

Forecasting Project
Process Documentation
V1.0

Online Retail Sales Data

Yu Song Ng
Wei Xiong Toh
Vibhu Krovvidi

Contents

Introduction	2
Data Extraction	4
Workflow	4
Data Processing	5
Data Overview	5
Data Diagnostics	5
Target Variables	6
Predictive Variables	10
Variable List	10
Pre-modelling	11
Training & Testing Split	11
Modeling	12
Facebook Prophet	12
Evaluation Metric	14
Team Information	15

Introduction

- The purpose of this project is to construct a robust forecasting solution that predicts **Gross Merchandise Value (GMV)** for an online retailer.
- The Gross Merchandise Value is a metric used in many retailers as a measure of sales and therefore revenue. It can be defined as:

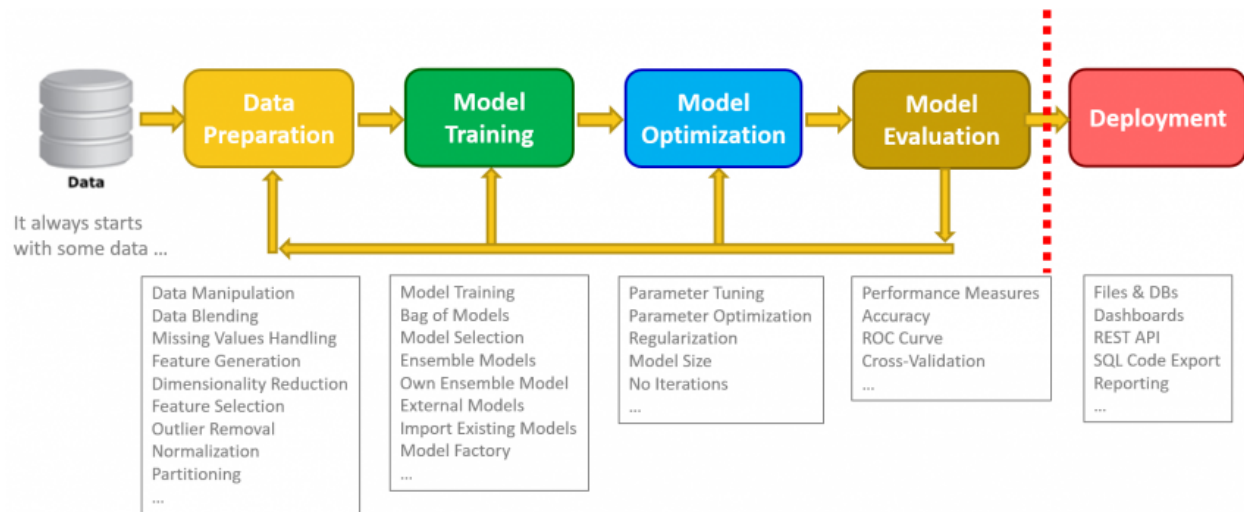
$$GMV = \sum^{day} (Price\ of\ product * Quantity\ of\ product\ sold)$$

- Being able to forecast GMV is critical in the operation of a business, especially a gift retailer such as the one being examined in this project.
 - Gifts have a seasonal element, with gift purchases peaking in the holiday season.
 - This seasonality means that the business will experience uneven sales year-round, resulting in a disproportionate reliance on winter sales to keep the business revenues up.
 - Predicting both on and off-season GMV will allow the company to plan ahead, prepare stock, marketing strategies, manpower requirements, and capital expenditures in a manner that ensures smooth business operations.
- In this project, we integrate GMV with several other macroeconomic and demographic variable factors that together can help create a robust and flexible forecasting solution.
 - These include the performance of the UK stock market, the unemployment rate in the UK and consumer spending, which could all be correlated with the willingness and ability of customers to purchase gifts.
- Our goal is not a one-off forecasting model but rather a sustainable solution with adequate change management and governance and ownership to allow the business to iterate and expand upon this solution over time. To this end, our project:
 - Standardizes a robust methodology (including analytical approach and data requirements) for the development of forecasting that will be leveraged in coming years.

- o Avoids black-box models in favor of more explainable models
- o Demonstrates the ability to adapt to changing environments
- Our client is the gift retailer cited in this dataset. Our objective is to provide 30-day forecasts based on historical data.

The purpose of this document is to provide a detailed technical overview of

1. Design specifications and parameters
2. Data inclusions and exclusions
3. Predictive variable creation process
4. Target variable definition
5. Iterative model building process
6. Metric evaluation process



Data Extraction

Workflow

From our preliminary rounds of data exploration, more than 90% of sales occurred within the UK. Hence, this project will focus on forecasting sales within the UK.

The following auxiliary data was aggregated from various sources to complement the original sales 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
FTSE 100 historical data	Indicator for overall economic performance. It has more granularity than GDP data.	https://www.investing.com/indices/uk-100-historical-data
Retail sales in the UK	Identify patterns in consumer spending	https://www.ons.gov.uk/businessindustryandtrade/retailindustry
Monthly unemployment rate in the UK	Another proxy for purchasing power	https://www.ons.gov.uk/employmentandlabourmarket/peoplenotinwork/unemployment/timeseries/mgsx/lms

Data Processing

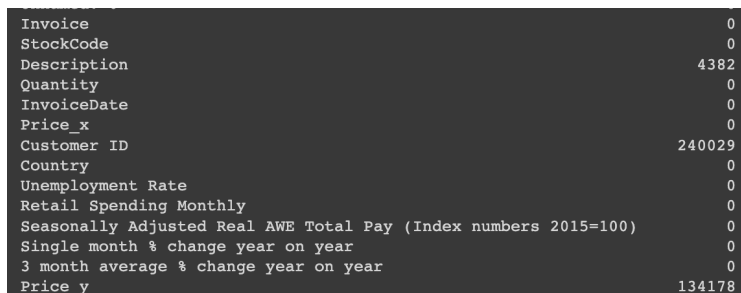
Data Overview

The dataset we use is a combination of the raw dataset and macroeconomic data sourced externally. The data is first scrutinized through a few exploratory checks that give an understanding of the status of the data available.

Data Diagnostics

A series of quality checks are performed on the data extract provided. These checks included:

- Number of records - 981330 rows and 15 columns
- Duplicate records if any - 0 duplicates
- Missing values in relevant fields:
 - o The image below shows the number of missing values for each variable:



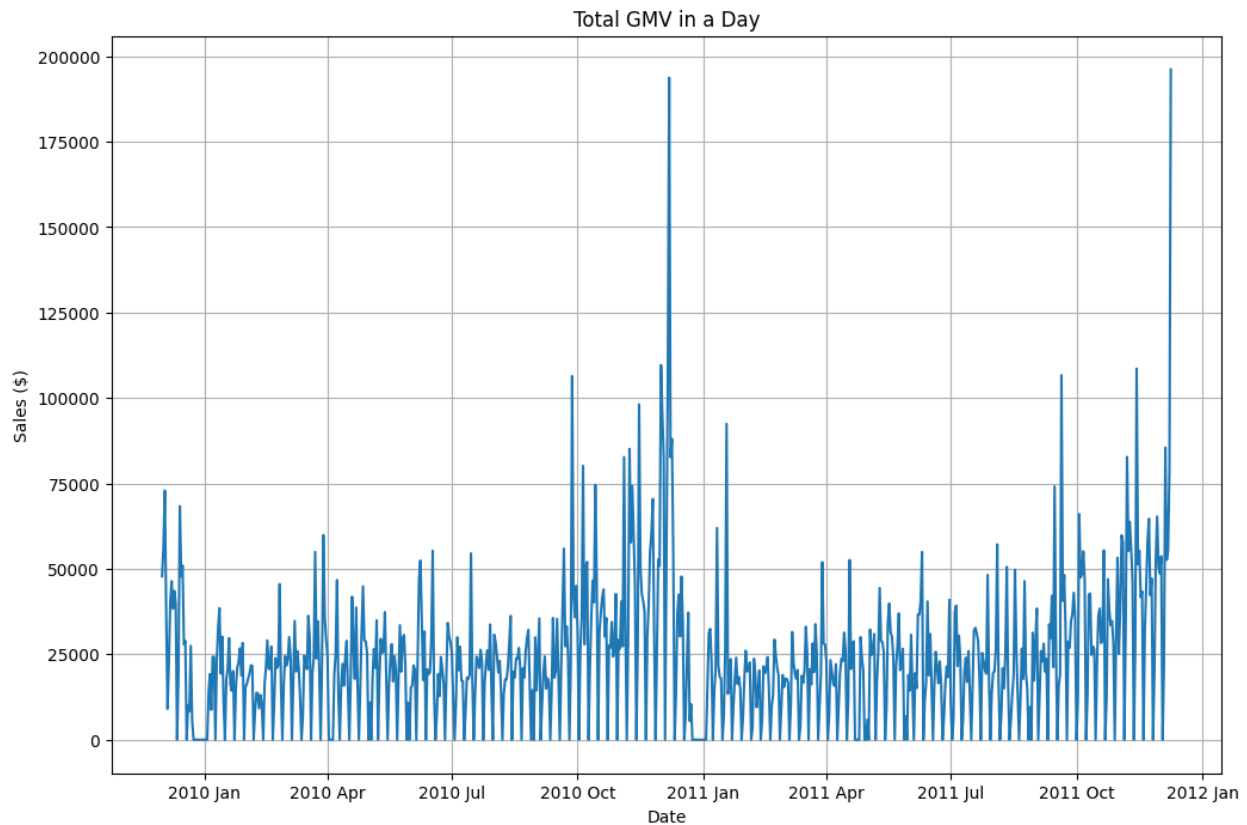
Invoice	0
StockCode	0
Description	4382
Quantity	0
InvoiceDate	0
Price_x	0
Customer ID	240029
Country	0
Unemployment Rate	0
Retail Spending Monthly	0
Seasonally Adjusted Real AWE Total Pay (Index numbers 2015=100)	0
Single month % change year on year	0
3 month average % change year on year	0
Price_y	134178

- Data period confirmation: Data ranges from 2009-12-01 07:45:00 to 2011-12-09 12:49:00.
- Sense check of Quantity column: The data includes cancellations and refunds which result in negative order quantities. The indicator for such entries are the presence of letters in the StockCode column as shown in the image:

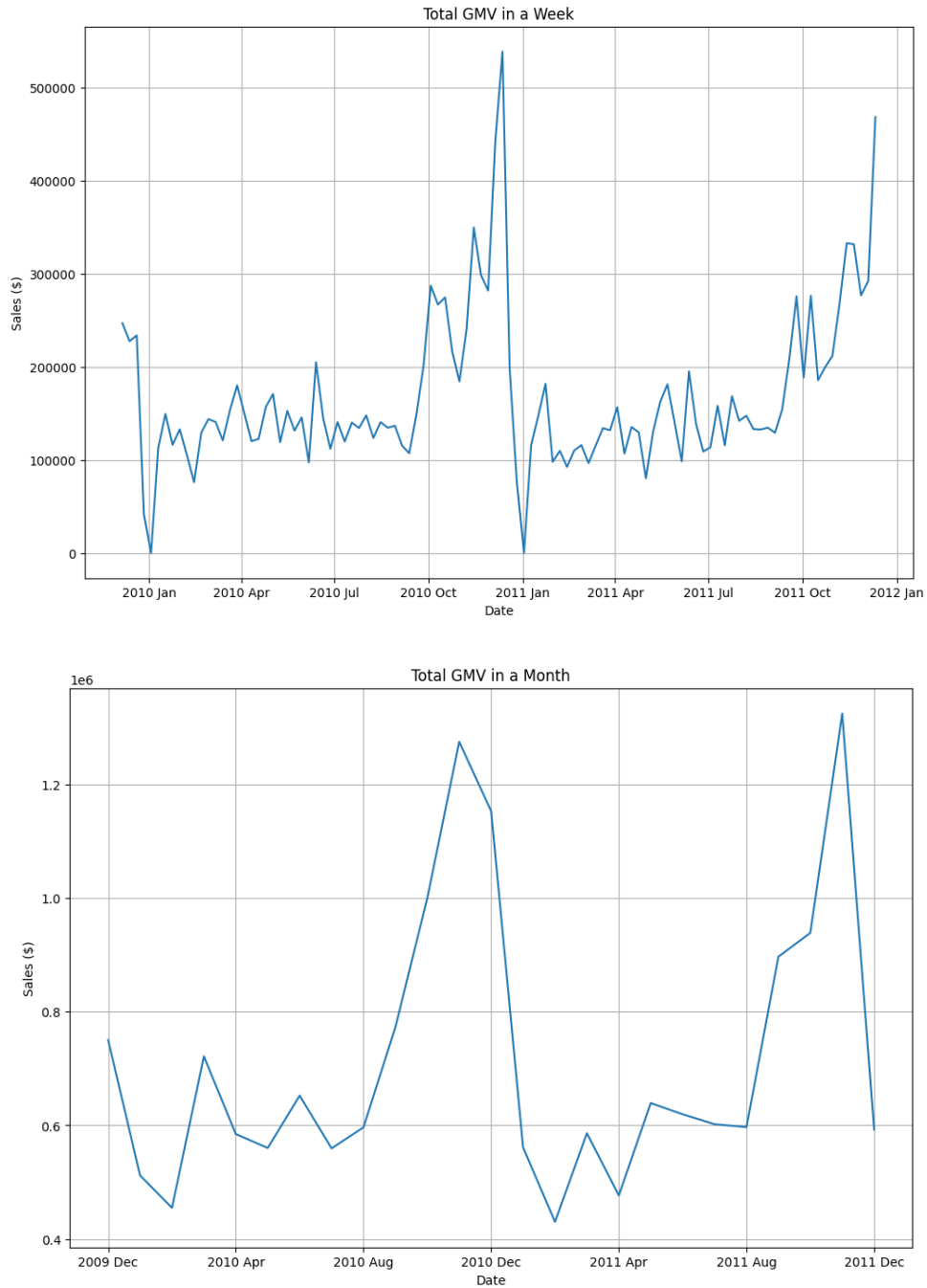
	Invoice	StockCode	Description	Quantity
157	C489459	90200A	PURPLE SWEETHEART BRACELET	-3
158	C489459	90200D	PINK SWEETHEART BRACELET	-3
159	C489459	90200B	BLACK SWEETHEART BRACELET	-3

Target Variables

The target variable selected to be forecasted is total sales. This can be aggregated in different time frames, such as daily, weekly or monthly. The visualization plots for the aggregated total sales over time are shown below.

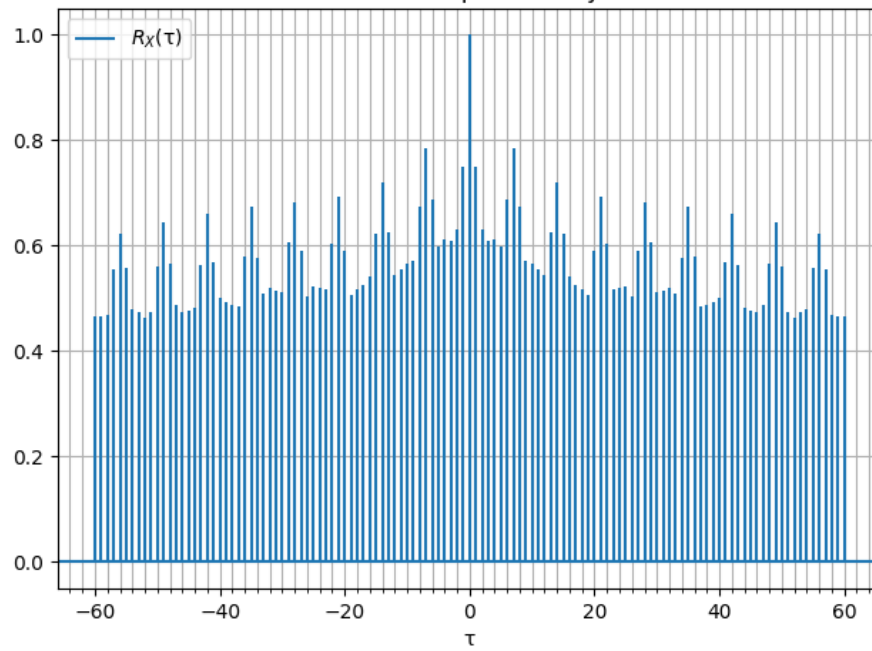


An interesting point to note in the daily sales graph is the fact that there are zero sales on all Saturdays and holidays (such as Christmas and Easter). This is likely due to operational reasons rather than the fact that there are truly no sales on those days.

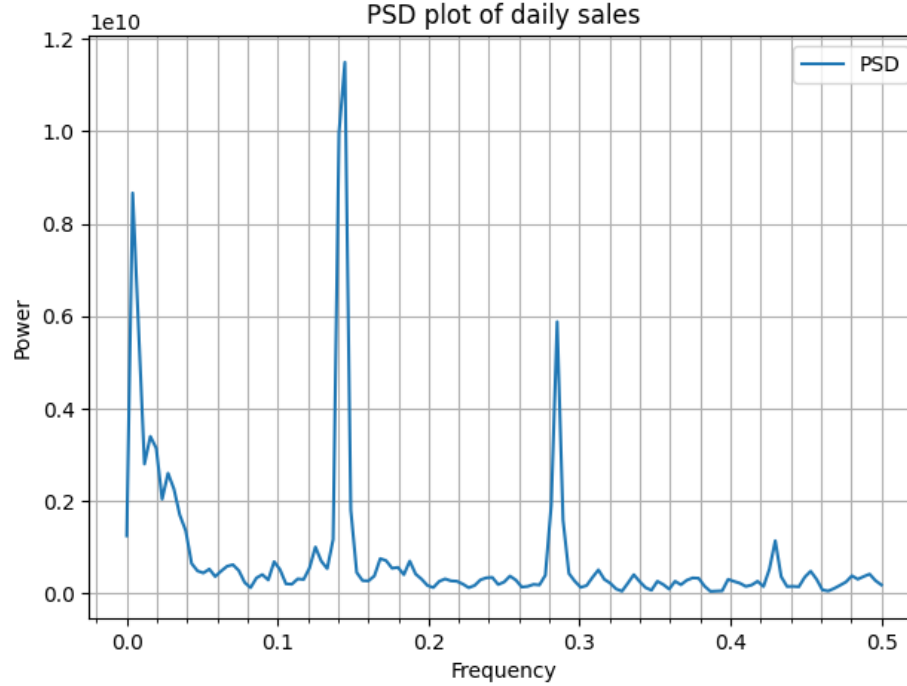


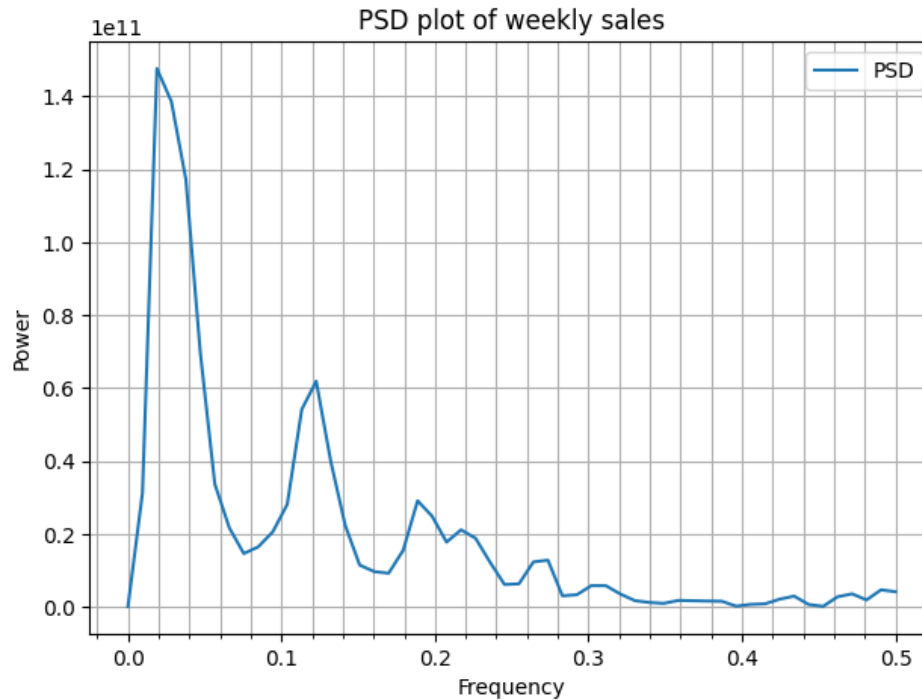
From these plots, the most obvious pattern one notices is the spike in sales during the November - December months each year. There also seems to be some seasonal patterns for other months as well. To identify all the frequencies at which there is a strong seasonal influence, the autocorrelation and power spectral density plots are done for the daily aggregate.

Autocorrelation plot of daily sales



PSD plot of daily sales





The autocorrelation plot shows strong autocorrelation at time intervals of 7, which equates to one week apart. This is to be expected because we can expect the day of the week to have some impact on the buying behavior of customers.

The PSD plot for daily sales shows several spikes at different frequencies. The largest spike is at around $f=0.142$, which corresponds to a time period of 7 days, which is to be expected. There is also a significant spike at around $f=0.003$, which corresponds to 365 days (1 year), and this was discussed before. The smaller peaks at $f=0.284$ and $f=0.43$ correspond to the multiples of the largest spike. In the weekly sales' PSD, the largest spikes occur at $f=0.019$ and $f=0.12$, which correspond to a period of a year and 2 months respectively. Therefore, we can infer that the data has seasonality in the yearly, bimonthly, and weekly time periods.

Data Cleaning

To not influence model prediction as well as the calculation of error metrics, zero values from Saturdays and holidays were removed from the dataset. The zero-valued data points can be added in after doing the prediction.

Moreover, the extremely high sales data near the Thanksgiving and Christmas periods will also affect model performance. The 2 most extreme values were replaced with the 7 day moving average. The forecasted model can then be adjusted to give an estimate of peak holiday sales.

Predictive Variables

A predictive variable is a variable used in algorithmic solutions to predict the target variable. During our analysis, we categorized predictive variables into two categories:

1. Direct variable – These variables were directly from the dataset that was provided by direct customers
2. Derived variable – These variables were created by manipulating the direct variables

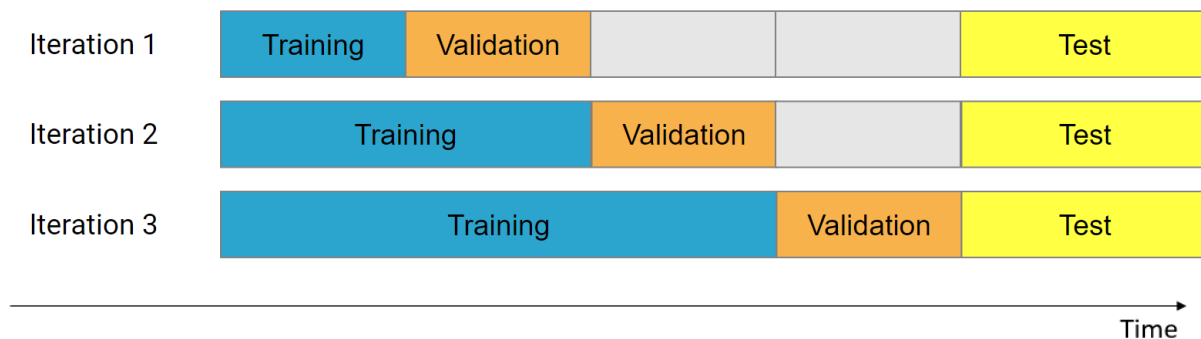
Variable List

1. InvoiceStockCode: Invoice number
2. Description: Brief description of product
3. Quantity: How many units sold in order
4. InvoiceDate: Date of invoice. Used as index for time series
5. Price_x: Price of each unit of a certain product
6. Customer ID: Unique identifier for customer
7. Country: Origin country of purchase
8. Unemployment Rate: UK unemployment
9. Retail Spending Monthly: Average per capita retail expenditure in the UK
10. Seasonally Adjusted Real AWE Total Pay (Index numbers 2015=100): Average total pay per capita for the UK
11. Single month % change year on year: % change in average per-capita total pay (monthly)
12. 3 month average % change year on year: % change in average per-capita total pay (3 month window)
13. Price_y: FTSE 100 closing price for Invoice Date.

Pre-modelling

Training & Testing Split

Cross validation was used to train the models.



As displayed in the figure above, the last 25% of the data was put aside as a test set to evaluate the performance of the final model after training. The other 75% goes through training and validation in three segments to tune the model hyperparameters. At each iteration, more data is available for training, and the model was re-trained and re-tested using the same ratio split.

Evaluation Metrics

The evaluation metrics of choice are Median Absolute Percentage Error (MAPE), $\frac{1}{n} \sum_{t=1}^n \left| \frac{A_t - F_t}{A_t} \right|$, where A_t is the actual value and F_t is the forecasted value. MAPE has the advantage of being easier to interpret, but it penalizes over-forecasts more than under-forecasts and therefore has no theoretical upper bound. To account for this, the upper bound is limited to 100%.

Modeling

Introduction

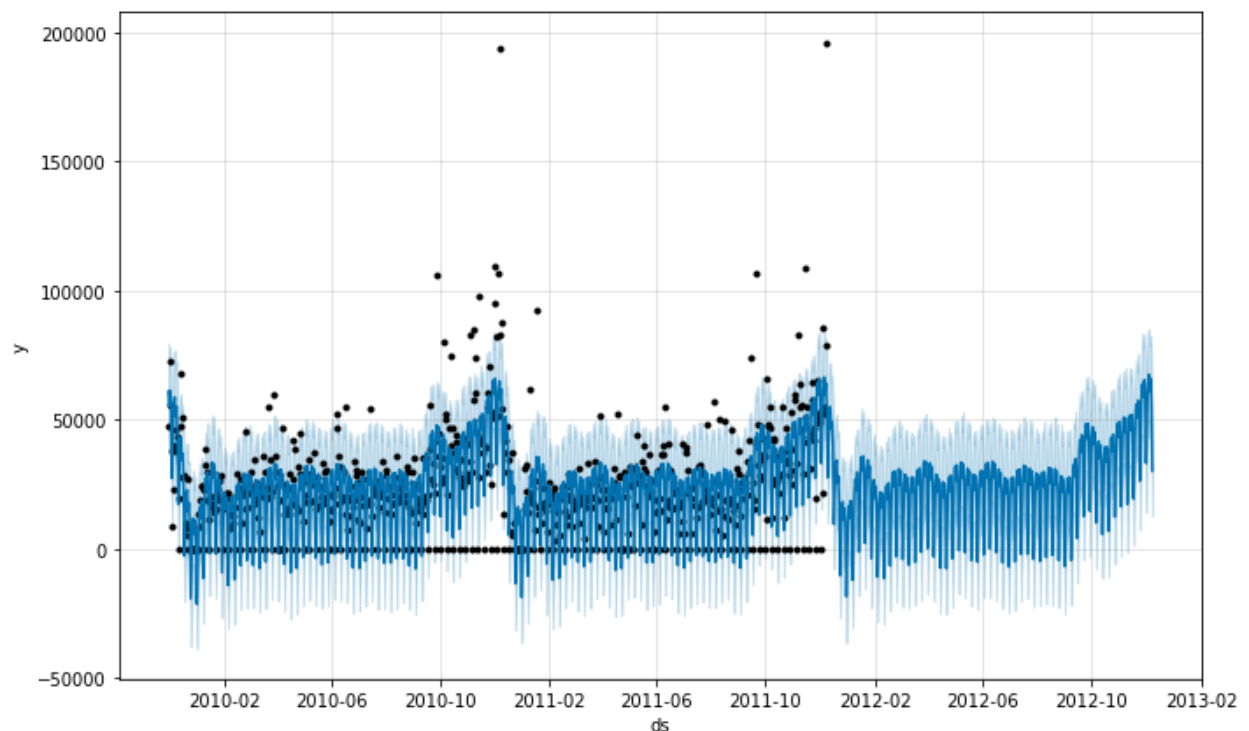
The first stage of our modeling focuses on the use of Facebook Prophet to create a basic, baseline model that does not rely on any predictive variables and looks at a purely autocorrelated model. We then move on to predict exogenous variables using ARIMA. The Prophet model is run on the original data without any modifications as a benchmark.

Facebook Prophet

Facebook prophet is a time-series solution that offers the ability to predict time-series data into the future using additive and multiplicative elements. It offers great flexibility in adding regressors, holidays, and other terms to improve the model performance and offers a convenient way of evaluating model performance as well.

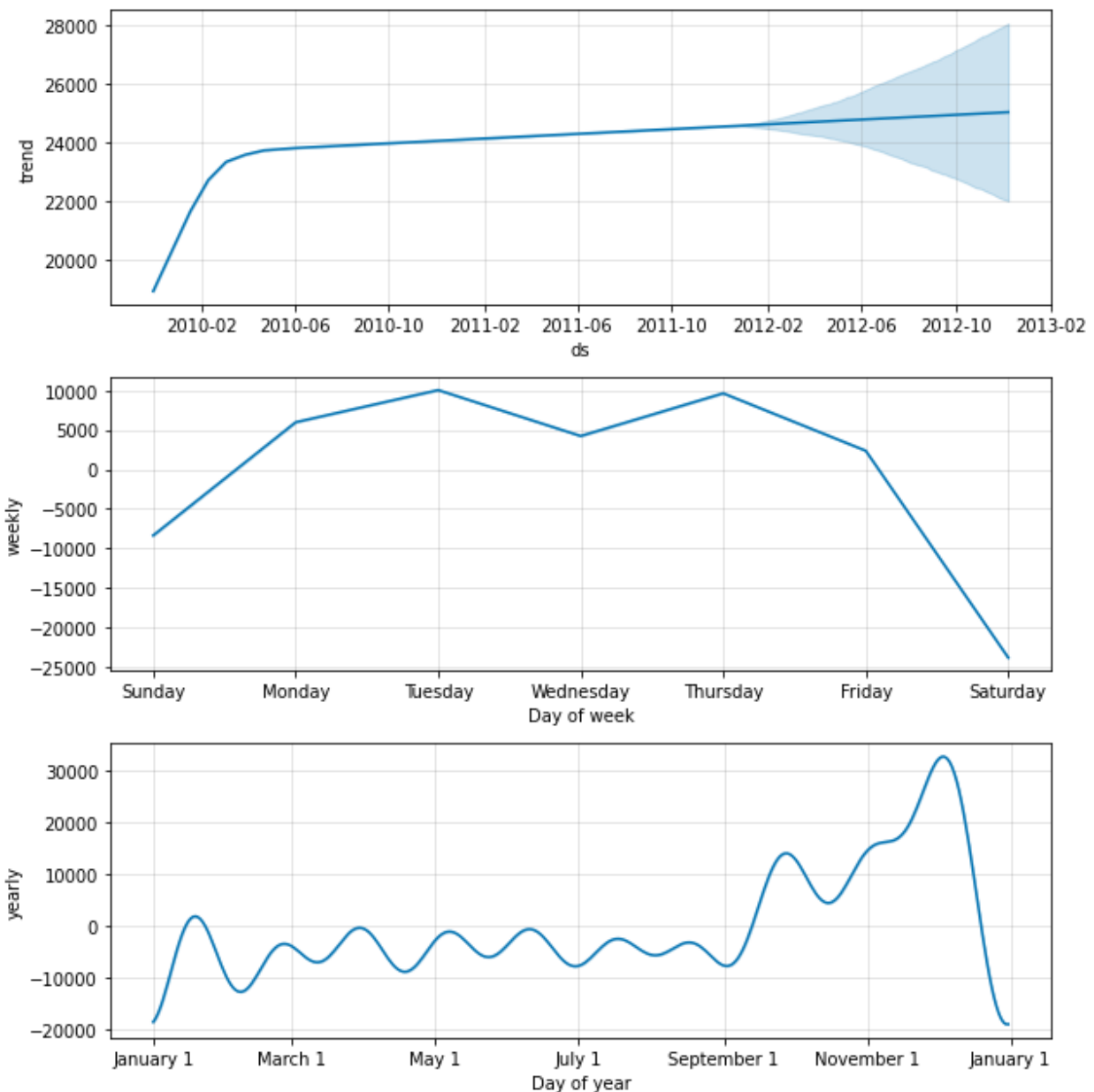
For our model, we decided to use Facebook Prophet to predict 1 year ahead and 30 days ahead. As the forecast horizon increases, accuracy undoubtedly decreases, but the ability to see these forecasts will enable us to visually identify if the Facebook Prophet model is correctly interpreting and recognizing the time series information.

A 1-year Prophet forecast is shown below:



The chart shows data points as black dots. The light blue lines show the confidence intervals at each timestamp. The dark blue line shows the trend fit of the model.

Facebook Prophet also gives us a breakdown of the time series data and the different components such as trend, seasonality, and cyclicalty:



As seen, there is an increasing trend in the data. Looking at the weekly data, we see a weekly seasonality with peaks on Tuesday and Thursday. Within the year as well, there is seasonality

with gift purchases and therefore GMV rising to its peak around December - a result of the holiday season.

Facebook Prophet allows us to see how the error metrics change as the forecast horizon increases. The table showing these changes is seen below:

	horizon	mse	rmse	mae	mdape	smape	coverage
0	3 days	9.505600e+07	9749.666423	7472.177857	0.220744	0.779423	0.833333
1	4 days	1.222866e+08	11058.328235	9140.819322	0.269818	0.528216	0.833333
2	5 days	7.349713e+08	27110.353585	20208.901465	0.350705	0.698255	0.666667
3	6 days	6.557151e+08	25606.935275	16442.474545	0.269818	0.338951	0.833333
4	7 days	6.379654e+08	25257.976149	15573.781728	0.251790	0.326258	0.833333

The data is interesting since it shows that the error metrics are not following a linear trend of increase as the horizon increases. Instead, it shows the greatest error on a 5 day forecast, with reduction in errors over the preceding and subsequent 2 days. This suggests that the model has correctly grasped the weekly seasonality components and is instead comparatively struggling to identify variations on top of the seasonality.

An important corollary to the fact that the GMV values on Saturdays and holidays are all equal to zero in the dataset is that both the MAPE and MdAPE will be skewed higher due to the calculation of both error values. Both MAPE and MdAPE calculate the absolute value of the forecast error divided by the original value of GMV. This results in extremely high error values on these occasions. The distribution of the actual forecasted percentage error will hence include these values too. Therefore, an alternative approach to evaluate our model will have to be used or we will have to exclude the data points that have zero revenue. The latter option was chosen for the subsequent models because MAPE gives a consistent metric to compare the performance of forecasting models.

ARIMA

The second model used was an ARIMA model with grid search. Firstly, the cleaned data (with data points with no sales removed and outliers replaced with a seven day moving average) was tested for stationarity with the Augmented Dickey-Fuller test, and gave a p-value of 0.0184. This shows that we can reject the null hypothesis that the data is non-stationary, and no differencing terms, d , are required. The limits for the number of autoregressive terms, p , and number of lagged forecast errors, q , were set to 14 and 7, respectively. Using the cross-validation methodology, the optimal p and q that reduces the MAPE at every fold was picked. The MAPE was evaluated for each iteration and the results are as shown:

Iteration	Best Model	MAPE
1	ARIMA(1,0,1)	18.4%
2	ARIMA(5,0,6)	31.8%
3	ARIMA(14,0,0)	45.4%

We evaluated the three models on the test data.

Order: (1, 0, 1)

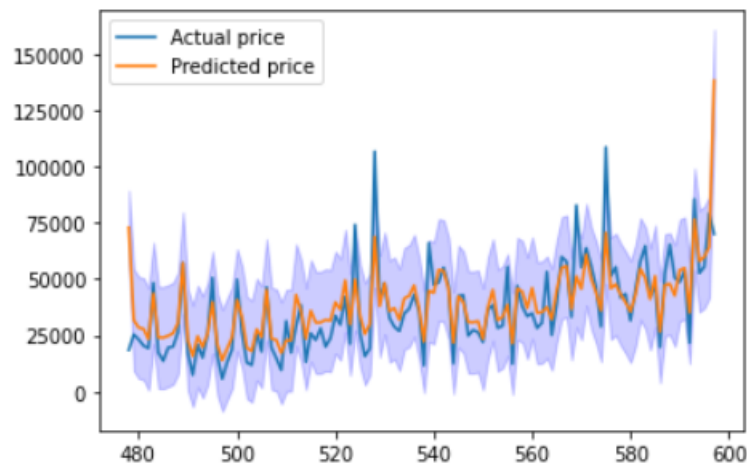
Overall MAPE: 0.28157947144872575 for 120 days

MAPE in period 1 - 30 days: 0.4253801065154002

MAPE in period 2 - 30 days: 0.3099937634422517

MAPE in period 3 - 30 days: 0.2091455867526762

MAPE in period 4 - 30 days: 0.18179842908457475



Order: (5, 0, 6)

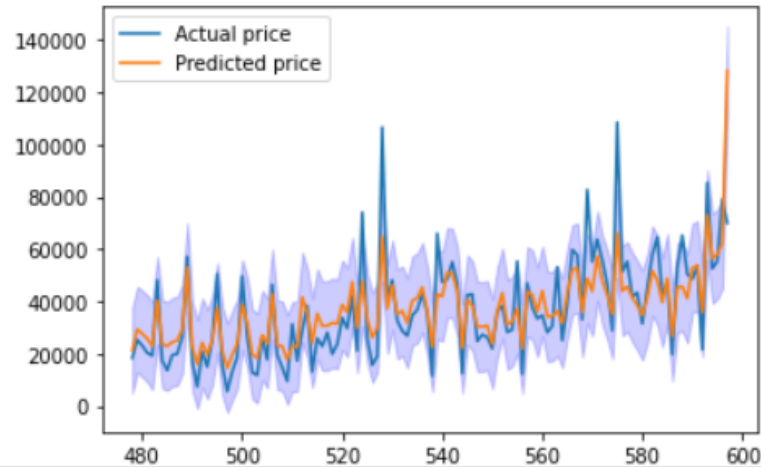
Overall MAPE: 0.26060177595703554 for 120 days

MAPE in period 1 - 30 days: 0.3309174037739068

MAPE in period 2 - 30 days: 0.30833817094718907

MAPE in period 3 - 30 days: 0.21160382827533386

MAPE in period 4 - 30 days: 0.19154770083171221



Order: (14, 0, 0)

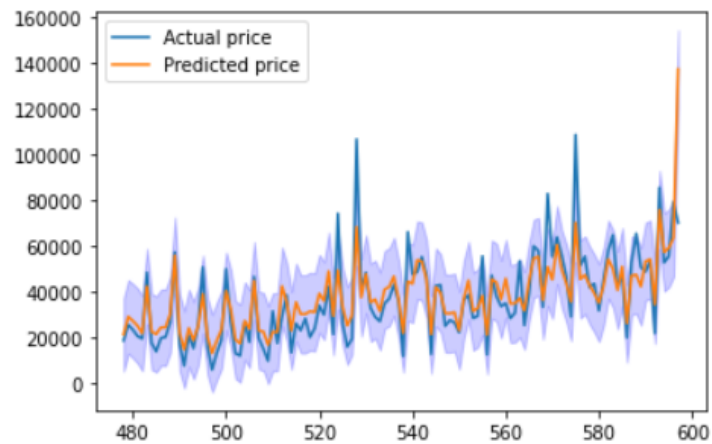
Overall MAPE: 0.23742503147894814 for 120 days

MAPE in period 1 - 30 days: 0.27769740448019514

MAPE in period 2 - 30 days: 0.29107874726730043

MAPE in period 3 - 30 days: 0.2005479526769388

MAPE in period 4 - 30 days: 0.1803760214913582



From the results above, the best model produced from the cross-validation process is ARIMA(14,0,0).

The MAPE for each section of the test set was evaluated and is as follows:

Time Period	MAPE
First 1/4	27.8%
Second 1/4	29.1%
Third 1/4	20.1%
Fourth 1/4	18.0%
Overall	23.7%

Modeling Limitations

MAPE values of 10% and less are desired in the industry but given the limitations of the data we were unable to achieve this.

Firstly, there were too few data points provided - only two years' worth of sales data. Visually, there are clearly seasonal cycles for the holiday period at the end of each year, but with the test set taking up one of those peaks, the models used were unable to capture the information. A simple solution for future extensions would be to obtain more data covering more years so that the models used can capture the seasonality. If this is not possible due to data limitations, we could either use more robust models that can account for such issues, or manually remove the seasonal differences by trial and error.

Moreover, treating the high peaks as outliers is not ideal. Getting more data over a larger number of years will allow the model to track the fluctuations more precisely.

Finally, the dataset lacks other variables to predict on - such as the company's operational, financial and logistics data. This results in omitted variable bias, which adds error to the predictions that no models can remove.

Team Information

Wei Xiong Toh - wt2354@columbia.edu

Vibhu Krovvidi - vk2500@columbia.edu

Yu Song Ng - yn2436@columbia.edu