Sales and Profit Prediction Report

Executive Summary

This report presents a comprehensive analysis of sales data and the development of predictive models for sales and profit forecasting. The project demonstrates how machine learning techniques can be applied to predict business metrics based on historical sales data. Two models were developed: a Random Forest Regressor for profit prediction and an XGBoost Regressor for sales prediction. Both models achieved reasonable accuracy, enabling reliable forecasting for business planning.

1. Introduction

1.1 Project Objective

The primary objective of this project was to analyze sales data and develop machine learning models that can accurately predict sales and profit based on various input features such as quantity, discount, category, segment, and shipping mode.

1.2 Dataset Overview

The analysis was performed on a retail sales dataset containing information about orders, products, customers, and financial metrics. The dataset includes various attributes such as order dates, ship dates, product categories, customer segments, geographic information, and sales performance metrics.

2. Data Understanding and Preprocessing

2.1 Data Exploration

Initial exploration of the dataset revealed its structure and key characteristics:

- The dataset contains both numerical and categorical features
- Key numerical variables include Sales, Profit, Quantity, and Discount
- Key categorical variables include Ship Mode, Segment, Category, and Sub-Category

- Geographic information includes Country, Region, State, and City
- Temporal data includes Order Date and Ship Date

2.2 Data Preprocessing Steps

The following preprocessing steps were performed to prepare the data for analysis and modeling:

1. **Missing Value Analysis**: Examined the dataset for missing values
2. **Outlier Detection**: Used IQR method to identify outliers in numerical columns
3. **Duplicate Removal**: Identified and removed duplicate entries based on Order ID, Product ID, and Order Date
4. **Feature Creation**:
    - Created 'Lead Time' feature by calculating the difference between Ship Date and Order Date
    - Extracted year from Order Date
    - Created year-month combination from Order Date
5. **Feature Reduction**: Removed unnecessary columns such as Product ID, Customer Name, Row ID, Postal Code, Country, and Ship Date

3. Exploratory Data Analysis

3.1 Sales Analysis

- **Top Selling Products**: Identified and visualized the top 10 products by sales
- **Yearly Sales Trends**: Analyzed sales performance over different years
- **Category Performance**: Analyzed sales by product category
- **Regional Performance**: Examined sales distribution across different regions
- **Subcategory Analysis**: Identified top performing product subcategories

3.2 Profit Analysis

- **Category Profitability**: Identified which product categories generate the highest profit
- **Shipping Mode Analysis**: Analyzed how different shipping modes affect profitability
- **Regional Profitability**: Examined profit distribution across different regions
- **Top Profitable Products**: Identified the top 5 most profitable products
- **Monthly Profit Trends**: Analyzed profit trends on a monthly basis

3.3 Key Insights

- Technology category generates the highest revenue and profit
- Performance varies significantly across different regions
- There is a relationship between quantity purchased and both sales and profit
- Discount has a noticeable impact on profitability
- Clear seasonal patterns are visible in monthly sales and profit trends

4. Feature Engineering

4.1 Derived Features

Several new features were created to improve model performance:

- **Profit Margin**: Calculated as Profit divided by Sales
- **Discount Impact**: Calculated as Discount multiplied by Sales
- **Log Transformations**: Applied logarithmic transformation to Sales and Profit to handle skewness
- **Winsorized Profit**: Applied winsorization to handle outliers in the Profit column

4.2 Categorical Encoding

One-hot encoding was applied to categorical variables:

- Category
- Segment
- Ship Mode

5. Model Development

5.1 Model Selection

Two different algorithms were selected for the prediction tasks:

- **Random Forest Regressor** for profit prediction
- **XGBoost Regressor** for sales prediction

5.2 Data Splitting

The dataset was split into training (80%) and testing (20%) sets to evaluate model performance.

5.3 Model Training

Both models were trained on the prepared dataset:

- The Random Forest model was trained with default parameters
- The XGBoost model was trained with 100 estimators and a learning rate of 0.1

6. Model Evaluation

6.1 Performance Metrics

Both models were evaluated using standard regression metrics:

- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- R-squared (R^2)

6.2 Results

The performance metrics showed:

- Both models achieved reasonable predictive accuracy
- The models captured the general trends in the data
- The distribution of prediction errors shows that most predictions are close to actual values
- Some outliers in the data led to larger prediction errors in certain cases

6.3 Visual Evaluation

Scatter plots comparing actual vs. predicted values for both Sales and Profit show that the models performed well overall, with predictions generally following the ideal diagonal line.

7. Implementation

7.1 User Input System

A user input system was developed to allow for real-time predictions:

- Users can input Quantity, Discount, Category, Segment, and Ship Mode
- The system processes these inputs and applies the same feature engineering as during training
- The trained models then generate sales and profit predictions

7.2 Model Deployment

Both models were saved using joblib for deployment:

- The Random Forest model for profit prediction was saved as "rf_profit_model.pkl"
- The XGBoost model for sales prediction was saved as "xgb_sales_model.pkl"

8. Conclusion and Recommendations

8.1 Summary

This project successfully developed predictive models for sales and profit forecasting based on historical data. The analysis provided valuable insights into factors affecting sales and profitability, while the predictive models enable data-driven decision making.

8.2 Business Implications

The models and analysis can be used to:

- Forecast revenue and profit for business planning
- Understand the impact of discounts on profitability
- Identify high-performing product categories and regions
- Optimize pricing and discount strategies

8.3 Future Work

Several potential improvements could enhance the current system:

- Incorporate more advanced feature engineering techniques
- Experiment with different machine learning algorithms
- Implement hyperparameter tuning to optimize model performance
- Develop a more user-friendly interface for the prediction system
- Include time series forecasting components for temporal predictions

9. Appendix

9.1 Data Dictionary

- **Order ID**: Unique identifier for each order
- **Order Date**: Date when the order was placed
- **Ship Date**: Date when the order was shipped
- **Ship Mode**: Shipping method (Standard Class, Second Class, First Class, Same Day)
- **Segment**: Customer segment (Consumer, Corporate, Home Office)
- **Category**: Product category (Furniture, Office Supplies, Technology)
- **Sub-Category**: Product sub-category
- **Product Name**: Name of the product
- **Sales**: Total sales revenue
- **Quantity**: Number of units ordered
- **Discount**: Discount applied to the product
- **Profit**: Profit generated from the sale

9.2 Model Features

Features used in the final models include:

- Quantity
- Discount
- Profit_Margin
- Discount_Impact
- Log_Sales
- Encoded categorical variables (Category, Segment, Ship Mode)

9.3 Model Parameters

- **Random Forest (Profit Prediction)**:
    - Number of estimators: 100 (default)
    - Random state: 42
- **XGBoost (Sales Prediction)**:
    - Number of estimators: 100
    - Learning rate: 0.1
    - Random state: 42