IEOR E4574 Prediction:

A Real-World Application

Sales Forecasting

Wenxin Zhang (wz2585)

Yajie Zhang (yz3876)

Ran Liao (rl3303)

# Contents

### Introduction

### Data Source

### Data Processing

### Target Variables

### Predictive Variables

### Pre-modeling

### Modeling

### Algorithmic Solution Results

# Introduction

The purpose of this project is to construct a robust forecasting solution that predicts Gross Merchandise Value (GMV) for online retailers based on the results generated by Group 2 (Members include: Yu Song Ng, Wei Xiong Toh, and Vibhu Krovvidi).

By analyzing their project, although they got a low value of MAPE with their model, we noticed multiple problems. To correct their approaches and improve the model, several changes were made. In addition to correction, we also included another external dataset with the aim of further optimizing the predictions. Therefore, the purpose of this document is to identify and explain the changes we made to the project, showcase the optimization we did, and compare the results with those of Group 2.

Changes and optimizations can be summarized into three sections:

* Data Processing
  + Remove duplicate records to avoid double counting
  + Remove cancellations and refunds as they are irrelevant to GMV
  + Add visualization based on the original dataset
* Feature Engineering
  + Add weather-related features (only for neural network approach)
  + Remove feature *average wage*
  + Remove feature *quantity*
* Modeling
  + Facebook Prophet (removing quantity and poorly correlated feature)
  + ARIMA (removing quantity and poorly correlated feature)
  + SARIMAX(newly added)
  + Neural Network (newly added)

To make this project comparable with the previous version, we kept the same target variable and used the same metric to evaluate our models, which is MAPE in out-sample tests.

# Data Extraction

The dataset we used in most models is composed of the original dataset and external features used by Group 2. For the Neural Network, we included one more external data source, which is weather information, to improve the performance of the model further.

**Original Dataset**

The original dataset used in this project was collected from UCI Machine Learning Repository, which is called [Online Retail II Data Set](https://archive.ics.uci.edu/ml/datasets/Online+Retail+II). It was stored as a CSV file and was downloaded in a zipped folder. The data are stored locally and are processed and integrated using Python scripts.

**External Dataset**

| Auxiliary Data | Rationale | Source |
| --- | --- | --- |
| Average earnings (wages) in  the UK | Proxy for purchasing power | <https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/earningsandworkinghours/datasets/averageweeklyearningsearn01> |
| Retail sales in the UK | Identify patterns in consumer spending | https://www.ons.gov.uk/busi nessindustryandtrade/retaili ndustry |
| Monthly unemployment rate in the UK | Another proxy for purchasing power | <https://www.ons.gov.uk/employmentandlabourmarket/peoplenotinwork/unemployment/timeseries/mgsx/lms> |

**New External Dataset**

| Auxiliary Data | Rationale | Source |
| --- | --- | --- |
| Daily weather conditions in the UK  (newly added) | Proxy for user propensity for online shopping | <https://www.kaggle.com/datasets/emmanuelfwerr/london-weather-data?resource=download> |

# 

# 

# 

# 

# 

# Data Processing

## Data Overview

The dataset we used is a combination of the raw dataset and macroeconomic dataset, similar to Group 2; while for the neural network model, not only did we use the existing features, but we also added a weather information dataset.

## Data Diagnostics

We first did a sanity check on the data diagnostics provided by Group 2. To summarize, these are all the checks they did, and we agreed.

| Number of records: | 981330 rows and 15 columns |
| --- | --- |
| Missing values | No missing value for attributes that will be used later as models’ features |
| Date Period | From 2009-12-01 to 2011-12-09 |
| Quantity column | Negative quantities deemed as potential cancellations and refunds that we don’t want to account for |
| Country column | Majority of sales are UK-based |

Again, most of the online retail data were collected from the transactions that ship domestically. As the graph shows below, with the y-axis representing the log of the number of transaction records, the UK has the dominantly largest percentage of the total records.

Chart, bar chart, histogram

Description automatically generated

Here are some problems with the data that we further identified.

| Duplicate records if any | duplicate records between 2010-12-01 and 2010-12-09 in both sheets of the original retail dataset |
| --- | --- |
| Price column | Non-positive prices were suspected as potential cancellations, refunds, and giveaways |

For the weather dataset, we did the same quality check, and the results are shown as below:

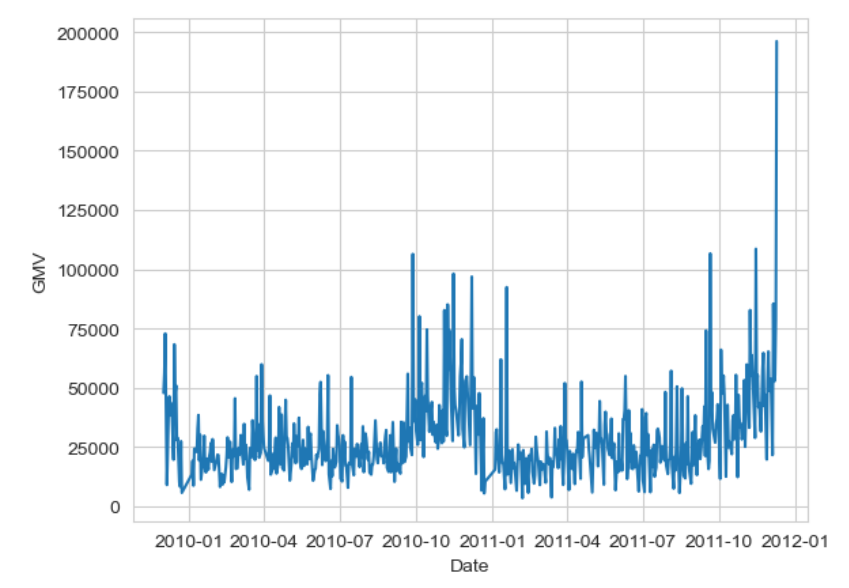
| Number of records: | 15341 rows, 10 columns |
| --- | --- |
| Duplicate records if any | 0 duplicate records |
| Missing values | 302 missing value come from snow\_depth after the date filtered for the period 2009-12-01 to 2011-12-09 |
| Date range | From 1979-01-01 to 2020-12-31 |

## Target Variables

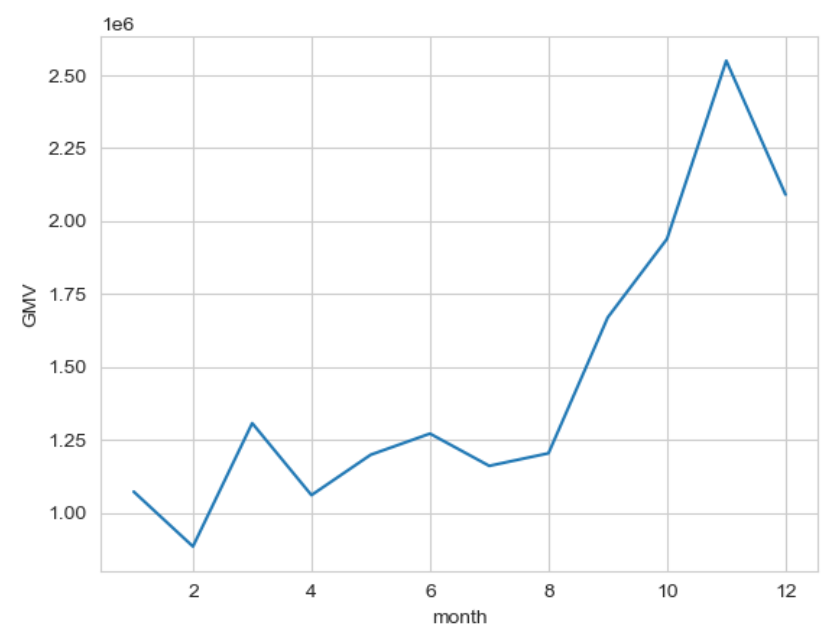
The target variable of this project is Gross Merchandise Value (GMV) for online retailers. GMV is a metric that is commonly used in analyzing financial performance. It is defined as

GMV = Sales Price of Goods x Number of Goods Sold

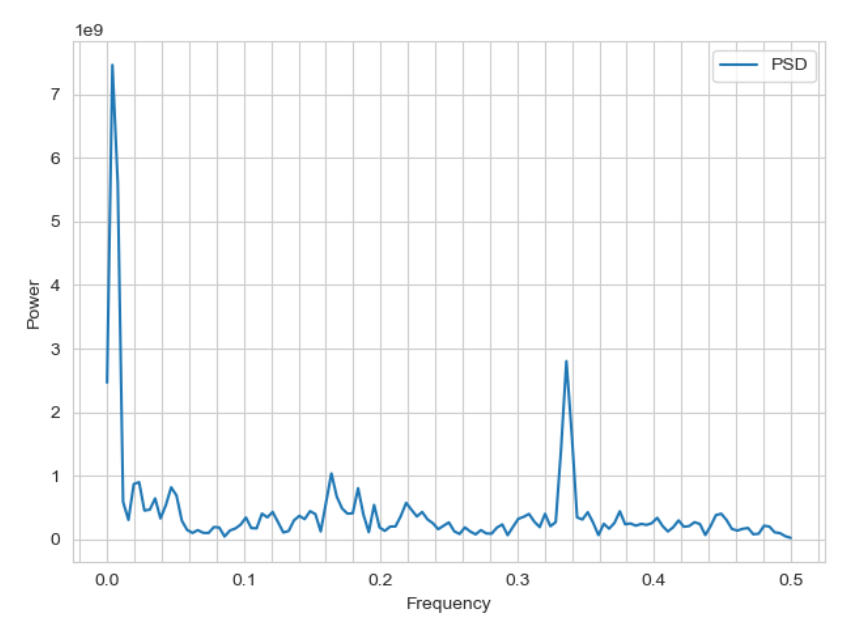
Instead of using the original time unit, we aggregated the records based on the *InvoiceDate*, so that each record in the dataset represents a daily total GMV of online retailers.

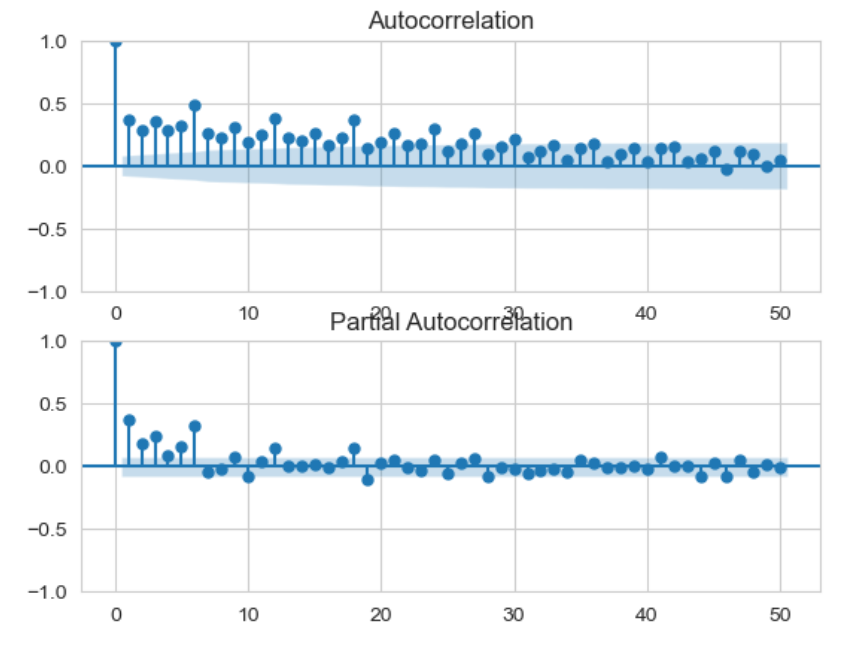


An interesting difference to note here is that there are no zero sales with our data cleaning compared to what Group 2 got. An additional thing to note is that there is an elevated spike in sales around December 2011.



From these plots, the most obvious pattern is the spike in sales around November each year, which matches what we found above. To identify all the frequencies at which there is a strong seasonal influence, the autocorrelation and power spectral density plots are done for the daily aggregate.





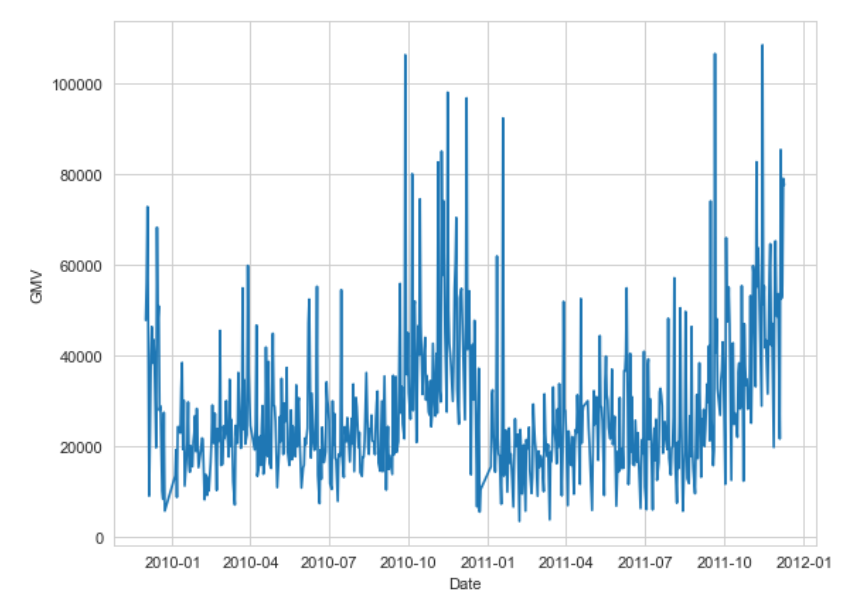
Both autocorrelation and the PSD plot show some signs of seasonality. The autocorrelation plot shows autocorrelation at time intervals of 6, which is about a week apart. This is expected because we can expect the day of the week to impact customers' buying behavior. The PSD plot for daily sales shows several spikes at different frequencies, at around f=0.003 and 0.007, which corresponds to a time period of 128 and 256 days, suggesting a seasonality in quarterly sales.

## Data Cleaning

To handle the previously mentioned problems, we took several processing steps:

1. Remove duplicate records within initial retail spreadsheets
2. Filter data into sales in the UK only
3. Filter out the sales with negative quantities and non-positive prices
4. Replace outlier with 7-day moving averages
5. Aggregate the sales data by day

After data preparation, our target variable, GMV, is shown below.



## Predictive Variables

All of the predictive variables come from external datasets. For our first three models, only the features from Group 2 are considered. For neural networks, weather conditions were added as new features.

## Variable List For All Models

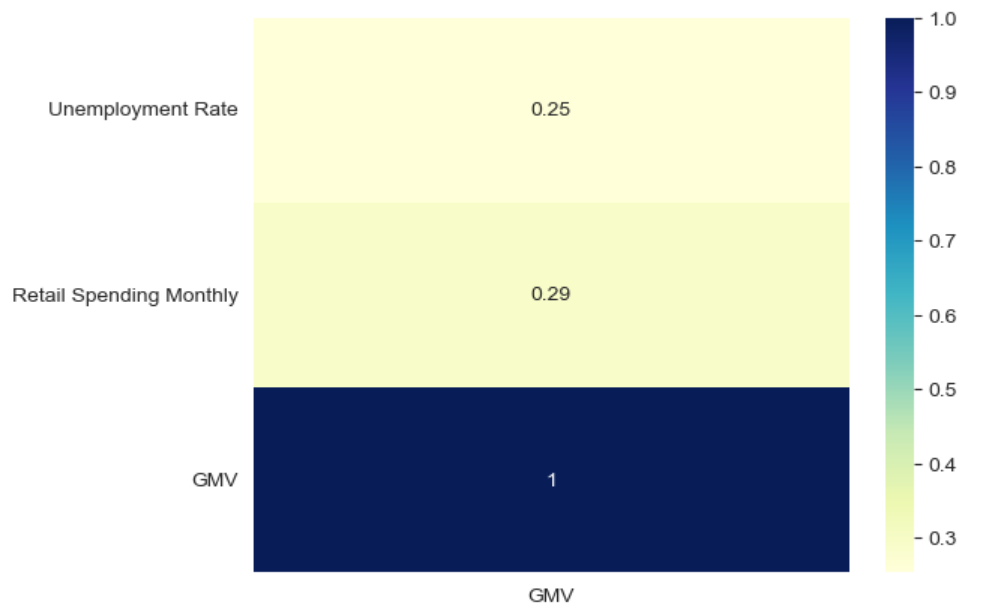
1. InvoiceDate: Date of invoice, used as an index for time series. (separated into year, month, day, and weekday for neural network)

2. **Unemployment Rate**: UK unemployment rate in Invoice Date

3. **Retail Spending Monthly**: Average per capital retail expenditure in the UK

4. Average Wage: Average wage per capita in each month in the UK

According to the correlation between features and target variables, we consider only the unemployment rate and retail spending monthly as predictive variables for the first three models since they have a correlation with GMV greater than 0.1 as shown.



## Additional Variable List For Neural Network

1. Cloud\_cover: cloud cover measurement in oktas

2. Sunshine: sunshine measurement in hours (hrs)

3. Global\_radiation: irradiance measurement in Watt per square meter (W/m2)

4. Max\_temp: maximum temperature recorded in degrees Celsius (°C)

6. Mean\_temp: mean temperature in degrees Celsius (°C)

7. Min\_temp: minimum temperature recorded in degrees Celsius (°C)

8. Precipitation: precipitation measurement in millimeters (mm)

9. Pressure: pressure measurement in Pascals (Pa)

In particular, as opposed to Group 2, we removed the feature *quantity*. The reason for this is that our target feature is calculated based on the value of *quantity* and *price*. Therefore, we believe that including the feature *quantity* introduces data leakage to the prediction despite its effect on improving the model’s performance. What’s more, quantities wouldn’t be available for sales forecasting in real life and therefore deemed as an implausible approach.

# Pre-modeling

## Train & Test Split

We set the first 85% of data as a training set while the remaining 15% of the data was put aside as a test set to evaluate the performance of the model after training.

# Modeling

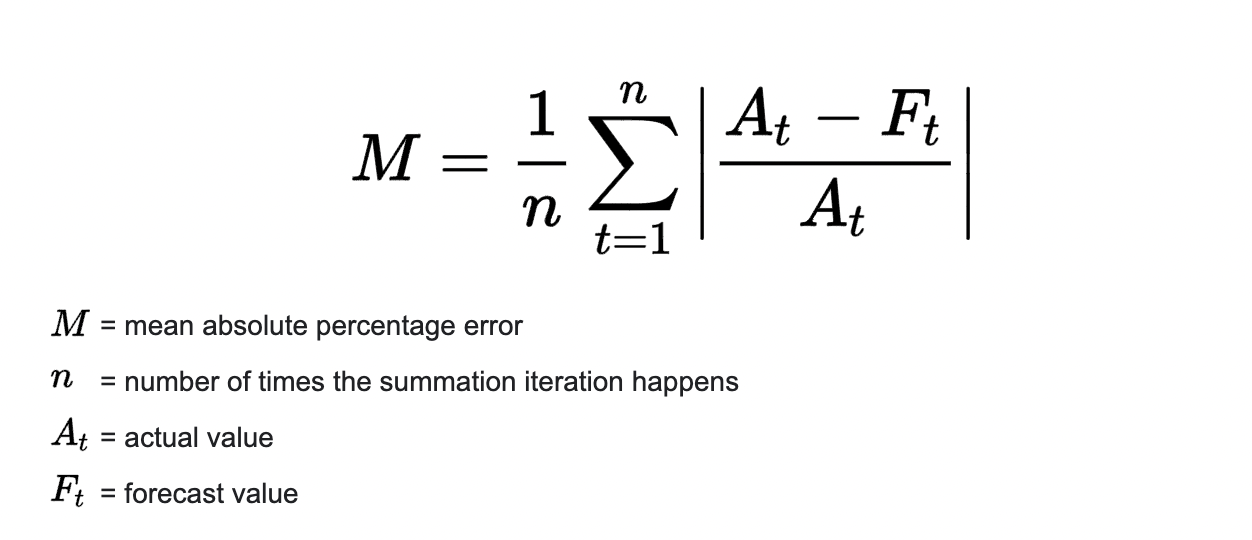
After the data was cleaned and new features were generated, we built 3+1 models based on the training data and tested the performance of the test model. The last model is built on top of the models built previously with the aim of further reducing the prediction errors.

## Models

* **ARIMA Model using Auto-ARIMA**
  + Order of the AR term (p).
  + Order of the MA term (q).
  + Order of the differencing (d).
  + AIC was selected for model selection.
* **SARIMAX Model Using Auto-ARIMA**
  + SARIMAX Model considers the seasonal and exogenous effect in addition to the ARIMA model.
  + Order of the AR term (p).
  + Order of the MA term (q).
  + Order of the differencing (d).
  + P: Seasonal autoregressive order.
  + D: Seasonal difference order.
  + Q: Seasonal moving average order.
  + m: The number of time steps for a single seasonal period.
  + AIC was selected for model selection.
* **Facebook Prophet with Exogenous Features**
  + Facebook Prophet decomposes a time series into trend, seasonality, and holiday effects.
  + Added the same exogenous features as the SARIMAX model
* **Neural Network**
  + Neural network is a series of algorithms that endeavors to recognize underlying relationships in a set of data through a process that mimics the way the human brain operates.
  + Number of hidden units
  + Number of epochs
  + Number of layers
  + Learning rate

## Evaluation Metrics

In this project, we use MAPE as our evaluation metric. MAPE represents the mean or average of the absolute percentage errors of forecasts. Its function is shown below:



# 

# 

# 

# 

# 

# 

# 

# 

# 

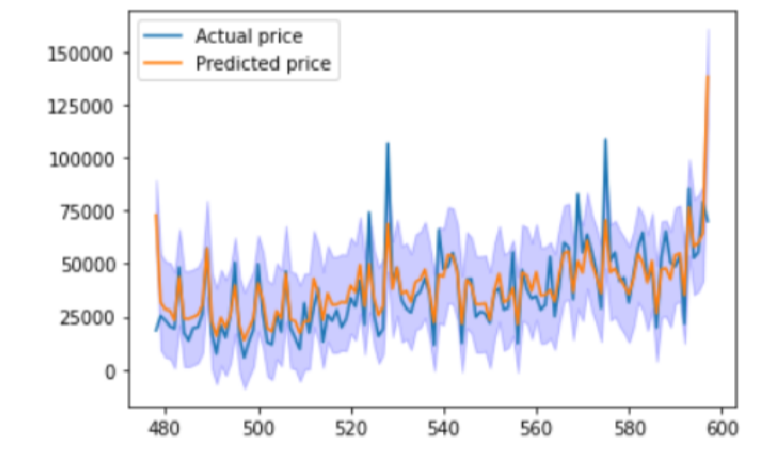
# Algorithmic Solution Results

| Time Period/  MAPE | Group 2’s model | Facebook Prophet | ARIMA | SARIMAX |
| --- | --- | --- | --- | --- |
| First 1/4 | 27.8% | 33.8% | 37.9% | 47.0% |
| Second 1/4 | 29.1% | 22.4% | 28.4% | 19.0% |
| Third 1/4 | 20.1% | 29.7% | 36.6% | 42.2% |
| Fourth 1/4 | 18.0% | 42.1% | 43.3% | 40.2% |
| Overall | 23.7% | 36.3% | 39.6% | 40.3% |

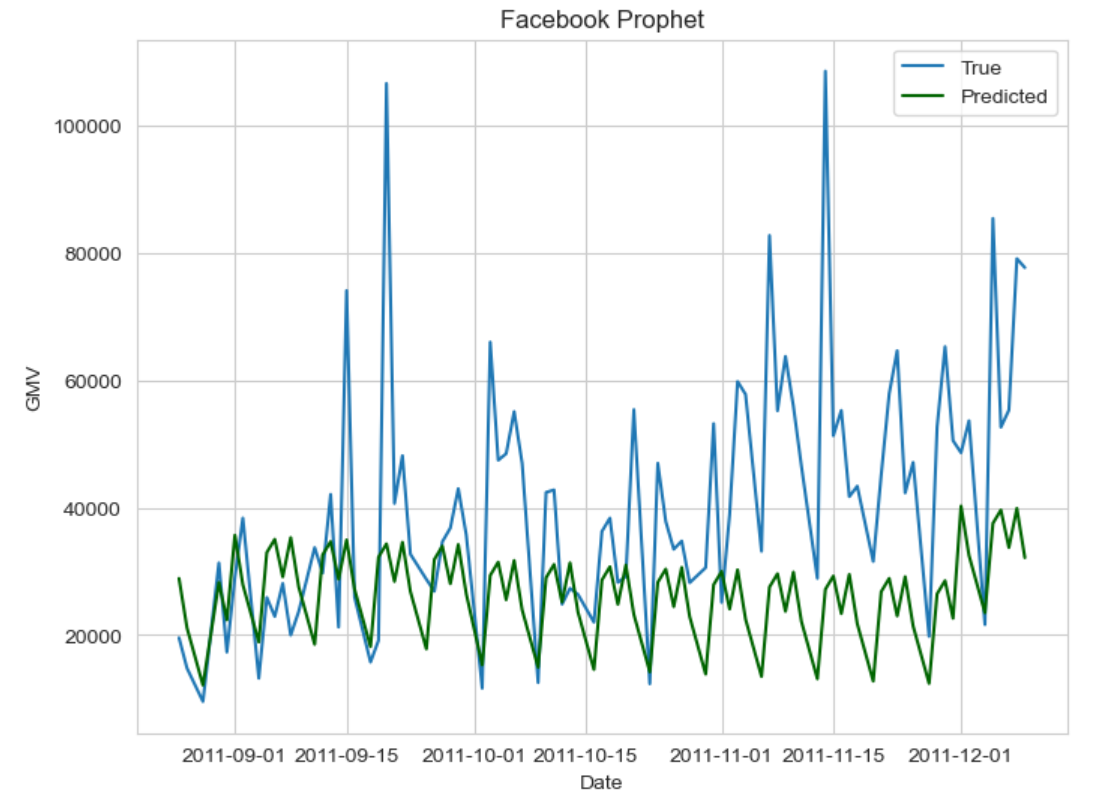
Again, recall they used quantity as a feature which overestimates the efficacy of their model. **We have also tried removing the quantity feature manually from their own model and got a value of 49.4% for MAPE. So, all of our models have done a better job of predicting the GMV.** Out of the three models we trained, Facebook Prophet outperforms the others. This conclusion is further supported by the visualization of prediction errors in equally-sized four regions of the forecasting time period. Lastly, the visualizations of predicted values vs. actual GMV shown below help capture the bigger picture.

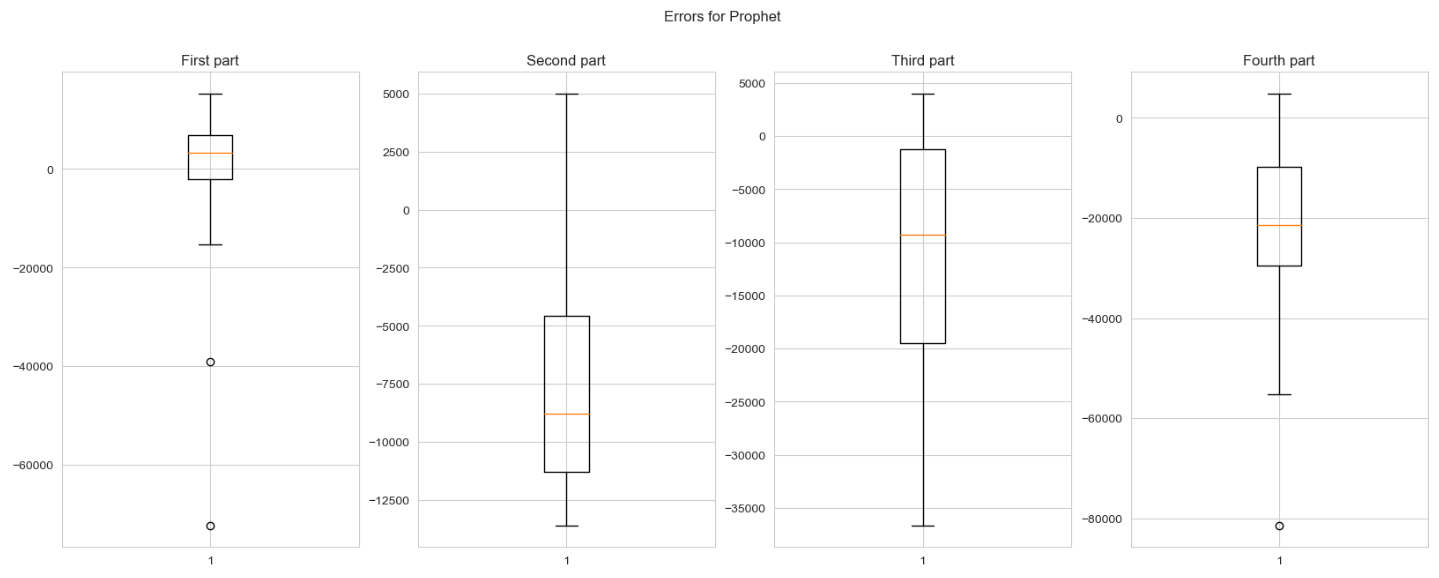
## Predicted vs. Actual

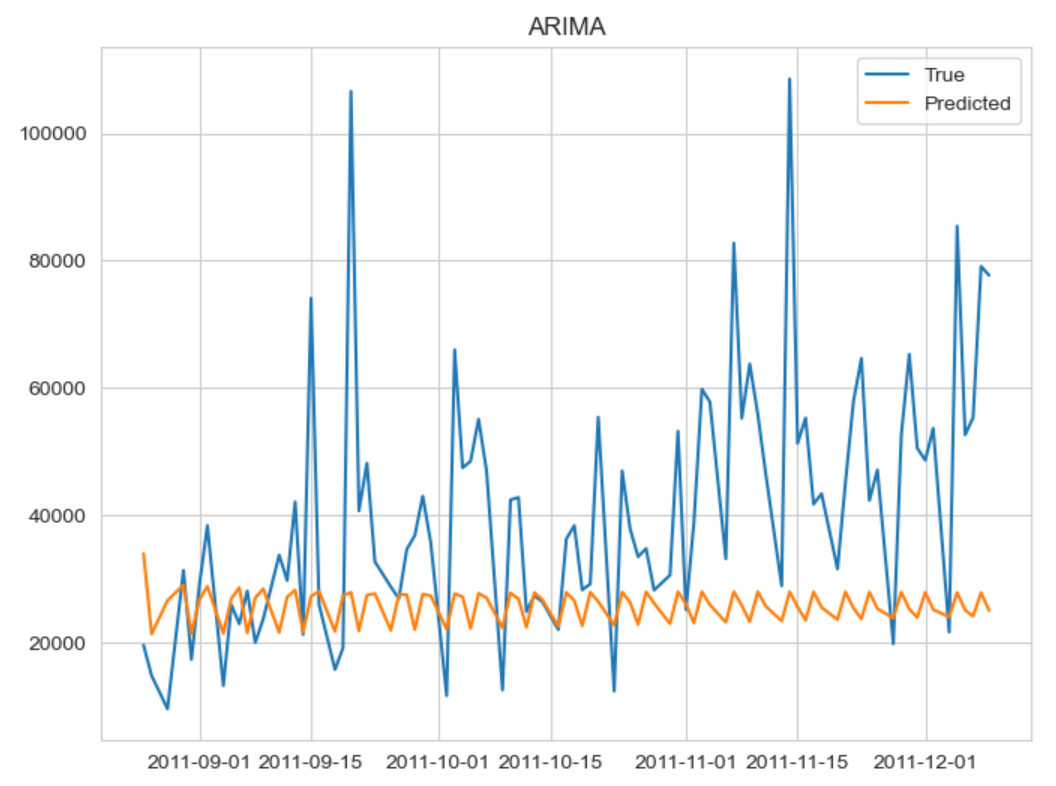
### Group 2 (with *quantity)*

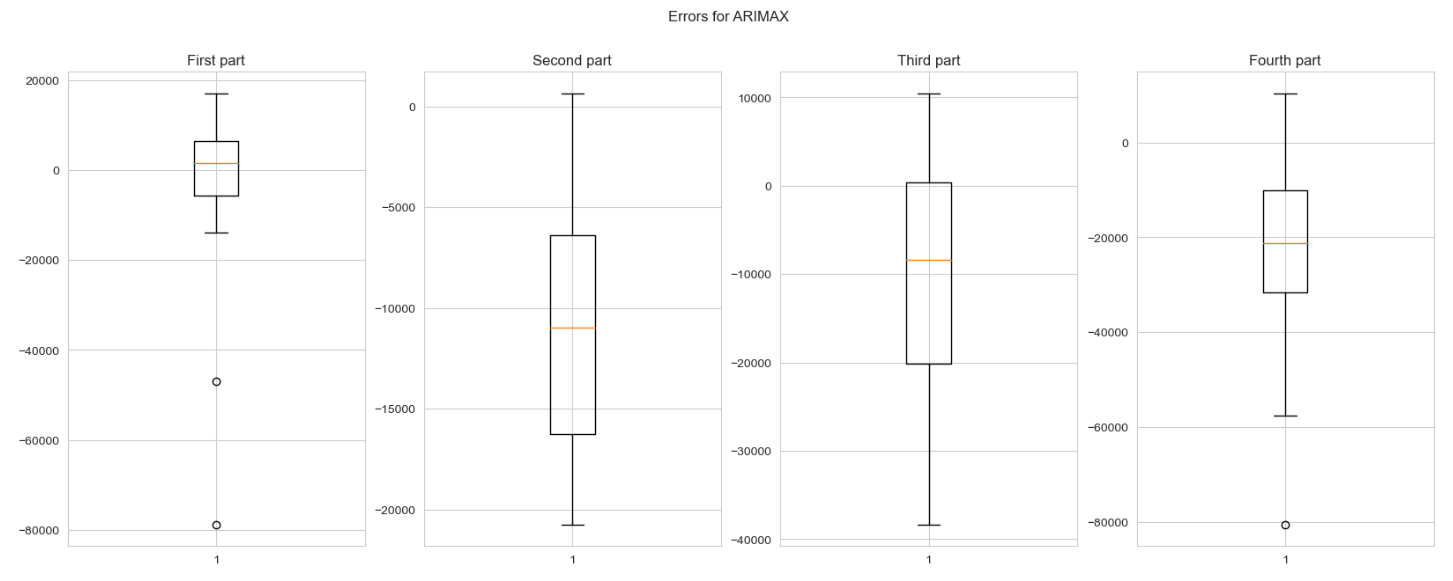


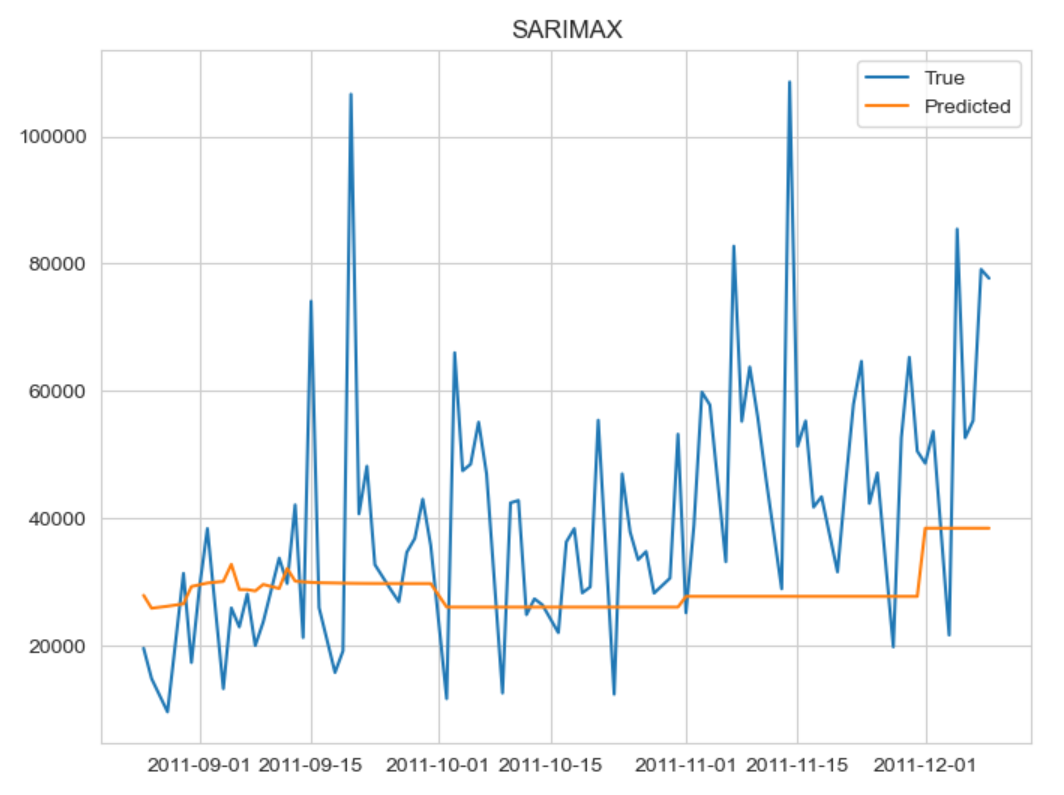
### Our models (without *quantity*)

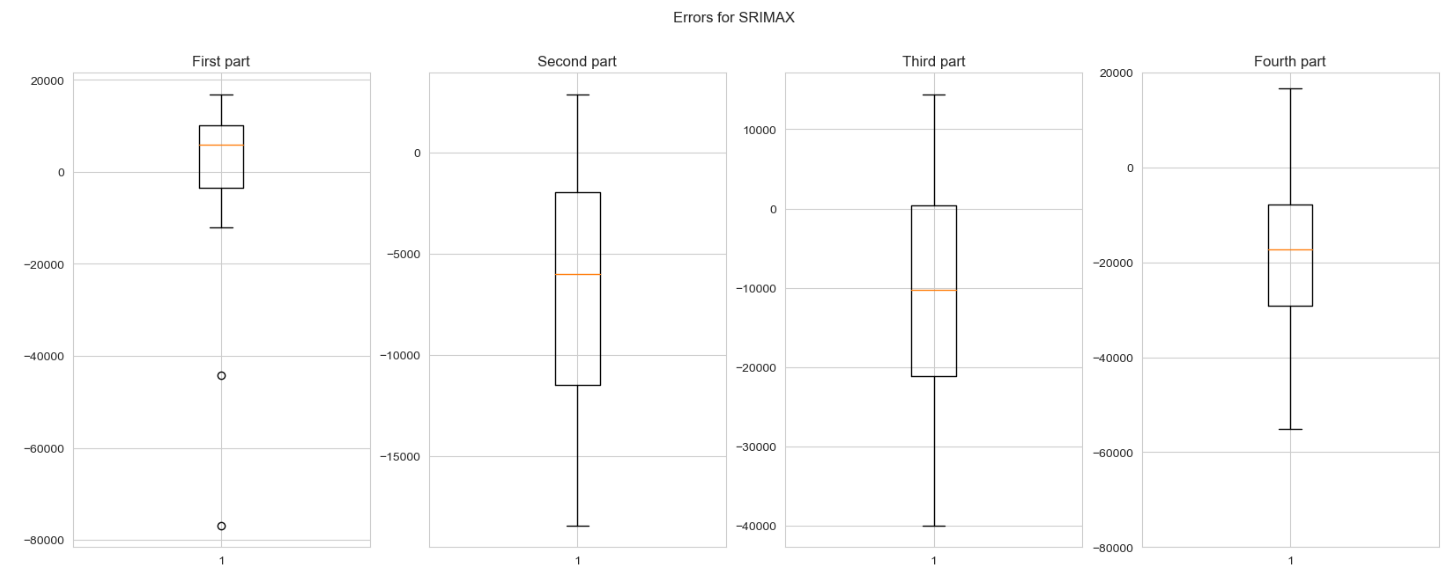












Generally speaking, the Facebook Prophet’s predictions seem to be the most consistent relatively. The spread of prediction errors tends to be bigger and bigger as we proceed from region 1 to region 4 for all modeling techniques, which makes sense as we would be less certain further in time (notice the scale of boxplots can be different within and across groups).

As we mentioned, we have tried an additional approach with a neural network to further improve the predictions by using another external dataset: weather information, which will be discussed below.

## Neural Network

A neural network is a series of algorithms that endeavors to recognize underlying relationships in a set of data through a process that mimics the way the human brain operates. In this project, we used it to predict the GMV in the future. To train the neural network, we used the rolling window with a window size equal to 20 and a test window size equal to 5. It means that we used the first 20 days as the small training set to predict the GMV of the next 5 days. Based on the loss calculated from predictions for the next 5 days, the parameters are updated. This process is repeated every 25 days. After all data in the dataset is run, we finish one epoch and move to the next epoch. There are 1000 epochs in total. After the training is finished on the training dataset, the model will be applied to the testing dataset. Notice that the size of the training dataset and testing dataset is similar as before with 85% and 15% respectively.

The Neural Network method is applied to two datasets. One is the dataset with external features added by Group 2. The other was the dataset with both external features added by Group 2 and weather information features. For both of them, after tuning the parameter, the setting of parameters are the same, which is shown below:

| **Parameter** | **Value** |
| --- | --- |
| Number of hidden units | 128 |
| Number of epoch | 1000 |
| Number of layers | 3 |
| Learning rate | 5e-3 |

The table below shows the best result of this model based on the two different datasets.

| **Dataset** | **MAPE** |
| --- | --- |
| Dataset with weather features | 27.3% |
| Dataset without weather features | 34.6% |

Based on the table, we can conclude that the dataset with weather features performs better compared to the original dataset.

# Next steps

If we have more data in the future, the models are likely to perform better in capturing the seasonal trend. Under the current restricted sample size, we may consider other datasets just as the weather dataset, for further improvement to our models’ performance. Also, the way we handle outliers could have a potential impact on the performance, and there may be better options than simply replacing by moving averages.

What’s more, we currently use the time series forecast framework to model the total business revenue. Alternatively, we can model the time series for each product separately to provide a more granular view of the sales situation. The forecasted values are insightful in that it provides statistical evidence to help the retail companies make better decisions in aspects including marketing, pricing and promotion, etc. In actual use cases, we can make another machine learning model to better comprehend the relationship between the sales volume and other factors such as price and marketing costs.