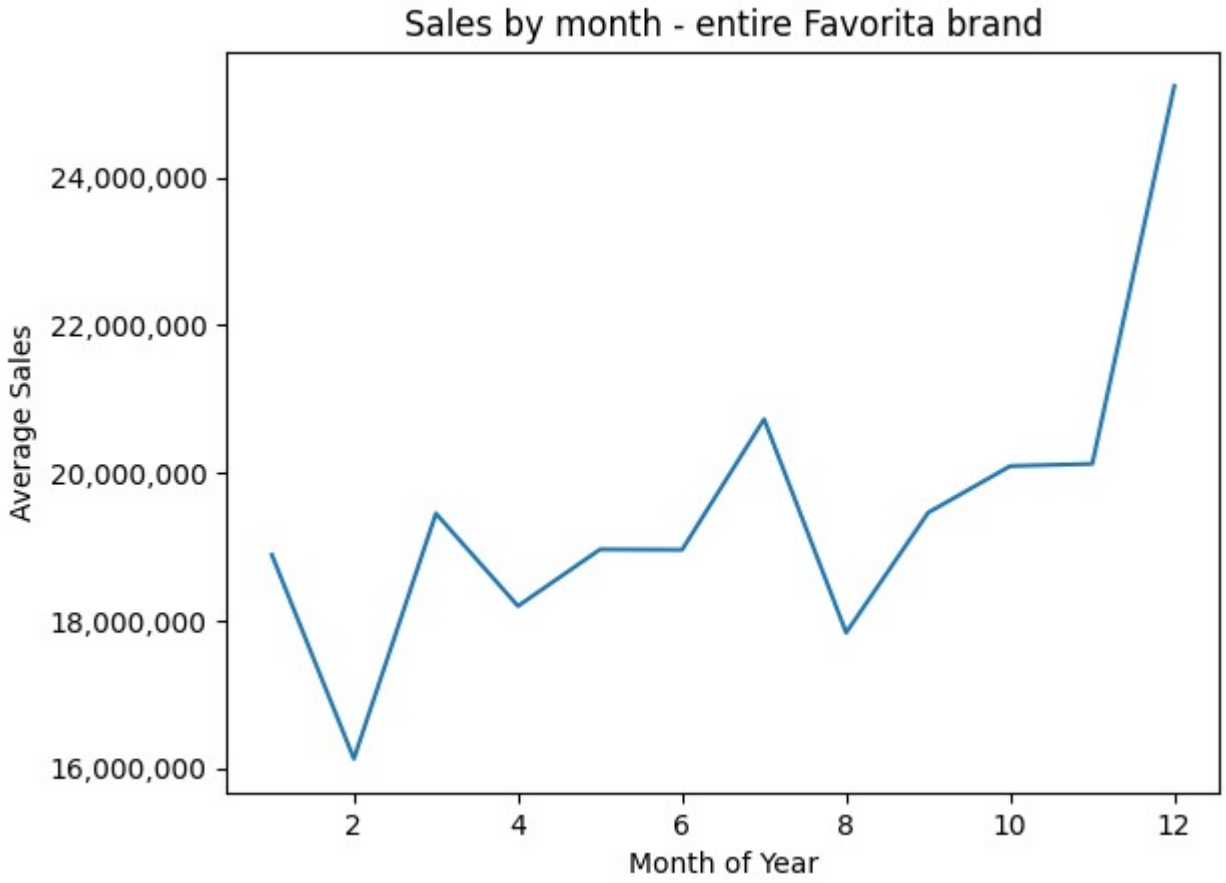
Our data:

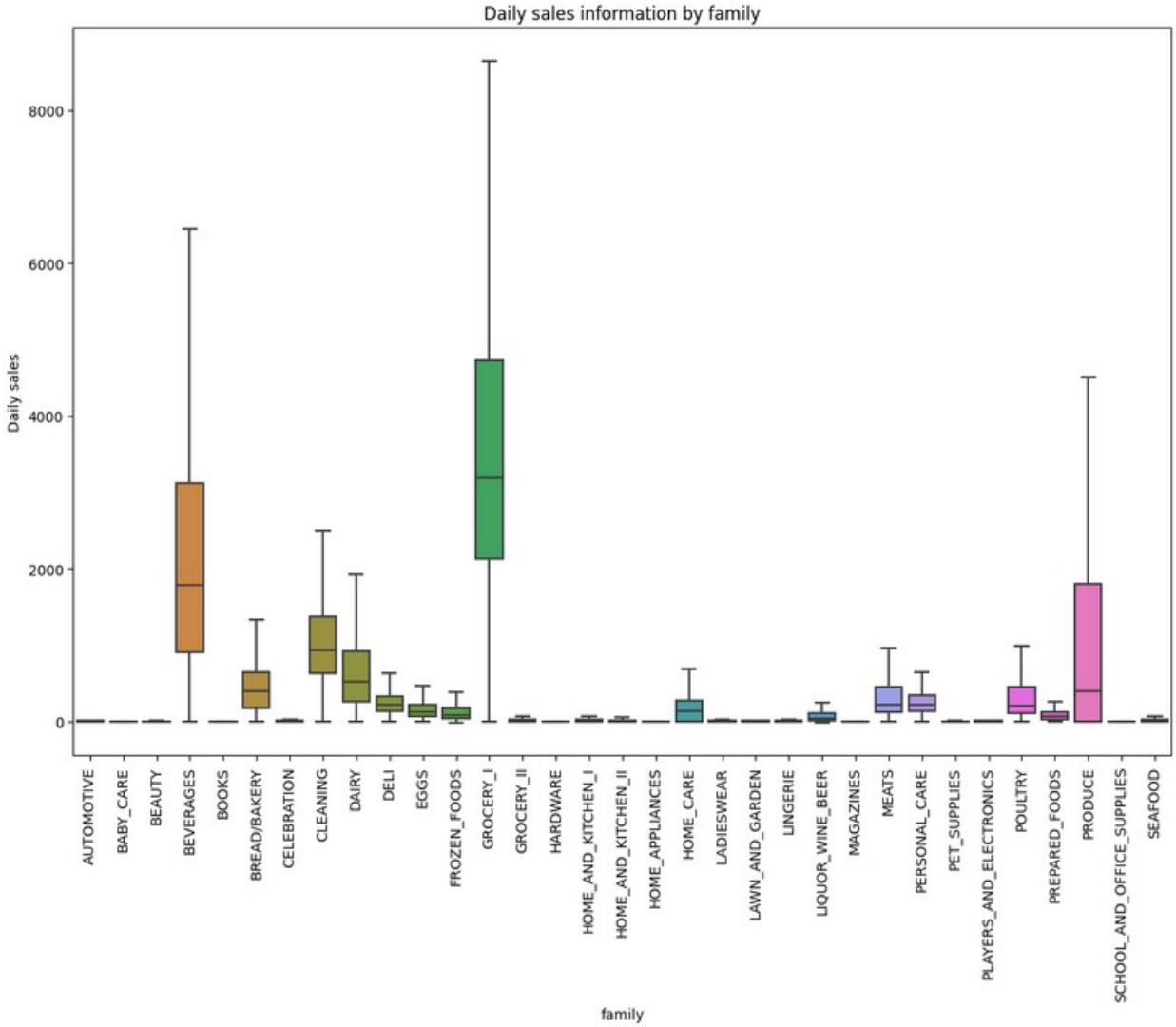
- 3,000,888 rows of data

- 54 stores

- 33 product categories

- 1782 time series





Feature engineering:

- Time gaps and missing data imputed and filled

- Multiple data sources merged:

  > Transactions

  > Oil price

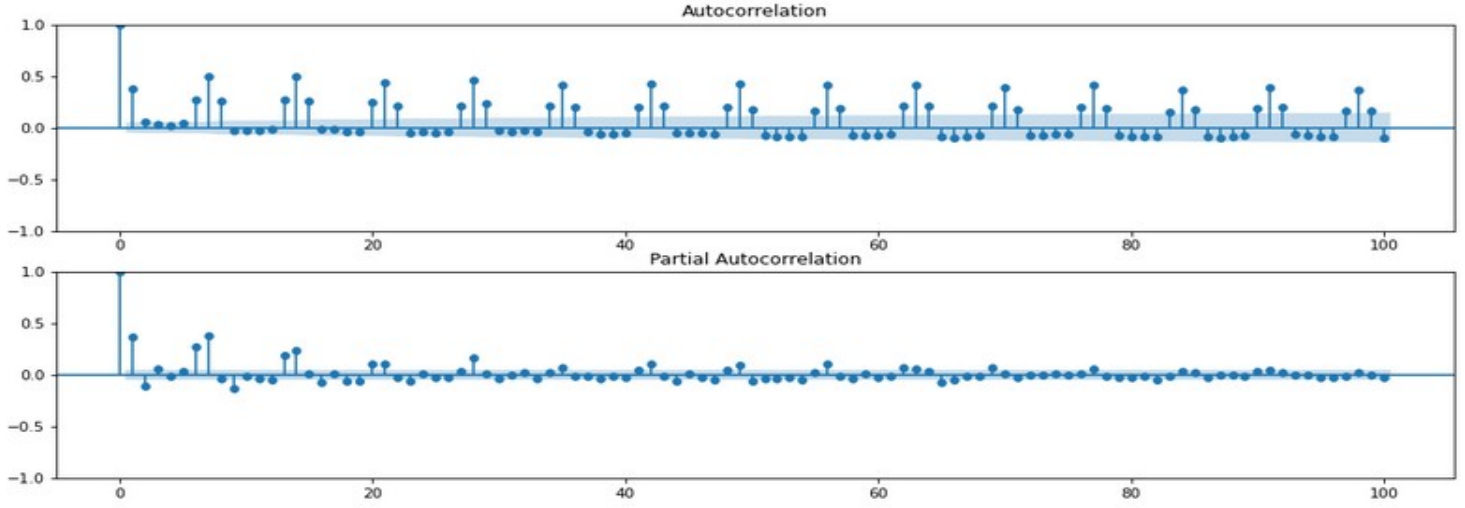
  > Holidays and events

Baseline Modeling

Goal of modeling: Accurate forecasts.

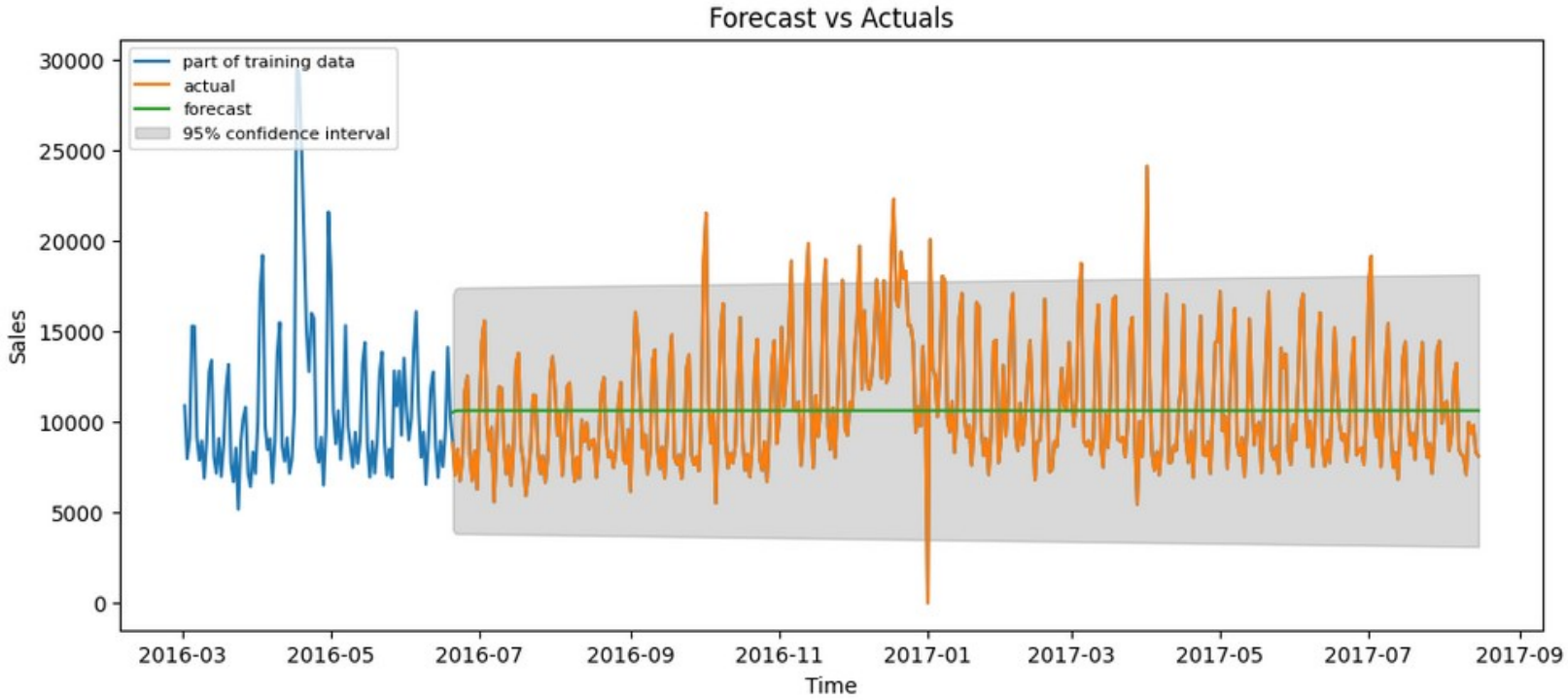
Performance metrics of choice: R^2 and MAPE

Reason: R^2 functions as a proxy “similarity score,” while Mean Absolute Percent Error gives a good idea of how wrong each forecast is, by percent.



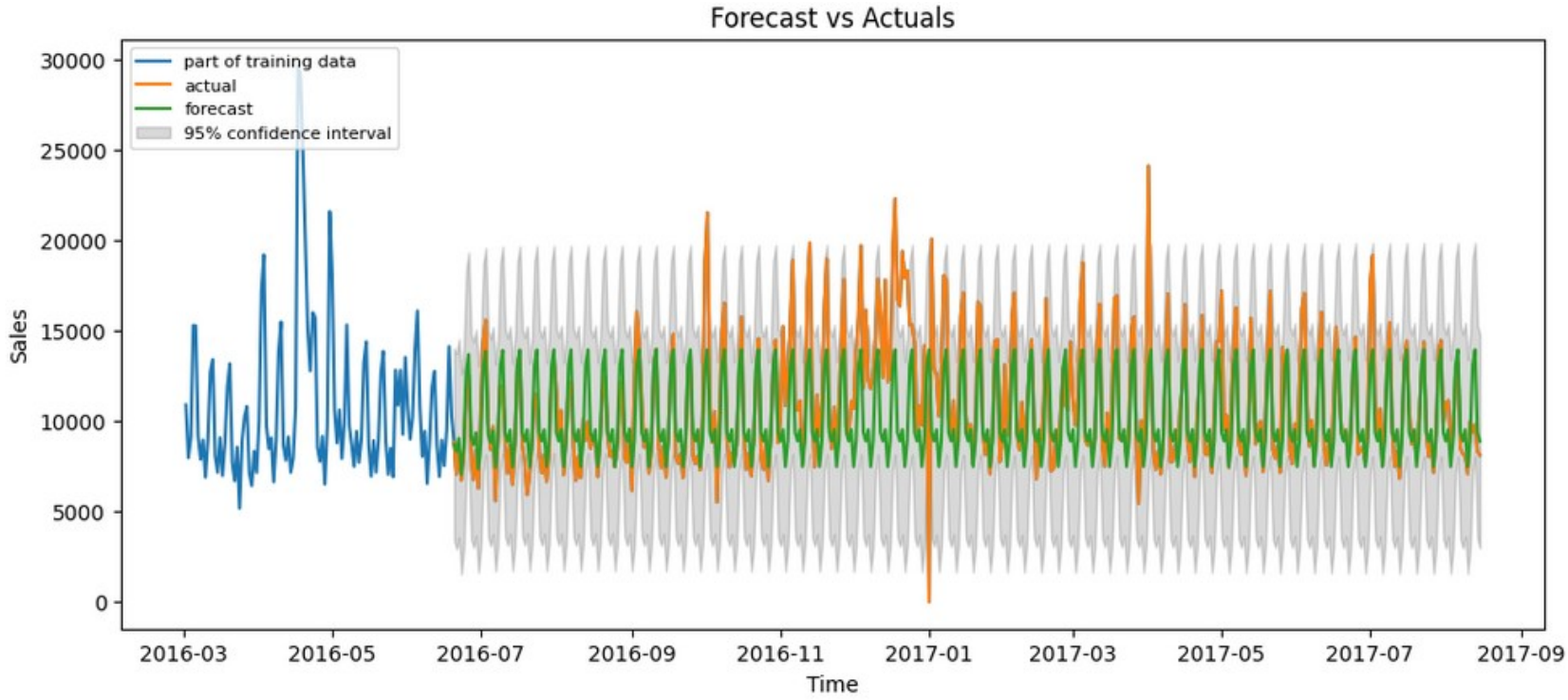
Simplest model: ARIMA with no tuning

Poor performance
R^2: 0.002
MAPE: 24.96



Next model: ARIMA weekly seasonality

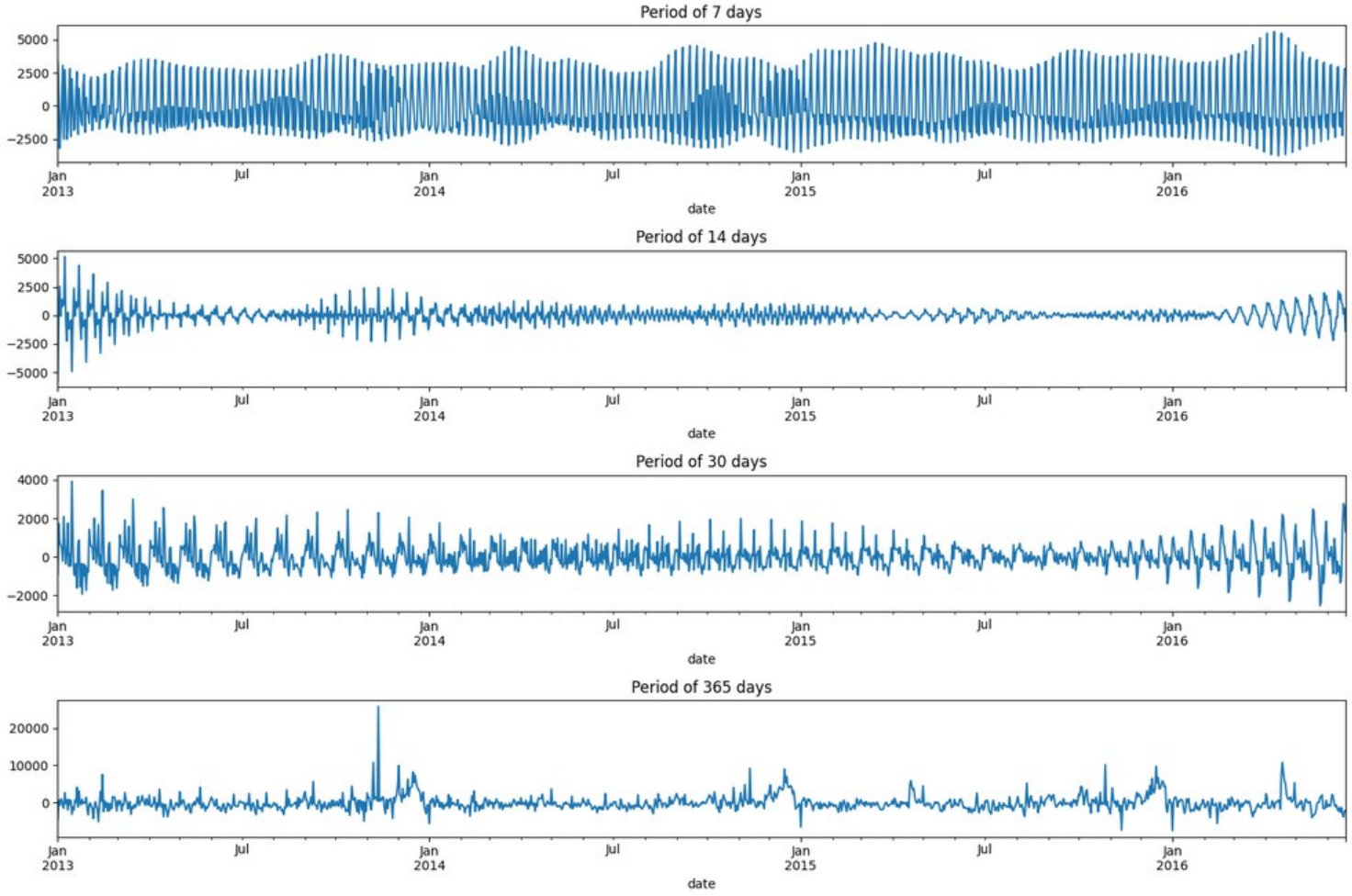
Fair performance
R^2: 0.51
MAPE: 13.93



Our data has at least 4 seasonal periods.

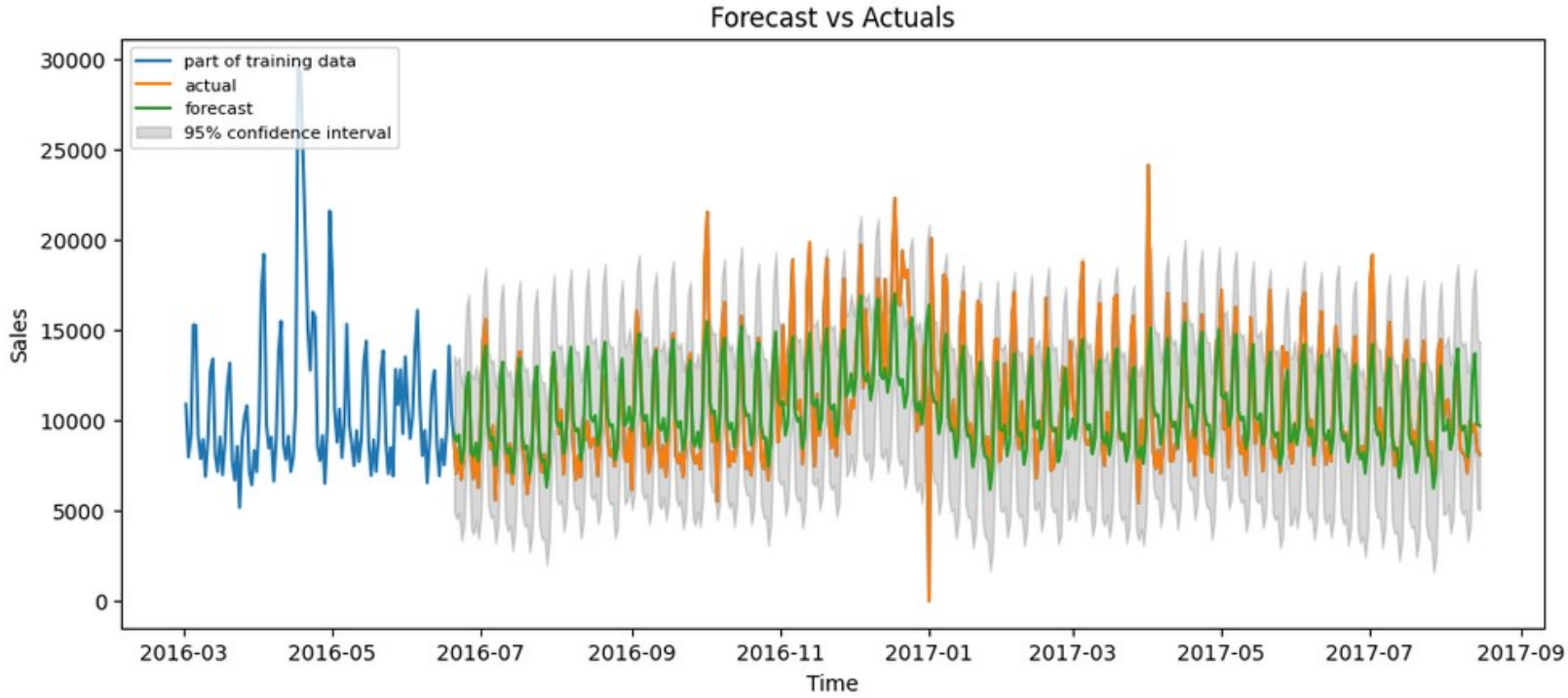
Issue:
ARIMA can only handle one seasonality at a time

Solution:
TBATS algorithm handles many seasonal periods



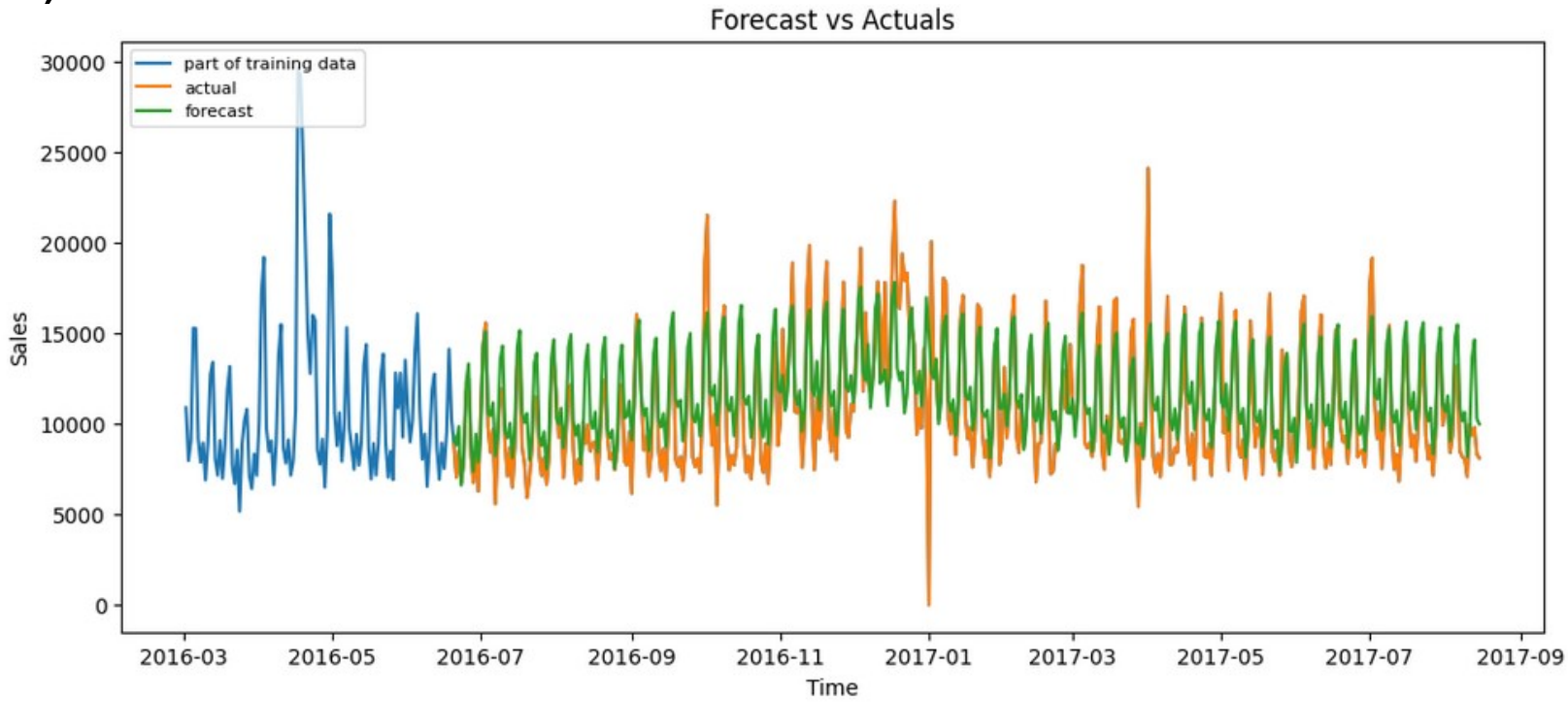
Next model: TBATS with seasonal periods of annual, monthly, bimonthly, and weekly

Great performance
R^2: 0.613
MAPE: 13.81



Next model: SARIMA with exogenous variables

Great performance
R^2: 0.649 (best)
MAPE: 16.76



Last model: TBATS/SARIMA with exogenous variables Ensemble

Great performance
R^2: 0.631
MAPE: 13.61 (best)

Issue: Runtime for any TBATS-inclusive model would take over 3 days to run all 1782 time series.

Solution:
Exogenous-inclusive SARIMA only takes 3 hours and is similarly accurate (with a better R^2 score).

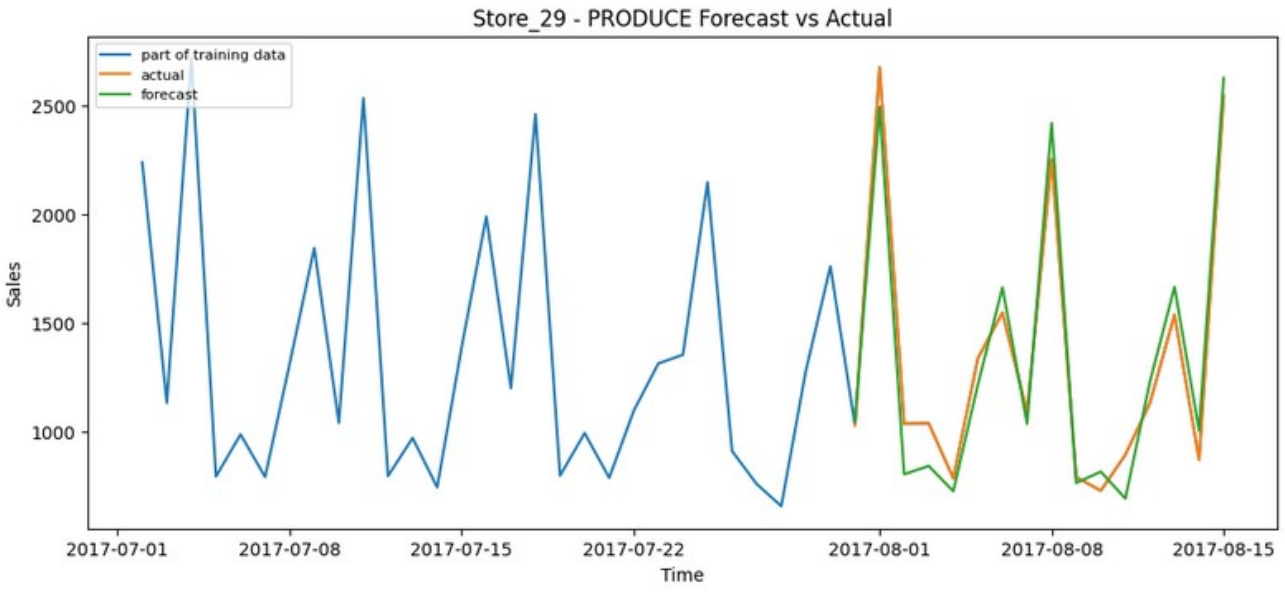
Model comparison:

Selected model >>>

	model_names	MAPE	R^2
4	ARIMA with exogenous variables	16.76	0.64902
7	ARIMA/TBATS combo	13.61	0.63149
3	TBATS2	13.81	0.61295
5	ARIMA with fourier exogenous only	16.82	0.5666
2	TBATS	15.42	0.5561
1	Seasonal_ARIMA	13.93	0.51081
6	ARIMA with promo exogenous only	33.94	0.4344
0	Basic_ARIMA	24.96	0.00222

Findings

Overall, my model performed very well, with many of the series having R^2 values greater than 0.90.



	Store/Product	MAPE	R^2
493	Store_22 - SCHOOL_AND_OFFICE_SUPPLIES	18.09	0.96659
723	Store_29 - PRODUCE	9.23	0.96396
1066	Store_39 - EGGS	52.61	0.95296
822	Store_31 - PRODUCE	7.85	0.95250
789	Store_30 - PRODUCE	12.41	0.95079
1152	Store_40 - PRODUCE	10.00	0.94494
1297	Store_45 - EGGS	7.56	0.93279
1705	Store_7 - LIQUOR_WINE_BEER	15.27	0.92791
855	Store_32 - PRODUCE	10.97	0.92781
1542	Store_51 - MEATS	9.58	0.92661
835	Store_32 - EGGS	25.96	0.92587
750	Store_3 - MEATS	9.87	0.92564
327	Store_18 - PRODUCE	13.36	0.92527
816	Store_31 - MEATS	8.67	0.91414
121	Store_12 - LIQUOR_WINE_BEER	48.85	0.91150
358	Store_19 - POULTRY	20.53	0.90318
420	Store_20 - MEATS	10.58	0.90048
228	Store_15 - PRODUCE	12.43	0.89608
261	Store_16 - PRODUCE	10.80	0.89121
783	Store_30 - MEATS	14.09	0.89065

Findings

Based upon the unbalanced sales totals across product categories, I decided to weight the final overall model metrics by sales volume.

Final model metrics:

	Before weighting:	After weighting:
Whole-project MAPE:	53.2	18.14
Whole-project R^2:	0.327	0.484

Conclusions

Our model was a success:
- Favorita daily sales data are able to be accurately predicted at least 16 days into the future using a tuned seasonal ARIMA model
- Consumable product categories have the highest accuracy forecasts.
- Highly efficient: both presice and fast

Future ideas:
-Use different techniques for modeling forecasts that use more of the exogenous variables.
    > Regression, Random Forest, XGBOOST, LGBM
- Use a larger forecasting horizon:
    > Try to forecast 2 months, 6 months, or 1 year.

Conclusions

Based on my findings, I can make the following concrete recommendations:

1. The seasonal cycles are strong in the food and beverage data. Favorita can prevent over/under stocking perishable goods by stocking such inventory with the assumption that our model's forecast will become reality, ensuring that customers do not encounter empty shelves, and that Favorita will not lose out on profit with extra items left to rot, unpurchased.

2. The weekly cycle appears to be the strongest seasonal period, with Sundays as a peak sales day for year-round. Since the foot-traffic is already present, offering special Sunday-only promotions on slower-selling non-perishable items could help to sell less-predictable, overstocked inventory. This would allow the stores to allot space more efficiently.

3. Local holidays correlate strongly with decreased sales numbers. Playing into these holidays with well-advertised "special holiday" promotions could help to minimize losses on those days.

Thank you to AJ, my mentor,
for his guidance throughout this project.