DS 5500
CAPSTONE: APPLICATIONS IN DATA SCIENCE
WALMART SALES FORECASTING

GITHUB - HTTPS://GITHUB.COM/GOURANG97/INFO-VIZ

Gourang Patel
Masters in Data Science
patel.gou@northeastern.edu

Hitashu Kanjani
Masters in Data Science
kanjani.h@northeastern.edu

Sanjan Vijayakumar
Masters in Data Science
vijayakumar.sa@northeastern.edu

Sagar Singh
Masters in Data Science
singh.sag@northeastern.edu

ABSTRACT

E-commerce is a huge part of the economy and is vital to businesses that sell their products or services online. Sales forecasting is the process of estimating future revenue by predicting the amount of product or services a sales unit will sell in the next iteration. This is vital to e-commerce giants and may help increase large chunks of revenue. In this project, we use hierarchical sales data from Walmart, the world's largest company by revenue, to forecast daily sales for the next 28 days.

INTRODUCTION

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 businesses 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.

DATASET

We have collected Walmart Sales Forecasring Dataset from M5 Forecasting Data Competition. The dataset contains 5 year historical sales from 2011- 2016 for various products and stores. Data is hierarchically organized: stores are divided into 3 states, and products are grouped by categories and sub-categories

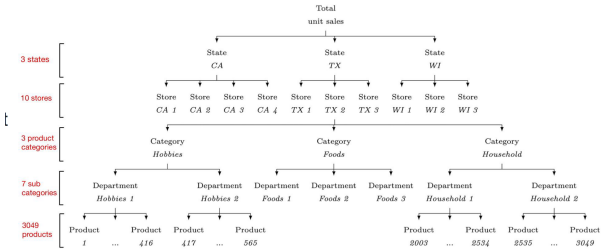

FIGURE 1. DATA OVERVIEW

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 data frame. 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 2, we observed that California has the highest sales, while Texas and Wisconsin have almost

identical sales.

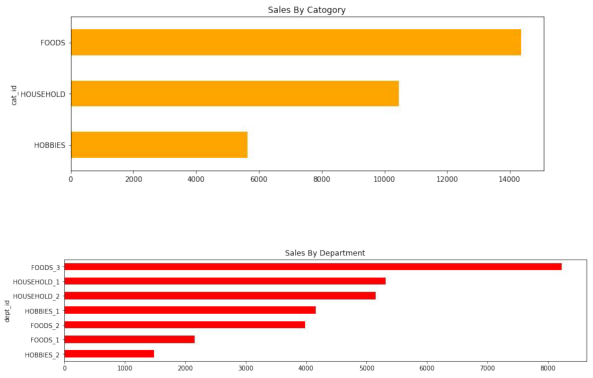

FIGURE 2. Sales Results for Walmart Data

FIGURE 3. Store Distribution results for Walmart Data

In *Fig*3 we wanted to identify the Sales at Category and Department 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.

METHODS

Feature Engineering

As the dataset we are using is too large to be processed using pandas dataframes and the computational architecture we had, we brainstormed on the approaches for processing our dataset. The possible approaches were to use Pyspark dataframes or cloud architecture. Using pyspark dataframes was able to solve our issue, but SparkML doesn't provide good support with the Forecasting Models. This inturn would require us to convert our spark dataframes to pandas dataframes, which again led us to running out of memory issues. Using cloud architecture could be a possible solution as well, but it was computationally very expensive for us. Therefore, we used downcasting as our approach to transform our data and deal with the memory issue.

Downcasting Downcasting is a type refinement that can be defined as the act of casting a reference of a base class to one of its derived classes. Our approach was downcasting our int64 and float64 objects to int8 and int16; float16 and float32 respectively which in turn helped us resolve out of memory issues while handling the data and performing various transformations on the table. The performed downcasting helped us optimize and save around 70

In time-series datasets another feature engineering approach which comes round the way is checking for stationarity. As if the data doesn't meet the stationarity checks we can't leverage the data to perform forecasting methods. To handle this issue we performed various tests to analyze the stationarity of the data in hand.

Stationarity Checks Stationary time series data do not depend on time. Time series are stationary if they don't have trend or seasonal effects. A model cannot forecast on non stationary time series so one of the most commonly used statistical tests to determine stationarity is the ADF (augmented Dickey Fuller) test. ADF is done to check the number of differencing used on the ARIMA model for forecasting. The ADF test is fundamentally a statistically significance test which means that there is a hypothesis test with null and alternate hypotheses.

The unit root is a characteristic of a time series that makes it non stationary. A Dickey-Fuller test is a unit root test that tests the null hypothesis that $\alpha = 1$ in the following model equation. α (alpha) is the coefficient of the first lag on Y. Below is the Figure 1 stating p value and test statistic before differencing for FOODS category

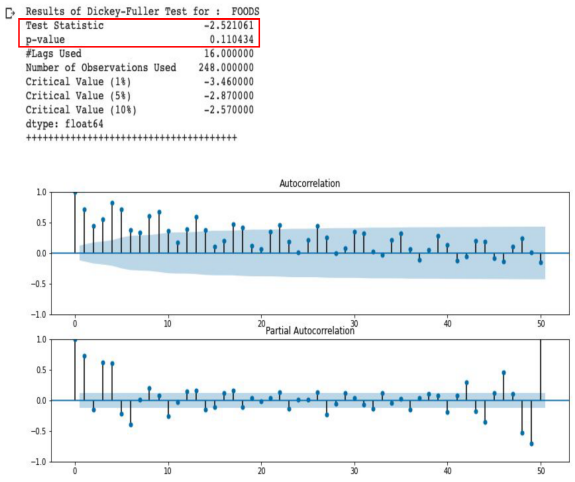

FIGURE 4. ADF TEST BEFORE DIFFERENCING

We observe that the p value is greater than 0.05 so there is no reason to reject the null hypothesis . Our null hypothesis is that the series is non stationary. We now tried differencing the time series by the order of 1 and again performed the ADF test.Below is the Figure 2 stating p value and test statistic after differencing for FOODS category

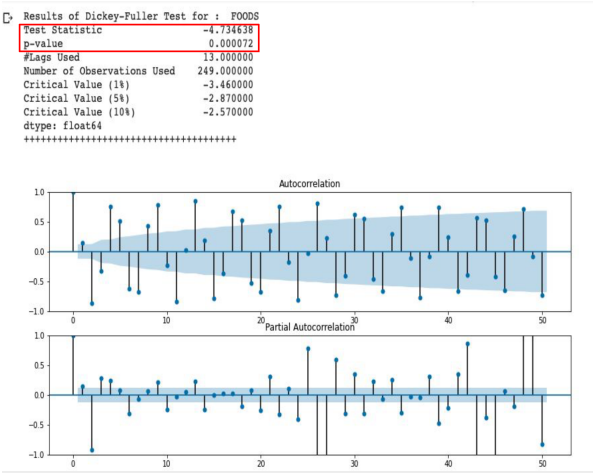

FIGURE 5. ADF TEST AFTER DIFFERENCING

After the first difference we notice that the statistical test and p value are very low than the critical value and significant value (0.05). We can now reject the null hypothesis and infer that the time series is stationary. After performing the stationarity test, another thing we need to check is the trend charts from the tests. As when we analyze them if we observe the variance within the seasonal decomposed chart. The trend chart helps us evaluate whether the data is of the additive or multiplicative form. If the data is of the form additive which was in the case of FOODS category we would simply differentiate it and use it as the model input. But in the case of HOBBIES and HOUSEHOLD, as observed in the figure below, it seems to be multiplicative. Therefore we planned to transform the HOBBIES and HOUSEHOLD categories by performing the log transformation and hence made it linear or additive w.r.t time, so as to use them for further modeling.

Introducing Exogenous Variables Exogeneous variables can be defined as the variables whose cause is external to the model and whose role is to explain other variables or outcomes in the model. In forecasting models, exogenous variables help to improve the explainability of a forecasting model and also enhance the results of the model. We incorporated the exogenous variables in our most recent update of the model, and it improved our model performance to a great extent.

We had calendar data which contains all the information about various events within a calendar year, we merged the calendar dataset with our main dataframe, and used information like events within a particular year and holidays. This information helped the model to explain various sales spike on a particular day or various downtimes as well. We incorporated columns like events, weekday, weekends and paydays.

MODELING

We planned on using couple of standard Forecasting Models for our Phase 1 namely Auto ARIMA, ARIMA and SARIMAX. Below is the explained apporach and results from the model on our Walmart Sales Forecasting Dataset.

Auto ARIMA Usually, in the basic ARIMA model, we need to provide the p,d, and q values which are essential. We use statistical techniques to generate these values by performing the difference to eliminate the non-stationarity and plotting ACF and PACF graphs. In Auto ARIMA, the model itself will generate the optimal p, d, and q values which would be suitable for the data set to provide better forecasting. The determination parameter is a low AIC and BIC score.

ARIMA MODEL The ARIMA (Auto Regressive Moving Average) model is a very common time series-forecasting model. The AR represents the autoregressive part which is denoted by p . The I represents the Integrated part which is the number of non seasonal differences needed for stationarity and is denoted by d. The MA is the moving average part which is the number of lagged forecast errors in the prediction equation and donated by q. The figure attached below showcase the best result from the FOODS category. The other two categories will be incorporated in the appendix.

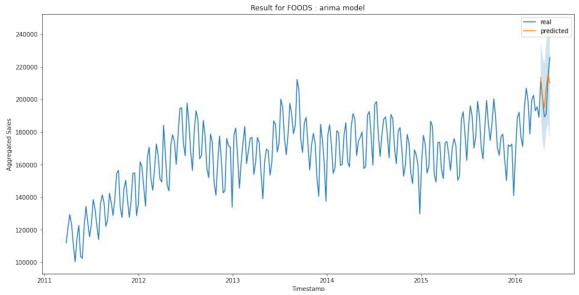

FIGURE 6. FORECASTING USING ARIMA WITHOUT GRID SEARCH

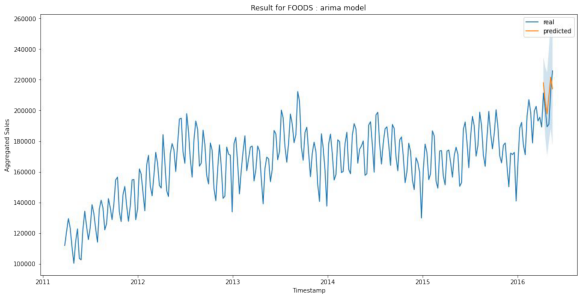

FIGURE 7. FORECASTING USING ARIMA WITH GRID SEARCH

As observed on the dataset and the above figure the ARIMA model did not perform well as it does not take seasonality into account. The RMSE score for ARIMA model was 6536 in FOODS category, 667 for HOBBIES category and 1651 for HOUSEHOLD category. After performing grid search on the ARIMA model , the RMSE score improved by 36.5%(for the FOODS category). For the HOUSEHOLD category, there was a poor fit despite the grid search and there was little to no trend capture for the HOBBIES category. The walk forward validation was used for back testing the variables.

SARIMAX MODEL A seasonal ARIMA model comes into play when the series has seasonal trends. SARIMAX is different from SARIMA as we are also incorporating the exogenous variables as discussed above. While working with the dataset we observed seasonal trends in all the three categories as the sale numbers increased in the last quarter for all the years and saw a sudden drop following that. This helped us conclude the evidence of seasonality within our dataset. The SARIMAX model allows the inclusion of exogenous variables with the seasonality parameters in the ARIMA model.

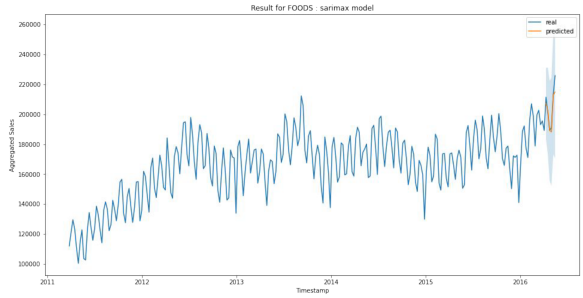

FIGURE 8. FORECASTING USING SARIMAX FOR FOODS CATEGORY

In *Fig*5 we observe, The RMSE further improved by 20.1%

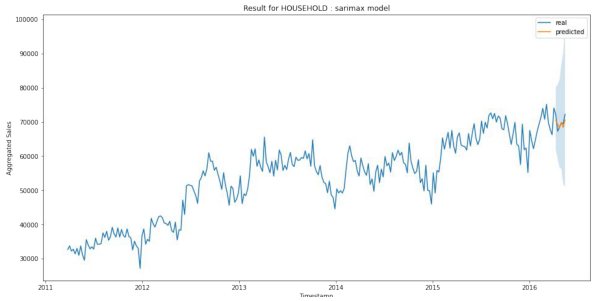

FIGURE 9. FORECASTING USING SARIMAX FOR HOUSEHOLD CATEGORY

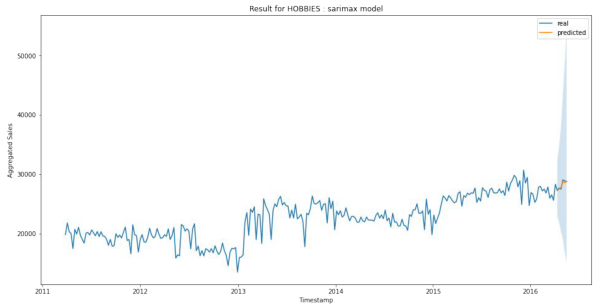

FIGURE 10. FORECASTING USING SARIMAX FOR HOBBIES CATEGORY

considering the effects of exogenous variables and seasonality for the FOODS category. In *Fig*6, we have plotted the results of HOUSEHOLD category on SARIMAX Model, we observed there was an improvement of 38% over the best ARIMA model.

In *Fig*7, we have plotted the results of HOBBIES category on SARIMAX Model, we observed that, there was a 53% improvement over the best ARIMA model and the overall trend was captured more accurately.

KEY LEARNINGS

A good strategy needs to be adopted for data preprocessing for large-scale datasets. Platforms like Apache Spark can be used for data parallelism, or cloud based architectures can be adopted to gain higher computation power. In our project, a simpler solution was achieved using downcasting to reduce memory footprints.Performing tests to check for and making necessary transformations to ensure data stationarity is key in time-series forecasting.Techniques like Gridsearch and Walk-forward validation help hypertune baseline traditional time-series models like SARIMAX. Another interesting aspect was inclusion of exogenous variables which enhanced our results to a great extend.

STATEMENT OF CONTRIBUTIONS

Listed below are the contributions of each team member towards Phase 1 of the project.

Gourang Patel: Data collection, Initial data cleaning and preprocessing, Baseline modeling

Hitashu Kanjani: EDA, Statistical tests (ADF test), Baseline modeling

Sanjan Vijayakumar: EDA, Tests to identify trend and seasonality, Data aggregation and model initiation steps

Sagar Singh: Data aggregation and model initiation steps, Grid search implementation, Addition of downcasting and walk-forward validation

REFERENCES

[1] Eklund, J., and Kapetanios, G., 2008. "A Review of Fore-casting Techniques for Large Data Sets".National Institute Economic Review,203, January, pp. 109–115.

[2] MOFC,2020.TheM5Competition.Onthewww,June.URL https://mofc.unic.ac.cy/m5-competition/.

[3] Chambers, J. C., Mullick, S. K., and Smith, D. D.,1971.How to Choose the Right Forecasting Technique. Tech. rep., Harvard Business Review, July. URL http://hbr.org/1971/07/how-to-choose-the-right-forecasting-technique.

Appendix

Code Repository: https://github.com/Gourang97/INFO-VIZ