Capstone Project-1

The Walmart Store

Table of Contents-

Problem Statement

A retail store that has multiple outlets across the country are facing issues in managing the inventory - to match the demand with respect to supply. You are a data scientist, who has to come up with useful insights using the data and make prediction models to forecast the sales for X number of months/years.

Project Objective

The objective of this project is to provide the retail store chain with data-driven insights that can be used to improve inventory management and maximise sales. By analysing the given data, we will identify key trends and factors that impact sales and suggest strategies that can be used to improve performance

Dataset Information

The walmart.csv contains 6435 rows and 8 columns.The features in the dataset are described as follows:

●Store: The 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 the fuel in the region

● CPI: Consumer Price Index

● Unemployment: Unemployment rate

Data Pre-processing steps and Inspiration

Before performing any analysis, the first step was to check for missing or null values in the dataset. Since there were no missing values, we proceeded with analysing the data. The inspiration for the analysis was to find correlations between variables and identify patterns in the data. We began by analysing the relationship between weekly sales and different variables like holidays, temperature, fuel price, CPI, and unemployment rate. We also analysed the date to understand the trends in weekly sales across different stores.

Choosing the algorithm for the project

In this section, we will evaluate the performance of several different regressors on our data. We will use root mean squared error (RMSE) as our evaluation metric. RMSE is a measure of the difference between the predicted values and the true values. It is calculated as the square root of the mean squared error (MSE), where MSE is the average of the squared differences between the predicted and true values. Lower values of RMSE indicate better performance.

We will fit and evaluate the following regressors:

- Linear Regression
- Decision Tree Regressor
- Random Forest Regressor
- Support Vector Regressor, etc.

Motivation and Reasons for choosing the algorithm

We will fit each of these regressors to our training data and make predictions on the test set. Then, we will calculate the RMSE of the predictions and compare the results to choose the best regressor.

To ensure that the original dataset is not modified during the modeling process and to facilitate debugging if needed, we will create a copy of the preprocessed dataset before fitting our various models. This will help to preserve the integrity of the original data and allow us to refer to it if any issues arise during the modeling process.

Assumptions

The forecasting models are based on past sales data, and it is assumed that the trends and patterns in the past will continue into the future. Additionally, it is assumed that there will be no significant changes in the economic, social, or political factors that may affect sales during the next 12 weeks.

Model evaluation and Techniques

In this subsection, we will create a function that will train multiple regressors and compare their performance using the root mean square error (RMSE) metric. We will use the RMSE values to compare the performance of the various regressors and determine which model has the lowest error and is therefore the best fit for our data.

Inferences from the same

After evaluating min RMSE value ,we find that the best model will be **Random Forest Regressor.**

Future possibilities of the project

The analysis and predictive models created in this project can be used to optimize inventory management and maximise sales. Additionally, the models can be updated regularly with data to ensure that they remain accurate and relevant.

1.Using the above data, come up with useful insights that can be used by each of the stores to improve in various areas

1. Holiday weeks have a noticeable impact on weekly sales. Stores could consider adjusting their staffing levels and inventory to account for increased demand during holiday weeks.

2. Temperature appears to have some correlation with weekly sales. Stores in areas with consistently high or low temperatures may want to adjust their marketing and sales strategies to account for this.

3. Fuel prices may also be a factor in weekly sales. If fuel prices are high, customers may be less likely to travel long distances to shop at a particular store. Stores could adjust their advertising and promotions to encourage local customers to visit more frequently.

4. CPI and unemployment rates may also have an impact on weekly sales, although these relationships may be more complex and require further analysis to fully understand.

5. Stores could use the provide data to compare their performance against other stores in the same region or across the company as a whole. This could help identify areas where they are lagging behind and opportunities for improvement.

6. Stores could also use the data to identify trends in weekly sales over time. This could help them adjust their strategies to account for changing customer preferences and market conditions. For example, if sales are consistently declining inn a particular product category, stores could consider reducing inventory levels and shifting resources to more profitable areas.

Forecast the sales for each store for the next 12 weeks

1. Prepare the data: We need to aggregate the sales data at the store level and convert the “Date” column to a time series index. We can also check if there are missing values, outliers, or any other data quality issues that need to be addressed.

2. Visualize the data: We can plot the sales data to see if there are any trends, seasonality, or other patterns that we need to take into account when building our forecasting model.

3. Split the data: We need to split the data into a training set and a test set. We can use the historical sales data as the training set and the most recent 12 weeks as the test set.

4. Build the forecasting model: We can use different regressors to forecast the sales for each store. We can use the training set to fit the model and tune the hyper parameters. We can then use the model to make predictions for the test set.

5. Evaluate the model: We can evaluate the performance of the model by comparing the predicted sales to the actual sales in the test set. We can use metrics such as Mean Absolute Error (MEA), Root Mean Squared Error (RMSE), Mean Squared Error (MSE) to quantify the accuracy of the model.

6. Make predictions: Once we have built and validated our forecasting model, we can use it to make predictions for the next 12 weeks of sales for each store. We can also monitor the performance of the model and adjust it if necessary based on new data or changes in the business environment.

----------END----------

Capstone Project-2

The Online Retail Store

Table of Contents-

Problem Statement

A retail store that has multiple outlets across the country are facing issues in managing the inventory to match the demand with respect to supply.We have to come up with useful insights using the data and make prediction models to forecast the sales for X number of months/years.

Project Objective

• finding useful insights about the customer purchasing history

• Segment the customers based on their purchasing behaviour.

Dataset Information

The online_retail.csv contains 387961 rows and 8 columns.

Feature Name	Description
Invoice	Invoice number
StockCode	Product ID
Description	Product Description
Quantity	Quantity of the product
InvoiceDate	Date of the invoice
Price	Price of the product per unit
CustomerID	Customer ID
Country	Region of Purchase

Data Pre-processing Steps and Inspiration

Before diving into insights from the data, we checked null values and shape of data, Quantity of the products having negative value as well as duplicate entries were removed from the data. The data contained 5268 duplicate entries (about ~1%). As per the data, if the invoice number code starts with the letter 'c', it indicates a cancelled order.

Choosing the Algorithm for the Project

For this project, I used Kmeans algorithm.

Motivation and Reasons For Choosing the Algorithm

In this project, I have to segment the customers based on their purchasing behaviour .So Kmeans works well because this one generate clusters and no of clusters are calculated with the help of Elbow test.

Assumptions

• No Influential Outliers: Influential outliers are extreme data points that affect the quality of the logistic regression model.

• No Multicollinearity: Multicollinearity is a problem because it creates redundant information that will cause the results of model to be unreliable.

• Observation Independence: observations should not come from a repeated measure design.

●The dataset is representative of the overall customer population

●The data is reliable and accurate

●The purchase behaviour of customers is consistent over time

Model Evaluation and Techniques

• Model evaluation is the process of using different evaluation metrics to understand a machine learning model's performance, as well as its strengths and weaknesses.

• The most popular metrics for measuring classification performance include accuracy, precision, confusion matrix, log-loss, and AUC (area under the ROC curve).

Inferences from the Same

The customers are well Segmented based on their purchasing behaviour.

Future Possibilities of the Project

●Integration of additional data sources like demographic information

●Implementation of recommendation systems for personalized product recommendations

●Development of predictive models for forecasting sales and product demand

1) Useful Insights about Customer Purchasing History

●Analysing the quantity of the products sold can provide insights about the customer's preferred products and their demand.

●Examining the price of the products sold can reveal customer's purchasing power and price sensitivity.

●Studying the frequency of invoices generated for each customer can help in identifying the most loyal customers.
●Analysing the sales trend over time can provide insights about the seasonality effect on customer purchase patterns.

●Examining the customer retention rate can provide insights about customer satisfaction and loyalty.

●Analysing the geographical distribution of customers and their purchases can help identify new target markets and optimize delivery services.

2) Customer segmentation based on purchasing behaviour

●Customer Lifetime value: Identifying high-value customers based on their purchase history and analysing their behaviour over time to identify patterns and insights.

●Cohort Analysis: Dividing customers into groups based on their purchase history and analysing their behaviour over time to identify patterns and insights.

●Behavioural segmentation: Grouping customers based on their behaviour such as products purchased, time of purchase, price sensitivity, etc.

●Demographic segmentation: Analysing customer data based on demographics such as age, gender, location, etc.

----------END----------