Project statements

Project 1

Problem Statement:

Background: Mr. Adrian, the store manager at Target's Scotts Valley store in Minneapolis, has recently concluded a pilot study in collaboration with Target's Workforce Management (WFM) team. This study aimed to anonymously track and record absenteeism trends among his 36-member staff over the past year.

Problem: The absence of Team Members (TMs) significantly impacts Target's operational efficiency, resource allocation, and ability to meet guest needs effectively. Consequently, Mr. Adrian is keen to investigate and understand the underlying factors contributing to TM absenteeism.

Objectives:

1. **Identify Key Factors:** Mr. Adrian seeks to determine what causes TMs to miss extended hours through data-driven analysis.
2. **Resource Deployment:** Understanding these factors will enable better planning for hiring needs and resource deployment.
3. **Policy Implementation:** Insights from the data will help implement policies that can mitigate absenteeism.
4. **Workforce Scheduling:** Leveraging insights on absentee patterns, Target aims to optimize work schedules by distributing higher-risk TMs across different shifts for balanced coverage.

Approach: Mr. Adrian has reached out to his Data Science partners for assistance in this endeavor. The expectation is for the team to provide unbiased findings strictly based on data analysis, ensuring recommendations are objective and actionable.

Desired Outcomes:

1. **Data-Driven Insights:** Derive clear insights into the reasons behind TM absenteeism.
2. **Actionable Recommendations:** Formulate strategic recommendations that can be translated into effective policies and practices.
3. **Optimized Scheduling:** Develop a scheduling strategy that minimizes disruption by balancing risk across shifts.
4. **Enhanced Efficiency:** Ultimately, improve operational efficiency and customer satisfaction through effective management of human resources.

Dataset: Absenteeism_at_work.csv

Attribute Information:

- TM_ID – Team Member ID; a surrogate key to identify a Team Member
- Reason for absence – the identified reason why the TM was absent from work
- Month of absence – <self-explanatory>
- Day of the week – <self-explanatory>
- Seasons – <self-explanatory>
- Transportation expense – the amount in dollars
- Distance from Residence to Work – in miles
- Service time
- Age – <self-explanatory>
- Workload Average/day
- Hit target – % of target achieved in terms of workload avg.
- Disciplinary failure (yes=1; no=0)
- Education - (high school (1), graduate (2), postgraduate (3), master and doctor (4))
- Social drinker (yes=1; no=0)
- Social smoker (yes=1; no=0)
- Pet (number of pets at home)
- Weight – <self-explanatory>
- Height – <self-explanatory>
- Body mass index – <self-explanatory>
- Absenteeism time in hours

Project 2

Problem Statement:
Guests are the driving force or the backbone of any retail company like Target. It is imperative to understand Guest purchasing behavior since most marketing strategies are designed around it. In this project we take a subset of guests and try to create guest segments and their response to campaigns Here we have taken a subset of 2240 Target guests and related information over a 3 month period for various metrics like Demographic information , Sales in multiple product categories,Response to campagins etc
Objective: The task is to understand customer shopping behaviour :-
1. find best ways to segment guests and label them accordingly, this will help the company create marketing and discount strategies for each segment
2. Analyse response to campaigns for each predicted segment and the likelihood of responding to the campaign

Data Provided: A dataset containing information of 2240 guests for 3 months with the following information :

Guest_id - Unique identifier for each individual in the dataset.
Year_Birth - The birth year of the individual.
Education -The highest level of education attained by the individual.
Marital_Status - the marital status of the individual.
Income - the annual income of the individual.
Kid_le_12 - The number of young children in the household.
Kid_teen -The number of teenagers in the household.
Dt_Customer - The date when the customer was first enrolled or became a part of the company's database.
Recency -The number of days since the last purchase or interaction.
Amt_Wines -The amount spent on wines.
Amt_Fruits - The amount spent on fruits.
Amt_Bake -The amount spent on Bake products.
Amt_Deli -The amount spent on Deli products.
Amt_Sweets - The amount spent on Sweet products.
Amt_Meats - The amount spent on Meat products.
No_Deals - The number of purchases made with a discount or as part of a deal.
No_Web - The number of purchases made through the company's website.
No_Catalog - The number of purchases made through catalogs.
N0_Store - The number of purchases made in physical stores.
N0_Web_Visits_Month - The number of visits to the company's website in a month.
Cmp3_accept - Binary indicator (1 or 0) whether the individual accepted the third marketing campaign.
Cmp4_accept - Binary indicator (1 or 0) whether the individual accepted the fourth marketing campaign.
Cmp5_accept - Binary indicator (1 or 0) whether the individual accepted the fifth marketing campaign.
Cmp1_accept - Binary indicator (1 or 0) whether the individual accepted the first marketing campaign.
Cmp2_accept - Binary indicator (1 or 0) whether the individual accepted the second marketing campaign.
Complain - Binary indicator (1 or 0) whether the individual has made a complaint.
Z_CostContact - A constant cost associated with contacting a customer.
Z_Revenue - A constant revenue associated with a successful campaign response.
Response - Binary indicator (1 or 0) whether the individual responded to the marketing campaign

customer_segmentation_data.csv

Project 3

We have dataset for the year 2013, for 1559 products across 10 stores in different cities. Where the dataset consists of 12 attributes like Item Fat, Item Type, Item MRP, Outlet Type, Item Visibility, Item Weight, Outlet Identifier, Outlet Size, Outlet Establishment Year, Outlet Location Type, Item Identifier and Item Outlet Sales. Out of these attributes response variable is the Item Outlet Sales attribute and remaining attributes are used as the predictor variables.

The data-set is also based on hypotheses of store level and product level. Where store level involves attributes like: city, population density, store capacity, location, etc and the product level hypotheses involves attributes like: brand, advertisement, promotional offer, etc.

Variable - Details

- Item_Identifier- Unique product ID
- Item_Weight- Weight of product
- Item_Fat_Content - Whether the product is low fat or not
- Item_Visibility - The % of total display area of all products in a store allocated to the particular product
- Item_Type - The category to which the product belongs
- Item_MRP - Maximum Retail Price (list price) of the product
- Outlet_Identifier - Unique store ID
- Outlet_Establishment_Year- The year in which store was established
- Outlet_Size - The size of the store in terms of ground area covered
- Outlet_Location_Type- The type of city in which the store is located
- Outlet_Type- Whether the outlet is just a grocery store or some sort of supermarket
- Item_Outlet_Sales - Sales of the product in the particulat store. This is the outcome variable to be predicted.

Train_uc.csv

Project 4

Problem Statement:

An e-commerce company wants to improve inventory management and optimize its stock levels by accurately forecasting future sales for each product category using historical sales data. The dataset consists of order details including order date, ship date, ship mode, customer information, geographical data, product details, and sales amount. The goal is to develop a time series model that can predict daily sales for the next 90 days for each product category.

Dataset: timeseries_sales.csv

Data Dictionary:

- order_id: A unique identifier for each order. This helps in distinguishing different orders within the dataset.
- order_date: The date when the order was placed by the customer. This is crucial for time series analysis as it represents the timestamp for sales data.
- ship_date: The date when the order was shipped to the customer. This can be useful for analyzing delivery times and shipping performance.
- ship_mode: The method by which the order was shipped (e.g., standard, express). Understanding shipping modes can help in analyzing their impact on sales and customer satisfaction.
- customer_id: A unique identifier for each customer, which can be useful for analyzing repeat customers and customer segmentation.
- country: The country where the customer resides or where the shipment is destined. This helps in geographic segmentation of sales data.
- city: The city where the customer resides or where the shipment is destined. This adds another layer to geographic sales analysis.
- state: The state or region where the customer resides or where the shipment is destined, which further refines geographic analysis.
- postal_code: The postal code associated with the customer's location, providing precise geolocation information which could be used to identify patterns at a very granular level.
- region: A higher-level geographical categorization (e.g., North America, Europe) that can help in broad segment analysis and understanding regional demand trends.
- product_id: A unique identifier for each product listed on the e-commerce website which helps track individual product performance over time.
- category: The main category to which a product belongs (e.g., Electronics, Furniture). Key for segmenting data and forecasting at a category level.
- sub_category: A more specific classification under each main category (e.g., within Electronics: Smartphones, Laptops). Useful for granular analysis within main categories.
- product_name: The name of the product as listed on the e-commerce platform; while useful for descriptive purposes, typically not used directly in numerical forecasting models.
- sales_$: The revenue generated from selling a particular product or group of products within an order, denominated in dollars ($). This is your target variable for forecasting models as it reflects sales performance directly.

Objectives:

The primary objectives of this forecasting project are:

1. **Data Preparation**:
    - Ensure data integrity by cleaning and preprocessing the given dataset.
    - Convert relevant columns to appropriate data types (e.g., dates).
2. **Exploratory Data Analysis (EDA)**:
    - Understand trends, seasonality, and patterns within the historical sales data.
    - Visualize sales over time to identify any notable behaviors or anomalies in different categories.
3. **Feature Engineering**:
    - Aggregate daily sales data for each product category.

- Create additional features that might improve prediction accuracy such as lagged values, moving averages, day of the week indicators, holiday seasons etc.

4. **Model Selection**:

   - Evaluate various time series forecasting models including classical models, and advanced machine learning models.

5. **Model Building and Training**:

   - Train selected models on historical sales data of each product category.
   - Fine-tune model parameters to achieve optimal performance.

6. **Model Evaluation**:

   - Assess the performance of the models using appropriate metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), or Mean Absolute Percentage Error (MAPE).
   - Compare these metrics across different models to select the best-performing one for each category.

7. **Forecasting**:

   - Generate sales forecasts for each product category over a 90-day horizon.

8. **Visualization**:

   - Visualize predicted vs actual sales to interpret model predictions clearly.
   - Provide insights using line charts or bar graphs demonstrating potential high-performance categories.