**Capstone Project:**

**Sales Forecasting**

**Problem Statement:**

A Retail Store with multiple outlets nationwide.

They need help in managing the inventory - to match the supply-demand.

**Dataset Information:**

You are provided with the Weekly\_Sales data for their various outlets.

The **Walmart.csv** contains 6435 rows and 8 columns.

| **Feature Name** | **Description** |
| --- | --- |
| **Store** | **Store number** |
| **Date** | **Week of Sales** |
| **Weekly\_Sales** | **Sales for the given store in that week** |
| **Holiday\_Flag** | **Indicates if it is a holiday week** |
| **Temperature** | **Temperature on the day of the sale** |
| **Fuel\_Price** | **Cost of fuel in the region** |
| **CPI** | **Consumer Price Index** |
| **Unemployment** | **Unemployment Rate** |

**Objectives:**

1. You are provided with the **Weekly\_Sales** data for their various outlets.

* Use
  + **Handle the missing values**
  + **Exploratory Data Analysis (EDA)**
  + **Outlier Analysis**
  + **Statistical Analysis**
* To come up with **various insights** that can give them a clear perspective on the following:
  1. If the **Weekly\_Sales** are affected by the **Unemployment** Rate, and
     + If **YES** - which **Store**s are suffering the most?
  2. If the **Weekly\_Sales** show a **seasonal trend**,
     + When and what could be the reason?
  3. Does **Temperature** affect the **Weekly\_Sales** in any manner?
  4. How is the **Consumer Price Index** (**CPI**) affecting the **Weekly\_Sales** of various stores?
  5. **TOP** Performing **Store**s according to the historical data.
  6. The **WORST** performing **Store**, and
     + How significant is the difference between the **HIGHEST** and **LOWEST**-Performing **Store**s?

1. Use predictive modelling techniques to forecast the sales for each **Store** for the next 12 weeks.

**1. Introduction:**

- Start with a brief introduction that explains the problem statement and provides an overview of the project goals.

- Define the project's objective to analyze the weekly sales data and forecast future sales for each store.

**2. Data Understanding:**

- Describe the dataset provided, including the number of rows and columns.

- Provide a table that explains the features (columns) in the dataset, including their descriptions.

**3. Data Preprocessing:**

- Perform exploratory data analysis (EDA) to gain insights into the data.

- Conduct outlier analysis to identify and handle any outliers in the dataset.

- Handle missing values using appropriate techniques such as imputation or deletion.

**4. Analysis and Insights:**

a. Impact of Unemployment Rate on Weekly Sales:

- Use statistical analysis to determine if there is a correlation between the unemployment rate and weekly sales.

- Identify stores that are most affected by the unemployment rate.

**b. Seasonal Trend in Weekly Sales:**

- Analyze the weekly sales data to identify if there is a seasonal trend.

- Determine the timing and reasons behind any observed seasonal variation.

**c. Impact of Temperature on Weekly Sales:**

- Investigate the relationship between temperature and weekly sales.

- Determine if there is a significant impact of temperature on sales.

**d. Influence of Consumer Price Index (CPI) on Weekly Sales:**

- Analyze the effect of the Consumer Price Index on the weekly sales of different stores.

- Determine the relationship between CPI and sales.

**e. Top Performing Stores:**

- Identify the top-performing stores based on historical data.

- Provide insights into the factors that contribute to their success.

**f. Worst Performing Store and Performance Comparison:**

- Identify the worst-performing store based on historical data.

- Calculate and discuss the significance of the difference between the highest and lowest-performing stores.

**5. Predictive Modeling:**

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

- Explain the methodology used for the forecasting and discuss the accuracy of the predictions.

**6. Conclusion:**

- Summarize the key findings and insights from the analysis.

- Highlight any recommendations or actions that can be taken based on the analysis.

- Emphasize the value of the project in helping the retail store manage inventory effectively.

**7. Project Deliverables:**

- Compile all the analysis, insights, and predictions into a comprehensive report document.

- Include all the code used for data preprocessing, analysis, and modelling in a separate script file.

- Create a zip file containing the report document, script file, and any additional supporting files or resources.

By following these steps, you will be able to create a professional project solution and report file for your Walmart sales forecasting project.

1. Data Cleaning and Preparation:

- Check for any duplicate rows or columns in the dataset and remove them if necessary.

- Convert the "Date" column to a proper date format for easier analysis.

- Handle any missing values in the dataset using appropriate techniques such as mean imputation or deletion.

- Perform outlier analysis to identify and handle any extreme values in the dataset.

2. Exploratory Data Analysis (EDA):

- Calculate summary statistics for each numerical feature in the dataset.

- Plot histograms, box plots, and scatter plots to visualize the distributions and relationships between variables.

- Analyze the distribution of weekly sales, unemployment rate, temperature, CPI, and other relevant features.

- Identify any patterns or anomalies in the data that may require further investigation.

3. Statistical Analysis:

- Conduct correlation analysis to determine the relationship between the weekly sales and the unemployment rate.

- Identify the stores that are most affected by the unemployment rate using statistical techniques such as hypothesis testing.

- Perform linear regression analysis to assess the impact of other variables such as temperature and CPI on weekly sales.

4. Seasonal Trend Analysis:

- Use time series analysis techniques like decomposition and autocorrelation to identify seasonal patterns in the weekly sales data.

- Determine the timing and reasons behind any observed seasonal variation, such as holidays or promotions.

5. Impact of Temperature and CPI on Weekly Sales:

- Perform statistical tests (e.g., t-tests or chi-square tests) to determine if temperature or CPI significantly affects weekly sales.

- Calculate correlation coefficients and p-values to quantify the strength of the relationships.

- Use visualization techniques like scatter plots or line plots to illustrate the relationships between temperature, CPI, and sales.

6. Identifying Top and Worst Performing Stores:

- Aggregate the weekly sales data by store and calculate the total sales for each store.

- Rank the stores based on their historical sales data to identify the top-performing stores.

- Similarly, identify the worst-performing store based on historical data.

- Calculate the difference in sales between the highest and lowest-performing stores and assess its significance.

7. Predictive Modeling:

- Split the dataset into training and testing sets for modelling.

- Choose an appropriate predictive modelling technique, such as linear regression or time series forecasting methods.

- Train the model using the training set and evaluate its performance on the testing set.

- Use the trained model to forecast sales for each store for the next 12 weeks.

8. Reporting:

- Compile all the analysis, insights, and predictions into a comprehensive report document.

- Include appropriate visualizations, tables, and charts to support your findings.

- Summarize the key findings, insights, and recommendations in a clear and concise manner.

- Provide a step-by-step explanation of the methodology used and the code/scripts written for the analysis and modelling.

- Create a zip file containing the report document, script file, and any additional supporting files or resources to submit to your client.

By following this detailed plan and roadmap, you should be able to complete your Walmart sales forecasting project successfully.

1. Set up Google Colab:

- Go to the Google Colab website (https://colab.research.google.com).

- Sign in with your Google account or create a new one.

- Create a new notebook by clicking on "New Notebook" or import your existing notebook.

2. Import and Load the Dataset:

- Upload the "walmart.csv" dataset to your Google Drive or directly to the Colab notebook.

- Mount your Google Drive to access the dataset if it's uploaded there. Otherwise, load the dataset directly from your local files.

3. Data Cleaning and Preparation:

- Write code in your Colab notebook to clean and preprocess the dataset.

- Handle duplicate rows or columns, format the date column, and handle missing values.

- Perform outlier analysis and remove any extreme values.

4. Exploratory Data Analysis and Statistical Analysis:

- Use Python libraries like Pandas, NumPy, and Matplotlib/Seaborn to conduct exploratory data analysis.

- Plot histograms, box plots, scatter plots, and other relevant visualizations.

- Calculate summary statistics and perform statistical tests or correlations.

5. Seasonal Trend Analysis and Impact Analysis:

- Use time series analysis techniques and statistical tests to identify seasonal trends.

- Analyze the impact of variables like unemployment rate, temperature, and CPI on weekly sales.

- Visualize relationships using appropriate plots and interpret the results.

6. Identifying Top and Worst Performing Stores:

- Aggregate the weekly sales data by store and calculate total sales.

- Rank the stores based on their historical sales data to identify the top-performing stores.

- Similarly, identify the worst-performing store based on historical data.

7. Predictive Modeling:

- Split the dataset into training and testing sets.

- Choose a predictive modelling technique like linear regression or time series forecasting.

- Implement and train the model in Colab using Python libraries like scikit-learn or stats models.

- Evaluate the model's performance and make predictions for the next 12 weeks.

8. Reporting:

- Create a markdown cell in your Colab notebook and write the report, including analysis, insights, and recommendations.

- Insert appropriate visualizations, tables, and code snippets to support your findings.

- Ensure your report is clear, concise, and well-structured.

- Download your Colab notebook as an HTML or PDF file to include in your final submission.

Alternative: If you are unable to use Google Colab's free version, you can consider using other cloud-based alternatives like Kaggle Kernels or IBM Watson Studio. These platforms provide similar functionality for data analysis and provide computing resources for running your code. Additionally, you can also set up a local development environment using Jupyter Notebook or any other Python IDE on your computer.

Remember to regularly save your work and back up your notebook to avoid losing any progress.

To build a project solution for the **Walmart sales forecasting project**, you can follow the below steps:

**1. Import and load the dataset:**

Read the "walmart.csv" dataset into your preferred programming language (e.g., Python) and load it into a pandas DataFrame for further analysis.

**2. Perform statistical analysis:**

Calculate measures of central tendency, dispersion, and correlation between variables to gain initial insights into the dataset.

Use descriptive statistics functions and visualization techniques to identify any trends or relationships.

**3. Exploratory Data Analysis (EDA):**

**Explore the dataset's**

characteristics,

distributions, and

relationships between variables.

Use various **data visualization techniques** such as

histograms,

scatter plots,

box plots, etc.,

To **understand and analyze** the data better.

**4. Outlier analysis:**

Identify and handle any outliers in the dataset.

Remove or treat the outliers based on their impact on the analysis and the domain knowledge.

**5. Handle missing values:**

Check for missing values in the dataset and handle them appropriately.

This can include imputing missing values or removing rows or columns based on the analysis requirements.

## **6. Answer specific objectives:**

#A.

1. **Determine** if the weekly sales are affected by the unemployment rate
   1. Explore the **correlation between** weekly sales and the unemployment rate.
2. **Identify** which stores are suffering the most:
   1. Identify stores with the highest negative correlation.

#B.

1. **Analyze if the weekly sales show a seasonal trend and identify when and what could be the reasons:**

Analyze the pattern of sales over time, identify any seasonality using time series analysis techniques, and investigate the potential reasons for the observed trends.

c. Investigate if temperature has any impact on the weekly sales:

Analyze the relationship between temperature and weekly sales. Use statistical techniques such as correlation analysis to determine the impact of temperature on sales.

d. Examine how the Consumer Price Index (CPI) affects the weekly sales of various stores:

Analyze the relationship between CPI and weekly sales. Determine the impact of changes in the CPI on sales.

e. Identify the top-performing stores based on historical data:

Analyze and compare the sales performance of different stores over time. Identify the stores with consistently high sales.

f. Determine the worst-performing store and assess the significance of the difference between the highest and lowest-performing stores:

Analyze and compare the sales performance of different stores over time. Identify the store with consistently low sales and assess the significance of the difference compared to the highest-performing store.

7. Predictive modeling for sales forecasting: Utilize predictive modeling techniques (e.g., Time Series Forecasting, Regression Analysis, etc.) to develop a model that can forecast sales for each store for the next 12 weeks. Train the model using historical sales data and evaluate its performance using appropriate evaluation metrics.

8. Prepare the report file: Document your analysis, methodology, findings, and insights in a professional report format. Include relevant visualizations, tables, and explanations for each objective. Organize the report logically and ensure that it is easy to follow. Compress all the necessary files (code, datasets, report) into a zip file before submitting it to your client.

By following these steps, you will be able to build a comprehensive project solution for the Walmart sales forecasting project and prepare a professional report that meets the client's requirements.

\*\*Objective 1: Analyzing the Impact of Variables on Weekly Sales\*\*

- a. Impact of Unemployment Rate:

- Checklist:

- Calculate the correlation between Unemployment and Weekly Sales.

- Identify the stores with the highest negative and positive correlation.

- Expected Output:

- Correlation coefficient between Unemployment and Weekly Sales.

- Stores suffering the most (highest negative correlation) and prospering the most (highest positive correlation) due to Unemployment.

- b. Seasonal Trend in Weekly Sales:

- Checklist:

- Explore the time series pattern of Weekly Sales to identify any seasonal trend.

- Use suitable techniques such as decomposition or autocorrelation to analyze the data.

- Expected Output:

- Time-series plots showing the trend, seasonality, and residuals.

- Conclusion on whether there is a significant seasonal trend and the reason behind it.

- c. Impact of Temperature on Weekly Sales:

- Checklist:

- Calculate the correlation between Temperature and Weekly Sales.

- Analyze the relationship between Temperature and Weekly Sales using scatter plots.

- Expected Output:

- Correlation coefficient between Temperature and Weekly Sales.

- Scatter plots showing the relationship between Temperature and Weekly Sales.

- d. Impact of Consumer Price Index (CPI) on Weekly Sales:

- Checklist:

- Calculate the correlation between CPI and Weekly Sales.

- Analyze the relationship between CPI and Weekly Sales using scatter plots.

- Expected Output:

- Correlation coefficient between CPI and Weekly Sales.

- Scatter plots showing the relationship between CPI and Weekly Sales.

- e. Top Performing Stores:

- Checklist:

- Calculate the total sales for each store.

- Identify the stores with the highest total sales.

- Expected Output:

- List of top-performing stores based on historical data.

- f. Worst Performing Store:

- Checklist:

- Calculate the total sales for each store.

- Identify the store with the lowest total sales.

- Calculate the difference in sales between the highest and lowest-performing stores.

- Expected Output:

- The worst performing store.

- Difference in sales between the highest and lowest performing stores.

\*\*Objective 2: Predictive Modeling for Sales Forecasting\*\*

- Checklist:

- Select an appropriate predictive modeling technique (e.g., time series forecasting).

- Split the dataset into train and test sets.

- Train the model using the train set.

- Evaluate the model's performance using suitable evaluation metrics.

- Make predictions for the next 12 weeks' sales for each store using the trained model.

- Expected Output:

- Sales forecast for each store for the next 12 weeks.

1. \*\*Data Exploration and Preparation\*\*

- Load the dataset (Sales data from Walmart).

- Perform basic exploratory data analysis (EDA) to understand the structure and characteristics of the data.

- Handle missing values, outliers, and any data quality issues.

- Check for data types and convert them if needed.

- Extract relevant features that could potentially impact sales forecasting, such as date, store ID, department, etc.

2. \*\*Time Series Analysis\*\*

- Verify the time series properties of the data (stationarity, seasonality, trend, etc.).

- Implement necessary transformations (e.g., differencing) to achieve stationarity.

- Train-test split the data, ensuring chronological order is maintained.

3. \*\*Baseline Model\*\*

- Develop a simple baseline model (e.g., naive forecast or moving average) to establish a performance benchmark.

- Evaluate the baseline model using appropriate metrics (e.g., mean absolute error or root mean squared error).

4. \*\*Model Selection\*\*

- Explore and choose an appropriate forecasting model based on the characteristics of the data and problem requirements. Consider models such as ARIMA, SARIMA, Prophet, or machine learning models like random forest or gradient boosting.

- Split the training set further into train and validation sets to perform model evaluation and hyperparameter tuning.

5. \*\*Model Training and Evaluation\*\*

- Fit the chosen model on the training data.

- Evaluate the model's performance using appropriate metrics on the validation set.

- Iterate and refine the model by adjusting hyperparameters and testing different variations.

- Validate the model's performance on the test set to assess its generalization ability.

6. \*\*Final Model and Forecasting\*\*

- Select the best-performing model based on the evaluation results.

- Re-fit the chosen model on the entire dataset (train + validation).

- Make forecasts on the unseen future data using the final model.

- Evaluate the forecasting performance using appropriate metrics.

7. \*\*Presentation and Documentation\*\*

- Prepare visualizations (plots, charts) to present the forecasted sales and model performance.

- Summarize the findings, insights, and limitations of the project.

- Document the methodology, assumptions made, and steps followed during the project.

- Provide recommendations for future improvements or alternative approaches.

Note: It's important to tailor the process to the specific details and goals of your Walmart sales forecasting project. This outline serves as a general guide, and you may need to adapt it based on your specific requirements and the characteristics of your dataset.

To represent the steps professionally in a flowchart or checklist format, you can use visual diagramming tools or spreadsheet applications. Be sure to include relevant details, decision points, and arrows indicating the flow of the process. You can also consider segregating the flow into multiple sections or subflows to make it more understandable.

Additionally, while the provided outline focuses on time series analysis and forecasting, keep in mind that there are alternative approaches to forecasting, such as machine learning-based regression or classification models, that might be applicable in certain scenarios. It's important to critically assess the nature of your problem and select the most suitable approach accordingly.

To make the code and its representation look more professional, you can follow coding best practices such as using meaningful variable names, adding comments where necessary, organizing code into functions or classes, using appropriate data structures, providing clear and concise documentation, and adhering to a consistent code style (PEP 8 guidelines). Visualizations can be enhanced by utilizing libraries like Matplotlib or Seaborn to create informative and visually appealing plots.

Remember, the success of any project depends on the careful exploration, analysis, and interpretation of the data, as well as thoughtful modeling choices and rigorous evaluation.

In addition to the above checklists, there are some cautions to keep in mind during the Walmart sales forecasting project:

1. Avoid using too many variables in the analysis: Using too many variables can increase the complexity of the analysis and lead to overfitting. It's essential to identify the most relevant variables that have a significant impact on the sales forecast and focus on those.

2. Be cautious of multicollinearity: High correlations between variables can lead to issues with multicollinearity. It's important to identify and address these issues to ensure the reliability of the analysis.

3. Take care of missing values: Filling missing values with the wrong technique can cause bias in the analysis. It's essential to carefully evaluate and select the appropriate missing value imputation technique.

4. Be careful while handling seasonal trends: Seasonality can have different periods (e.g., daily, weekly, monthly), making it crucial to choose a relevant period for modeling. Additionally, trends can differ based on the data used and the time frame. Ensure you identify the relevant trend and seasonality period to avoid incorrect modeling.

5. Select appropriate evaluation metrics: Selecting unsuitable evaluation metrics can provide a false sense of performance. Ensure you choose the appropriate metrics for evaluating each objective and model.

After completing each step, check to ensure the expected output is in line with the project's objectives. Additionally, sanity checks (e.g., comparing with previous results, comparing with expected values) of each objective should be performed to ensure the output is accurate. Lastly, refine the methodology for improved results in future projects.

To make the code more professional, ensure it adheres to the PEP 8 guidelines. It is also a good practice to include comments in the code to enhance code readability. Also, use separate functions for each objective and method used to improve code modularity and organization. Lastly, include code documentation in the form of a readme file detailing the purpose, methods, and assumptions made in the code.

By following these cautions and expected outcomes, you will be able to develop a robust and effective solution for the Walmart sales forecasting project and produce accurate results.