Walmart Sales Forecasting

Gourang Patel, Sanjan Vijayakumar, Hitashu Kanjani, Sagar Singh
Github: https://github.com/Gourang97/INFO-VIZ

October 6, 2021

Context

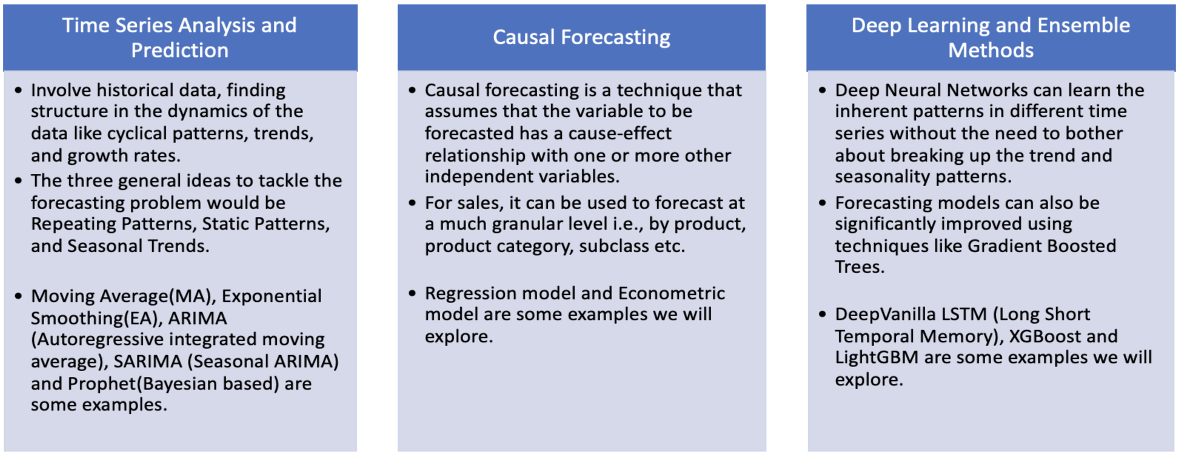
Ecommerce has been an ever-growing industry with retail revenues projected to grow to 4.9 trillion US dollars in 2021. With this tremendous growth, sales forecasts will help every business understand changing customer demands, manage inventories as per the demands thus reducing the financial risks, and create a pricing strategy that reflects demand. Companies will be able to take strategic steps on their short - term and long - term performances and can decide their decision metrics.

This project will present the right methodologies to analyze time-series sales data and predict 28 days ahead point forecasts for the company to take strategic decisions based on the predictions. Additionally, the project's goal includes making recommendations on inventory management based on the 28 days forecast. Sales forecasting is a very crucial research area with companies heavily investing and proposing advanced Methods like FaceBook's Neural Prophet, Amazon's DeepAR Model, Dilated convolutional neural networks. We plan to leverage the traditional time series forecasting methods as well as the modern forecasting methods and analyze the time-series sales data for Walmart.

Proposed Plan

Making predictions about the future is called extrapolation in the classical statistical handling of time series data. We plan to address this problem statement by using M5 forecasting competition 2020. This is Walmart's dataset with information about various products sold in the US into 3 different states California, Texas, and Wisconsin. The dataset involves the unit sales of 3049 products classified in 3 product categories(Food, Household, Hobbies) and 7 product departments across a timespan of 5 years starting from 2011 to 2016. The Data also includes explanatory variables such as sell prices, promotions, days of the week, and special events that typically affect unit sales and could improve forecasting accuracy.

For the forecasting and predictions on sales data, we are planning to use 3 fundamental approaches to attain best possible results:



Time Series Analysis and Prediction

- Involve historical data, finding structure in the dynamics of the data like cyclical patterns, trends, and growth rates.
- The three general ideas to tackle the forecasting problem would be Repeating Patterns, Static Patterns, and Seasonal Trends.
- Moving Average(MA), Exponential Smoothing(EA), ARIMA (Autoregressive integrated moving average), SARIMA (Seasonal ARIMA) and Prophet(Bayesian based) are some examples.

Causal Forecasting

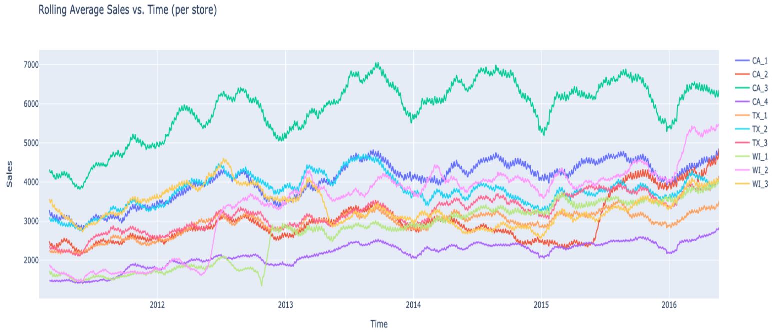
- Causal forecasting is a technique that assumes that the variable to be forecasted has a cause-effect relationship with one or more other independent variables.
- For sales, it can be used to forecast at a much granular level i.e., by product, product category, subclass etc.
- Regression model and Econometric model are some examples we will explore.

Deep Learning and Ensemble Methods

- Deep Neural Networks can learn the inherent patterns in different time series without the need to bother about breaking up the trend and seasonality patterns.
- Forecasting models can also be significantly improved using techniques like Gradient Boosted Trees.
- DeepVanilla LSTM (Long Short Temporal Memory), XGBoost and LightGBM are some examples we will explore.

Preliminary Results

Before starting with the forecasting, we started with some exploratory analysis to understand the data. Since it was spread across 3 tables, we first merged them into a single dataframe. We then marched towards analysing the sales across the three regions California, Texas and Wisconsin for the entire timeline of 5 years. As depicted in Figure 1 and 2, we observed that California has the highest sales, while Texas and Wisconsin have almost identical sales.



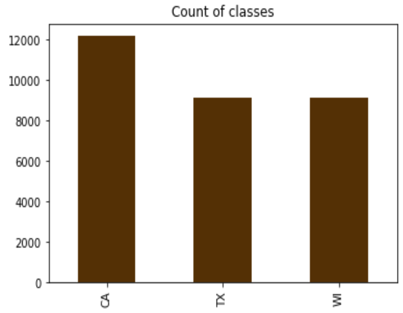
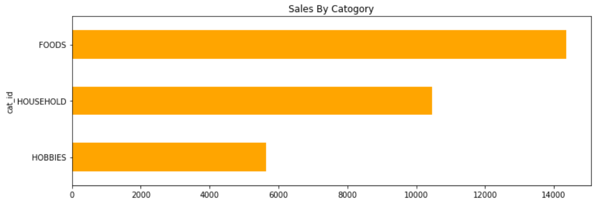


Fig 1: *Rolling Average Sales vs Time*

Fig 2: *Aggregates sales*

Next we wanted to identify the Sales at Category(Fig 3) and Department(Fig 4) level. We found that Food has the highest sales followed by Household and Hobbies. We further observed that FOODS_3 had the highest sales while HOBBIES_2 had the lowest sales.



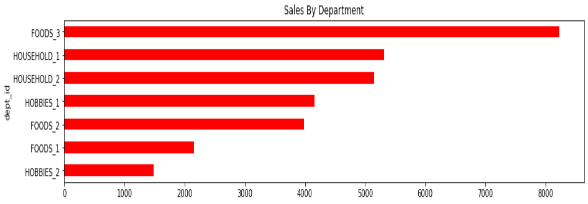


Fig 3: *Sales at Category Level*

Fig 4: *Sales at Department Level*

Project Timelines

We intend on completing this project for Phase 1 of the course. Having a well-established timeline helps us achieve this systematically.

Date	Milestone
Week of 10/12	Completing the data cleaning and pre-processing tasks, performing an exploratory data analysis
Week of 10/19	Identifying top performing products, feature engineering and finalizing the transformed data and key features for models
Week of 11/02 [Phase 1 submission]	A model analysis including a review of traditional time-series models vs modern approaches, to finalize the baseline models. We plan to submit the baseline models and results by Phase 1
Week of 11/16	Working on advanced forecasting methods like deep learning based approaches and causal forecasting methods.

Week of 11/30	Finalizing the models and results from advanced forecasting methods. Building a web-based application and hosting the data and model on cloud architecture
Week of 12/14 [Phase 2 Submission]	QC and QA on the results of the models. Work on presentation and final report submission.

Risks and Mitigation Strategies

When it comes to Sales Forecasting, accuracy plays an important role to inform decision making. If a prediction is too optimistic, businesses may overinvest in products and personnel, resulting in squandered funds. Companies may underinvest if the prediction is too low, resulting in a shortage of goods and a poor customer experience. However, high accuracy remains challenging for two reasons:

I. Traditional forecasts struggle to incorporate large amounts of previous data, resulting in the omission of significant past signals that get lost in the noise.
II. Traditional forecasts rarely include related but independent data (Exogenous Variables) that can provide important context (for example, price, holidays/events, stock-outs, marketing promotions, and so on). As a result of this, most forecasts fail to accurately predict the future without the full history and context.

To mitigate the above risks, we will be leveraging models like SARIMA, and various Deep Learning and Ensemble based models that can account for exogenous variables and handle both large data and multi step forecasting. For implementation, the potential risks would be:

Potential Risks	Mitigation Strategy
Time series model works best for stationary data, however real world time series are frequently non-stationary	Perform stationarity tests - Augmented Dickey-Fuller Test
Prediction window is an essential component for model reliability. Important questions to consider are - How long into the future should we forecast?, How often do we need to generate forecasts?	Following the guidelines of M5 competition, our forecast will be for 28 days, which may be later refined based on results
Model Evaluation is not similar to classification or regression tasks. We must ensure to avoid any information propagating backwards, that can influence training and show over optimistic results.	Train Test split should be on time period and use backtesting technique for validation
Cross Validation for time series is different from classification where the timestamp of the prediction is not important. The k-fold CV which randomizes data points can induce serious Look-Ahead bias.	Blocked K-fold CV (Snijders, 1988), Modified K-fold CV (McQuarre & Tsai, 1998) and HV-Blocked K-fold CV (Racine, 2000) will be preferred.

References

1. M5 Competition: https://mofc.unic.ac.cy/m5-competition/
2. Harvard review: http://hbr.org/1971/07/how-to-choose-the-right-forecasting-technique
3. Forecasting: http://journals.sagepub.com/doi/abs/10.1177/00279501082030011201