CAPSTONE PROJECT

Walmart

TABLE OF CONTENT

Problem Statement

- A nationwide retail chain with multiple outlets is grappling with a financial crisis and inventory management issues.
- The primary challenge lies in aligning demand with supply, impacting operational efficiency.
- Geographic diversity further complicates demand-supply dynamics.
- The company seeks data-driven solutions to optimise inventory, considering technology adoption.
- Inefficient management incurs higher operational costs and poses risks of revenue loss, highlighting the strategic importance of effective inventory management.

Problem Statement

Exploratory Data Analysis:

- Conduct comprehensive exploratory data analysis (EDA) on weekly sales data from multiple outlets.
- Analyse statistical measures, identify outliers, and handle missing values.
- Evaluate the impact of unemployment rates on weekly sales, focusing on stores facing significant challenges.
- Detect and interpret seasonal trends in sales, determining their timing and underlying causes.
- Investigate the relationship between temperature and weekly sales.
- Assess how the Consumer Price Index (CPI) affects sales performance across different stores.
- Identify top-performing and worst-performing stores based on historical data.
- Quantify the performance gap between the highest and lowest performers.

Forecast Sales

- Utilise predictive modelling techniques to forecast sales for each store over the next 12 weeks.

Data Description

Features Overview:

The provided data appears to be a tabular dataset with several columns. Here's an explanation of each column:

- Store: This feature represents the unique identifier for each store in the dataset. Each store is assigned a distinct number for identification purposes.

- Date: Date indicates the specific date of the weekly purchase recorded in the dataset. It allows tracking sales over time and identifying any temporal patterns or trends. Format appear to be DD-MM-YYYY (day-month-year)

- Weekly_Sales: Weekly_Sales denotes the total sales made during a particular week. It serves as the target variable in predictive modeling tasks, aiming to predict future sales based on other features.

- Holiday_Flag: Holiday_Flag is a binary indicator representing whether a particular week includes a holiday or not, which has an influence on the purchase. It takes the value 1 if there is a holiday during that week and 0 otherwise.

- Temperature: Temperature denotes the temperature recorded in the region. It provides information about weather conditions, which may influence consumer behaviour and affect sales.

- Fuel_Price: Fuel_Price indicates the price of fuel in the region where each store operates. Changes in fuel prices can impact transportation costs and ultimately affect retail sales.

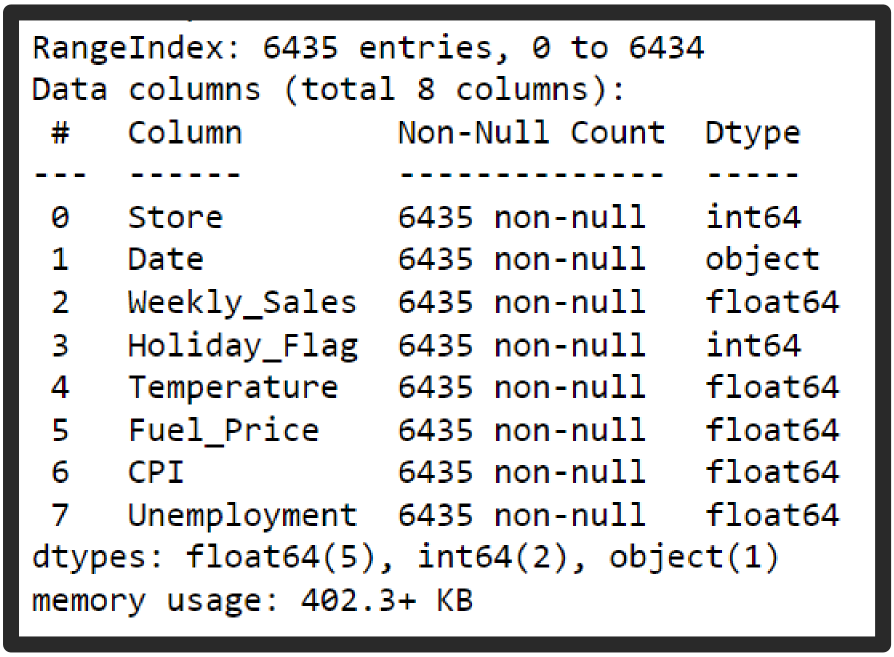
- CPI (Customer Price Index): CPI reflects the Customer Price Index, which measures the average change over time in the prices paid by consumers for a basket of goods and services. It helps in understanding the purchasing power of consumers in the region.

- Unemployment: Unemployment represents the unemployment rate in the region where each store is situated. High unemployment rates may lead to lower consumer spending and subsequently impact sales performance.

Each row in the dataset represents a specific week's sales data for a particular store, including information such as date, sales figures, holiday status, temperature, fuel price, CPI, and unemployment rate. The goal is to analyse the factors influencing weekly sales and make predictions/ forecast based on the patterns observed in the data.

Metadata information:

- ✓ There are no null records in data.
- ✓ Total records are 6435 and number of fields are 8.
- ✓ Have to format the Date column form object to Datetime.



```
RangeIndex: 6435 entries, 0 to 6434
Data columns (total 8 columns):
 #   Column        Non-Null Count  Dtype  
---  ------        --------------  -----  
 0   Store         6435 non-null   int64  
 1   Date          6435 non-null   object 
 2   Weekly_Sales  6435 non-null   float64
 3   Holiday_Flag  6435 non-null   int64  
 4   Temperature   6435 non-null   float64
 5   Fuel_Price    6435 non-null   float64
 6   CPI           6435 non-null   float64
 7   Unemployment  6435 non-null   float64
dtypes: float64(5), int64(2), object(1)
memory usage: 402.3+ KB
```

Data Preprocessing Steps And Inspiration

The preprocessing of the data included the following steps:

Library Imports:

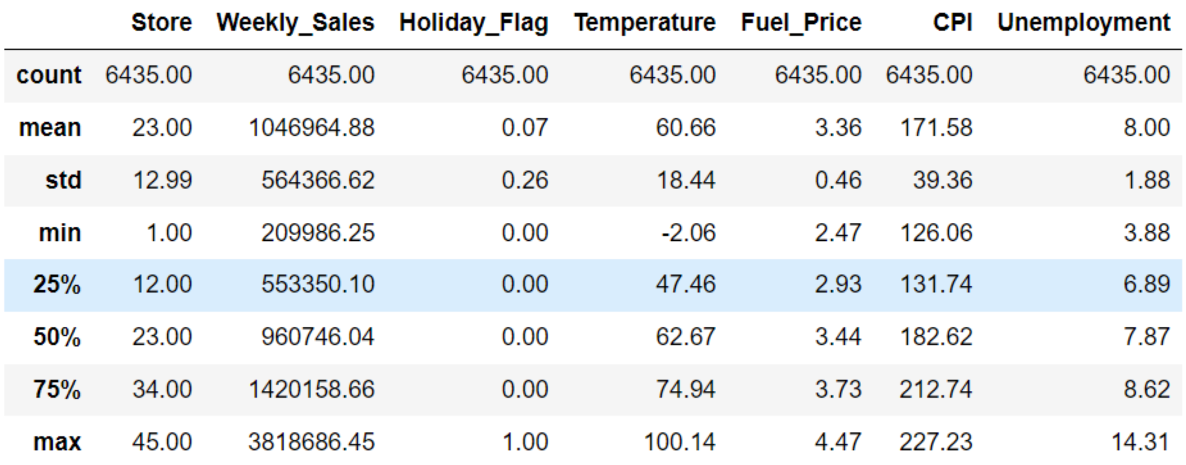
- ➢ The required libraries, including pandas, numpy, matplotlib, statsmodels, and the specific ARIMA-related functions, are imported.

Reading Data and Understanding:

- ➢ The dataset is read from the 'Walmart.csv' file into a Pandas DataFrame (data).
- ➢ An overview of the dataset's columns and the first few rows is displayed.

Exploratory Data Analysis (EDA):

- ➢ Basic information about the data is displayed using data.info() and data.shape.
- ➢ The presence of null values and duplicate rows is checked.
- ➢ Initial visualizations are created, including bar plots and a heatmap of correlations.



	Store	Weekly_Sales	Holiday_Flag	Temperature	Fuel_Price	CPI	Unemployment
count	6435.00	6435.00	6435.00	6435.00	6435.00	6435.00	6435.00
mean	23.00	1046964.88	0.07	60.66	3.36	171.58	8.00
std	12.99	564366.62	0.26	18.44	0.46	39.36	1.88
min	1.00	209986.25	0.00	-2.06	2.47	126.06	3.88
25%	12.00	553350.10	0.00	47.46	2.93	131.74	6.89
50%	23.00	960746.04	0.00	62.67	3.44	182.62	7.87
75%	34.00	1420158.66	0.00	74.94	3.73	212.74	8.62
max	45.00	3818686.45	1.00	100.14	4.47	227.23	14.31

The provided table is a summary of descriptive statistics for each numerical column in the dataset.

Impact of Sales on Other Parameters:

- Bar plots are created to visualize the impact of holidays and temperature on weekly sales.

Seasonal Trend Analysis:

- Time series decomposition is performed using the seasonal_decompose function to identify seasonal trends.
- The ADF (Augmented Dickey-Fuller) test is conducted to check the stationarity of the time series.

Correlation Analysis:

- Correlation analysis is performed between temperature and weekly sales using the Pearson correlation coefficient.

Top Performing Stores:

- The top-performing stores are identified based on total sales.
- Bar plots are created to visualize total sales by store.

Worst Performing Store and Difference Analysis:

- The worst-performing store is identified based on total sales.
- The difference in total sales between the best and worst-performing stores is calculated.
- Predictive Modelling for Sales Forecast:

Algorithm Selection for the Project

Exploratory Data Analysis (EDA):

Common statistical and visualization tools like pandas, matplotlib, and seaborn for initial data exploration and visualization.

These libraries provide descriptive statistics and graphical representations, aiding in understanding data patterns.

Time Series Analysis and Forecasting:

Employ the ARIMA (AutoRegressive Integrated Moving Average) model for time series analysis and short-term forecasting in the provided code.

ARIMA effectively captures temporal patterns, but other models like SARIMA, Prophet, or LSTM can be considered based on data complexity and accuracy requirements.

Correlation Analysis:

Use the Pearson correlation coefficient to measure linear correlation between variables, suitable for numerical variable relationships.

Consider alternative methods like Spearman's rank correlation for non-linear relationships or specific algorithms for feature selection if necessary.

Top Performing and Worst Performing Stores:

Basic statistical measures and Python functions (e.g., numpy and pandas) suffice for identifying top and worst-performing stores based on total sales.

Machine Learning (ML) Algorithms:

For predictive modelling of weekly sales using various features, consider machine learning algorithms such as linear regression, random forests, etc.

Enhance model performance through feature engineering, selection techniques, and hyperparameter tuning.

Motivation and Reasons for Algorithm Selection

Time Series Analysis with ARIMA:

Motivation:

The core objective is to analyse and forecast weekly sales data, inherently exhibiting time-dependent patterns.

Reasons:

ARIMA stands as a robust time series forecasting model adept at capturing temporal dependencies and trends.

It accommodates seasonality, a common characteristic in retail sales data, making it an apt choice.

ARIMA offers a straightforward yet effective method for short-term sales predictions.

Correlation Analysis:

Motivation:

The aim is to comprehend relationships between variables, such as temperature's correlation with weekly sales.

Reasons:

Employing the Pearson correlation coefficient facilitates quantifying linear relationships accurately.

It provides a precise measure of correlation strength and direction, aiding in insightful analysis.

Impact Analysis on Sales:

Motivation:

Unraveling the impact of external factors (e.g., unemployment rate, CPI, temperature) on weekly sales.

Reasons:

Provide a quantitative assessment of each variable's influence on sales.

Interpretation of coefficients aids in understanding the direction and magnitude of impact.

Feature Engineering and Selection:

Motivation:

Enhancing machine learning model performance by crafting meaningful features and selecting the most pertinent ones.

Reasons:

Feature engineering uncovers novel insights from existing data, bolstering model efficacy.

Feature selection mitigates overfitting and focuses on pivotal variables, refining predictive accuracy.

Assumptions

Stationarity:

Time series data is assumed stationary for simpler modelling and forecasting, as in ARIMA.

Linear Relationships:

Regression models assume linear relationships between variables, though nonlinear relationships may exist.

Independence of Observations:

Crucial assumption for statistical tests and modelling, often violated in time series data due to temporal dependencies.

Normality of Residuals:

Residuals in regression models are assumed normally distributed for hypothesis testing and confidence intervals.

No Perfect Multicollinearity:

Independent variables in regression are assumed not perfectly correlated to prevent unstable coefficient estimates.

Homoscedasticity:

Residuals exhibit constant variance across independent variable levels in regression models, avoiding biased standard errors.

Feature Linearity:

Machine learning models assume linear relationships between features and the target variable, though non-linear relationships may require complex models.

Absence of Outliers:

No significant outliers assumed to avoid disproportionate influence on model parameters.

Normality of Data:

Data assumed normally distributed for certain statistical tests, deviations may be acceptable in large samples due to the Central Limit Theorem.

Model Evaluation and Techniques

Exploratory Data Analysis (EDA):

Methods like shape, info(), isnull().sum(), and duplicated().sum() to assess dataset characteristics such as size, information, missing records, and duplicates.

Employs visualizations like bar plots and scatter plots to explore relationships between variables, like weekly sales and holiday flags, temperature, and store-wise sales during holidays.

Impact of Sales on Other Parameters:

Analyses sales impact on variables like holidays and temperature using bar plots to visualize average weekly sales during holidays and different temperature ranges.

Correlation and Statistical Tests:

Calculates and visualises the correlation matrix between variables using a heatmap.

Conducts statistical tests like the Augmented Dickey-Fuller test to assess stationarity in time series data.

Time Series Decomposition:

Utilizes seasonal_decompose function from statsmodels to conduct seasonal decomposition of time series data.

Visualizes resulting components (trend, seasonality, residual) to identify patterns.

Store-wise Analysis:

Analyzes correlation between weekly sales and unemployment rates for each store.

Top Performing Stores:

Identifies top-performing stores based on total sales and presents results using bar graphs.

Worst Performing Store:

Identifies worst-performing store based on total sales and calculates performance gap between highest and lowest performers.

Predictive Modeling (ARIMA):

Fits ARIMA model to historical sales data for each store.

Forecasts sales for next 12 weeks and visualizes results with confidence intervals.

Inferences for the Same

Based on the analysis, the insights which can be used by each of the stores to improve

- ➔ Seasonal Sales Promotions: Utilize holiday flags to plan and implement targeted sales promotions during peak shopping seasons, leveraging increased foot traffic and consumer spending patterns.

- ➔ Temperature-Dependent Product Assortment: Adjust product offerings based on temperature fluctuations to meet changing consumer preferences and demand for seasonal items, such as clothing, beverages, and outdoor equipment.

- ➔ Customer Price Index (CPI) Alignment: Align pricing strategies with changes in the Consumer Price Index to ensure competitive pricing while maintaining profit margins and customer satisfaction.

- ➔ Unemployment Rate Consideration: Tailor marketing campaigns and promotional activities to appeal to consumers affected by changes in the unemployment rate, such as offering value-oriented products or services and flexible payment options.

- ➔ Operational Efficiency Optimization: Streamline operational processes and optimize staffing levels to improve efficiency and reduce costs, ensuring smooth store operations and enhancing the overall customer experience.

- ➔ Inventory Management Enhancements: Implement advanced inventory management systems to optimize stock levels, reduce out-of-stock situations, minimize excess inventory, and improve overall inventory turnover rates.

- ➔ Customer Relationship Management (CRM): Implement CRM systems to collect and analyze customer data, enabling personalized marketing strategies, targeted promotions, and enhanced customer engagement and loyalty.

- ➔ Data-Driven Decision Making: Embrace data-driven decision-making processes by regularly analyzing sales data, customer feedback, and market trends to identify opportunities for improvement, optimize strategies, and drive informed business decisions.

- ➔ Performance monitoring: Establish key performance indicators (KPIs) to track poorly performing stores progress over time. Regularly review performance metrics and adjust strategies as needed to ensure continuous improvement and alignment with organizational goals

- ➔ Predictive Modeling (ARIMA): Employs ARIMA modeling to forecast sales for the next 12 weeks for each store. The forecasted values and confidence intervals are plotted, providing insights into the expected future sales trends.

➔ Conclusion: Overall, the analysis provides valuable insights into the factors influencing weekly sales, seasonal patterns, and the performance of individual stores. The use of statistical tests, correlation analysis, and time series modeling enhances the understanding of the dataset.

Future Possibilities

Feature Engineering: Explore additional sales-influencing features like promotional events or marketing campaigns for improved model accuracy.

Machine Learning Models: Experiment with advanced models like Random Forests or Neural Networks for capturing complex data relationships and enhancing forecasting.

Hyperparameter Tuning: Optimize ARIMA or ML model hyperparameters using grid or random search methods for better performance.

Ensemble Methods: Implement ensemble techniques to combine ARIMA predictions with other models for more robust forecasts.

Dynamic Forecasting: Develop a system for dynamic forecasting that continuously updates predictions with new data.

Cross-Validation: Employ cross-validation techniques to assess model robustness and generalizability.

Interactive Dashboard: Build an interactive dashboard with Plotly or Dash for dynamic sales data visualization.

Anomaly Detection: Integrate anomaly detection algorithms to identify unusual sales patterns or outliers.

External Data Sources: Incorporate economic indicators or social media trends for broader market insights.

Deployment: Deploy the forecasting model as a web service or integrate it into a business intelligence platform for real-time predictions.

User Feedback Integration: Gather user feedback to enhance the model based on practical insights and experiences.

Continuous Monitoring: Establish a monitoring system to track model performance and detect significant deviations.

Documentation and Knowledge Sharing: Thoroughly document project steps and insights for stakeholders and the data science community.

Scalability: Design the solution to handle growing data volumes, possibly utilizing distributed computing frameworks.

Conclusion

In conclusion, the Walmart Capstone Project entailed a comprehensive analysis of sales data, uncovering significant insights such as increased sales during holidays and the influence of temperature on sales variation. It successfully identified both top and worst-performing stores and utilized an ARIMA model for sales forecasting. Future directions include leveraging advanced modeling techniques and exploring additional features to further enhance forecasting accuracy. The project underscores the necessity of ongoing monitoring and provides actionable insights crucial for strategic decision-making within the retail sector.

Reference

- ➢ Libraries: Imported necessary Python libraries - pandas, matplotlib.pyplot, seaborn, numpy, and statsmodels.
- ➢ Dataset: Read 'Walmart.csv' dataset using pd.read_csv().
- ➢ Data Overview: Displayed the first few rows of the dataset with data.head() to provide a snapshot of columns and values.
- ➢ Exploratory Data Analysis (EDA): Utilized data.info(), data.shape, data.isnull().sum(), and data.duplicated().sum() for data overview, null checks, and duplicate row detection.
- ➢ Time Series Analysis: Conducted time series decomposition using seasonal decomposition (STL) and tested stationarity with the Augmented Dickey-Fuller test (adfuller).
- ➢ Store-Specific Analysis: Created bar plots to visualize holiday impact on weekly sales per store and calculated unemployment-sales correlation for each store.
- ➢ Time Series Forecasting (ARIMA): Applied ARIMA models to forecast sales for each store over the next 12 weeks and plotted forecasted values with confidence intervals.