Technical Report: Final Project
Sales Forecasting

Ayman Mushtaq Ahmad, Amisha Tiwari, Ronhit Neema
Khoury College of Computer Science
ahmad.ay@northeastern.edu

November 24, 2024

Contents

1 Introduction

The Sales Forecasting Project aims to accurately predict sales using machine learning techniques. This report documents the steps taken, including data preprocessing, feature engineering, model implementation, and evaluation.

2 Methodology

2.1 Data Loading and Exploration

The dataset was loaded using Pandas and explored to understand its structure:

- Null and duplicate values were identified and handled.
- Summary statistics of numerical features were generated.
- Categorical features were identified for encoding.

2.2 Feature Engineering

- Categorical variables were encoded using `LabelEncoder`.
- Features were standardized using `StandardScaler` for models requiring normalization.

2.3 Model Implementation

The following machine learning models were implemented:

- **Linear Regression:** A simple baseline model for comparison.
- **Lasso and Ridge Regression:** Regularized linear models to handle multi-collinearity.
- **Decision Tree Regressor:** Captures non-linear relationships.
- **Random Forest Regressor:** An ensemble method for improved accuracy.
- **Extra Trees Regressor:** Another ensemble model focusing on feature importance.

2.4 Evaluation

Models were evaluated using the R^2 score, which measures the proportion of variance explained by the model.

3 Results

- **Linear Regression:** $R^2 = 0.72$
- **Lasso Regression:** $R^2 = 0.68$
- **Ridge Regression:** $R^2 = 0.71$

- **Decision Tree:** $R^2 = 0.85$
- **Random Forest:** $R^2 = 0.89$
- **Extra Trees:** $R^2 = 0.88$

Random Forest achieved the highest accuracy, demonstrating its effectiveness in handling complex data relationships.

4 Discussion

The results highlight the benefits of ensemble methods for sales forecasting. While linear models provide a baseline, tree-based models excel in capturing non-linear patterns. Feature standardization and encoding were crucial for model performance.

5 Conclusion

This project successfully implemented multiple models for sales forecasting. We are working on working with the best model and proceed with the remaining aspects of our Project Scope.

6 References

A Appendix A: Code

B Appendix B: Additional Figures