Retail Forecasting

Name: Olena Panchenko
Email: hel.panchenko@gmail.com
Country: United Kingdom
Specialisation: Data Science
Link: https://github.com/helenapanchenko/DS_Internship_Glacier/tree/master/Week8

Problem Description

The large company is in the beverages business in Australia. They sell their products through various supermarkets and engage in heavy promotions throughout the year. Various factors like holidays and seasonality also influence their demand. They needed a forecast of each of the products at the item level every week in weekly buckets.

Data Understanding

Dataset Overview

The dataset contains time series data exhibiting various patterns, including trends, seasonality, and some without either. At the time, the company relied on custom-built, in-house software for forecasting, but it frequently generated unrealistic predictions. They sought to explore the potential of AI/ML-based forecasting as a replacement for their local solution.

Data Description

The data is contained in the file forecasting_case_study.csv. The dataset files are written as comma-separated values files with a single header row. Columns that contain commas (,) are escaped using double-quotes ("). These files are encoded as UTF-8.

Forecasting Data File Structure (forecasting_case_study.csv):

Product, text - represents names of studied products.
date, date - represents the dates range from February 5, 201, to December 27, 2020, and includes every Sunday within that period.
Sales, numeric - represents the total amount of items sold during one week.
Price discount (%), numeric - represents the discount applied to the original price.
In-Store Promo, numeric - indicates whether the company is running a special offer, discount, or promotion at the physical store.
Catalogue Promo, numeric - indicates whether special offers or discounts are being promoted through catalogues.
Store End Promo, numeric - indicates whether there is a special promotion at the end of the store (could be clearance sales, discount racks, or a specific promotion near checkout areas).
Google_Mobility, numeric - indicates whether a significant movement or mobility is recorded (suggesting that people are more mobile and possibly visiting stores more frequently), or movement is below normal levels.
Covid_Flag, numeric - represents a measure of the change in mobility.
V_DAY, numeric - indicates whether it is a Valentine's Day or not.
EASTER, numeric - indicates whether it is an Easter Day or not.
CHRISTMAS, numeric - indicates whether it is a Christmas Day or not.

Data Quality Assessment

Evaluating and validating our data

1. Missing Values (MVs):
   - There are no missing values in the dataset, indicating completeness of the data.

2. Duplicate Rows:
   - There are no duplicate rows in the dataset, which ensures data integrity and avoids redundancy in the analysis.

3. Outliers:
   - The dataset has some outliers in the Sales column. Outliers in this context could indicate extreme scores that deviate significantly from the norm. Further investigation may be necessary to understand the nature of these outliers and determine whether they should be treated or excluded from the analysis.

Overall, the dataset is complete and free from duplicates or missing values, the presence of outliers may require careful consideration and preprocessing before conducting further analysis or modelling.

The extent of Data Quality Problems and Potential Impact on Analysis:

1. Missing Values:

The absence of missing values indicates that the dataset is complete and fully populated, reducing the need for data cleaning and ensuring more reliable analysis and modelling.

2. Duplicates:
- The absence of duplicate rows in all datasets indicates good data integrity, reducing the risk of biased analyses or misleading results.

3. Outliers:
- With over 1,000,000 outliers in the `relevance` column of `genome-scores.csv` and `rating` column of `ratings.csv`, there could be a significant impact on analysis:
    - Outliers in the `Sales` column might skew the measurement of the number of items the company should deliver to the store. They could also influence statistical analyses and modelling efforts, potentially leading to biased results or unreliable predictions.

Overall Impact:
- The quality issue identified in the dataset (outliers) has the potential to significantly impact the accuracy, reliability, and interpretability of analyses and models.
- Failure to address this issue adequately could lead to inaccurate insights and unreliable predictions, undermining the effectiveness and trustworthiness of any derived systems or findings.
- Therefore, thorough outlier detection is essential to mitigate these issues and ensure the integrity of subsequent analyses.

Data Preprocessing Requirements

After observing and analyzing the dataset for this project, I will need to perform data cleaning and feature engineering. The preprocessing tasks I will undertake include:

• If any outliers are detected that could negatively impact my work, I will address them promptly. After assessing their impact on the data and project requirements, I will handle them by substituting outliers with the average or median values.

Exploratory Data Analysis (EDA) Plan

1. Data Overview:
• I'll begin by loading the dataset and examining its structure i.e. the number of rows, columns, and data types.
• I'll check for missing values and duplicates in each column of relevant dataframes and devise a strategy for handling them.

2. Univariate Analysis:
• For numeric data:
a. I'll perform summary statistics: by checking the values of the mean, median, mode, range, variance, and standard deviation.
b. I'll visualise distributions using histograms, box plots, and kernel density plots.

• For date data:
a. I'll examine the distribution of dates and any temporal patterns by plotting time series graphs.
b. I'll analyze trends over time, seasonality, and periodicity using line charts and seasonality decomposition techniques.

3. Bivariate Analysis:
- Here, I'll explore relationships between pairs of variables:
• For numeric-numeric pairs:
a. I'll calculate correlation coefficients (Pearson, Spearman) and I'll perform visualisations using scatter plots.
b. I'll use heatmaps to visualise correlation matrices.
• For numeric-date pairs:
a. I'll analyze trends and patterns over time using line charts to see how numeric variables change with different dates.
b. I'll investigate any temporal effects or seasonality by plotting numeric variables against time to uncover periodic trends or anomalies.

4. Data Visualization:
Here I'm going to use interactive visualisations to explore complex relationships and patterns in the data:
• I'm going to utilise libraries like matplotlib, and Seaborn for standard visualisations and plotly or bokeh for interactive plots.

5. Documentation and Reporting:
• I'll document findings, insights, and decisions made during the EDA process.

6. Iterative Process:
• EDA is an iterative process, so I'll revisit earlier steps as needed based on insights gained later in the analysis.
• I'll continuously refine analysis techniques and explore alternative visualisations to gain a deeper understanding of the data.

PROPOSED EDA AND VISUALIZATION TECHNIQUES

I'll be utilising the following techniques and libraries for our EDA and Visualization for general and case-specific analysis:

• General Exploratory Data Analysis:
a. I'll use the isna().sum() to check for missing values.
b. I'll use a duplicated() function to find duplicates.
c. I'll check for unique values using the nunique() function.

• Univariate Analysis:
a. I'll conduct summary statistics using the describe() function.
b. I'll use visualisations like boxplots, histograms, etc to check for the distribution of our data and to also check for outliers.

c. I'll also create functions for detecting outliers and handling them
d. I'll explore frequency distributions, calculate proportions and percentages using barplots, pie charts, etc

• Bivariate Analysis:
a. I'll explore relationships between variables using scatter plots
b. I'll create correlation matrices and use heatmaps to visualise them
c. I'll compare the distribution of numeric variables across different categories using box plots or violin plots.
d. I'll create pair plots to visualise relationships between two features.

• Data Visualization:
For our data visualisation, I'm going to use the following libraries;
matplotlib, seaborn, plotly and bokeh to create visually appealing and interactive visualisations.