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

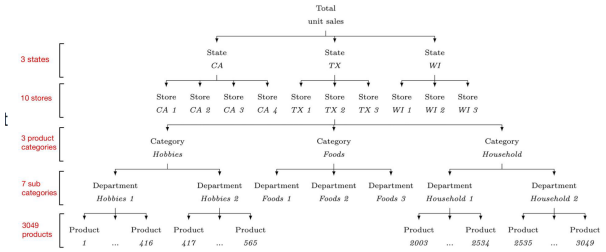

FIGURE 1. Data overview

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.

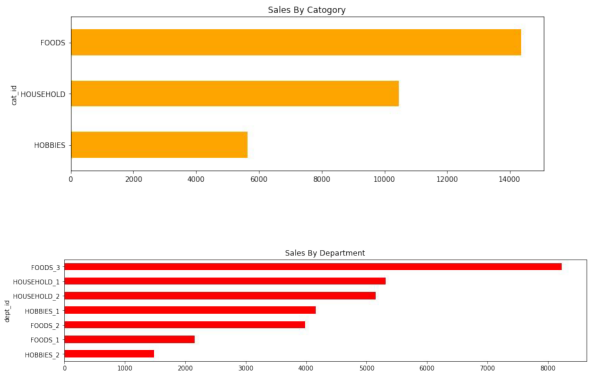

FIGURE 2. Sales results for Walmart data

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 department FOODS_3 had the highest sales while department 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 could have solved our issue, however SparkML doesn't provide good support with the Forecasting Models. Using this approach would also require us to convert our spark dataframes to pandas dataframes, which again caused 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% of our memory which was used while using the data frame without performing the transformation

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 with null hypothesis as that a unit root is present within the time series sample and is non stationary and alternate hypothesis would be its counterpart.

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 $\alpha = 1$ as per the model equation. α (alpha) is the coefficient of the first lag on Y. Below is the Figure 4 stating p value and test statistic before differencing for FOODS category

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 5 stating p value and test statistic after differencing for FOODS category

After the first difference we notice that the test statistic 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 stationary test, we do analysis on seasonal decomposition charts to understand the trends and stationary of the data in hand. 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, it's observed to be multiplicative during the analysis. Therefore we planned to transform the HOBBIES and HOUSEHOLD categories by performing the log transformation

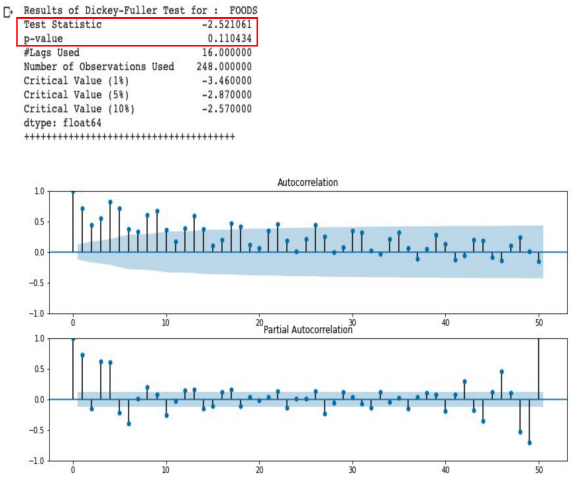

FIGURE 4. ADF test before differencing on FOODS Category

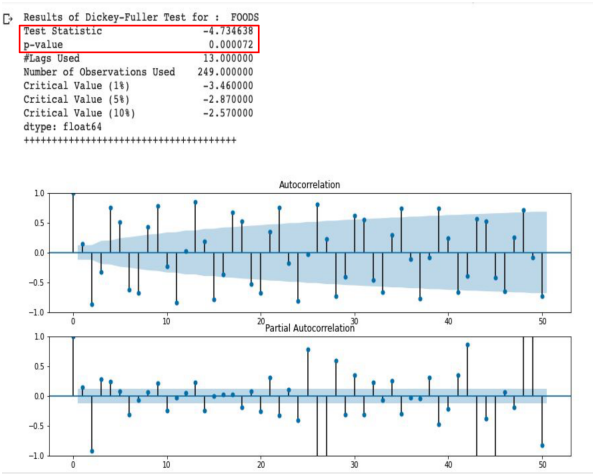

FIGURE 5. ADF test after differencing on FOODS Category

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 created columns like events, weekday, weekends and paydays and incorporated them with our main dataframe.

MODELING

PHASE 1

We built models on a few standard Forecasting Models like Auto ARIMA, ARIMA and SARIMAX for Phase 1. Below is the explained apporach and results from the model on our Walmart Sales Forecasting Dataset.

Auto ARIMA In basic ARIMA models we need to provide p,d, and q values which are essential to build the model. 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 generates optimal p, d, and q values itself 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 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 shows the best result from the FOODS category. Results from other categories can be found in the appendix.

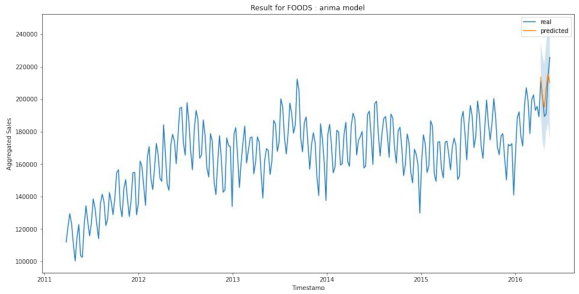

FIGURE 6. Forecasting using arima for foods category without grid search

As observed on the dataset and the above figure the ARIMA model did not perform well as it does not take seasonality into

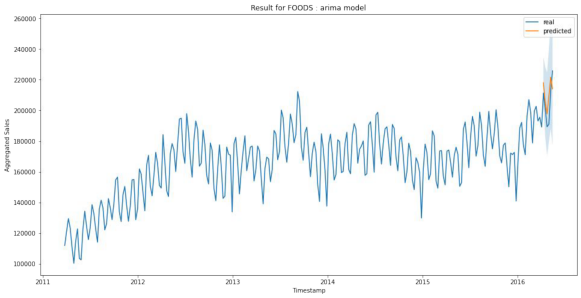

FIGURE 7. Forecasting using arima for foods category with grid search

account. The RMSE score for ARIMA model was 9802.53 in FOODS category. After performing grid search on the ARIMA model , the RMSE score improved by 36.5%. 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.

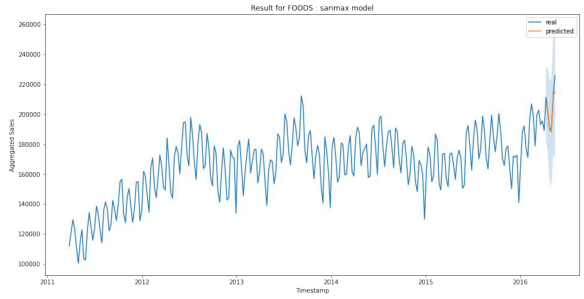

FIGURE 8. Forecasting using sarimax for foods category

In *Fig*8 we observe, The RMSE further improved by 20.1% considering the effects of exogenous variables and seasonality for the FOODS category. In *Fig*9, we have plotted the results of HOUSEHOLD category on SARIMAX Model(RMSE: 1314.61), we observed there was an improvement of 38% over the best ARIMA model.

In *Fig*10, we have plotted the results of HOBBIES category on SARIMAX Model(RMSE: 215.28), we observed that, there

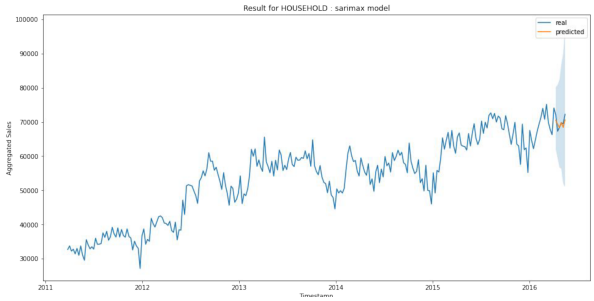

FIGURE 9. Forecasting using sarimax for household category

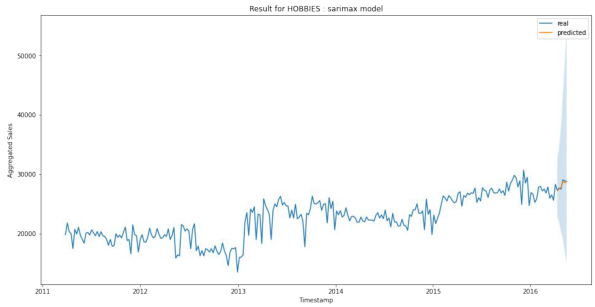

FIGURE 10. Forecasting using sarimax for hobbies category

was a 53% improvement over the best ARIMA model and the overall trend was captured more accurately.

PHASE 2

To build upon the benchmark set by the baseline models, we use advanced forecasting techniques in Phase 2.

Random Forest Random forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. Random Forest can be used for time series forecasting by transforming the dataset into a supervised learning problem. The model can then be evaluated using walk-forward validation, to prevent optimistically biased results. An added advantage of using modern forecasting techniques was being able to fit the model on daily data, as opposed to a weekly sampling we employed in Phase 1.

While Random Forest was able to capture the trend in all sections, the results were constantly under predicted. RMSE for the Random Forest model was 6514 in FOODS category. This is an improvement over the Phase 1 models as we now sampled data on a daily basis (weekly sampling in baseline models). *Fig*11 shows the forecasting results using Random Forest for the FOODS category. Refer the appendix for the remaining categories.

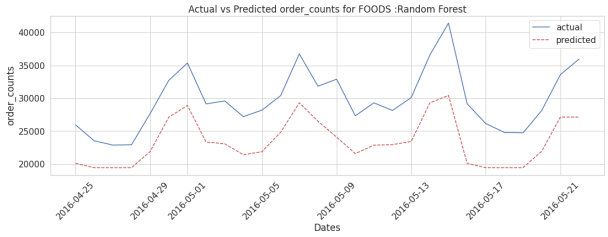

FIGURE 11. Forecasting using random forest for foods category

LightGBM LightGBM (Light Gradient Boosting Machine) is a gradient boosting framework that uses tree based learning algorithms. LightGBM is an advanced model with faster training speed, higher efficiency, and lower memory usage. It also supports parallel and distributed computing. We were able to model daily sampling data with LightGBM and also include exogenous variables like snap, event v/s no events, paydays and holidays.

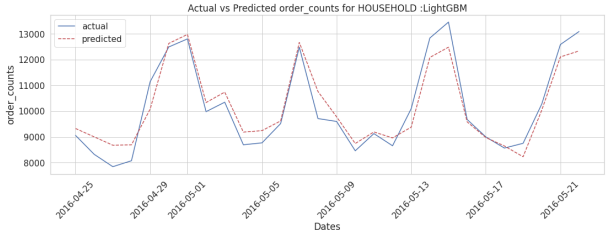

FIGURE 12. Forecasting using lightgbm for household category

The LightGBM models were able to capture daily trends better than the Random Forest models. The HOUSEHOLD category had an RMSE of 536.88 as shown in *Fig*12 which is an 80% improvement over Random Forest and a 59.1% improvement over the SARIMAX model. Refer the appendix for the remaining categories.

LSTM Long Short-Term Memory (LSTM) networks are a type of recurrent neural network capable of learning order dependence in sequence prediction problems. LSTM networks have the promise of learning long sequences of observations which make them suitable for time-series forecasting. We were able to model daily sampling data with custom stacked LSTM model and also included variables like snap, event v/s no events, paydays and holidays.

The LSTM models were significantly able to capture daily trends better than the LightGBM and Random Forest models. The FOODS category which performed the best as compared to its performance with all the other traditional models, had an

FIGURE 13. Forecasting using LSTM for foods category

RMSE of 4373.36 as shown in *Fig*13 which is an 32.86% improvement over Random Forest and a 7.06% improvement over the LightGBM model. Refer the appendix for the results from remaining categories.

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.

To add onto the key learnings from Phase 1, fitting ensemble models on daily sampled data improved forecasting to a great extent. LightGBM showed better performance than Random Forest, which continuously under-predicted. The main reason for Random-Forest to under-predicted were the hyper-parameters of the model were not appropriately tuned and also no transformations were applied to make the data stationary. Also there were significant number of zero sales in the day-wise data for all the categories which caused such under-prediction. We used appropriate methods to deal with such problems in LightGBM with a more tuned-set of parameters, and hence the results were improved The best model performance was achieved by the LSTM model which captured daily trends effectively. As LSTM is a complex model which is very well known to capture trends within the big data sets with complex parameters. With appropriate tuning of the hyper-parameters and also scaling the data, as smallers values works well with LSTM model, we are able to achieve better results as compared to any fore-mentioned models.

STATEMENT OF CONTRIBUTIONS

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

Gourang Patel: Data collection, Initial data cleaning and preprocessing, Baseline modeling, LSTM modeling, Streamlit model deployment

Hitashu Kanjani: EDA, Statistical tests (ADF test), Baseline modeling, Random Forest modeling, LightGBM Modeling

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

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

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

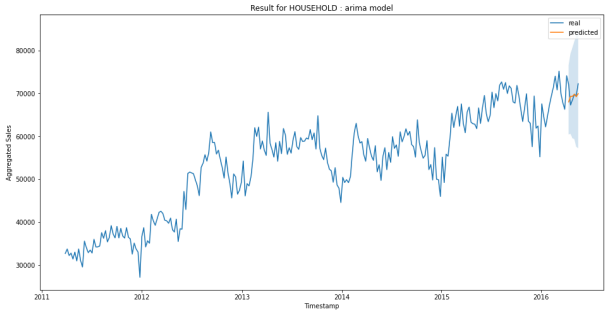

FIGURE 14. Forecasting using arima for household category

ARIMA The RMSE score for ARIMA model was 2532.63 for the HOUSEHOLD category.

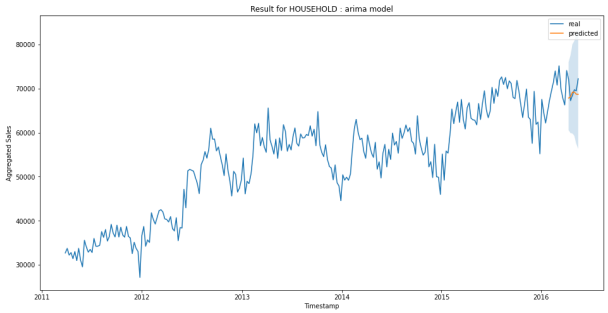

FIGURE 15. Forecasting using arima for household category with grid search

The RMSE was 2130.08 after gridsearch for the HOUSEHOLD category which is a minor 15.9% improvement over baseline ARIMA.

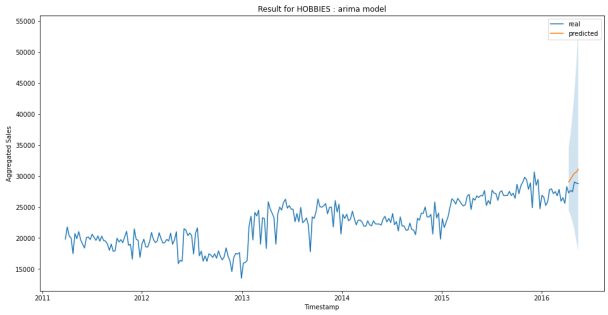

FIGURE 16. Forecasting using arima for hobbies category

The RMSE score for ARIMA model was 712.39 for HOBBIES category which became 466.83 after gridsearch.

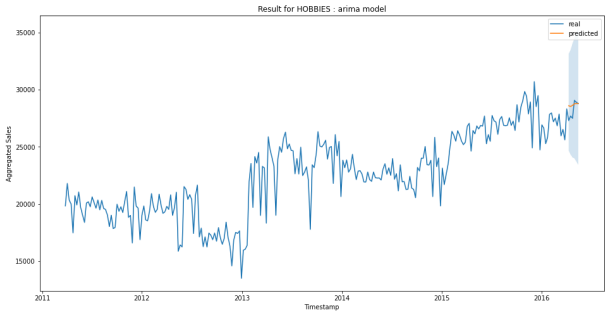

FIGURE 17. Forecasting using arima for hobbies category with grid search

Log Transformation was applied to account for variance.

There was little to no trend capture for the HOBBIES category despite grid search.

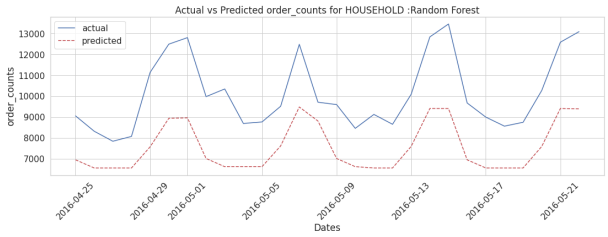

FIGURE 18. Forecasting using random forest for household category

Random Forest RMSE for the HOUSEHOLD category for the Random Forest model was 2704.61.

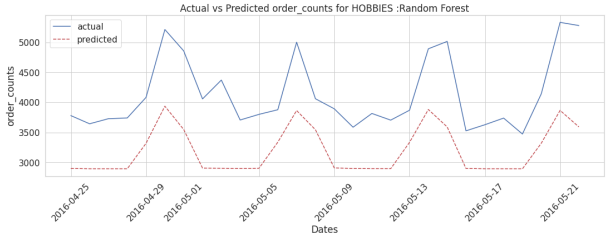

FIGURE 19. Forecasting using random forest for hobbies category

RMSE was 988.62 for HOBBIES category. Despite not capturing the trend perfectly, this was an improvement from our previous models as sampling was now done on a daily basis.

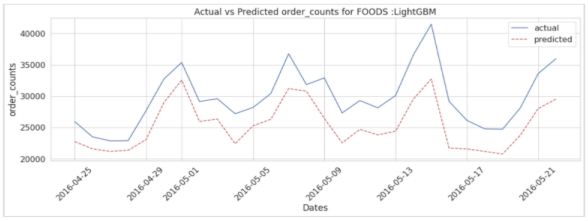

FIGURE 20. Forecasting using lightgbm for FOODS category

Light GBM The RMSE for the FOODS category for LightGBM was 4705.59, a 27.7% improvement over the Random Forest model and a 2.02% improvement over the SARIMAX model.

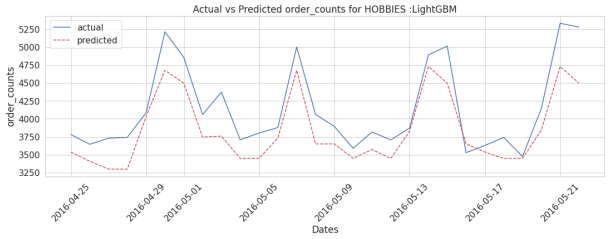

FIGURE 21. Forecasting using lightgbm for hobbies category

The HOBBIES category had an RMSE of 356.64, a 64% improvement from Random Forest.

LSTM Models The RMSE for the Hobbies category for LSTM was 646.36, a 34.6% improvement over the Random Forest model and there was no significant improvement from LightGBM model. But, a better transformation of the data or more hyperparameter tuning might lead to better results, and is still a point of more experimentation.

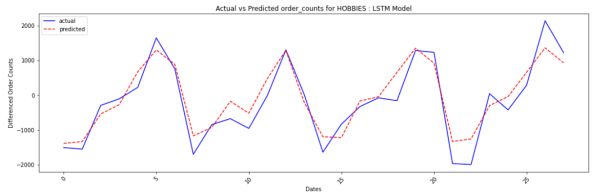

FIGURE 22. Forecasting using LSTM for Hobbies category

FIGURE 23. Forecasting using LSTM for Household category

The HOUSEHOLD category had an RMSE of 446.29, a 83.5% improvement from Random Forest. The results were almost similar as that of the LightGBM model with only 16.8% improvement.