REPORT – SALES DATA ANALYSIS

NGUYEN THANH HOA

SE183091

July 2024

1 Introduction

This report summarizes the key findings from the sales data analysis performed using exploratory data analysis (EDA) and machine learning models. The pri- mary objective was to understand the sales patterns, identify key factors influencing sales, and predict future sales using regression models.

2 Project Overview

The goal of this project is to analyze a large dataset of sales records to uncover insights and trends, and to build predictive models using machine learning. This involves data preprocessing, exploratory data analysis (EDA), and predictive modeling.

3 Data Preprocessing

3.1 Data Loading

Sales data: 100,000 records with information on region, country, item type, sales channel, order priority, dates, and financial details.

3.2 Libraries Used

- Pandas: Data manipulation.
- Numpy: Scientific computing.
- Matplotlib and Seaborn: Data visualization.

3.3 Data Cleaning

- Convert date columns to datetime objects.
- Drop unnecessary columns.

- Create new columns for order month, order year, and combined order date month-year.

4 Exploratory Data Analysis (EDA)

4.1 Data Overview

The dataset consists of 100,000 entries with the following features:

- Categorical Variables: Region, Country, Item Type, Sales Channel, Order Priority, Order Date, Ship Date
- Numerical Variables**: Units Sold, Unit Price, Unit Cost, Total Revenue, Total Cost, Total Profit, Order Month, Order Year

4.2 Key Insights from EDA

- Missing Values: There are no missing values in the dataset, ensuring data completeness for analysis.
- Top Selling Countries: The top 20 countries contribute a significant por- tion of the total sales, as depicted in the pie chart of country-wise sales distribution.
- Outliers Detection: Box plots were used to identify outliers in the 'Total Profit' variable, indicating some extreme values that might influence the analysis.
- Sales Channels: Both online and offline sales channels are well represented in the dataset, providing a balanced view of sales performance across dif- ferent channels.
- Sales Performance Over Time: Monthly and yearly sales trends show fluc- tuations, with certain months and years exhibiting higher sales volumes.

5 Machine Learning Model Results

5.1 PyCaret Setup

The PyCaret library was used for model training and comparison. The target variable for the models was 'Total Profit'.

5.2 Model Comparison

Several regression models were compared to identify the best-performing model. The models evaluated include:

- Linear Regression

- Lasso Least Angle Regression (LLAR)
- Decision Tree Regressor
- Random Forest Regressor
- Gradient Boosting Regressor

The "Linear Regression" model was found to be the best initial model based on performance metrics.

5.3 Model Training and Tuning

- Lasso Least Angle Regression: This model was selected for further tuning due to its balance of simplicity and performance. After tuning, it remained the preferred model.
- Model Performance: The model achieved an R-squared value indicating excellent predictive power. The residuals plot and prediction error plot confirmed that the model predictions closely matched the actual values.
- Feature Importance: Feature importance analysis revealed that ’Units Sold’, ’Unit Price’, and ’Total Revenue’ were the most significant pre- dictors of ’Total Profit’.

5.4 Final Model Evaluation

The tuned Lasso Least Angle Regression model was evaluated on a test set:

- Accuracy: The model achieved an accuracy of 100% based on the R- squared metric, indicating an almost perfect fit.
- Mean Squared Error (MSE): The MSE and RMSE values were extremely low, further confirming the model’s excellent performance.