Retail Forecast

Name: Olena Panchenko
Email: hel.panchenko@gmail.com
Country: United Kingdom
Specialisation: Data Science
Github repo:
https://github.com/helenapanchenko/DS_Internship_Glacier/tree/master/Week10

Problem Description **1**
EDA **1**
Introduction 1
Data Visualisation 2
Correlation Analysis 2
Multicollinearity Analysis 3
Model Selection 3
Conclusion 3

Problem Description

The large company is in the beverages business in Australia. They sell their products through various supermarkets and engage in heavy promotions throughout the year. Various factors like holidays and seasonality also influence their demand. They needed a forecast of each of the products at the item level every week in weekly buckets.

EDA

Introduction

This document outlines the Exploratory Data Analysis (EDA) performed on a forecasting_case_study dataset and recommends a final model for predicting item level. The purpose of EDA is to understand the data structure, examine potential relationships between variables, identify patterns or anomalies, and ultimately support the development of a suitable model for weekly demand forecasting at the item level. Data Overview and Data Description steps were performed earlier.

Data Visualisation

The initial data visualization phase provides insights into the distribution and variability of demand across different items, temporal trends, and the effects of promotional and seasonal factors. Key visualizations include:

Demand Trends Over Time: Plots show how demand fluctuates week by week, revealing clear seasonal patterns where demand rises and falls at regular intervals. These trends suggest that demand could be influenced by external factors such as weather, holidays, or planned promotions.

Demand Distribution by Product: Demand distribution across items shows high variability, with some items being high in demand and others not as much.

Promotional Impact Analysis: Comparing demand during promotional periods versus non-promotional periods indicates that promotions drive increases in sales for many items. This insight suggests that promotional data will be valuable as a feature in the forecasting model.

Seasonal and Holiday Effects: Demand spikes aren't evident around holidays.

Correlation Analysis

Correlation analysis reveals relationships between demand and various features, such as promotions, holidays, and item-level attributes. Key findings include:

The correlation analysis reveals that several variables, including V_DAY, EASTER, CHRISTMAS, WeekOfMonth, Month, and Quarter, show very low or near-zero correlations with Sales. This suggests they do not have a meaningful relationship with Sales and likely add minimal predictive power, making them candidates for removal.

Additionally, Quarter, Month, and Year exhibit high inter-correlations (above 0.9) due to overlapping information, which could introduce redundancy in the model. The Covid_Flag and Google_Mobility variables also show little to no correlation with Sales, although Covid_Flag is moderately correlated with Google_Mobility. Given that Covid_Flag may have had an impact on Sales during specific periods, it will be retained as a potential substitute for the Year variable to capture any pandemic-related effects.

Among other features, In-Store Promo and Store End Promo have a moderate correlation (around 0.5), indicating some relationship but not to the extent that it requires adjustment or removal.

Finally, for the Product columns (e.g., Product_SKU1, Product_SKU2, etc.), while correlations with Sales vary—some SKUs show high correlations, and others have low or even negative correlations—all are retained. Retaining these product-specific features allows the model to learn distinct sales patterns across SKUs and differentiate which products contribute more or less to overall sales.

Multicollinearity Analysis

Multicollinearity analysis ensures that features are independent enough to avoid redundancy and instability in the model. Using Variance Inflation Factor (VIF) calculations, features with high multicollinearity are identified. In our case, no columns were removed at this stage as all the scores were relatively low.

Model Selection

Based on the insights gathered, several machine learning models can be considered for forecasting demand at the item level. Although traditional time-series models are useful for temporal data, these models are not necessary if a time-independent approach is preferred. Below are some recommended non-time-series machine learning models suitable for demand forecasting based on this dataset's characteristics:

1. **Random Forest Regressor**: This model can handle complex relationships and nonlinear interactions between features like promotions and holidays. Random Forests are robust to overfitting, especially with a refined feature set, and can provide variable importance, helping interpret the effect of promotions or holidays.
2. **Gradient Boosting Machines (GBM)**, such as **XGBoost** or **LightGBM**: These models are powerful for capturing intricate patterns in data with minimal tuning. Given the correlation between demand and specific features like promotions, a gradient-boosting approach can focus on the most predictive features and yield high accuracy.
3. **Linear Regression with Regularization** (e.g., Ridge or Lasso Regression): For simpler demand patterns, a linear regression model with Lasso or Ridge regularization can control for any residual multicollinearity while providing interpretable coefficients. This may be suitable for items with more consistent demand patterns.
4. **Support Vector Machines (SVM) with RBF Kernel**: SVMs can be effective for forecasting demand, particularly if the relationship between features and demand is non-linear. The RBF kernel allows the model to handle complex interactions, making it suitable for cases with varying demand responses to promotions or other events.

Conclusion

The EDA process has yielded valuable insights into the structure and behaviour of the demand for the company's products. By identifying high-impact features such as promotions and seasonal indicators and removing multicollinear or redundant variables, the analysis has laid the groundwork for an effective demand forecasting model. Based on the insights from data visualization, correlation, and multicollinearity analysis, the recommended models include Random Forest, XGBoost, and Ridge or Lasso Regression, all of which can leverage the refined feature set to predict item-level demand accurately and interpretably. This approach should support the company's goal of improving weekly demand forecasts, enhancing inventory management, and meeting customer needs efficiently.