**Pharmaceutical Sales prediction across multiple stores**

**INTRODUCTION:**

* Predicting sales of pharmaceutical company across multiple stores using Machine Learning
* Machine Learning includes EDA(Exploratory Data Analysis) which is the cornerstone of successful machine learning projects.

**PURPOSE OF EDA:**

* EDA’s primary goal is to gain insights into customer behaviour across different stores.
* Connect the exploration of customer behaviour to broader business objectives, emphasizing its significance for strategic decision-making.

**SIGNIFICANCE OF EDA:**

* EDA plays significant role on machine learning in data preprocessing and cleaning the data
* It eliminates outliers, correct inconsistencies, and handle missing values.
* Ensure the audience understands that a clean dataset is foundational for reliable insights.

**EXPECTATION:**

* Encourage exploration beyond the provided questions, promoting a comprehensive analysis.
* Emphasize the need for clear and insightful visualizations to support the communication of findings.

**PROJECT:**

**STEP 1:** Import libraries needed for EDA.

**STEP 2:** I tried adding train sheet but it causes error as column 7 had mixed values. So I split the column 7(State Holidays) into 3 columns as State Holidays A, State Holidays B & State Holidays C.

**STEP 3:** Then I loaded train and test data into python as train\_data & test\_data.

# Task 1: Exploration of customer purchasing behaviour

# Distribution of promotions in training and test sets

# In this step we have compared promo on train and test set to understand the promotion distribution between both train and test sets are similar

# OBSERVATION:

# The promotion on both train and test sets are not similar according to the countplot chart.

# As the promotion on train and test sets are not consistent. We have to ensure the distribution of features is consistent between training and test data to build a model that generalizes well to unseen data.

# 2: Sales behaviour before, during and after holidays

# In this step we have compared sales behaviour of stores before, during and after holidays using lineplots.

# OBSERVATION:

# The sales behaviour of before, during and after holidays looks similar. Comparatively the sales on before, during and after holidays is little higher than the normal working days.

# 3: Seasonal purchase behaviours

# In this step we have compared sales on seasonal times using lineplots.

# OBSERVATION:

# The sales behaviour on all the seasonal times are looking same and the average sales on seasonal time is around 9000 as per lineplots.

# 4: Correlation between sales and number of customers

# In this step we have analysed the correlation between sales and the number of customers using scatter Plot.

# OBSERVATION:

# The correlation efficiency is 0.89 which indicates the very strong correlation between sales and the number of customers.

# 5: Comparison of sales with and without promotions

# In this step we are going to compare sales with promotion and without promotion using bar plot.

# OBSERVATION:

# Comparatively the sales on promotional times are high than the non-promotional times.

# 6: Sales performance for different stores during the promotional periods

# In this step we compare sales performance for different stores during the promotional periods using heatmap.

# OBSERVATION:

# Sales performance of different stores during the promotional time looks distinct as per the heatmap and there is no correlation or similarity between each store.

# 7: Average sales trends during store open hours

# For this step there is no time duration of closing and opening is available in the given sheet. I tried changing the format of date and pull the timing from date and also charted it in lineplot but the values are empty as the timings are not available.

# 8: Weekend sales for stores open all weekdays

# In this step we are going to see the sales behaviour on weekend sales for stores which are open on all weekdays including weekends using bar plot.

# OBSERVATION:

# In this we have observed that only on very few days the sales are high apart from that on all days the sales are average and limited to around 5000 on all weekends.

# 9: Assortment type impact on sales

# In this step we are going to compare the impact of sales according to the assortments for this we need to import store data and once added we are going to merge data to have sales and and assortment type on one area to analyse the assortment type impact on sales.

# For this we are going to use box plot to analyse the data.

# OBSERVATION:

# The boxplot indicates that assortment type b has wider range than assortment a & c.

# The assortment for a is low compare to b & c even though the sales in assortment a is higher than b & c which indicates the assortment type doesn’t have any impact on sales.

# 10: Competitor Distance impact on sales

# In this step we are going to analyse the sales impact due to competitor based on the distance using scatterplot.

# OBSERVATION:

# This scatterplot indicates that the stores with nearby competitors has more sales compare to the stores with long distance competitors.

# This indicates that the stores that are closer to their competitors has a competitive advantage and the chance of high Walkins because the customers visiting competitor brand can also visit the store.

# 11: Impact of New competitors on store sales

# In this step we are going to analyse the impact of new competitors on store sales using lineplot.

# OBSERVATION:

# This line plot shows that only around 600 stores are not impacted by the new competitors and their sales is also good.

# And around 1000 stores are impacted by the new competitors this shows new competitors has heavy impact on store sales.

# Task 2: Prediction of stores sales

# 2.1 Preprocessing

# For preprocessing the data we have to import standard scaler from the sklearn.preprocessing library

# After that I have checked the info of train sheet where there is no null values.

# I have created the columns for total holidays, month, day, year with the help of date column to preprocess the data.

# Then I created weekend column with the help of dayofweek column to understand better on weekend sales.

# And Daytoholiday and Daysafterholiday columns to sturdy the holiday period sales.

# To check with month data I have splitted month in to Beginningofmonth, midofmonth and Endofmonth columns to split the sale data to understand the sales for the month.

# I have split the year into 3 quarters to understand in which quarter the sale is better.

# CONCLUSION:

# In this task we have find the sales behaviour and various facts which affects the sales and also we have preprossed data and cleaned the data to analyse the sales better.

# 2.2 Building models with sklearn pipelines

# In this step I have build the models with necessary data’s which are required for sales prediction.

# I merged the train data and store data to get the necessary coloums such as 'StoreType', 'Assortment', 'CompetitionDistance', 'CompetitionOpenSinceMonth', 'CompetitionOpenSinceYear', 'Promo2', 'Promo2SinceWeek', 'Promo2SinceYear', 'PromoInterval','Sales','Sales\_MonthTotal\_StoreTotal'

# After building model with necessary coloumns I pipelined with standardscaler and randomforestregressor.

# 2.3 Choose a loss function

# In this we have checked the loss function between actual data and the trained data to see is there is any squared error between these two.

### **2.4 Post Prediction analysis**

# Since there is no square error out data’s accuracy is high so we move on to next step which is Post prediction analysis.

# In this step we have checked the important features that can provide the accurate prediction.

# From this step we noticed that store type, Assortments, Competition Distance are having more correlation than the other columns in data

# 2.5 Serialize models

# In this step we have serialized the model with timestamp using datetime to make the model precise so that we can track the prediction of various models

### **2.6 Building model with deep learning**

# In this step we are going to build LSTM model to predict the sales.

# I have used nlag as 30 so I can train data with 30 days sales behaviour

# At the end of this LSTM model I have checked the mean squared error which is 0.004 which means the LSTM model which we build is predicting the sales accurately.

# I have plotted the actual sales and the predicted sales using the scatter plot to check weather the model’s prediction is good.

# Observation

# In scatter plot the correlation between the actual sale and predicted sale are on the same range which means the model is predicting the sales successfully.

# 2.7 Using MLFlow to serve the prediction

# In this step we have log the LSTM model which we create to the mlflow.

# It is successfully loaded into the LSTM model.

# In the next step I tried to load the LSTM model but the path is not defined. I tried to find out the path where it is showing it is saved inside the path/to/artifact/directory

# But if I tried to load it from artifact it is showing as path not defined.