Capstone 2: E-Commerce Sales Prediction

Springboard – JUNG, JIYOON

Introduction

In the fast-paced world of eCommerce, accurately predicting future sales is crucial for maintaining a competitive edge, optimizing inventory management, and enhancing customer satisfaction. Recognizing the importance of data-driven decision-making, Olist, a prominent Brazilian eCommerce platform, has generously made its comprehensive order dataset publicly available. This rich dataset offers a unique opportunity to delve into the intricacies of online sales patterns and consumer behavior in Brazil. Leveraging this valuable resource, we aim to develop a sophisticated sales prediction model utilizing Prophet, a powerful and flexible forecasting tool developed by Facebook. By employing advanced time series analysis techniques, our goal is to provide Olist with actionable insights that can guide strategic planning, facilitate efficient resource allocation, and ultimately drive the company's growth in the highly competitive eCommerce landscape.

Approach

The initial phase of our analysis will involve data wrangling and exploratory data analysis (EDA) to uncover underlying sales patterns, trends in order sizes, product popularity rankings, geographical distribution of customers, and more. This preliminary exploration is essential for understanding the dynamics at play within Olist's order data.

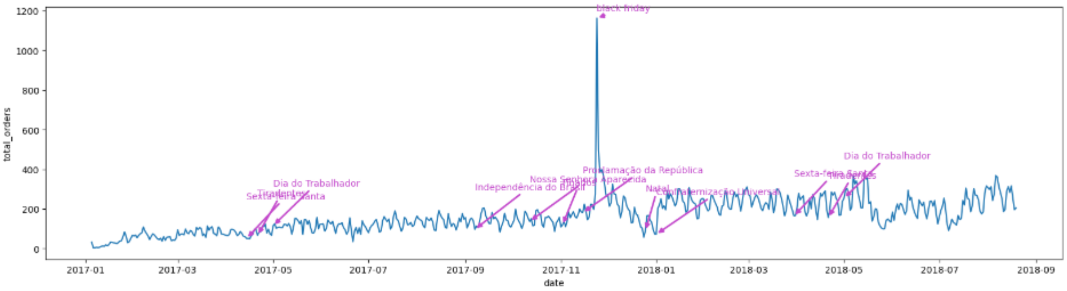
In the subsequent stage, we will employ the Min-Max scaler for data normalization, ensuring that our numerical variables are on a similar scale, which is pivotal for effective modeling. Additionally, we will apply the winsorization technique to address and mitigate the impact of outliers, further refining our dataset for precise analysis. The Root Mean Square Error (RMSE) will serve as our primary metric for assessing model performance, providing a clear measure of accuracy by quantifying the difference between predicted and actual sales values.

Finally, we will leverage the Prophet model, renowned for its robust handling of time series data with strong seasonal components, to forecast future sales. By comparing the RMSE values of our train and test datasets, we will determine the most accurate model configuration. This step is critical for selecting the optimal approach to sales prediction. Additionally, the findings will highlight areas for further research and exploration, potentially uncovering new opportunities for enhancing Olist's sales forecasting capabilities. Through this comprehensive analysis, we aim to equip Olist with the insights needed to navigate the complexities of the eCommerce market strategically and confidently.

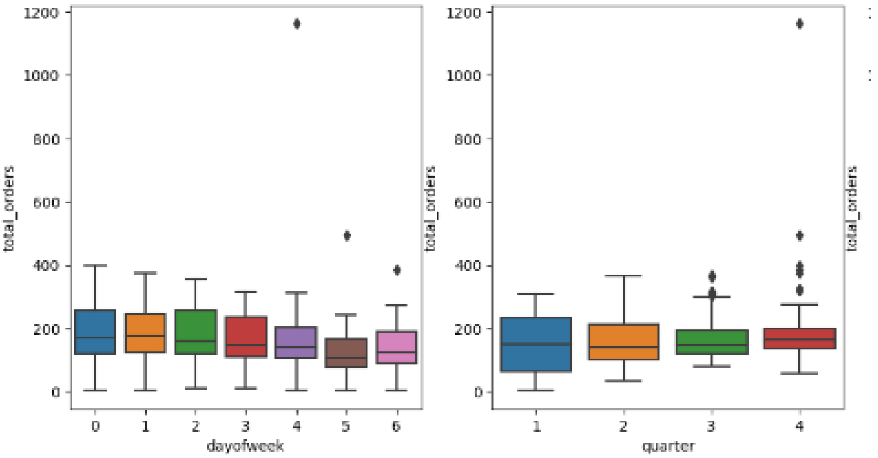
Finding

1. **Sales Trend**

   The "holiday" package provides information about holidays in Brazil. It reveals a sales spike in November 2017, particularly around the Black Friday season. Despite the presence of several other holidays throughout the dataset period, a strong correlation similar to that observed during Black Friday was not found. Overall, the sales trend indicates an increase from 2017 to 2018.

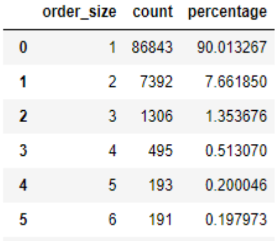


The sales are generally stronger on weekdays and lower on weekends. Typically, sales peak on Tuesdays and hit their lowest point on Saturdays. However, sales on Mondays fluctuate the most. When looking at this by quarter, we can see that the first quarter sales in 2017 and 2018 show the most variation, as indicated by the bar plot displaying the most comprehensive ranges. This can be partly explained by the upward trend in sales and the dataset range, which covers data from January 2017 to July 2018.

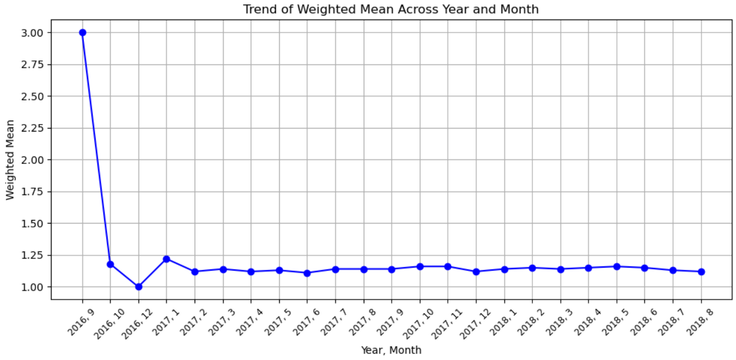


2. Order Size

The order size table indicates that approximately 90% of total orders consisted of single-product orders. The line chart illustrates that the average order size fluctuated between 1.00 and 1.25 during the 2017-2018 period, with no noticeable seasonal patterns in order size from month to month.

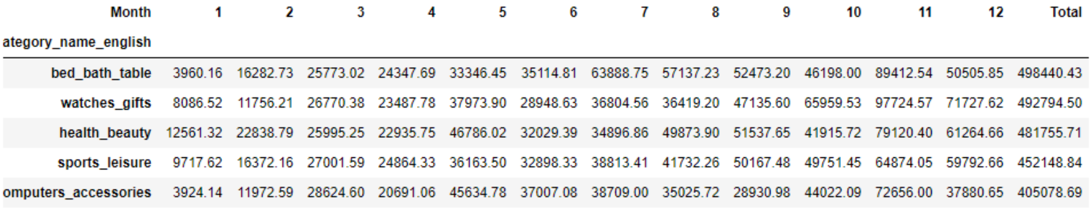


	order_size	count	percentage
0	1	86843	90.013267
1	2	7392	7.661850
2	3	1306	1.353676
3	4	495	0.513070
4	5	193	0.200046
5	6	191	0.197973

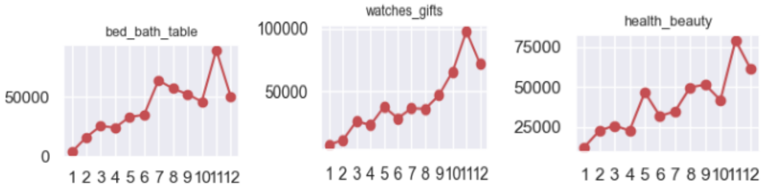


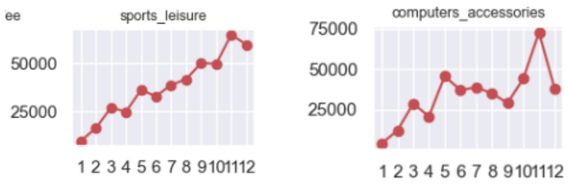
3. Top product category

The top five sales categories are bed bath table, watches gifts, health beauty, sports leisure, and computer accessories. These top five categories account for about 70% of total Olist eCommerce sales. Overall, the top five categories show a very similar trend over time; however, the bed bath table, health beauty, and computer accessories categories show a sudden sales increase in 2017 Q2.



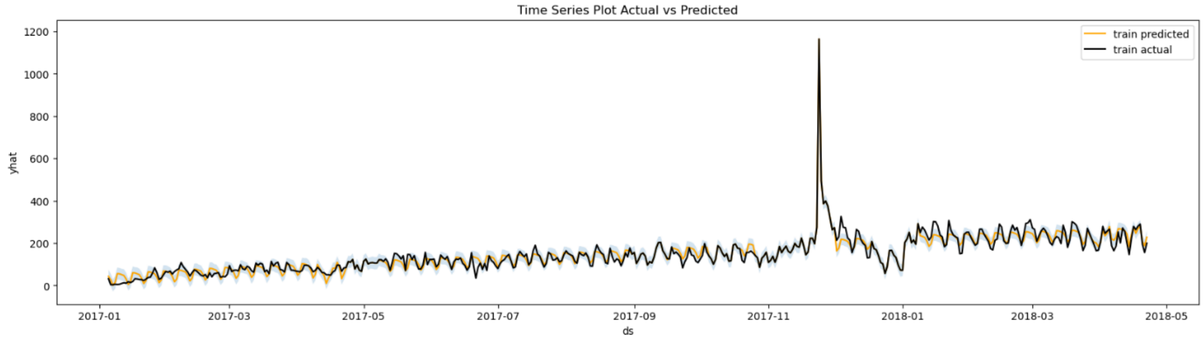
Month	1	2	3	4	5	6	7	8	9	10	11	12	Total
ategory_name_english													
bed_bath_table	3960.16	16282.73	25773.02	24347.69	33346.45	35114.81	63888.75	57137.23	52473.20	46198.00	89412.54	50505.85	498440.43
watches_gifts	8086.52	11756.21	26770.38	23487.78	37973.90	28948.63	36804.56	36419.20	47135.60	65959.53	97724.57	71727.62	492794.50
health_beauty	12561.32	22838.79	25995.25	22935.75	46786.02	32029.39	34896.86	49873.90	51537.65	41915.72	79120.40	61264.66	481755.71
sports_leisure	9717.62	16372.16	27001.59	24864.33	36163.50	32898.33	38813.41	41732.26	50167.48	49751.45	64874.05	59792.66	452148.84
omputers_accessories	3924.14	11972.59	28624.60	20691.06	45634.78	37007.08	38709.00	35025.72	28930.98	44022.09	72656.00	37880.65	405078.69

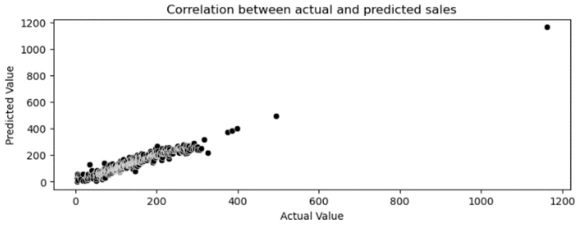


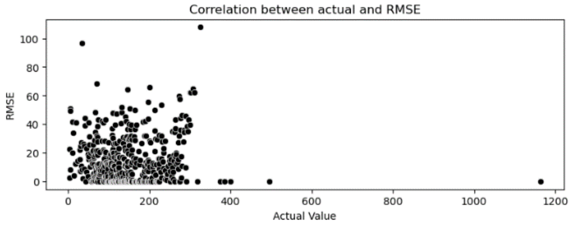


Model Selection

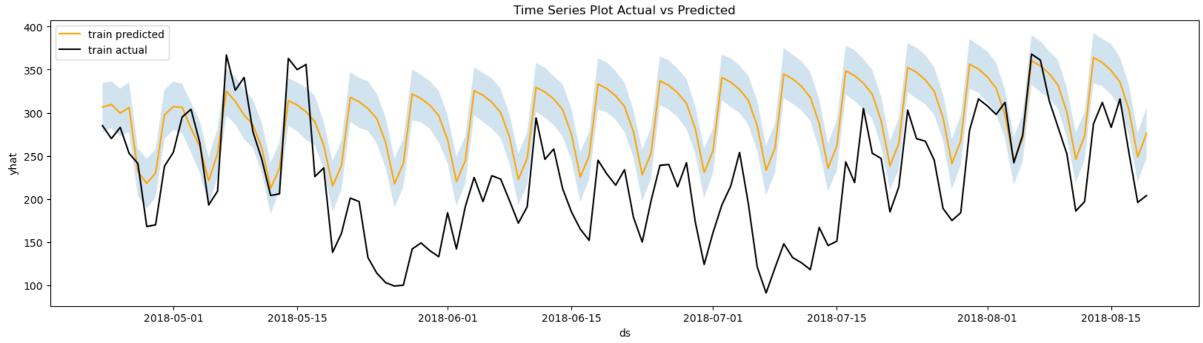
Our modeling revealed that optimal forecasting accuracy was achieved through a combination of **data scaling, hyperparameter tuning, and incorporating holiday contexts into the model**. Notably, the application of these techniques significantly enhanced the model's predictive capabilities, as evidenced by an impressively low Root Mean Square Error (RMSE) of 22.15 in our trend dataset. This indicates a strong correlation between our predicted sales and actual sales figures during the training phase, as illustrated in the accompanying graph. This finding underscores the efficacy of our chosen approach, demonstrating that by carefully calibrating our model and factoring in key variables such as holidays—which can significantly influence sales patterns—we can achieve a highly accurate forecast.

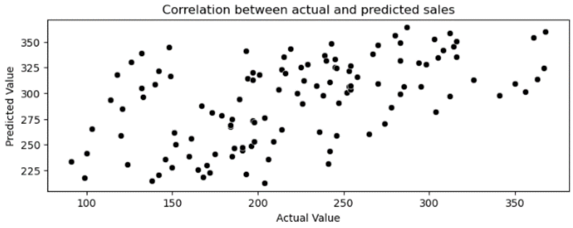


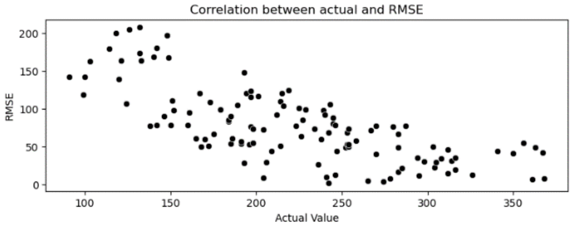




While our model demonstrated high accuracy for the train dataset, as highlighted by the RMSE evaluation scores, we encountered a challenge in maintaining this precision through Q2 of 2018 in the test dataset. During this period, the company faced unexpected sales drop not observed in 2017, with a significant drop in sales during Q2 of 2018. Consequently, this unforeseen fluctuation led to a diminished model accuracy for this specific timeframe, with the RMSE score for the period reaching 78.01. This score, when compared to the RMSE scores from other modeling efforts, which generally ranged between 75.00 and 85.00, indicates a discrepancy primarily attributed to the unique sales dip in 2018 Q2.

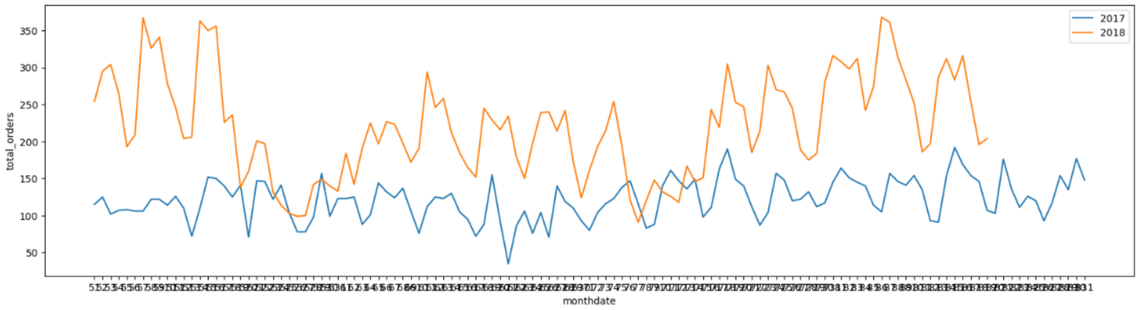






Further Research

The discrepancy observed in our final model's performance, particularly during Q2 of 2018, accentuates the need for ongoing model optimization to adeptly navigate and preempt unpredictable market dynamics, thereby achieving steadier accuracy across diverse periods. The unexpected sales dip observed in 2018's Q2 could potentially be elucidated by external influences, such as holiday seasons, which traditionally impact buying behaviors. This situation highlights the imperative to delve deeper into our dataset, identifying and incorporating additional features that significantly influence sales outcomes. Incorporating such variables into our model-building process is crucial, enabling a more comprehensive understanding and prediction of sales trends, ensuring the model's robustness and adaptability to the nuances of market fluctuations.


total_orders
monthdate