Retail Sales Prediction

EMMANUEL EJEU | 2024/HD05/21922U | 2400721922 | MCSC

Introduction

- In retail, being able to predict sales accurately is essential for making smart business decisions. This project focuses on building models to forecast sales by using a dataset that includes various factors affecting sales, such as discounts, marketing spend, and seasonal trends like holidays.
- This project will not only predict sales but also evaluate the performance of different models such as linear regression, decision trees, and gradient boosting methods, among others. The comparison will help identify the most effective model for improving sales forecasting accuracy in real-world retail settings.

Dataset used

Title

Retail Sales Data with Seasonal Trends & Marketing

Link of Dataset

https://www.kaggle.com/datasets/abdullah0a/retail-sales-data-with-seasonal-trends-and-marketing/data

Citation

Abdullah. (2024). Retail Sales Data with Seasonal Trends & Marketing [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/9349466

Other Reference used for learning

https://www.analyticsvidhya.com/blog/2022/07/step-by-step-exploratory-data-analysis-eda-using-python/

About Dataset

This dataset provides detailed insights into retail sales, featuring a range of factors that influence sales performance. It includes records on sales revenue, units sold, discount percentages, marketing spend, and the impact of seasonal trends and holidays.

Key Features of Dataset

The dataset provided is a retail sales dataset, containing 30,000 rows and 11 columns. Below are the key features in the dataset;

1. **Sales Revenue (USD):** Total revenue generated from sales.
2. **Units Sold:** Quantity of items sold.
3. **Discount Percentage:** The percentage discount applied to products.
4. **Marketing Spend (USD):** Budget allocated to marketing efforts.
5. **Store ID:** Identifier for the retail store.
6. **Product Category:** The category to which the product belongs (e.g., Electronics, Clothing).
7. **Date:** The date when the sale occurred.
8. **Store Location:** Geographic location of the store.
9. **Day of the Week:** Day when the sale took place.
10. **Holiday Effect:** Indicator of whether the sale happened during a holiday period.

Use Cases of the Dataset

1. **Predictive Modeling:** Build models to forecast future sales based on historical data.
2. **Marketing Analysis:** Evaluate the effectiveness of marketing spend and discount strategies.
3. **Seasonal Trend Analysis:** Examine how different seasons and holidays impact sales.
4. **Revenue Optimization:** Identify strategies to optimize pricing and marketing for increased revenue.

Key Objectives

- Build and compare different models to predict retail sales.
- Analyze how features like marketing, discounts, and holidays affect sales.
- Find the most effective model for forecasting future sales.

Problem: Accurately predicting retail sales

Research Questions

- **Main Research Question:** What factors drive the sales performance of stores in different geographic locations? How can we make relevant predictions based on these factors?

- Follow-up questions

1. What is the relationship between marketing spend, discount percentage, and units sold?
2. Are there specific product categories or store locations that perform better?
3. Does the day of the week or holiday effect influence sales performance?
4. What is the distribution of units sold, sales revenue, discount percentage, marketing spend, sales by day of the week, top store location, and sales by product category?
5. Marketing Spend vs. Units Sold: Does increased marketing lead to more units sold?
6. Discount Percentage vs. Sales Revenue: What impact does discounting have on sales revenue?
7. Holiday Effect vs. Sales Revenue: Do holidays drive higher sales?
8. Marketing Spend vs. Sales Revenue: Does increased marketing lead to more units revenue?
9. Sales Revenue by Day of the Week: Which days of the week generate the highest revenue?
10. Top Sales Revenue by Store Location: Which locations bring what revenue threshold?
11. Sales Revenue by Product Category: What are the most sold products?
12. Correlation Matrix (Heatmap): How do different factors relate to each other?
13. Heatmap of Sales Revenue by Day of the Week and Product Category?

About the data

Check for Duplication:

Store ID	1
Product ID	42
Date	731
Units Sold	32
Sales Revenue (USD)	2545
Discount Percentage	5
Marketing Spend (USD)	200
Store Location	243
Product Category	4
Day of the Week	7
Holiday Effect	2

dtype: int64

Data Types:

Store ID	object
Product ID	int64
Date	datetime64[ns]
Units Sold	int64
Sales Revenue (USD)	float64
Discount Percentage	int64
Marketing Spend (USD)	int64
Store Location	object
Product Category	object
Day of the Week	object
Holiday Effect	bool

dtype: object

There was no data inconsistency

Initial data exploration reveals the following key points

Initial data exploration reveals the following key points:

- **Observations:** The dataset consists of 30,000 rows and 11 columns.
- **Missing Values:** No missing values are present in the dataset.
- **Column Types:** The dataset includes a mix of integers, floats, objects, and boolean values.
- **Marketing and Discounts:** Most days had no marketing spend or discounts applied.
- **Maximum Values:** The highest marketing spend was $199, and the largest discount offered was 20%.
- **Store ID:** There is only one unique value in 'Store ID', so this column can be dropped.
- **Product ID:** While 'Product ID' is numerical, it contains 42 unique values.
- **Product Location:** There are 243 unique product locations in the dataset.

Exploratory Data Analysis

Observations from the visualization:

- 📊 **Numerical Distribution:** Numerical columns have a right-skewed distribution.
- 📅 **Days of the Week:** The distribution of days of the week is well balanced.
- 📉 **Discount Percentage:** The distribution of discount percentage is imbalanced, with most values being 0.
- 🛋️📦🍎👗 **Product Categories:** There are 4 product categories: Furniture, Electronics, Groceries, and Clothing.
- 💸 **Revenue by Product Categories:** Electronics and Clothing are correlated with higher revenues.
- 🎉 **Holiday Effect:** The holiday effect is mostly 'False' in 99.5% of the cases (which is expected, as holidays are not daily occurrences).
- 🎁 **Holiday Revenue:** Holidays tend to bring more revenue overall.
- 🔗 **Correlation:** The only highly correlated numerical variables are 'Units Sold' and 'Sales Revenue'.

Regression Findings

- We achieved **outstanding results**, with an **R^2 score of 99.8%**! It's also worth noting that a simple and highly interpretable model, the **Decision Tree**, performed exceptionally well, achieving 99.6%.
- While we should be pleased with these results, there's an important issue. By including the **"Items Sold"** feature when predicting **"Sales Revenue"**, we've introduced **data leakage**.
- These two features are directly correlated, and in a real-world scenario, we wouldn’t know the number of items sold in advance when predicting future sales. Hence, there is a need to remove this feature and repeat the process.

Linear Regression Model:

- Test R^2: 0.8629
- Test RMSE: 959.3323

SGD Regressor Model:

- Test R^2: 0.8622
- Test RMSE: 961.4637

Random Forest Model:

- Test R^2: 0.9967
- Test RMSE: 149.6964

ElasticNet Model:

- Test R^2: 0.0894
- Test RMSE: 2471.9898

K-Neighbors Regressor Model:

- Test R^2: 0.7782
- Test RMSE: 1219.8619

Decision Tree Model:

- Test R^2: 0.9961
- Test RMSE: 161.6051

SVR Model:

- Test R^2: -0.0478
- Test RMSE: 2651.6621

XGBoost Model:

- Test R^2: 0.9981
- Test RMSE: 113.9861

Gradient Boosting Model:

- Test R^2: 0.8913
- Test RMSE: 853.9157

Bagging Model:

- Test R^2: 0.9958
- Test RMSE: 167.4768

Removing one variable to remove data leakage

- Just by removing one variable (Items Sold) the R^2 score dropped significantly from 99.8% to 54.3%!
- That’s a huge difference, but now we can be confident there’s no data leakage. While the results aren't as impressive, they are far more reliable.
- R-squared (R2) and root mean square error (RMSE) are both metrics used to evaluate how well a linear regression model fits a dataset, but they differ in how they measure fit and what they quantify.
- While R2 tells you about correlation between two datasets, RMSE tells you about the difference between them.

Linear Regression Model:

- Test R^2: 0.5417
- Test RMSE: 1753.7249

SGD Regressor Model:

- Test R^2: 0.5426
- Test RMSE: 1751.9873

Random Forest Model:

- Test R^2: 0.4657
- Test RMSE: 1893.5792

ElasticNet Model:

- Test R^2: 0.0846
- Test RMSE: 2478.4418

K-Neighbors Regressor Model:

- Test R^2: 0.4335
- Test RMSE: 1949.8000

Decision Tree Model:

- Test R^2: 0.2210
- Test RMSE: 2286.3940

SVR Model:

- Test R^2: -0.0495
- Test RMSE: 2653.7860

XGBoost Model:

- Test R^2: 0.5330
- Test RMSE: 1770.1885

Gradient Boosting Model:

- Test R^2: 0.5043
- Test RMSE: 1823.8799

Bagging Model:

- Test R^2: 0.4476
- Test RMSE: 1925.3837

More Gaps that can be addressed

- Using other parameters to measure like the Date column by extracting new features such as year, month, day, etc.
- Dropping other variables that could be outliers if any
- Incorporate hyperparameter tuning for machine learning models to optimize performance.
- Explore Deep Learning approaches to potentially improve results further.

SHAP and LIME

Implementing SHAP

SHAP Values (SHapley Additive exPlanations) break down a prediction to show the impact of each feature.

Step 1: Import Libraries and Prepare Data

Load and preprocess the data. Use a smaller background sample to improve SHAP runtime.

Step 2: Train a Model

We'll use RandomForestRegressor, which is compatible with SHAP's TreeExplainer for efficiency.

Step 3: Run SHAP Analysis

Use TreeExplainer: Specifically optimized for tree-based models.

Limit Samples: For both background and test samples.

Generate Global and Local Interpretations.

SHAP Explanation

- **SHAP Summary Plot**: The SHAP summary plot provides a global overview of feature importance across all samples. Each dot represents a Shapley value for a feature in a single observation. This plot helps you understand which features are most influential for the model overall. For example, if "Marketing Spend" and "Discount Percentage" are near the top, it suggests they have the largest effect on sales predictions.
- **SHAP Dependence Plot**: Each SHAP dependence plot shows the relationship between a feature's value and its SHAP value (impact on prediction). The dependence plot helps identify non-linear relationships and interactions between features. For instance, if the plot for "Discount Percentage" has a high spread of SHAP values, it suggests that discounts significantly affect predictions, especially when combined with other factors like "Units Sold."

SHAP Explanation

- **SHAP Force Plot**: The force plot provides a breakdown of feature contributions for a single prediction, showing how each feature pushes the prediction higher or lower compared to the model's expected value. This plot helps interpret individual predictions. For example, if "Marketing Spend" has a large red bar, it means a high marketing spend pushed the prediction significantly higher for that instance.

Implementing LIME

- **Load and Preprocess Data:** Load the data and perform basic cleaning.
- **Train a Model:** Use RandomForestRegressor as our model.
- **LIME Explanations:** Use LIME to explain predictions for multiple instances in the test set.
- **Save LIME Explanations:** Save explanations to HTML files for easy viewing and sharing.
- **LIME Explanations for Multiple Instances:** By running LIME on multiple instances (e.g., looping through the first 5 samples), you can observe patterns across instances.
- **Comparison Across Predictions:** Analyzing LIME explanations across multiple predictions helps identify common influential features for certain types of predictions.

Interpreting LIME Results

LIME Local Explanation for a Single Prediction

- **Interpretation**: Each LIME explanation shows feature contributions to a single prediction by locally approximating the model around that instance.
- **Feature Contributions**: The LIME plot displays each feature’s contribution to the prediction, either positively or negatively.
- **Order of Features**: Features with the most influence are listed at the top, with orange bars showing positive contributions and blue bars showing negative contributions.
- **Key Insights**: This explanation provides a local understanding of why a specific prediction was made. For instance, if “Units sold" and "Product Category" have the longest bars, they were the most significant factors in that particular prediction.

How to Use These Insights

- **Identify Key Drivers**: Use SHAP's summary and dependence plots to understand the most influential features globally and how they interact.
- **Refine Strategies**: If "Marketing Spend" and "Discount Percentage" consistently drive sales predictions, you might want to focus marketing efforts on optimizing these factors.
- **Investigate Individual Predictions**: LIME helps explain outliers or unexpected predictions, allowing you to understand and potentially correct unexpected behaviors in the model.
- These tools together provide both a big-picture and instance-level view, giving you actionable insights into how your model interprets the retail dataset.
- The ultimate aim of optimizing retail prices is to charge a price that helps you make the most money and attracts enough customers to buy your products. It involves using data and pricing strategies to find the right price that maximizes your sales and profits while keeping customers happy.