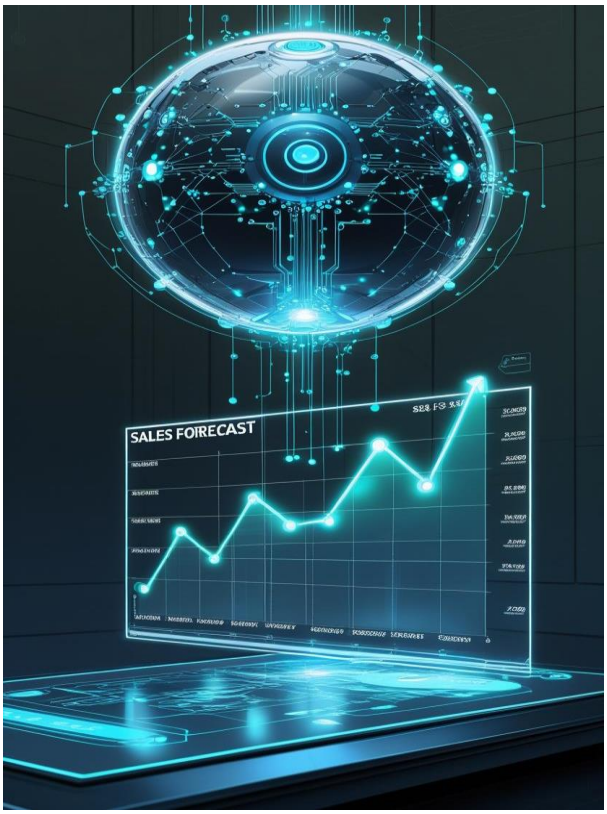

WAREHOUSE AND RETAIL SALES FORECASTING

Submitted to Gateshead College in completion of Skills for Success Bootcamp

ABSTRACT

This project aimed to develop a robust, accurate, and interpretable pipeline for forecasting monthly warehouse sales, leveraging historical sales data. By following industry best practices in data cleaning, feature engineering, and machine learning, we produced a model that supports improved inventory planning and business strategy.

Nnaemeka Stephen Nnamani

AI and Machine Learning BootCamp – Level 5

Introduction

In this project, we delve deep into the Warehouse and Retail Sales (Liquor and Alcohol) dataset from the Department of Liquor Control (DLC) in Montgomery County, MD in the United States, the dataset is a public dataset available at the dataMontegomery Repository. This dataset contains a list of sales and movement data by item and department appended monthly. Our primary objective is to amplify the efficiency of sales in either the warehouse or retail stores and provide insight to suppliers where to focus their distribution for maximum sales.

This project aimed to develop a robust, accurate, and interpretable pipeline for forecasting monthly warehouse sales, leveraging historical sales data. This segmentation will allow us to understand the top products and where they sell more. Building upon this, we intend to develop a forecasting system that will predict warehouse sales for top suppliers within each segment of products, ultimately enhancing supply-chain efficiency and fostering increased sales.

Project Objectives

- **Forecast warehouse sales** at the monthly and supplier level.
- **Identify key drivers of sales** (supplier, item type, seasonality).
- **Deliver actionable predictions** to guide operational and purchasing decisions.

Data Description

Source: Warehouse and retail sales transaction records (307,645 rows, 9 columns) is a dataset from Montgomery County Department of Liquor Control in the state of Alabama in the United of America. The dataset is updated monthly and is available at dataMontegomery public repository. It contains a list of sales and movement data by item and departments.

The dataset has the following features:

Features	Description
YEAR	Calendar Year
MONTH	Month
SUPPLIER	Supplier Name
ITEM CODE	Uniquely assigned code to distinguish each item
ITEM DESCRIPTION	Description of each item
ITEM TYPE	The type of the item (Wine, Beer, Liquor, and Tote
RETAIL SALES	Cases of product sold from DLC dispensaries
RETAIL TRANSFERS	Cases of products transferred to DLC dispensaries.
WAREHOUSE SALES.	Cases of product sold to MC licensees.

Approach and Methodology

We explored overall sales trends, top suppliers/items, seasonality, and data quality (duplicates, missing values, and outliers). TITO'S HANDMADE VODKA - 1.75L came out on top for the top item sold by retailers while CORONA EXTRA LOOSE NR - 12OZ is the top item sold from the warehouse. E & J GALLO WINERY is the top retailer and CROWN IMPORTS is the top warehouse seller. Overall warehouse sales are much higher than retail sales from our time series trends. Retail sales and Retail transfer are strongly correlated.

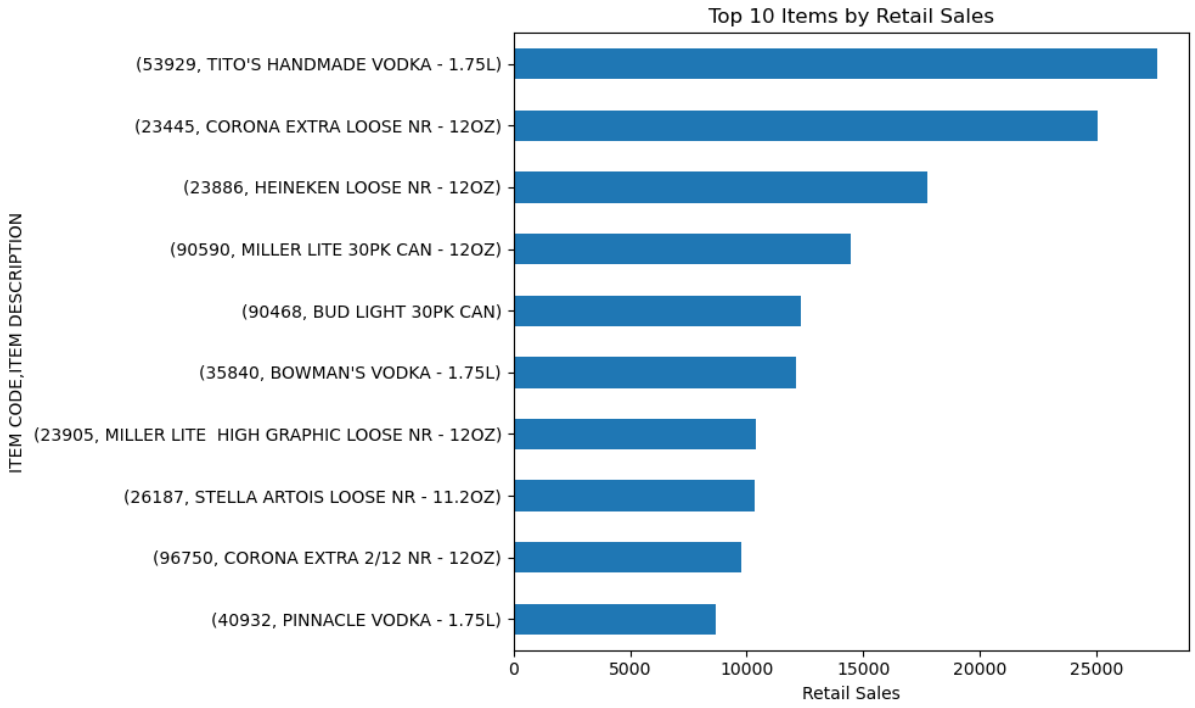

Fig. 1.0: Top 10 Items Sold by Retailers

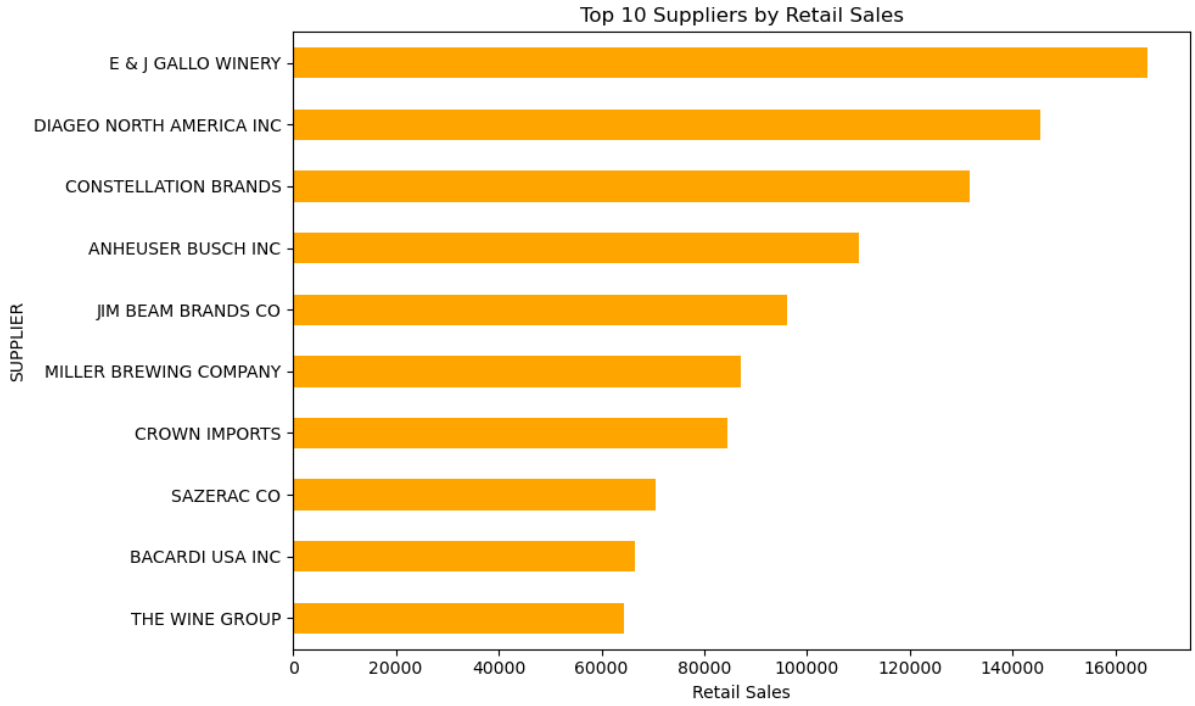

Fig. 2.0: Top 10 Retail Suppliers

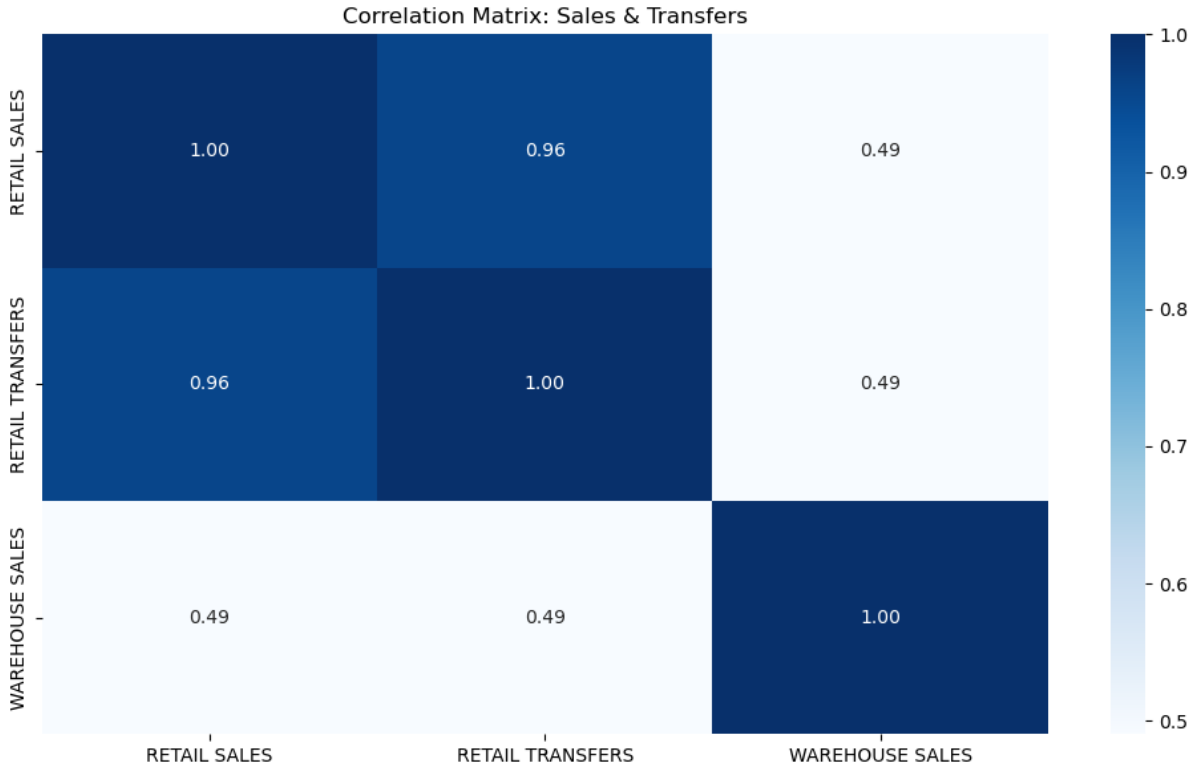

Fig. 3.0: Correlation Analysis

The dataset has no duplicate entries but has some missing and negative values which were removed. We aggregated sales by month, supplier, and item type, using one-hot encoder we encoded categorical data (supplier and item type) for our model.

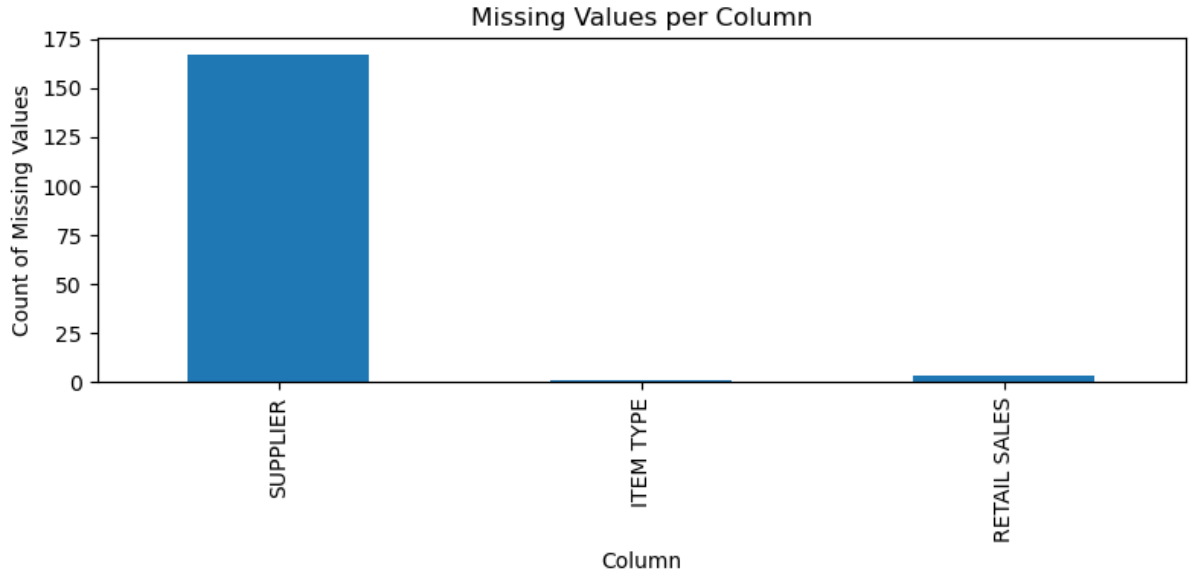

Fig. 4.0: Missing Values

Model Selection and Evaluation

Compared Linear Regression and Random Forest models and Random Forest performed better than Linear Regression. The Random Forest model was tuned with GridSearchCV for optimal accuracy and generalizability. We evaluated our model using RMSE, MAE, and R^2 on both training and test data.

We also performed a visual inspection of our model plotting the predicted vs actual warehouse sales.

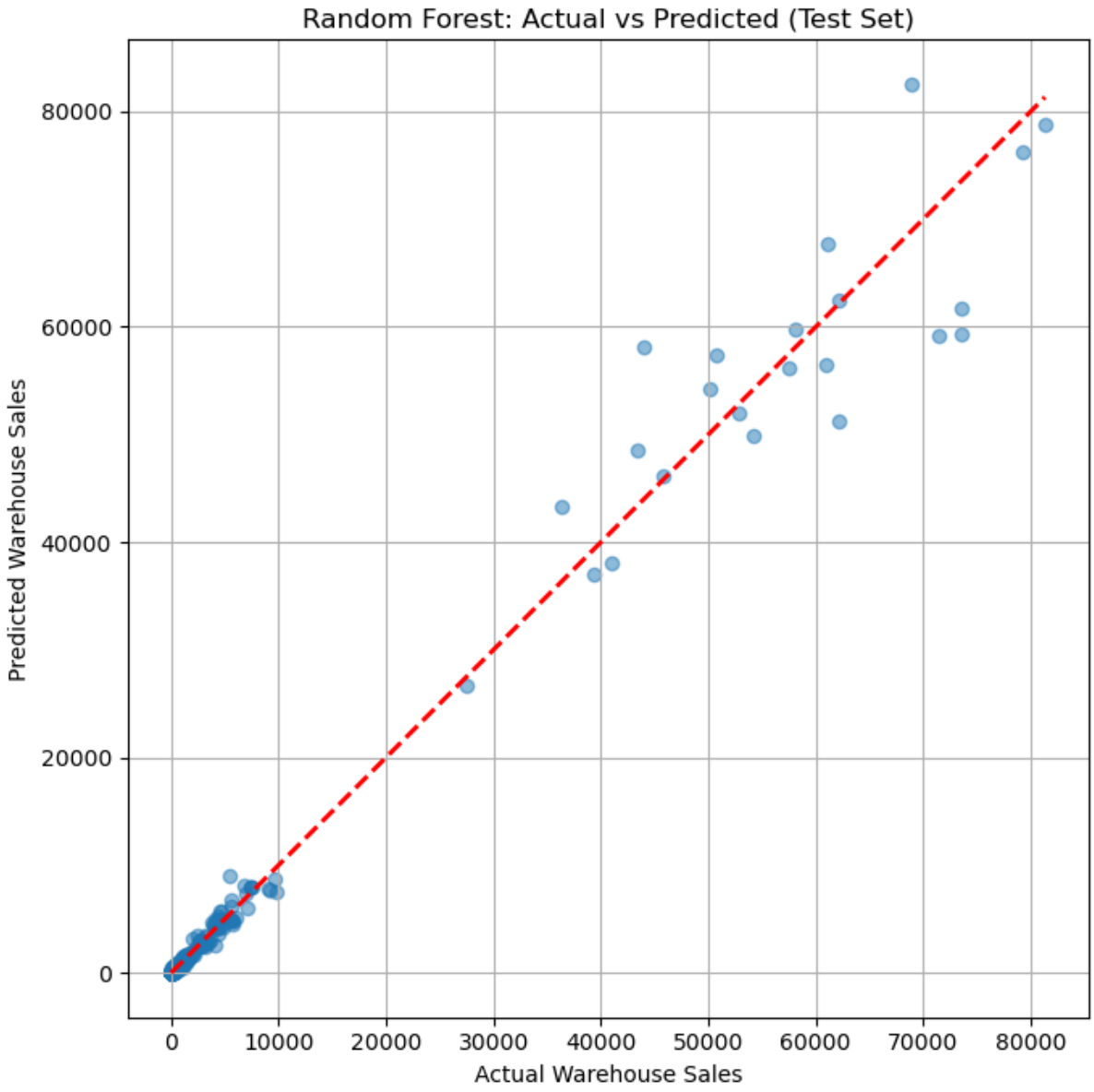

Fig. 5.0: Actual vs Predicted Warehouse Sales

Key Results

The tuned Random Forest Regressor provided the best model with test R^2 = 0.98 and test RMSE = 809, the model shows no significant overfitting; test and train metrics are both excellent. Feature Importance; model highlights the influence of specific suppliers, item types, and time (month, year) on sales.

Conclusion

We have been successfully explored our dataset to uncover trends in the retail and warehouse sales, discover and handle anomalies in our dataset. We engineered some new features by converting the categorical features to numerical feature for our model consumption. We successfully trained and compared Linear Regression and Random Forest

models and the Random Forest gave the best result and even better result when tuned with GridSearchCV.

In the end our model could predict future warehouse sales with up to 98% accuracy with no significant overfitting on both the train and test data.

Recommendations

- Adopt the tuned Random Forest model for ongoing warehouse sales forecasting.
- Retrain regularly as new data becomes available.
- Monitor for data quality issues and address them promptly in future data ingestions.
- Consider incorporating additional business features (promotions, store locations, external events) in future modelling.
- Integrate predictions into operational dashboards for real-time decision support.