Big Data Project Report

Ahmed Essam, Ahmed Khaled, Sara Maher, Ibrahim Mahmoud

May 3, 2020

Abstract

This report gives a time series analysis of historical Walmart sales data about more than three thousand different products sold across three different states in the United States. The data is part of the M5 forecasting competition (Makridakis et al., 2020). In addition to the time series analysis we also describe our attempts at building forecasting systems on multiple levels and an analysis of its prediction.

Contents

1 Introduction

Sales forecasting is an important application. Especially for business as it enables companies to make better, more informed decisions. This is best done with enough information about previous sales at hand, and using it to predict future revenue and growth.

1.1 Problem Definition

We are given Walmart sales data from ten stores in three different states. Our project can be divided into two main problems: the first problem is to answer business-related questions on the dataset (data analysis), we consider some questions such as:

1. Which department is the most important by sales? Which department is the least important?

2. Which feast days generate the most sales? Which feast days generate the least?

3. Which items are the most profitable by total revenue? Where are these sold?

The second problem we tackle is to to predict sales twenty eight days into the future. We perform this on three levels: individual items, store, and category. We try three different models and report our results on them.

This document is organized as follows: we review the dataset next, perform exploratory data analysis in the succeeding section, and outline our efforts at developing a forecasting model (and its evaluation) in the last section.

1.2 Dataset Overview

The dataset is given by the *M5 competition* available on Kaggle[1] which is the current installment in the popular M competition series on forecasting (Makridakis et al., 2020). The dataset includes time series data of the sales of various Walmart store products divided hierarchically by the item level, department, product category, and geographical area. The dataset also includes explanatory variables such as price, promotions, day of the week, and special events (e.g. Valentine's Day, Orthodox Easter, and the Super Bowl, one of the largest sporting events in American Football). There are $3,075$ products classified in 3 product categories and 7 product departments. The products are sold across 10 stores in 3 different states (California, Texas, and Wisconsin). The total number of M5 series across the entire hierarchy is $42,840$. The dataset guide is given on the on the M5 competition website.

[1] https://www.kaggle.com/c/m5-forecasting-accuracy.

2 Data Analysis

2.1 Dataset Files

The dataset is divided into four files, each of which is described in Table 1.

Table 1: Dataset description

File name	Description	Shape
sales_train_validation.csv	Contains the sales data for each item.	30490×1919
calendar.csv	Contains calendar dates as well as any events that happen on each date.	1969×14
sell_prices.csv	Contains the sell price for each item divided by store and week.	6841121×4
sample_submission.csv	Contains sample submission forecasting for the competition.	60980×29

The columns in each file are described in Tables 2 to 4.

3 Exploratory Data Analysis

In this section, we analyze and visualize the data from 3 files; calendar.csv, sales_train_validation.csv, and sell_prices.csv.

3.1 Calender-Focused Analysis

The data starts at January 29, 2011 to June 19, 2016, with day-count up to 1969. The dataset contains 30 events, some of them are repeated annually (counting duplicates, there are 167 events). They are divided into: Cultural, National, Religious and Sporting events. With Religious events being the most frequent with 55 occurrence. We also have 10 SNAP days each month in each state. Columns descriptions are in Table 2.

3.2 Sales-Focused Analysis

The sales data covers 1913 days only, from *January 29, 2011* to *April 24, 2016*, unlike the calender which covered 1969 days. The data contains 30490 rows and 1919 columns, with sales data from 3 states; California (CA), Texas (TX) and Wisconsin (WI), with 4, 3, and 3 stores in each respectively. It also contains 4 categories of products; Hobbies, Household, Foods, with 2, 2, and 3 departments respectively. Columns descriptions are in Table 3.

Table 2: Calender details. SNAP stands for the Supplement Nutrition Assistance Program, which provides low income families and individuals with an Electronic Benefits Transfer debit card to purchase food products, this purchasing process is done monthly across 10 days.

File	Column name	Column description
calender.csv	`date`	The date in Year-Month-Day format.
	`wm_yr_wk`	Code the week year from the dataset starting date.
	`weekday`	The weekday.
	`wday`	The weekday's number, starting from Saturday.
	`month`	The month.
	`year`	The year.
	`d`	The incremental ID of the day in the dataset.
	`event_name_1`	Name of the event occurring in this day, if one exists.
	`event_type_1`	Type of the event occurring in this day, if one exists.
	`event_name_2`	Name of the second event occurring in that date,if one exists.
	`event_type_2`	Type of the second event occurring in that date, if one exists.
	`snap_CA`, `snap_TX`, `snap_WI`	Equals 1 if SNAP purchases are allowed on this date.

Table 3: Sales details

File	Column name	Column description
sales_train_validation.csv	`id`	With the id of the item codes as *item_id_store_id*_validation.
	`item_id`	Product ID.
	`dept_id`	Department ID.
	`cat_id`	Category ID.
	`store_id`	Store ID.
	`state_id`	State ID.
	`d_1`, `d_2`, . . . , `d_1913`	1913 columns, containing the total sold items each day.

3.2.1 Sales Exploration Figures

What is the product distribution like?
Products are distributed equally on the different stores by Walmart, without favoring any of the stores, as shown in Figure 1.

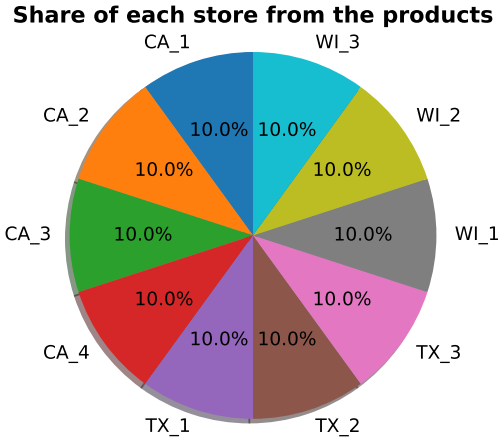
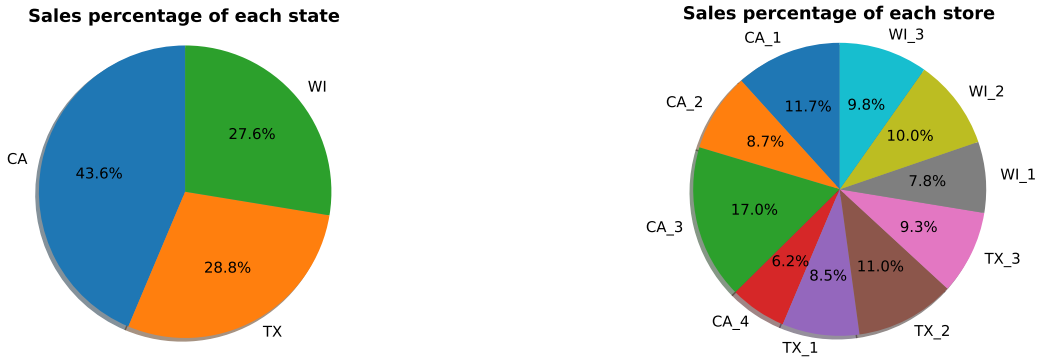


Figure 1: Share of each store from the products

How good is this distribution scheme?
To know that, we need to check the total sales and revenue. So we add a total sales column to the data. From Figure 2, we find that no, not all states are equally profitable, California provides 1.5 of the sales of each of Texas and Wisconsin. CA_3 is especially profitable, amounting to 17% of the total sales, while CA_4 is the least profitable, with 6.2% of the total sales.

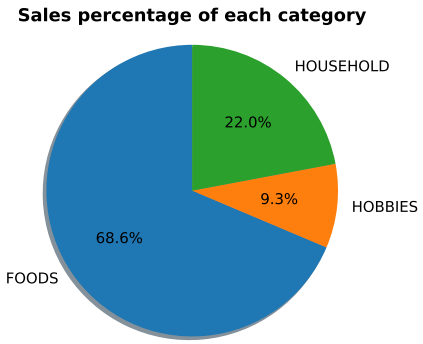


(a) Sales percentage of each state

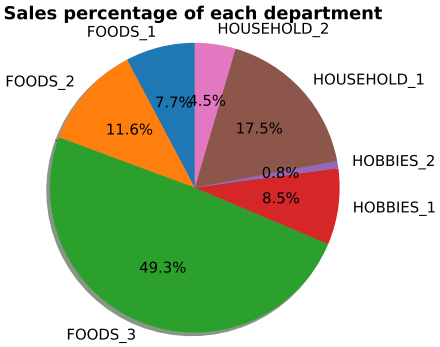
(b) Sales percentage of each store

Figure 2: Sales percentage for the states and stores

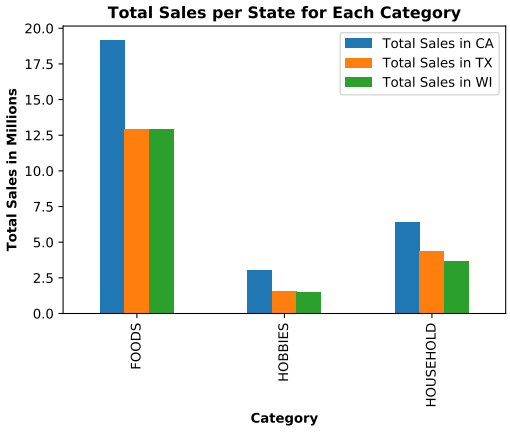
What is the most popular category?
We found that FOODS is the most popular category, with a staggering 68.6% of the total sales, and the most popular department is FOODS_3, with 49.3% of the sales. While the least popular category and department is HOBBIES and HOBBIES_2, with 9.3% and 0.8% of the total sales respectively. This is shown in Figure 3. We notice that CA_3 has the highest percentage of HOUSEHOLD_2 sales, while WA_2 has the highest percentage of FOODS_2 sales, and WA_3 has the highest percentage of FOOD_3 sales.



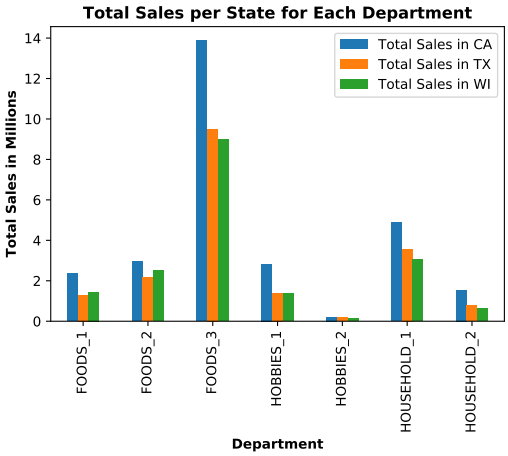
(a) Sales percentage of each category



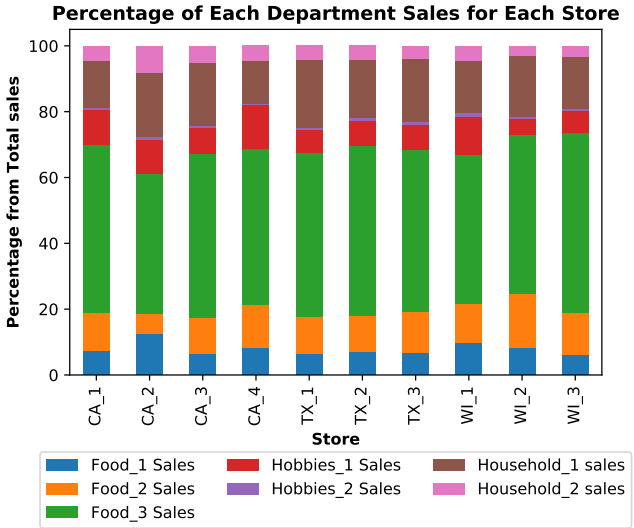
(b) Sales percentage of each department



(c) Total Sales per State for Each Category



(d) Total Sales per State for Each Department

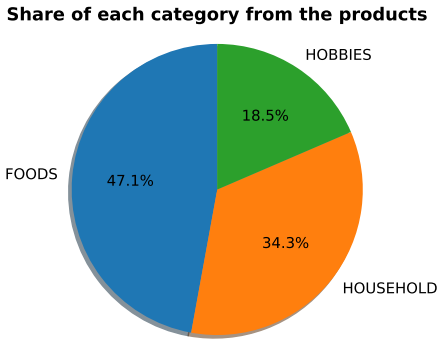


(e) Percentage of Each Department Sales for Each Store

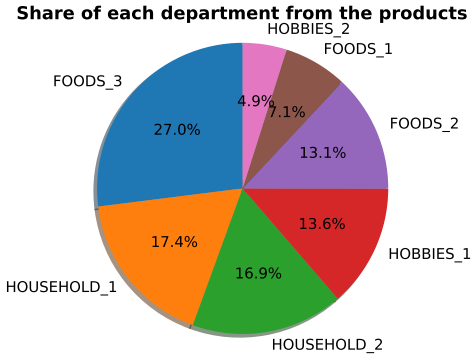
Figure 3: Total sales for the categories and departments

Is the current stocking scheme good?

Yes, Walmart stocks up on a variety of products of the most popular category and department very well, as shown in Figure 4.



(a) Share of each category from the products

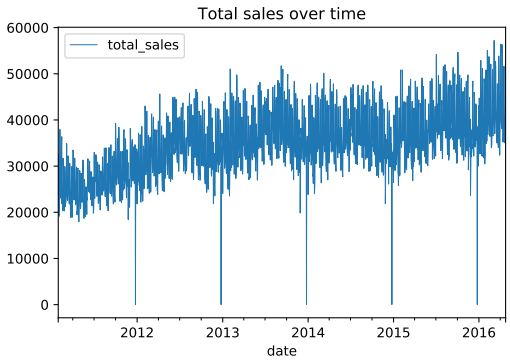


(b) Share of each department from the products

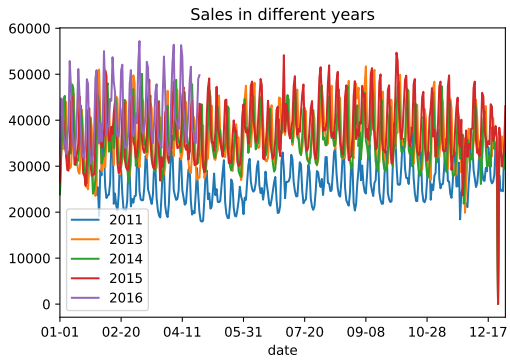
Figure 4: Share of each category and department from the products

How do sales fluctuate over time?

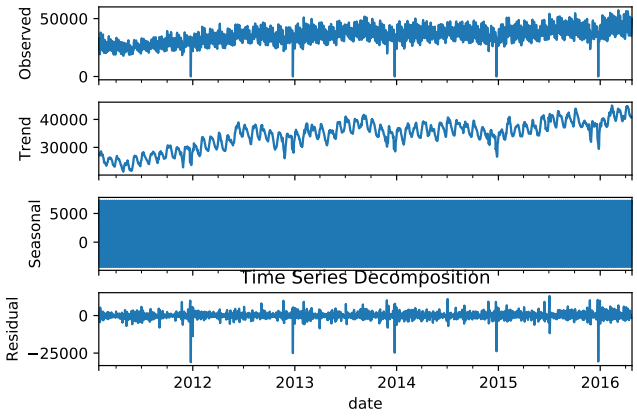
Generally, they increase throughout the covered period as shown in Figure 5a, especially from 2011 to 2012 as shown in Figure 5b. The observed sales data has a trend and it is highly seasonal, as shown in Figure 5c. For California and Texas, the seasonality is similar. while it differs for Wisconsin. The trend increase in Texas is the least and the slowest, these are shown in Figure 6



(a) Total sales over time

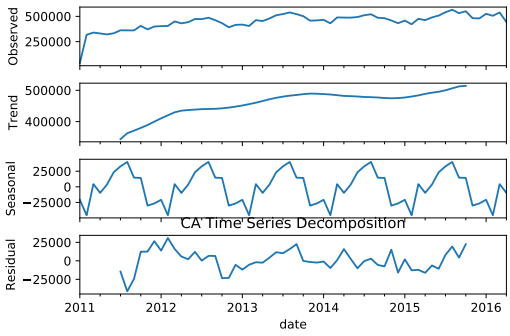


(b) Sales in different years

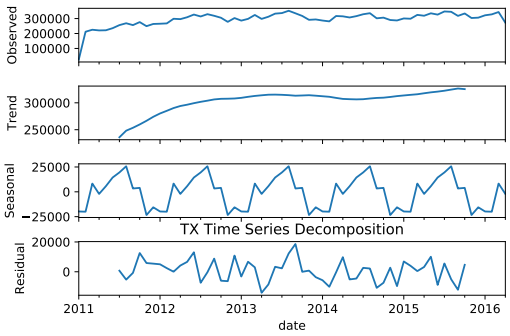


(c) Time Series Decomposition

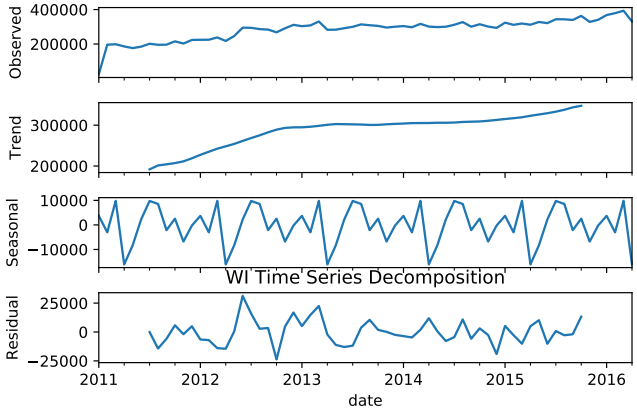
Figure 5: Sales as a time series



(a) CA Time Series Decomposition



(b) TX Time Series Decomposition



(c) WI Time Series Decomposition

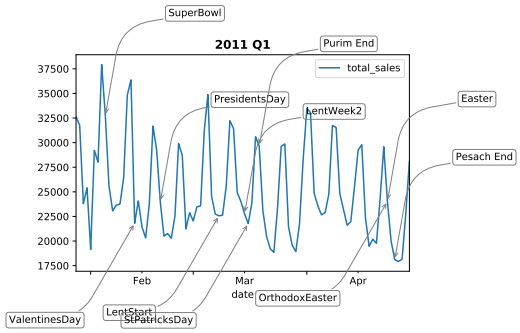
Figure 6: Time Series decomposition of sales at each state

How do events affect sales?

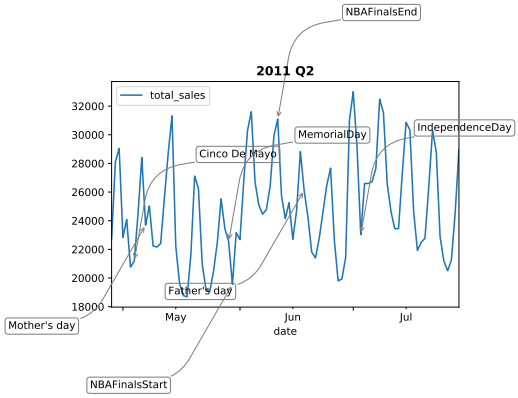
By zooming in on one year specifically, from January 29, 2011 to January 29, 2012, shown in Figure 7, we find that, surprisingly, sales don't go through a substantial spike on most events days. Still we find that from all the events categories, days of sporting events have the highest sales, while national event days have the lowest ones, as shown in Figure 8b. On the contrary, the highest sales day given an event is on the Labour Day, followed by the SuperBowl day, and the lowest is Christmas day[2], followed by Thanksgiving as shown in Figure 8a. [3]

[2]on which the sales are zero, so these Walmart stores probably close

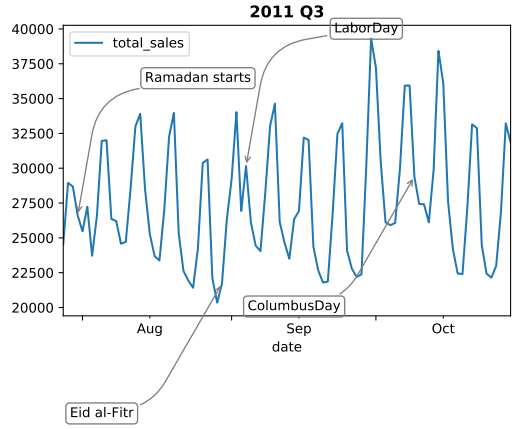
[3]We skipped events_2 in some of the calculations as it contributed very little to the total events (4 out of the total 167, counting duplicates)



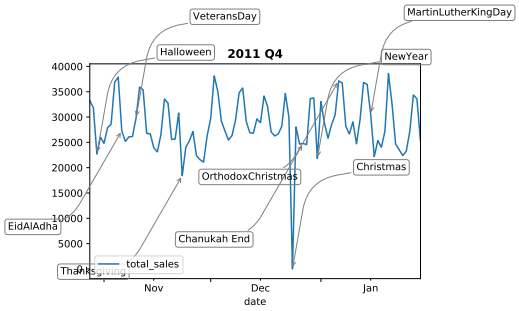
(a) First Quarter



(b) Second Quarter

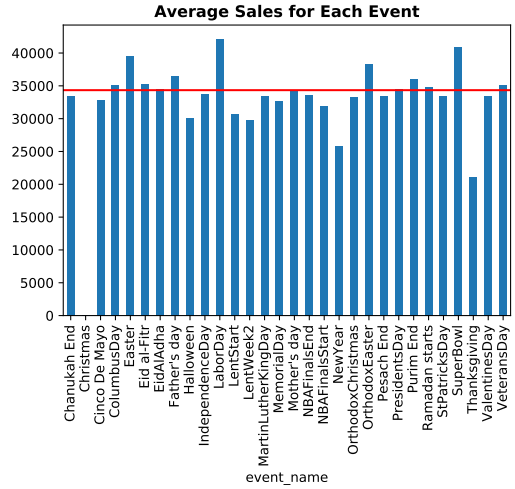


(c) Third Quarter

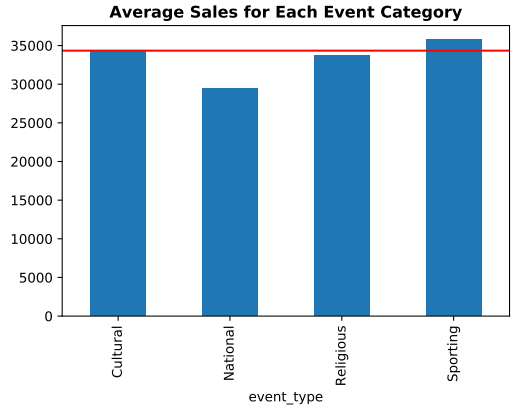


(d) Fourth Quarter

Figure 7: Events occurring in 2011



(a) Average Sales for Each Event

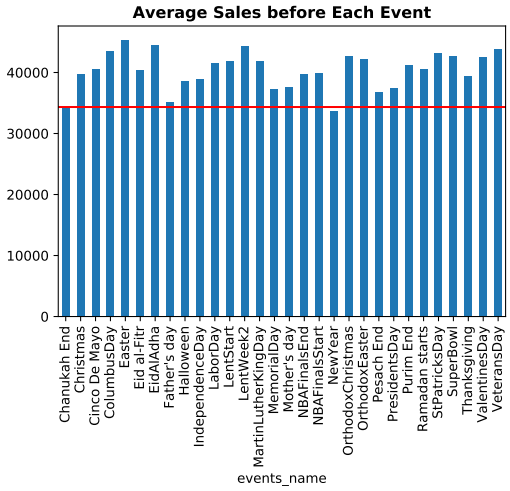


(b) Average Sales for Each Event Category

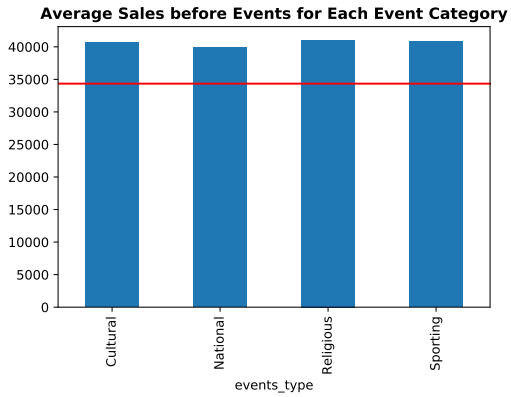
Figure 8: Average Sales given an event

Also by analyzing the sales on the weekend before the events, which are a high point for

sales usually as shown in Figure 11, we found that there is an increase to the average, as shown in Figure 9. The weekend before Easter is the most profitable weekend before an event, while the least profitable one is New Year. For the categories, the most profitable are religious events, followed by sporting events, with the least profitable being national events. And this contrast makes sense, as people would usually shop for watching a match on the same day, it demands only a short-term arrangement, while they would shop for celebrating a religious feast, which is a family event, in advance.



(a) Average Sales before Each Event

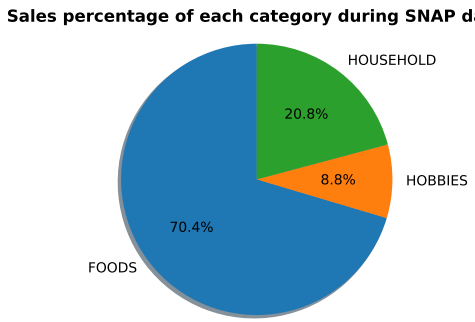


(b) Average Sales before Events for Each Event Category

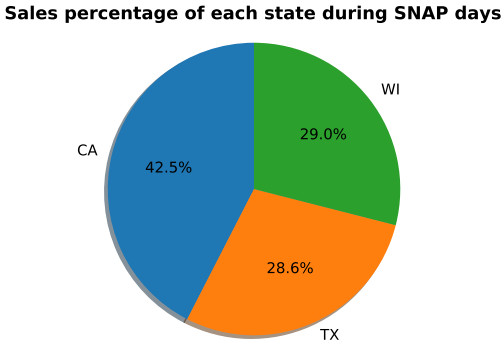
Figure 9: Average Sales on the weekend before an event

How are sales affected by SNAP days?

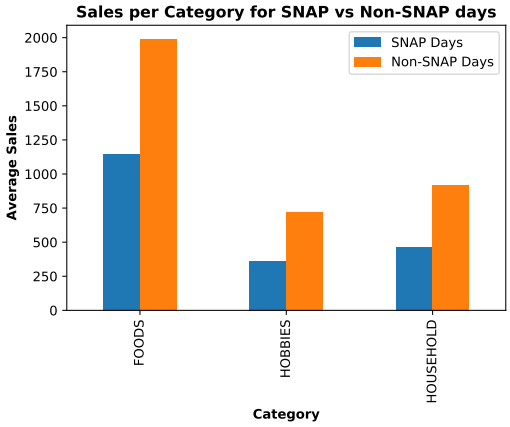
SNAP days make up only third of the month and by analyzing, it was found that the sales on SNAP days comprise about a third of each months sales of each category, except for FOODS, of which they comprise 36.5%. We also found that the highest sales percentage on SNAP days is in Wisconsin, with 37.4%, and the least is in California, with 34.65%. FOODS is still themost popular category, with 70.4% of the total sales, and California is still the most profitable state, but Wisconsin's share increases from 27.2% to 29.0%. This is shown in Figure 10.



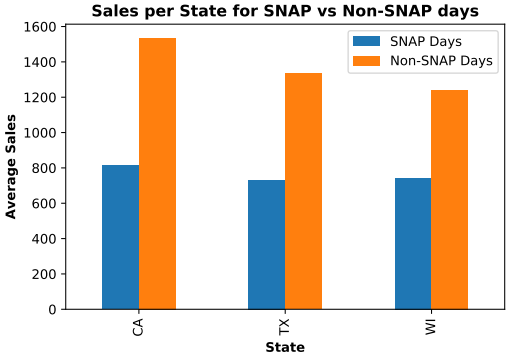
(a) Sales percentage of each category during SNAP days



(b) Sales percentage of each state during SNAP days



(c) Average Sales per Category for SNAP vs Non-SNAP days



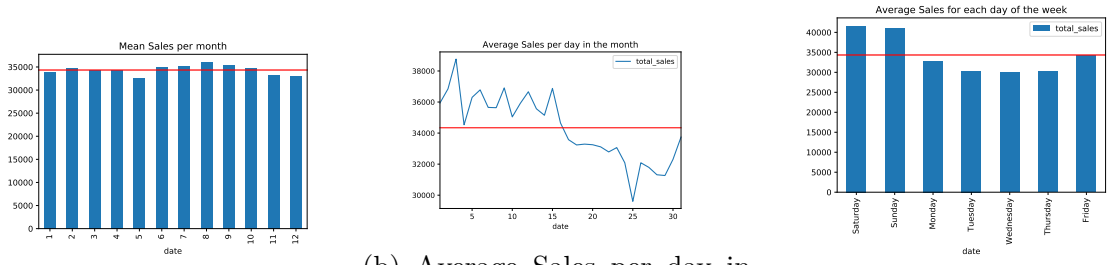
(d) Average Sales per State for SNAP vs Non-SNAP days

Figure 10: SNAP days sales

When are the highest sales?

As shown in Figure 11, August is the highest average sales month, which makes sense as the time series has a peak in the seasonality each August [4], this could be due to summer vacations, so both parents and non-parents have extra time for shopping. Each month, the sales are higher in the first 15 days of the month, which makes sense as it is the time where people are more prone to buy things, after cashing the monthly paycheck, and most of the sales happen on the weekend.

[4] the highest overall month is March



(a) Mean Sales per month

(b) Average Sales per day in the month

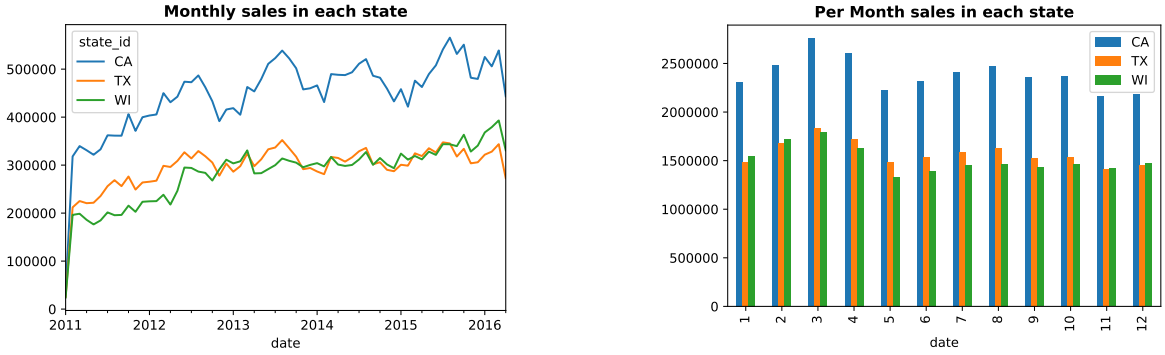
(c) Average Sales for each day of the week

Figure 11: Average Sales given a certain month, day, or weekday

Zooming in on states, we find that the absolute month with the maximum sales for each state:

- California: August, 2015.
- Texas: August, 2013.
- Wisconsin: March, 2016.

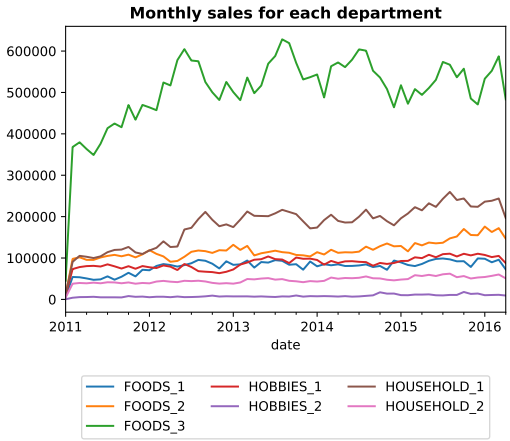
With California showing the maximum increase over the years. The month with the maximum sales in each state was the same as the one with the maximum total sales, March. This is shown in Figure 12. This is also the case for departments, shown in Figure 13.



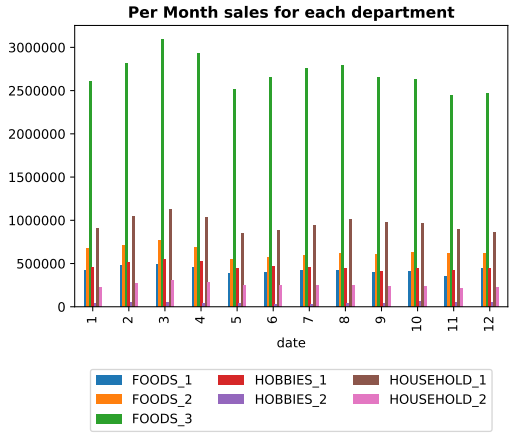
(a) Monthly sales in each state

(b) Per Month sales in each state

Figure 12: Total monthly sales given a state



(a) Monthly sales for each department



(b) Per Month sales for each department

Figure 13: Total monthly sales given a department

3.3 Prices-Focused Analysis

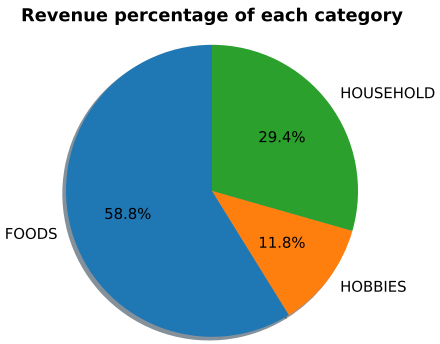
The sell prices data has the weekly price for each product from January 29, 2011 to June 19, 2016, so the price data is for all the calender file weeks, unlike the sales data. We found that the prices of products are not the same in different stores, even at the same day. Also, that the product with the greatest price change is HOUSEHOLD_2_406 sold in WI_2 store, having a minimum price of 3.26 USD and a maximum price of 107.32 USD. Moreover, There are 8247 products, in different stores, that don't undergo any price change through the 5 years; with 3497 household items, 2907 foods items, and 1843 hobbies items. Some of these products' prices haven't changed in any of the stores, like: FOODS_3_154, HOUSEHOLD_2_322, HOUSEHOLD_1_538, FOODS_3_309 and HOBBIES_2_014. Columns descriptions are in Table 4.

Table 4: Prices details

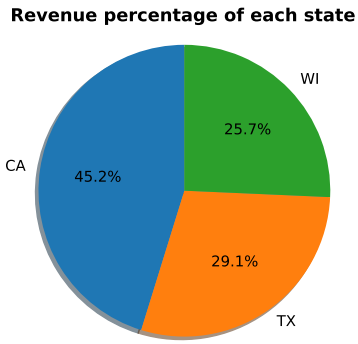
File	Column name	Column description
sell_prices.csv	`store_id`	store ID.
	`item_id`	The product ID coded as in Table 3 .
	`wm_yr_wk`	Code the week month and year from the dataset starting date.
	`sell_price`	Selling price of each product, for each week, in each store.

3.3.1 Prices Exploration Figures

What are the most profitable states? and categories?
Comparing the total revenue with the total sales, we find that they match, with FOODS being the highest grossing category, and California being the most profitable state. This is shown in Figure 14.



(a) Revenue percentage of each category



(b) Revenue percentage of each state

Figure 14: Revenue percentage

What are the most profitable items?
The top 5 items that generate the most revenue are FOODS_3 items, and they are[5]:

- FOODS_3_120, sold in CA_3 with revenue 197541.66.
- FOODS_3_090, sold in CA_3 with revenue 173741.55.
- FOODS_3_586, sold in TX_2 with revenue 171385.16.
- FOODS_3_120, sold in CA_1 with revenue 151387.02.
- FOODS_3_586, sold in TX_3 with revenue 137892.80.

More details can be found in the `data_exploration.ipynb` notebook.

[5] in million USDs

Table 5: Root Mean Square Error for item-level prediction by method.

Model Used	Root Mean Square Error
ARIMA	1.9555
ExpSmoothing	1.9338
Prophet	2.0887

4 Forecasting

All the results of this section can reproduced by running the `Forecasting.ipynb` notebook.

4.1 Description of Models Used

We use mainly three different models:

1. ARIMA models: studied in the course.

2. Exponential smoothing models: given an input time series $(x_t)_{t=1}^{T}$ we set $y_0 = x_0$ and predict a forecasting sequence $(y_t)_{t=1}^{T+k}$ by

$$y_t = \alpha x_t + (1 - \alpha)\, y_{t-1},$$

   where $\alpha > 0$ is a smoothing parameter that can be estimated from the data.

3. Facebook Prophet: A toolkit from Facebook which does forecasting based on an additive noise model which tries to model yearly, weekly, and daily seasonality as non-linear additive noise. We turn off yearly seasonality for our experiments.

4.2 Item-Level Predictions

To come up with item-level predictions, we judged that the dataset is too large to use at once. We used random subsampling and chose a small subset of items (only 30 items) to run experiments on, and created a train-test split. The training set contained 30 items with 28×4 days of observation data for each. The test set contained item sales for twenty eight days after the last sold item. We calculated the root mean square error over all items for each model used over a forecast length of 28 days, and it is tabulated as Table 5, where we can see that both ARIMA and ExpSmoothing led to very similar performance and were better than Prophet.

To gain more insight into the difference between the methods, we plot the predicted time series for four randomly chosen items in Figures 15 to 17. We see that ARIMA and ExpSmoothing both seem to capture the general trend relatively well, while Prophet is too conservative and predicts too close to the mean.

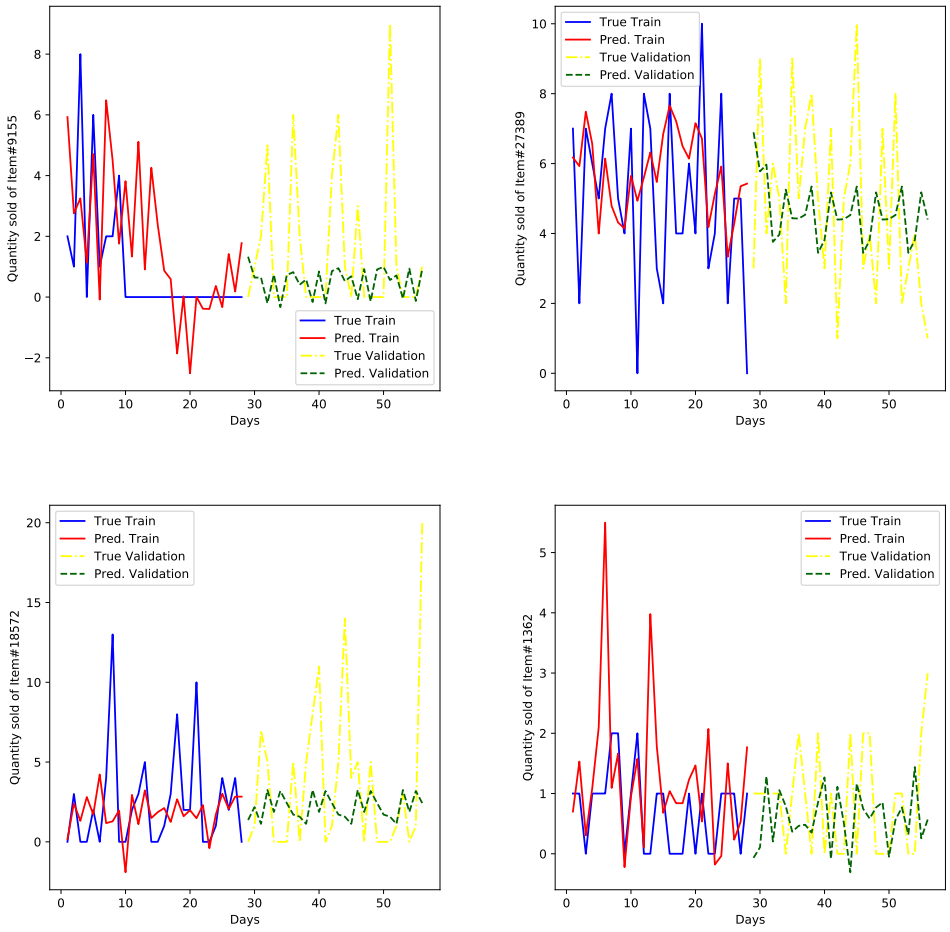


Figure 15: Item sales as predicted by ARIMA against the ground truth on four randomly chosen items.

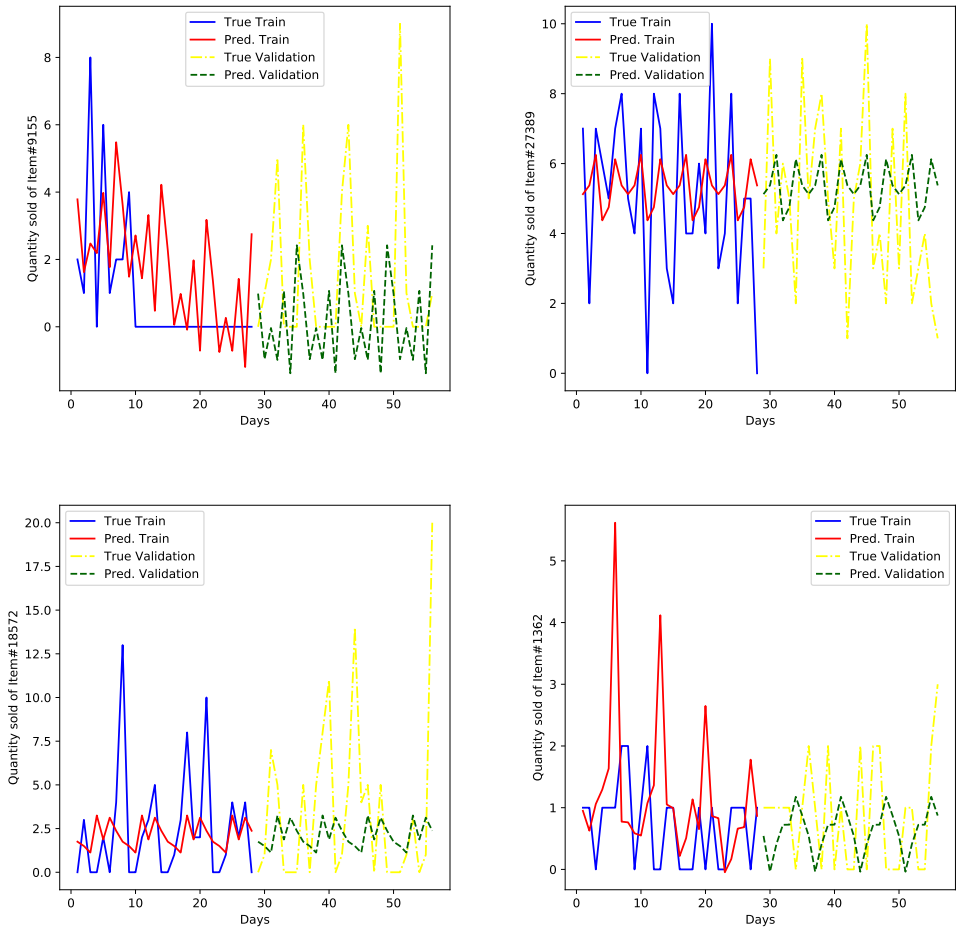


Figure 16: Item sales as predicted by ExpSmoothing against the ground truth on four randomly chosen items.

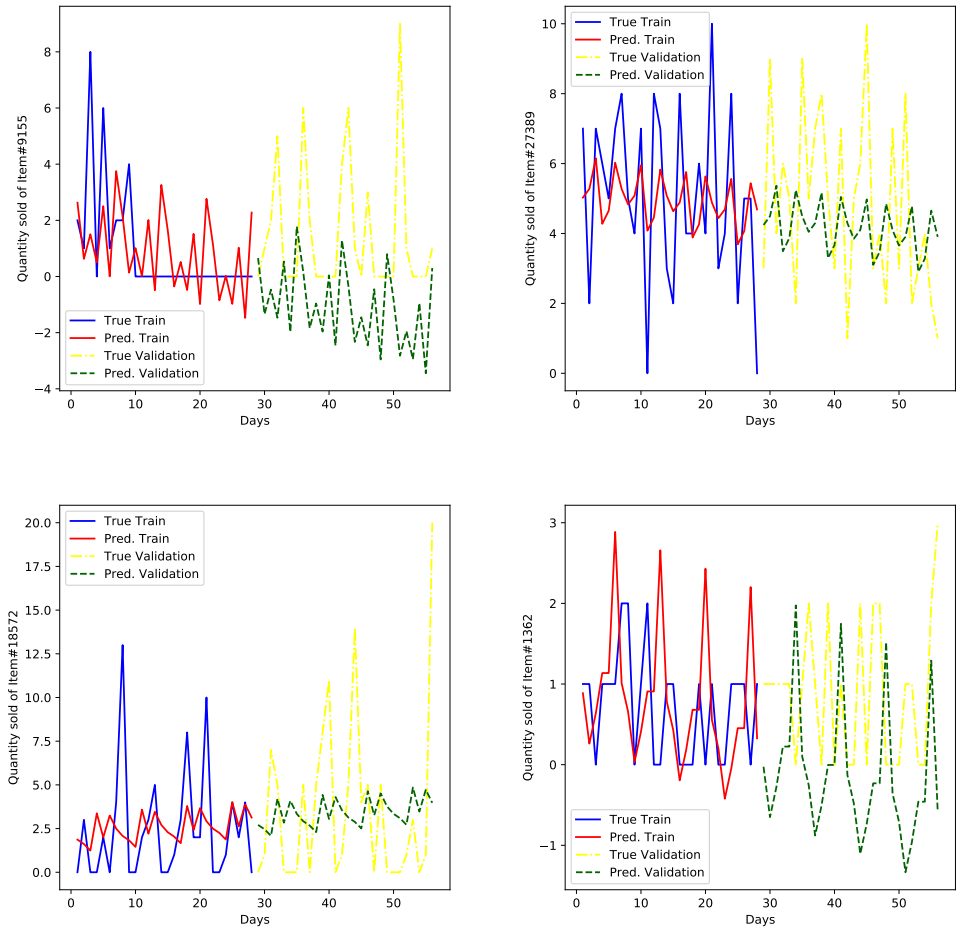


Figure 17: Item sales as predicted by Prophet against the ground truth on four randomly chosen items.

4.3 Category and State-Level Predictions

There are three main categories in this dataset: Foods, Hobbies, and Household items, and all items are sold in three states: California, Texas, and Wisconsin. We run forecasting by using the same three models from the previous section but this time we try to predict the volume of sells in each item category as well as in each state. As the number of item categories (3) and the number of states (3) is manageable, we were able to train using the entire history of observations for this task. The root mean squared errors are tabulated in Tables 6 and 7. Surprisingly, we see that Prophet performs best on this task, and this is because the quantity of data is too large for a single smoothing parameter to capture well (as in ExpSmoothing) and because Prophet models seasonality by a stronger model than ARIMA. Compared to the items (where we used only 28×4 days), here we use the entire history (about 1900 days). We plot the time series predictions for the last 28 days of the training set and forecast 28 days into the future in Figures 18 to 20 for categories and Figures 21 to 23 for states. We observe that the closest fit is Prophet, followed by ARIMA and ExpSmooth.

Table 6: Root Mean Square Error for category-level prediction by method.

Model Used	Root Mean Square Error
ARIMA	6.8457
ExpSmoothing	7.0749
Prophet	6.4010

Table 7: Root Mean Square Error for state-level prediction by method.

Model Used	Root Mean Square Error
ARIMA	5.5871
ExpSmoothing	5.9941
Prophet	5.4918

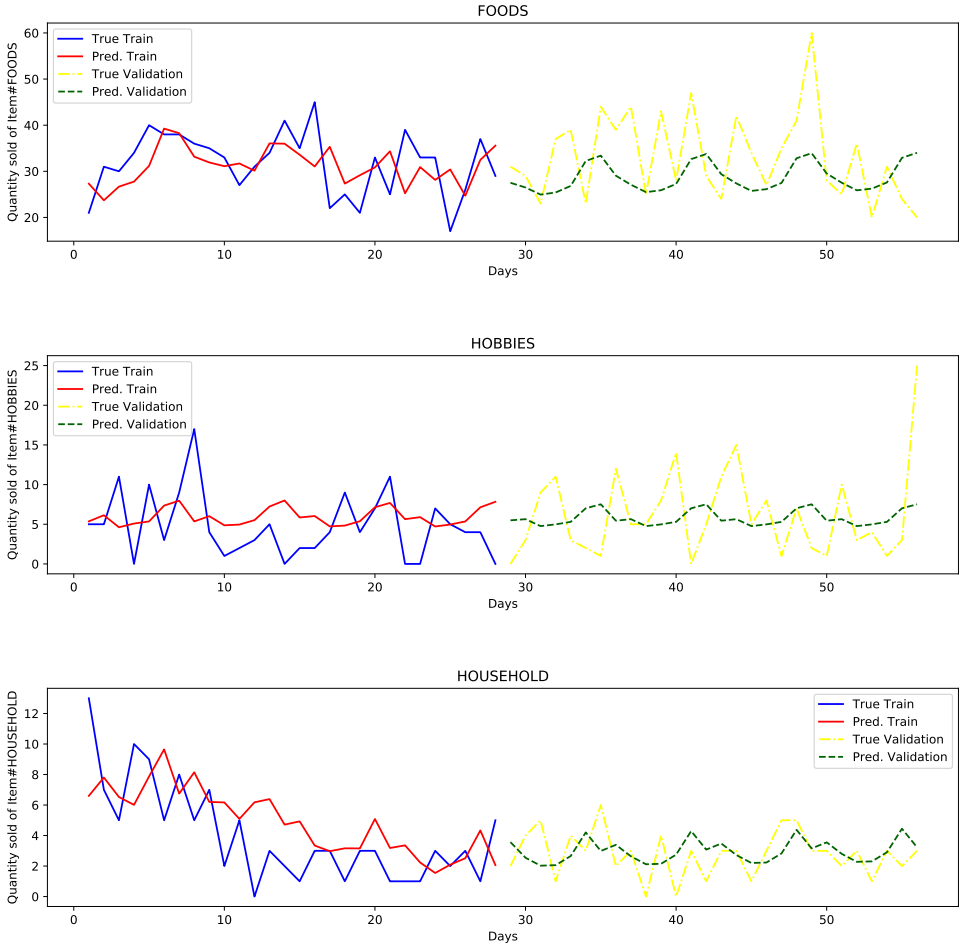


Figure 18: Item sales per category as predicted by ARIMA against the ground truth.

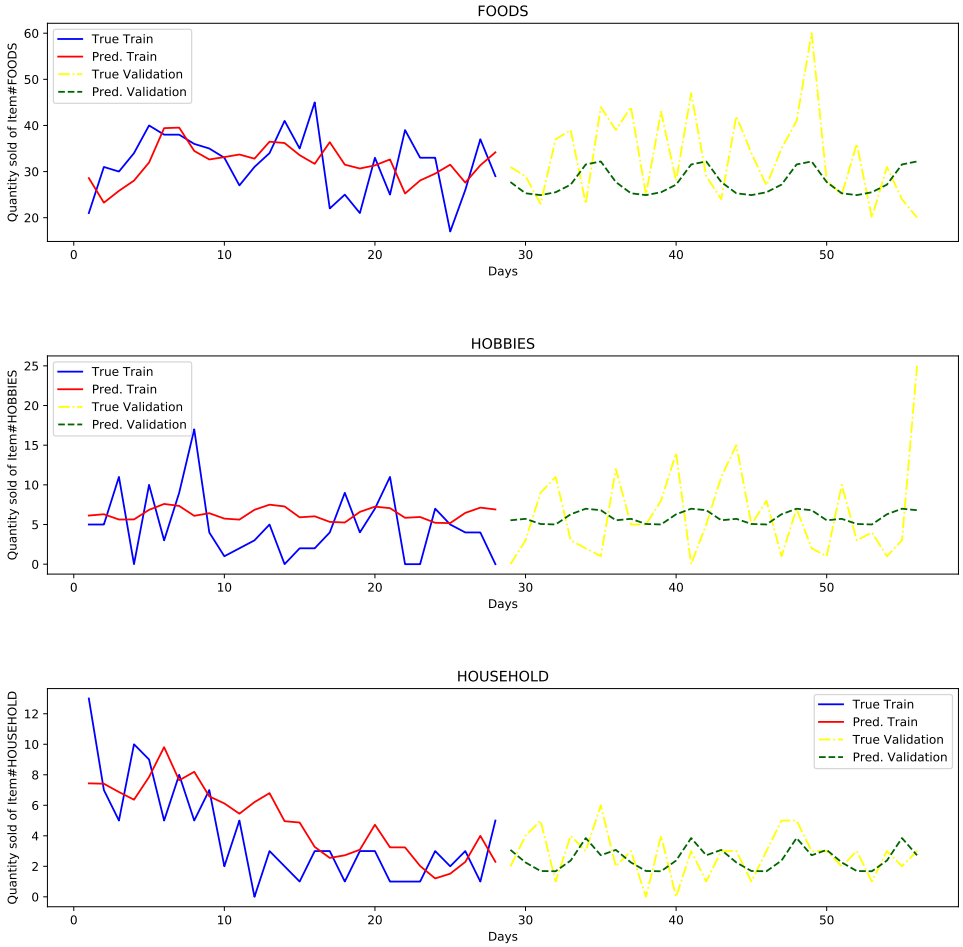


Figure 19: Item sales per category as predicted by ExpSmoothing against the ground truth.

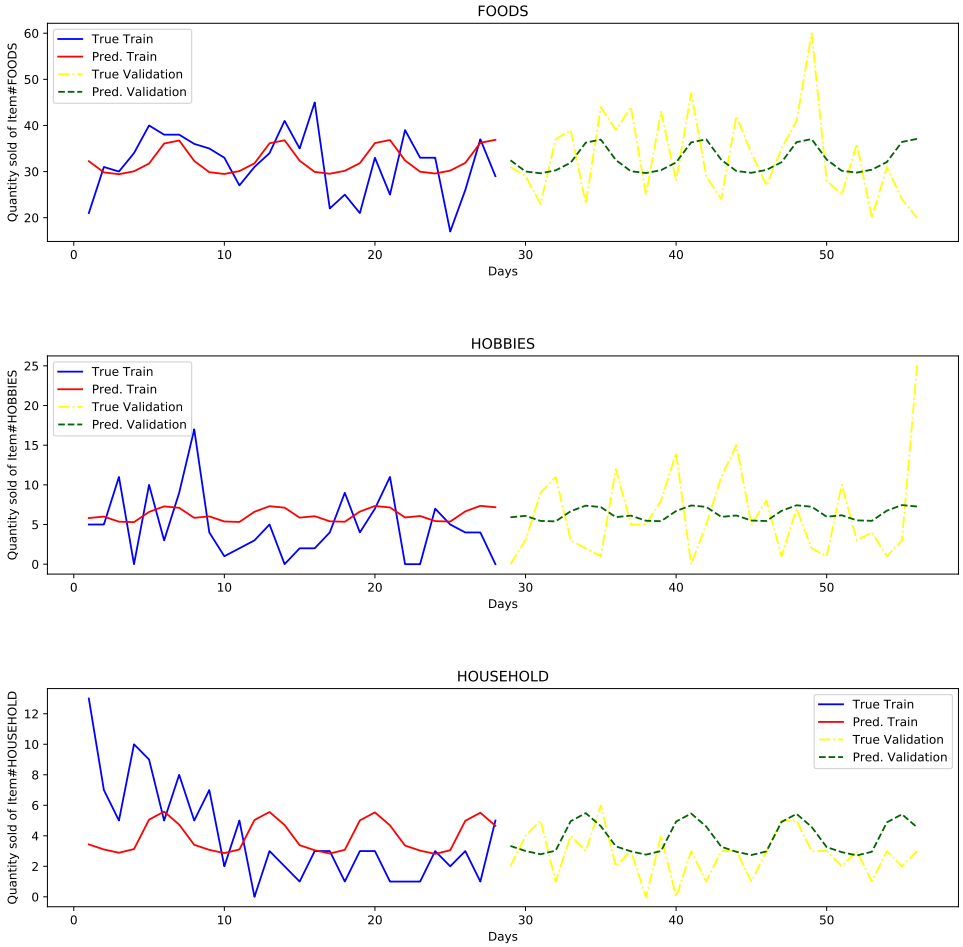


Figure 20: Item sales per category as predicted by Prophet against the ground truth.

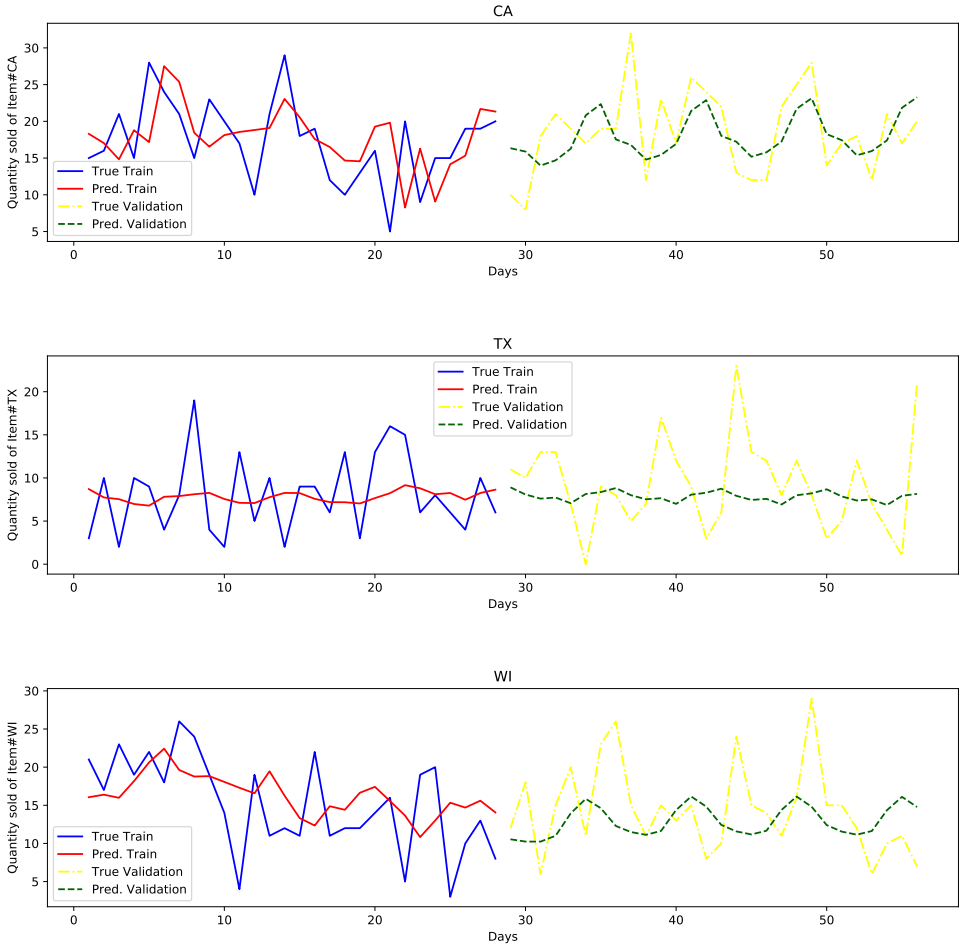


Figure 21: Item sales per state as predicted by ARIMA against the ground truth.

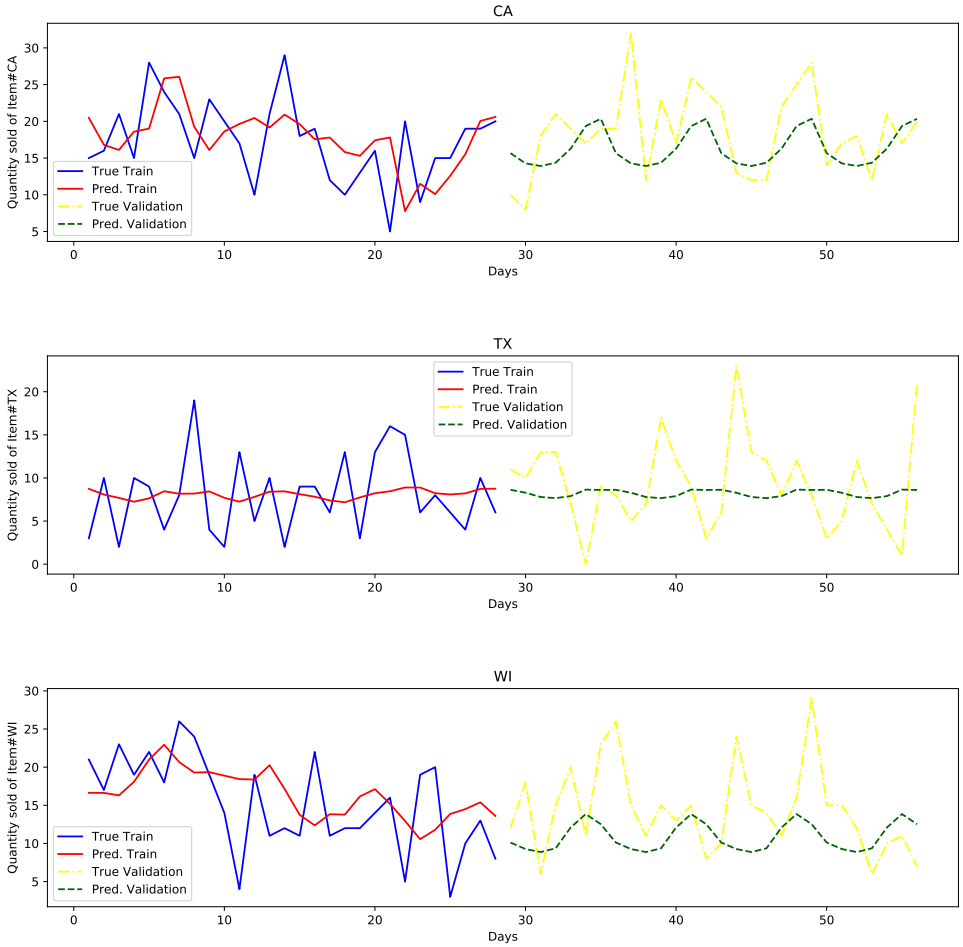


Figure 22: Item sales per state as predicted by ExpSmoothing against the ground truth.

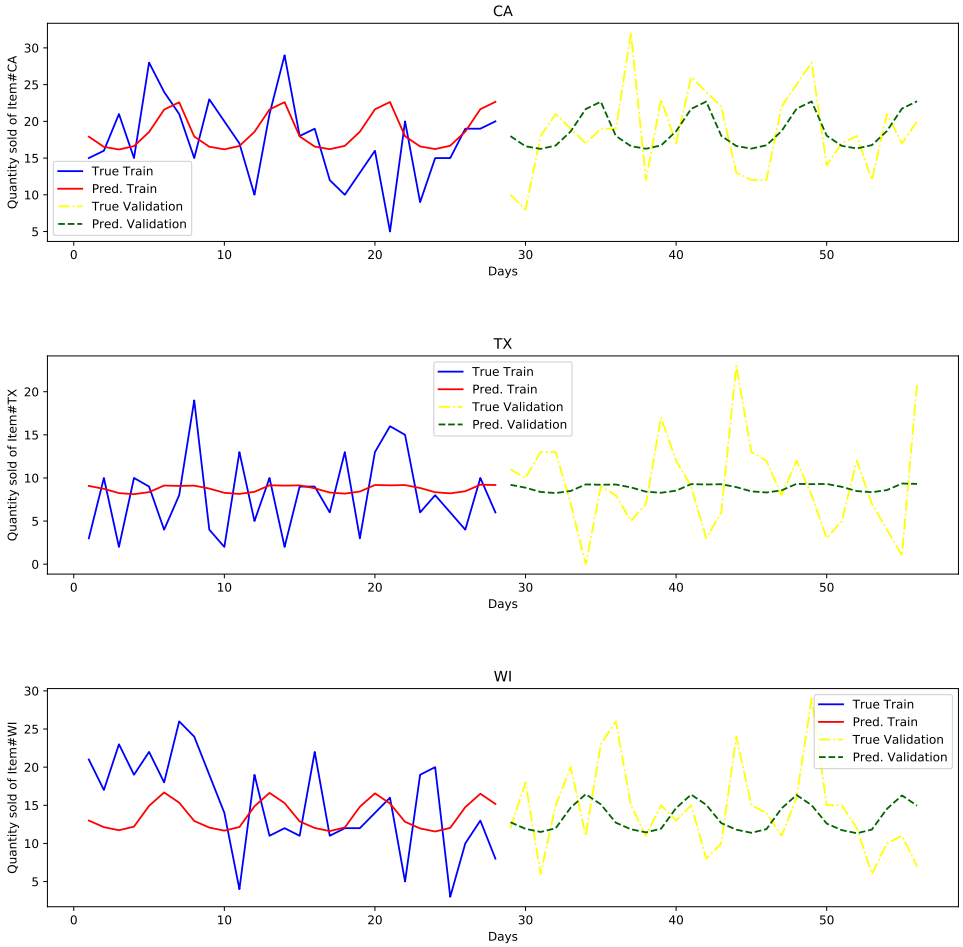


Figure 23: Item sales per state as predicted by Prophet against the ground truth.

4.4 Other approaches and future Work

We tried to implement prediction using recurrent neural networks, but it was too difficult and we could not get it working on time. We also did not use any ensembles, when eighty percent of the winning models from last year's competition (the M4 competition) were ensembles. In future work on time series forecasting (in this competition or otherwise), we will consider using ensembles.

References

Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. The M4 Competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, 36(1):54 – 74, 2020. ISSN 0169-2070. doi: https://doi.org/10.1016/j.ijforecast.2019.04.014. M4 Competition.