Retail Store Sales Forecasting Using Machine Learning

COMSATS UNIVERSITY ISLAMABAD

Muhammad Fakhir Zahoor

1/08/25

Tools: *Scikit-learn, Python, pandas , gradio*

Table of Contents

Abstract: 3

Introduction: 3

Problem Statement: 3

Methodology: 3

Model Implementation: 4

Key Steps: 4

Evaluation Metrics: 4

Visualization: 5

Correlation Heatmap: 5

Actual vs predicted Plot: 5

Architecture Diagram: 5

Conclusion: 6

References: 6

Abstract:

This project presents a machine learning-based sales forecasting model designed for a retail environment. Leveraging a comprehensive dataset containing approximately 73,100 records, the model uses key features such as inventory level, demand forecast, product category, store and region identifiers, discount, and price to predict unit sales. Advanced preprocessing techniques—including one-hot encoding for categorical variables and feature engineering—were employed to ensure the dataset was ready for modeling.

A **Random Forest Regressor** was selected due to its robustness and high performance on tabular data. The model was evaluated using standard regression metrics such as **Mean Absolute Error (MAE), Root Mean Squared Error (RMSE**), and **R-squared (R^2)** to measure its accuracy and generalizability. The model demonstrated strong predictive capabilities with minimal error, making it suitable for practical use.

To make the model easily accessible, a **Gradio**-based web interface was developed, allowing users to input product and store-related features and receive real-time predictions for unit sales. This interface simplifies interaction with the model, making it a valuable tool for retail decision-makers and analysts.

Introduction:

Sales forecasting is a critical component in the retail industry, playing a key role in inventory control, pricing strategies, and demand planning. Accurate forecasts help businesses maintain the right stock levels, reducing the risks associated with overstocking or understocking. This not only optimizes storage and supply chain operations but also contributes to better profit margins and improved customer satisfaction. Effective sales forecasting empowers retailers to make informed decisions, respond quickly to market trends, and enhance overall operational efficiency.

Problem Statement:

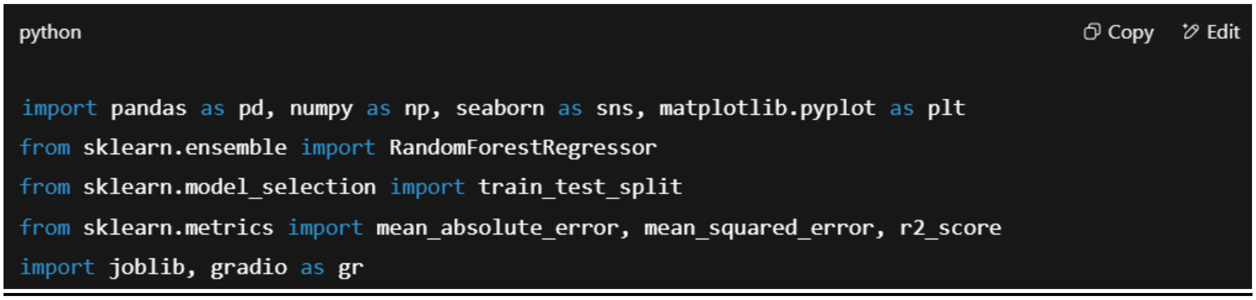
Retail businesses often struggle to accurately predict future sales due to unpredictable factors like holidays, weather, or competitor pricing. This project aims to build a model that accurately forecasts Units Sold using historical data and multiple influencing factors.

Methodology:

The dataset used for this project was titled **retail_store_inventory.csv**, containing 73,100 rows and 15 columns. Initial data cleaning revealed no duplicate or missing values, ensuring the dataset's integrity. The **Date** column was removed as it did not contribute meaningfully to the prediction task.

During feature engineering, several categorical features were identified, including Store ID, Product ID, Category, Region, Weather Condition, Holiday/Promotion, and Seasonality. These variables were transformed using one-hot encoding to make them suitable for machine learning models. The predictive model chosen was the **RandomForestRegressor** from the **sklearn.ensemble** library. The data was split into training and testing sets using an 80-20 ratio. The model was trained with 100 estimators and a fixed random state of 42 to ensure reproducibility. Finally, the trained model was deployed using a user-friendly Gradio interface, allowing real-time sales predictions based on user input.

Model Implementation:



```python
import pandas as pd, numpy as np, seaborn as sns, matplotlib.pyplot as plt
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score
import joblib, gradio as gr
```

➢ Key Steps:

- Encoded data using pd.get_dummies()
- Split features (X) and target (y)
- Fit model and save using joblib
- Created a Gradio UI for user inputs and predictions

Evaluation Metrics:

Metrics	**Value**
MAE(Mean Absolute Error)	7.26
MSE(Mean Squared Error)	73.27
RMSE(Root MSE)	8.56
R^2 Score	0.99

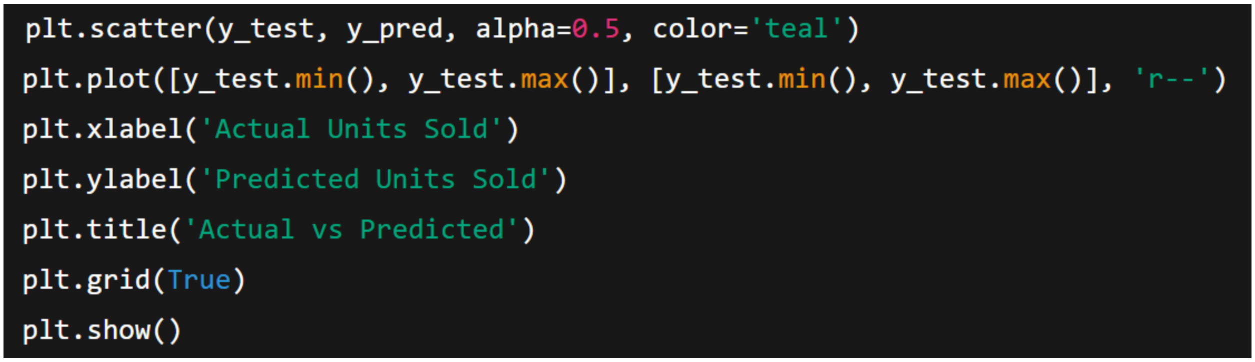
The R² score of 0.99 indicates the model captures 99% of the variance in the target variable, showing excellent predictive power.

Visualization:

- ### Correlation Heatmap:

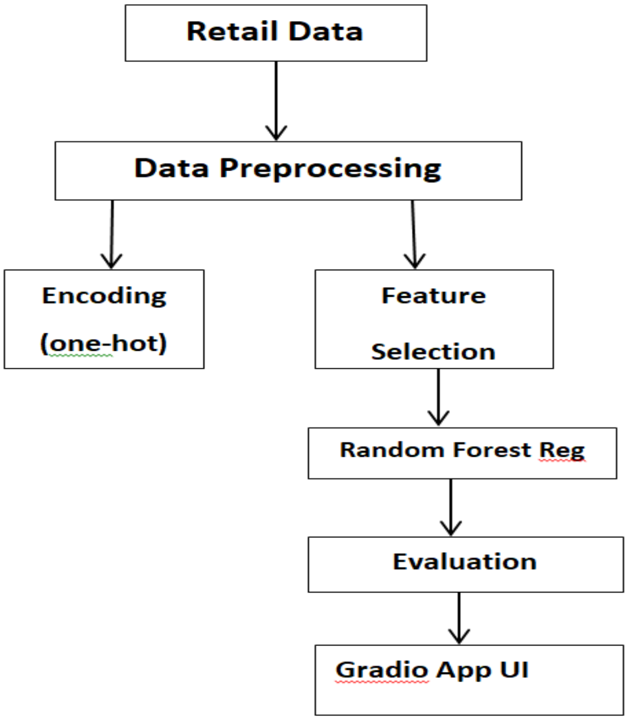
  Shows strong correlation between Units Sold and Demand Forecast, Inventory Level.

- ### Actual vs predicted Plot:



```
plt.scatter(y_test, y_pred, alpha=0.5, color='teal')
plt.plot([y_test.min(), y_test.max()], [y_test.min(), y_test.max()], 'r--')
plt.xlabel('Actual Units Sold')
plt.ylabel('Predicted Units Sold')
plt.title('Actual vs Predicted')
plt.grid(True)
plt.show()
```

Architecture Diagram:



Conclusion:

This project demonstrates how machine learning can significantly enhance sales forecasting in retail. With a highly accurate Random Forest model ($R^2 = 0.99$), and a user-friendly Gradio app, this tool can help business stakeholders make informed decisions regarding inventory and promotions.

References:

1- Scikit-learn. (n.d). *Machine Learning in python*. Retrieved from https://scikit-learn.org

2- Gradio. (n.d). *Build Machine Learning interfaces.* Retrieved from https://gradio.app

3- Waskom, M. L. (2021). *Seaborn: Statistical Data Visualization. Journal of Open Source Software, 6(60), 3021.*