M5 Forecasting – Accuracy

Intermediate status report

Team: Economic Science Squad

Ismail Aigunov iaaygunov@edu.hse.ru
Xianghao Li xli@edu.hse.ru

Introduction

Xianghao Li Mainly responsible for Introduction and Data.

Ismail Aigunov Mainly responsible for Exploratory and Descriptive Analysis

Repository: https://github.com/XHaoLi/Project-MLDM

The M5 Competition

Background: The Makridakis Open Forecasting Center (MOFC) at the University of Nicosia conducts cutting-edge forecasting research and provides business forecast training. The MOFC is well known for its Makridakis Competitions, the first of which ran in the 1980s. M5 competition is the fifth iteration.

The **objective** of M5 competition is using hierarchical sales data from Walmart to forecast daily sales for the next 28 days.

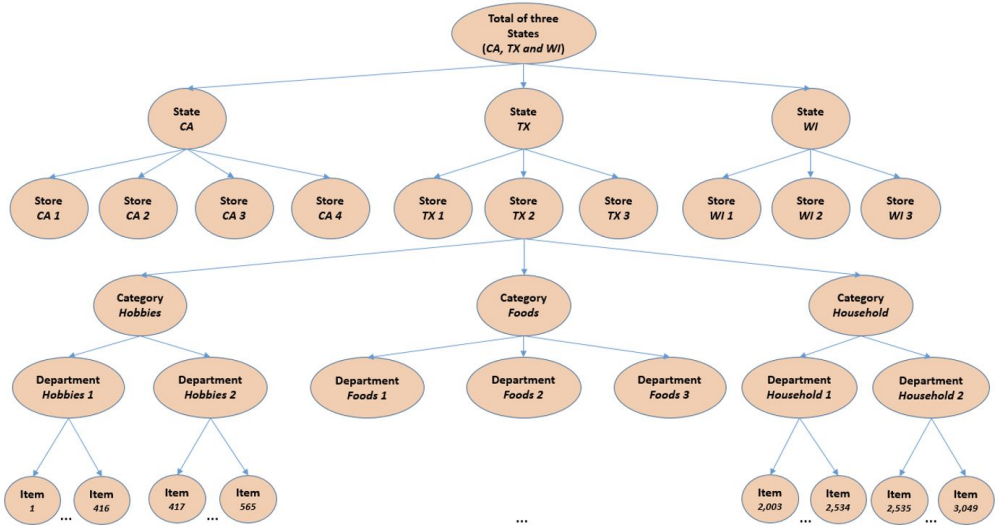
The **data**, covers stores in three US States (California, Texas, and Wisconsin) and includes item level, department, product categories, and store details.

M5 Forecasting - Accuracy: https://www.kaggle.com/competitions/m5-forecasting-accuracy/overview

Data

The M5 dataset made available by **Walmart**, involves the unit sales of various products sold in the USA.

Organized in the form of **grouped time series**, involves the unit sales of **3,049 products**, classified in **3 product categories** (Hobbies, Foods, and Household) and **7 product departments**. The products are sold across **ten stores**, located in **three States** (CA, TX, and WI).



Data

The historical data range from **2011-01-29** to **2016-06-19**. The products have a (maximum) selling history of 1,941 days / 5.4 years (**test data of h=28 days not included**).

The dataset consists three (3) files: ***"calendar.csv"***, ***“sell_prices.csv”*** and ***“sales_train.csv”***.

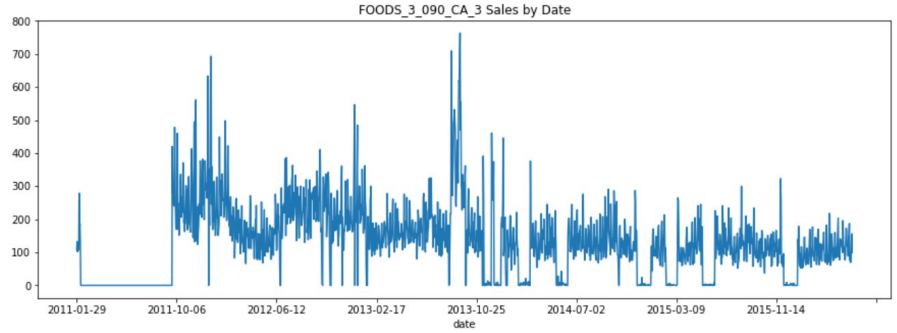
- ***"calendar.csv"*** - information about the dates the products are sold.
- ***“sell_prices.csv”*** - information about the price of the products sold per store and date.
- **“*sales_train.csv*”** - the historical daily unit sales data per product and store.

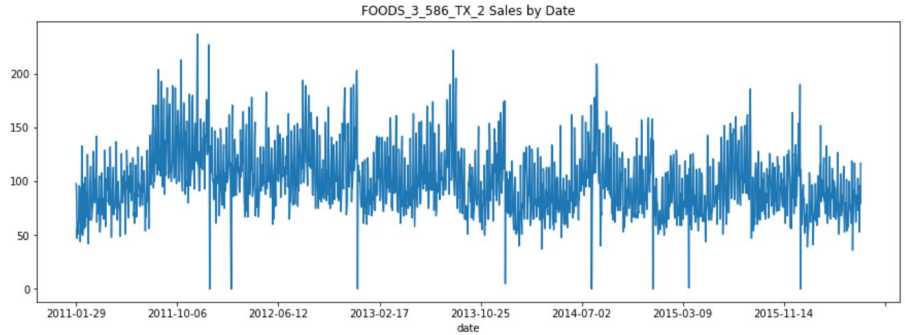
Details can be viewed in the file "Data.ipynb"

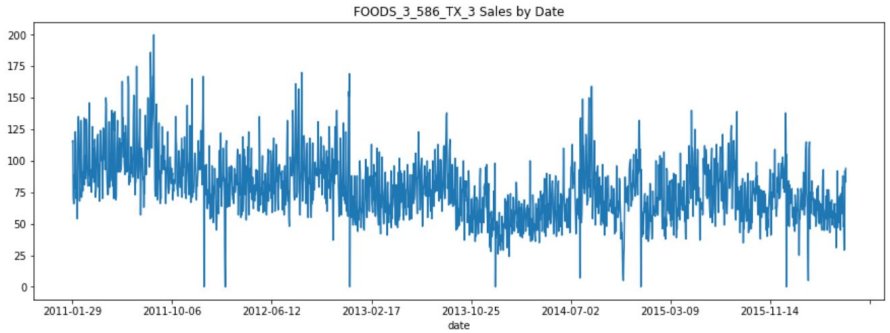
https://github.com/XHaoLi/Project-MLDM/blob/main/Data.ipynb

Exploratory and Descriptive Analysis

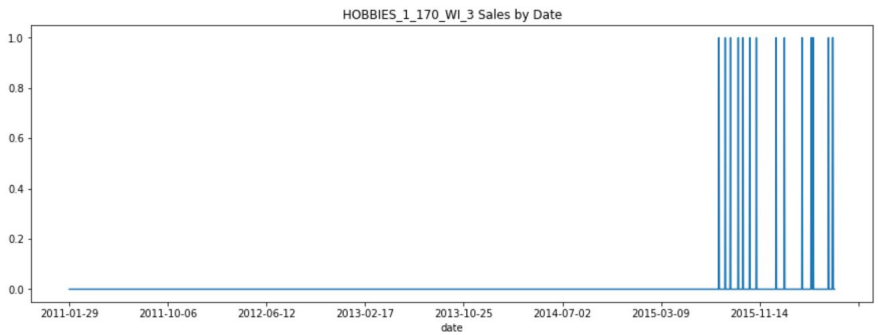
Top 3 units by Sales volume

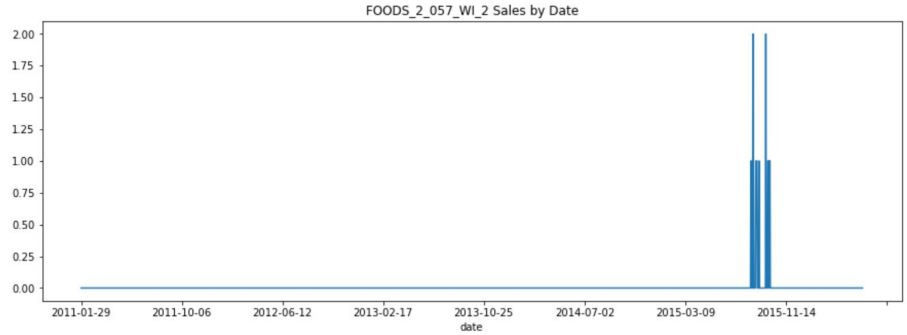


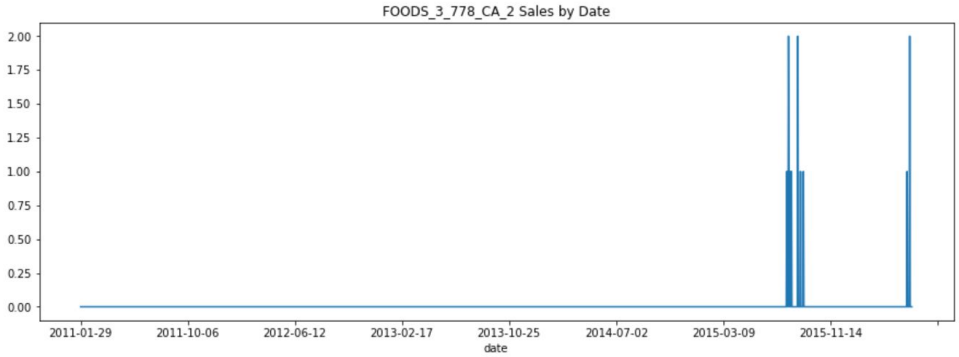




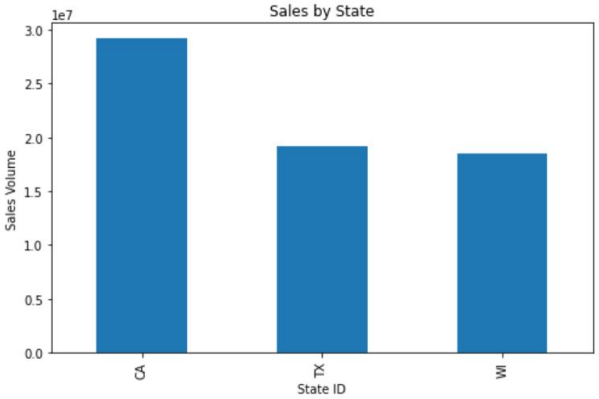
Last 3 Units by Sales Volume



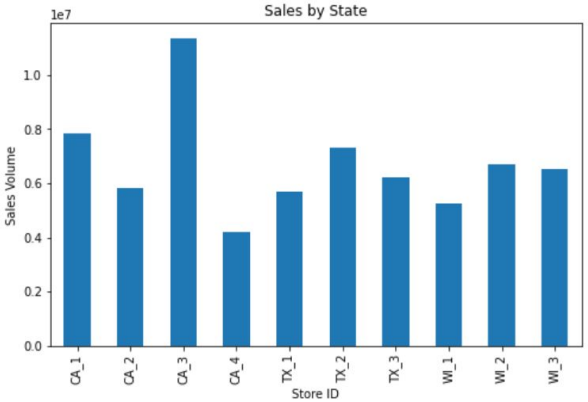


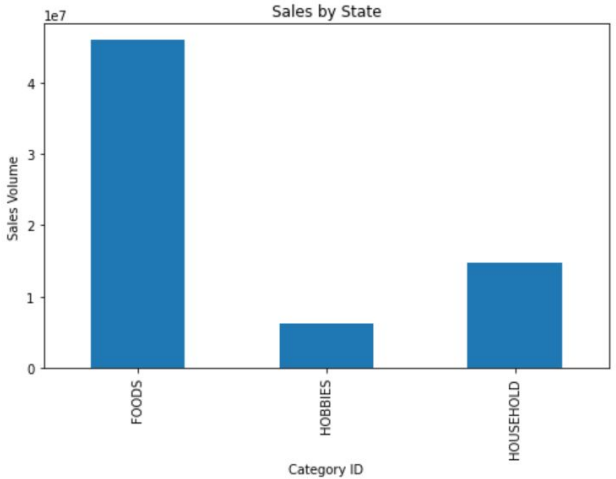


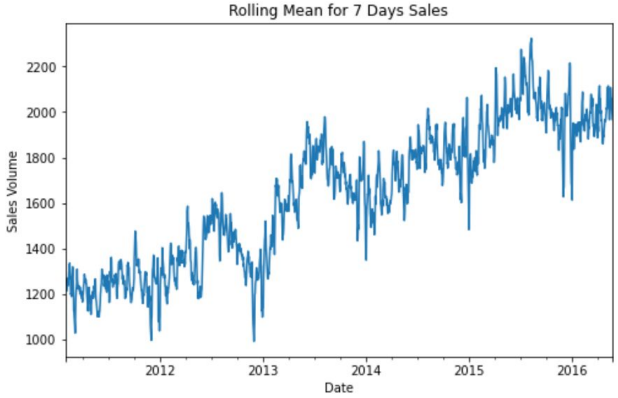
Sales Volume by State, Stores and Categories

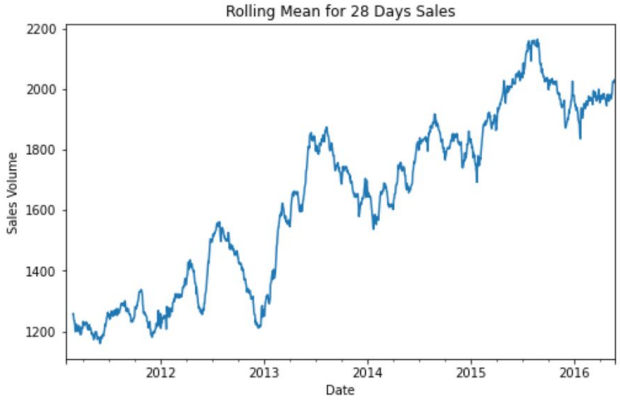


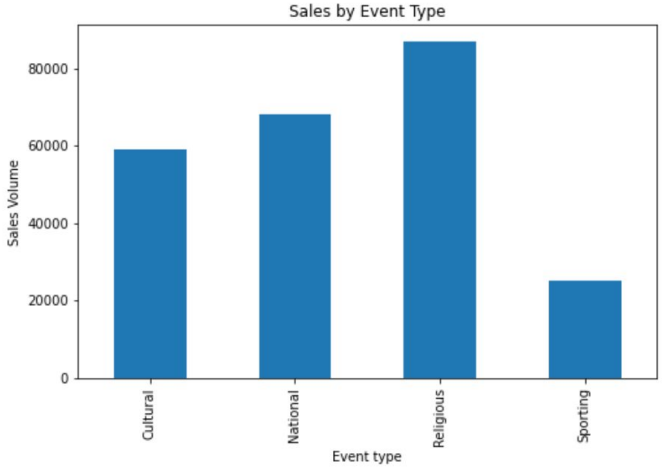
Total amount of sales account for 66 927 173 units



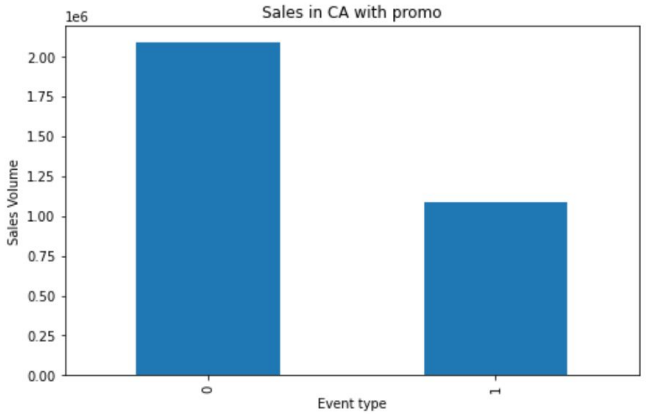


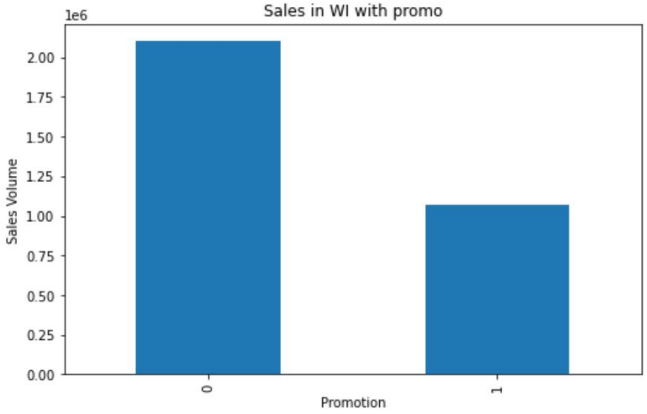

Rolling Mean for 7 Days Sales
Sales Volume
2200
2000
1800
1600
1400
1200
1000
2012
2013
2014
2015
2016
Date

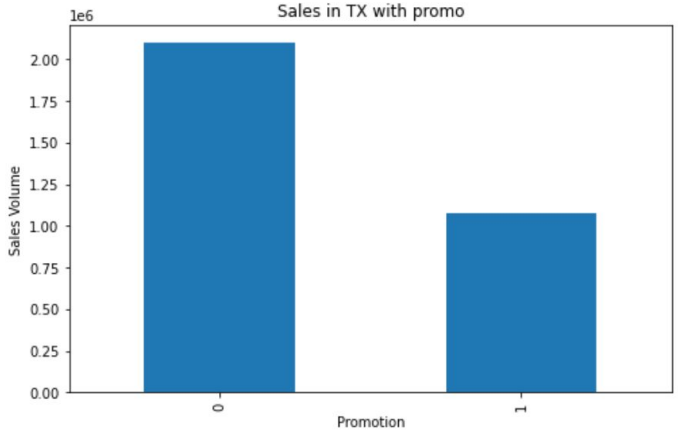

Rolling Mean for 28 Days Sales
Sales Volume
2200
2000
1800
1600
1400
1200
2012
2013
2014
2015
2016
Date


Sales by Event Type
Sales Volume
80000
60000
40000
20000
0
Cultural
National
Religious
Sporting
Event type

Promo Sales


Sales in CA with promo
1e6
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Sales Volume
0
1
Event type


Sales in WI with promo
1e6
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Sales Volume
0
1
Promotion


Sales in TX with promo
1e6
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Sales Volume
0
1
Promotion

Main insights about the data

- Series values substantially differ across units of stock range
- Food category is the most popular across all states and stores
- Rolling mean both for 7 and 28 days contains upward trend
- Events and holidays may affect sales, type of event matters
- Promo activities affect sales

Plans for the Final Project

Initialize two models based on different approaches:

- ARIMA
- XGBoost

Tune hyperparameters for each one

Introduce new features based on data insights