Data Glacier Data Scientist Internship

Batch: LISUM23: 30

Week10: Deliverables

Project: Retail Forecasting

Team member’s details:

Group Name: Retail_forecasting

Name	Richa Mishra	Shalu Kumar	Madoka Fujii	Shushun Ren
Email	mricha828@gmail.com	ss4676@njit.edu	mdkfji@gmail.com	shushunr@umich.edu
Country	United States	United States	United States	United States
College/Company	NJIT	NJIT	Sumitomo Mitsui Trust Bank	Umich
Specialization	Data Science	Data Science	Data Science	Data Science

Problem Description:

This major Australian beverage corporation operates within the beverage industry. Their product distribution spans across multiple supermarket chains, and they actively conduct robust promotional campaigns year-round. The demand for their products is subject to fluctuations driven by factors such as holidays and seasonal trends. They require a weekly item-level forecast for each of their products, categorized into weekly intervals.

Data Understanding:

The data describes Sales transitions from 2017-02-05 to 2020-12-27 that is before Covid-19 and after Covid-19 including any specific events with binary value. We can find its starting binary values of Covid_Flag from February 9, 2020. And Google Mobility starts recording from February 9, 2023 as well. There are 6 products and SKU1, SKU2, SUK3, SKU4, SKU5 have 0 sales between 11/22/2020 and 12/27/2020. On the other hand, during the time period, SKU6 does not have any records.

What type of data have you got for analysis?

No missing value (NA) in all variables.

-Product: string: 6 products: SKU6 has 198 counts while others have 204 counts.

-date: date: from 2017-02-05 to 2020-12-27

-Sales: integer

-Price Discount (%): double

-In-Store Promo: integer: dummy variable from one-hot encoding

-Catalogue Promo: integer: dummy variable from one-hot encoding

-Store End Promo: integer: dummy variable from one-hot encoding

-Google_Mobility: double: until Feb 2, 2023, it was not recorded. After that it started recording and was highly volatile.

-Covid_Flag: integer: dummy variable from one-hot encoding: From Feb 9, it became 1. Before that it was 0.

-V_DAY: integer: dummy variable from one-hot encoding

-EASTER: integer: dummy variable from one-hot encoding

-CHRISTMAS: integer: dummy variable from one-hot encoding

What are the problems in the data (number of NA values, outliers , skewed etc)?

Data Assessment Summary:

1. Zero Sales Observation:

- Based on our research, we have observed instances where sales data contains zero values for each product.
- We will investigate the context of these zero sales observations, as they could be legitimate data points, or they may require special handling.
- Understanding the reasons behind zero sales can help refine our analysis.

2. Outliers:

- Outliers have been identified in the data.
- We recognize that the presence of outliers can distort our analysis and results.

Proposed Approach:

1. Partitioning by Product:

- To gain a more granular understanding of the data and to address the unique sales characteristics of each product, we plan to partition the dataset by product category.
- This partitioning will enable us to perform data quality checks and analyses specific to each product, which can yield more meaningful insights.

2. Handling Missing Values:

- After partitioning the data by product, we will examine each product's dataset for missing values.
- Our goal is to implement data imputation techniques or strategies that are tailored to each product's sales behavior, thus mitigating the impact of missing data.

3. Outlier Treatment:

- For each product category, we will assess and address outliers individually.
- Techniques such as outlier removal, transformation, or the use of robust statistical methods will be employed to manage outliers effectively.
- Our aim is to ensure that our analysis and modeling are not unduly influenced by extreme data points.

4. Zero Sales Observation:

- Based on our research, we have observed instances where sales data contains zero values for each product.
- We will investigate the context of these zero sales observations, as they could be legitimate data points, or they may require special handling.

By adopting this systematic approach of partitioning the data by product, addressing missing values, managing outliers, and examining zero sales data, we aim to enhance the quality and relevance of our analysis, ultimately leading to more accurate and actionable insights.

Data Preprocessing:

1. Zero Values Removed:

   - Zero sales values have been successfully removed from the dataset.
   - This step helps ensure that our analysis focuses on meaningful sales data points.

2. Outliers Removed:

   - Outliers have been identified and removed from the dataset.

- The removal of outliers aids in creating a more robust and representative dataset for analysis.

By eliminating zero values and outliers, we are now working with a cleaner dataset that is better suited for our analytical goals.

What approaches are you trying to apply on your data set to overcome problems like NA value, outlier etc and why?

1. Partitioning by Product:

- To address the variability in sales for different products, we partitioned the dataset by product category.
- This partitioning allows for tailored analysis and treatment of issues specific to each product.

2. Handling Missing Values:

- Within each product category, we removed rows with missing sales values.
- This step ensures that our analysis focuses on complete and relevant sales data for each product.

3. Outlier Detection and Removal:

- Box plots were utilized to visualize the sales distribution for each product.
- The Interquartile Range (IQR) formula was applied to identify and remove outliers specific to each product category.
- Managing outliers ensures that our analysis and modeling are not unduly influenced by extreme data points.

4. Data Cleanliness Achieved:

- Following these steps, our dataset is now free of missing sales values and outliers.
- This cleanliness enhances the dataset's suitability for further analysis and modeling.

Next Steps:

- With clean data in hand, we can now explore further transformations and feature engineering to prepare the data for modeling.
- Additional steps might include normalization, encoding categorical variables, or creating new features to improve predictive performance.

Project life cycle along with deadline:

Project weeks	**Deadline**	**Lifecycle**
Week7	Aug 19, 2023	Problem statement, Pre-process
Week8	Aug 26, 2023	Data process, understanding
Week9	Sep 02, 2023	Data Cleaning, Merge, Review
Week10	**Sep 09, 2023**	**EDA, Final recommendation**
Week11	Sep 16, 2023	EDA presentation for business users
Week12	Sep 23, 2023	Model Selection and Model Building/Dashboard
Week13	Sep 30, 2023	Final Project Report and Code

Tabular data details: forecasting_case_study.xlsx:

Total number of observations	1218
Total number of files	1
Total number of features	12
Base format of the file	.xlsx
Size of the data	80KB