Project Report: BlinkIt Sales Analytics

Table of Contents

1. Introduction

The BlinkIt Sales Analytics Project aims to analyze historical sales data across BlinkIt's retail outlets. By identifying trends driven by various factors such as outlet tier, size, item type, and fat content, this project provides actionable insights to optimize sales strategies, inventory management, and resource planning. The project has culminated in the development of a Power BI dashboard and a Machine Learning model that enhances BlinkIt's retail strategy and complements existing reporting tools.

2. Project Overview

2.1 Objective

The primary objective of the BlinkIt Sales Analytics Project is to leverage historical sales data to predict sales trends across BlinkIt's outlets. This involves analyzing various influential factors including:

- Outlet Tier: Tier 1, Tier 2, Tier 3
- Outlet Size: Small, Medium, High
- Item Type: Snack Foods, Dairy, etc.
- Fat Content: Low Fat, Regular

The insights derived from this analysis will enable BlinkIt to optimize its sales strategies, inventory allocation, and resource planning effectively.

2.2 Data Sources

The analysis is conducted using the dataset which consists of 8,523 rows and 12 columns, including:

- Item Fat Content

- Item Identifier

- Item Type

- Outlet Establishment Year

- Outlet Identifier

- Outlet Location Type

- Outlet Size

- Outlet Type

- Item Visibility

- Item Weight

- Sales

- Rating

The raw dataset is processed for advanced predictions.

3. Data Cleaning and Preparation

3.1 Initial Data

The initial dataset, ‘BlinkIt Grocery Data.xlsx’ had several inconsistencies including missing values in ‘Item Weight’ and ‘Item Visibility’ and potential duplicates or outliers in ‘Sales’. A comprehensive data cleaning process was necessary to ensure the accuracy of the data for analysis and modelling.

3.2 Cleaning Process

To address the identified issues:

- Missing Values: Item Weight and Item Visibility were imputed using their respective medians to maintain dataset completeness.

- Outlier Treatment: Outlier values in Sales were capped using the Interquartile Range (IQR) method to maintain a skewness of 0.1272. This involved clipping values to the range of [clip(lower=Q1 - 1.5 * IQR, upper=Q3 + 1.5 * IQR)].

After cleaning the dataset was saved as 'cleaned_data.csv'.

3.3 Feature Engineering

From 'cleaned_data.csv' the 'featured_data.csv' was created, increasing the number of columns to 18. The new features included:

- Avg_Sales_per_Outlet: Average sales per outlet to contextually analyze performance.

- Years_Since_Establishment: Calculated as 2025 - Outlet Establishment Year to gauge outlet age.

- Visibility_per_Weight: A ratio of Item Visibility to Item Weight, adjusted for zero values.

- Sales_per_Outlet_Type: Average sales by outlet type to analyze performance trends.

- Visibility_x_ItemType & Weight_x_ItemType: Interaction effects normalized by dataset size, capturing key relationships for predictive modelling.

These features significantly enhanced the predictive power of the machine learning model.

4. Data Analysis and Modelling

4.1 Machine Learning Model Development

A Random Forest Regressor was developed using Scikit-learn (version 1.3.2) to predict sales outcomes. The model was trained on 'featured_data.csv' utilizing a total of 19 features, including:

- 13 one-hot encoded categorical variables.

- 6 numerical features.

The model hyperparameters were optimized, achieving the following performance metrics:

- R^2 : 0.6142 (explaining 61.42% of sales variance).

- MAE: 26.16 (indicating average prediction errors).

- MSE: 1528.41 (reflecting low variance in large errors).

5-fold cross-validation yielded a mean R^2 of 0.6069, confirming the model's generalization capability.

4.2 Trends and Insights

Key trends identified through the analysis include:

- Outlet Performance: Outstanding sales from OUT049 and OUT027, whereas OUT010 and OUT019 showed weaknesses, indicating potential stock or visibility issues.

- Outlet Size Gaps: Medium outlets underperformed the predicted sales by approximately $4,000 in total, indicating areas for improvement.

- City Tier Potential: Tier 3 outlets showed the highest predicted sales, suggesting a need to prioritize resources in this tier.

- Item Categories: Fruits and Vegetables along with Snack Foods predicted to lead in sales, indicating strategic inventory focus.

- Fat Content Trends: A significant shift towards Low Fat items was noted, with predicted sales of ~$750K compared to Regular items at ~$400K.

4.3 Visualizations

Interactive visualizations were created in Power BI from the processed data, presenting:

- Sales by outlet tier.
- Sales categorized by fat content.
- Additional charts for outlet sizes and item types.

Standalone Python scripts were additionally created using Matplotlib and Seaborn for visualizations such as:

- ‘outlet_sales_comparison.png’
- ‘predicted_sales_by_tier.png’

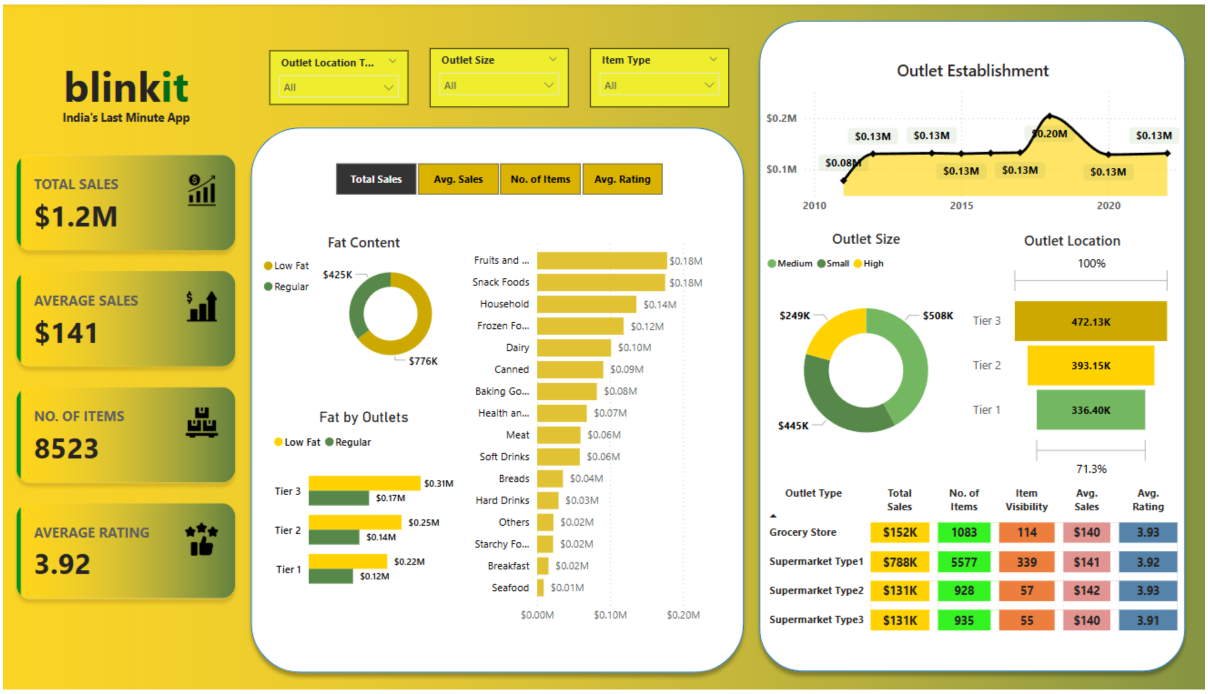
This offered predictive visuals comparing actual vs. predicted sales across different dimensions.

5. Implementation and Deployment

5.1 Power BI Dashboard Integration

A Power BI dashboard was constructed from the processed data, featuring 6 interactive charts that powerfully visualize historical sales trends. These charts highlight key insights, such as sales by outlet tier, size, and item type, enabling BlinkIt to explore patterns, identify performance drivers, and make informed strategic decisions based on robust, user-friendly visualizations.

Here is an attached image of the dashboard for the reference:



Outlet Type	Total Sales	No. of Items	Item Visibility	Avg. Sales	Avg. Rating
Grocery Store	$152K	1083	114	$140	3.93
Supermarket Type1	$788K	5577	339	$141	3.92
Supermarket Type2	$131K	928	57	$142	3.93
Supermarket Type3	$131K	935	55	$140	3.91

5.2 Standalone Prediction Tools

To complement the dashboard effectively:

- Prediction Reports: Sales predictions were generated in two formats:

- Detailed reports with approximately 8,523 rows per Item Identifier-Outlet Identifier pair.

- Aggregate reports by Outlet Identifier for strategic planning.

- Insight Graphs: Key insights were saved as PNG files, illustrating trends like predicted sales by item type.

The model (random_forest_final.pkl) is hosted on Google Drive, with link and setup steps accessible on GitHub (excluding the model itself due to size limits).

6. Conclusion

The BlinkIt Sales Analytics Project successfully analyzed historical sales data and predicted future performance using a Random Forest model with an R² of 0.6142. Interactive visualizations through Power BI and standalone reports enriched BlinkIt's strategic insights for inventory, promotions, and resource allocation. The project's materials are hosted on GitHub and Google Drive, ensuring accessibility and utility for ongoing analysis.

7. Limitations and Future Work

Model Limitations

- The R^2 value indicates that 38.58% of the variance remains unexplained, potentially due to external factors like promotions or seasonal variations not captured in the dataset.
- The MAE of 26.16 suggests room for improvement in prediction accuracy.

Future Work

Future iterations should consider integrating seasonal data, promotional information, and external factors to enhance prediction accuracy. Additionally, refining reports to focus on item-level insights and validating results against dashboard trends could further improve effectiveness.

8. References

- Libraries utilized: Scikit-learn (1.3.2), Pandas (2.1.4), NumPy (1.26.4), Matplotlib, Seaborn.
- GitHub Repository: https://github.com/KrishChopra69/BlinkIt-Analytics).
- Google Drive Model Link: https://drive.google.com/file/d/1EwigIAW-OZHuGLRtNSUjkAu27gJzINsW/view?usp=sharing