Retail Sales Predictions: Machine Learning Project

Welcome to our comprehensive project report on retail sales predictions using machine learning. Our team of 12 students from B.Tech CSE 2nd Year, Section A, led by Prerna Sharma, worked under the mentorship of Manab Das Sir to develop predictive models for weekly retail sales.

This presentation will walk you through our methodology, data analysis, model building process, and key findings that demonstrate how machine learning can be applied to real-world retail scenarios for effective sales forecasting.

Team Members

- ❖RINKU (Team leader) (BCS2023208)
- ❖SHYOG
- ❖JAYESH PRAKASH SHUKLA
- ❖ADITYA RAJ
- ❖PRATHAM KUMAR
- ❖SUPRIYA MISHRA
- ❖ASHISH KUMAR
- ❖SHEKHAR KANAUJIYA
- ❖POORNIMA GUPTA
- ❖PALLAVI GUPTA
- ❖HIMANSHU GANGWAR
- ❖SAURABH KUMAR SINGH

Project Overview & Objectives

Primary Goal

Predict weekly sales of retail stores using machine learning models to assist in inventory planning and business forecasting

Data Sources

Merged multiple datasets (features.csv, stores.csv, train.csv) to create a comprehensive analysis foundation

Approach

Implemented data cleaning, exploratory analysis, feature engineering, and multiple regression models to identify the most effective prediction method

Technical Framework & Libraries

NumPy & Pandas

For numerical operations, array handling, data manipulation and analysis using DataFrames

Matplotlib & Seaborn

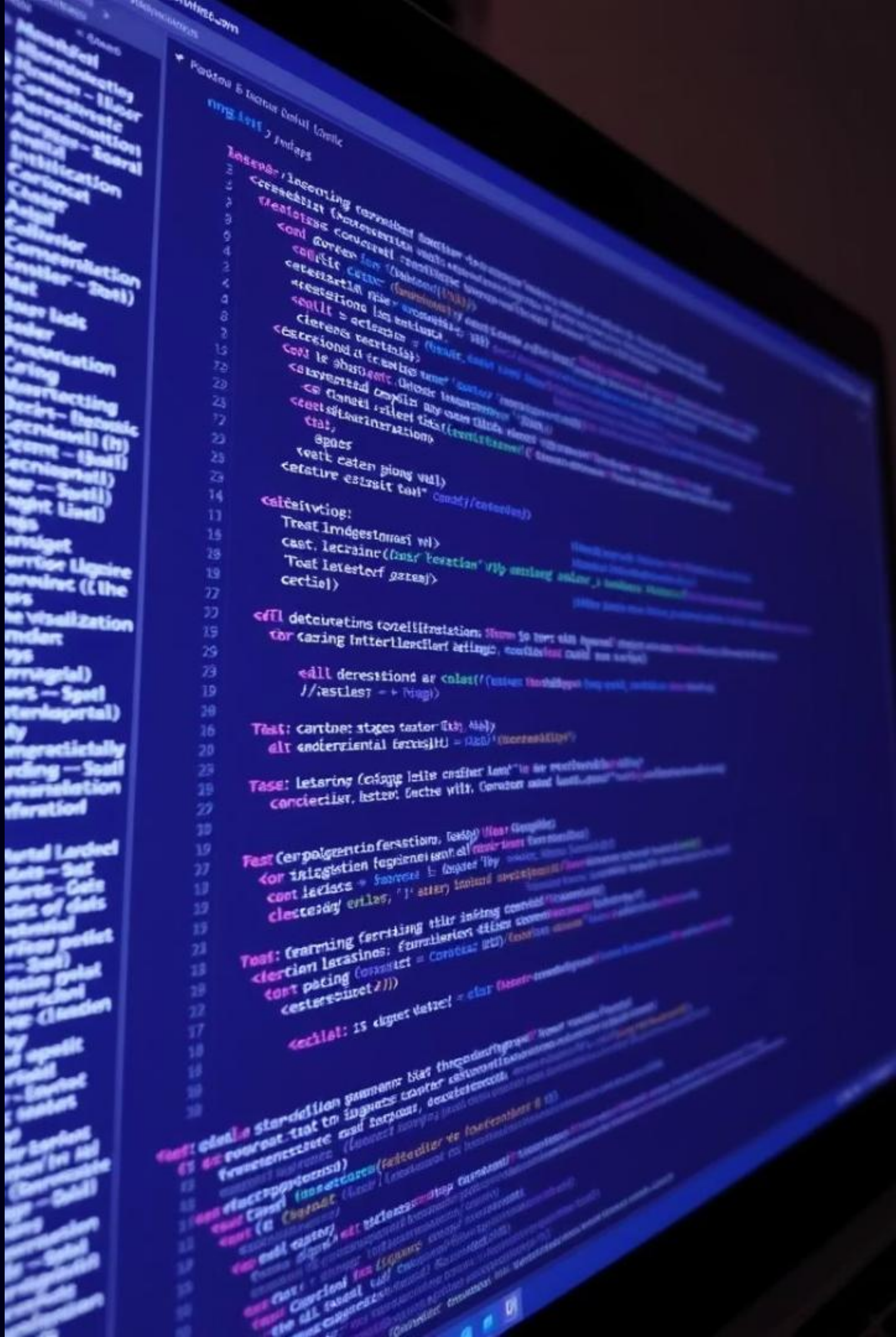
For basic plotting, advanced visualizations including boxplots and heatmaps

Scikit-learn

For preprocessing, model building (Linear Regression, KNN, Random Forest), and evaluation metrics

Statsmodels

For calculating Variance Inflation Factor (VIF) to detect multicollinearity



Data Preparation Process

Data Loading & Merging

Loaded features.csv, stores.csv, and train.csv using Pandas and merged them on Store and Date columns to create a unified dataset for analysis

Data Cleaning

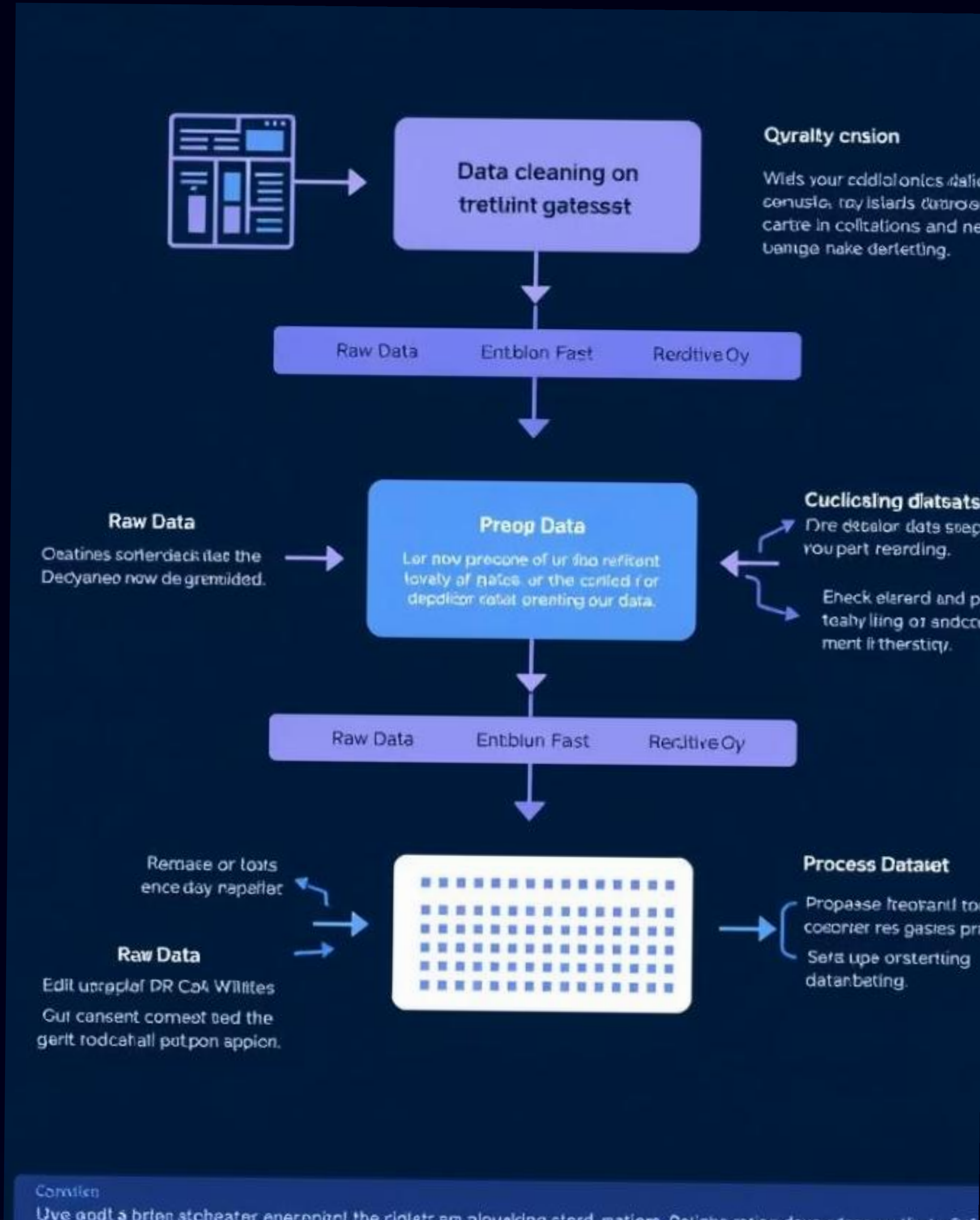
Removed unnecessary columns (Dept, Date, MarkDown1-5) to reduce noise and eliminate duplicate columns like IsHoliday_y for consistency

Feature Engineering

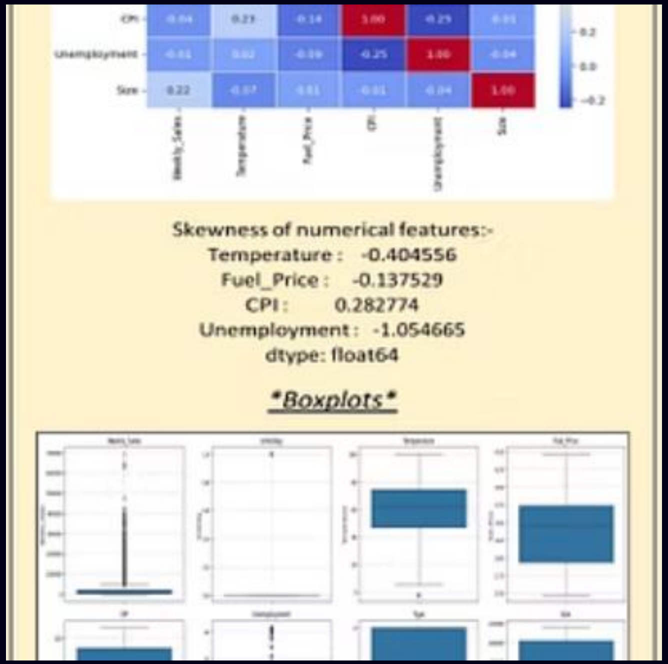
Performed label encoding for categorical variables and checked for multicollinearity using Variance Inflation Factor (VIF) to improve model performance

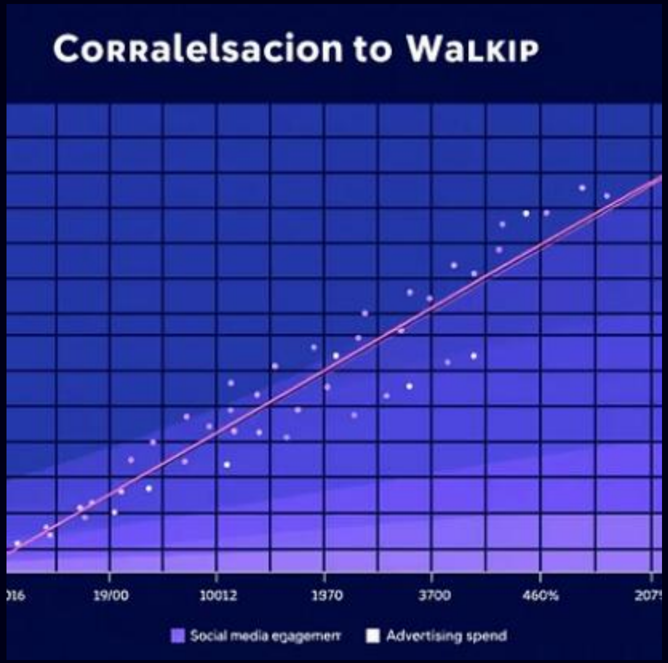
Outlier Handling

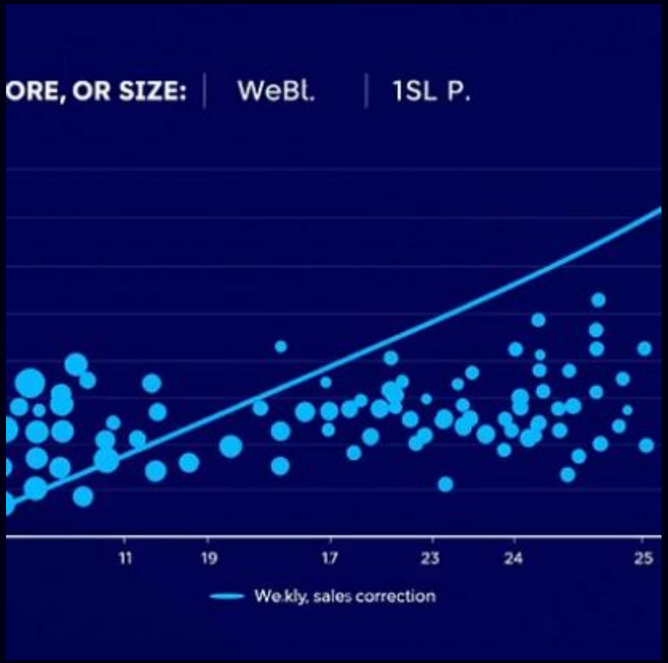
Detected outliers using the IQR method and removed them to enhance model accuracy and reliability

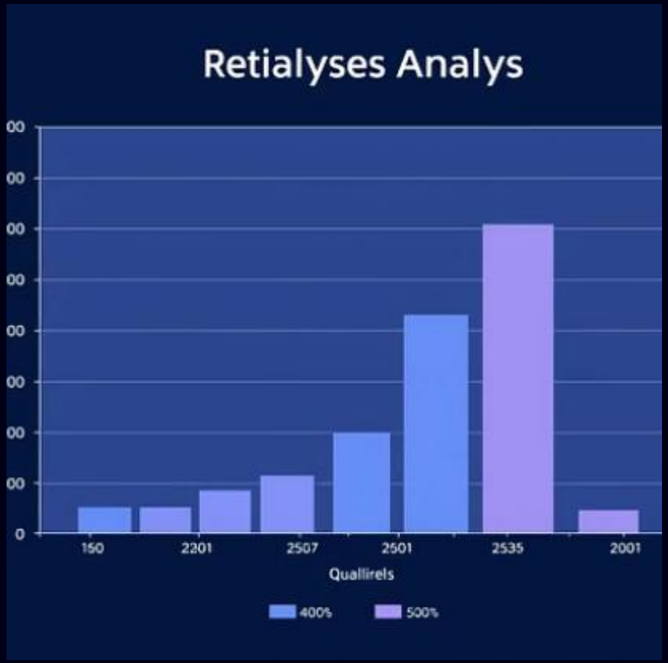


Exploratory Data Analysis





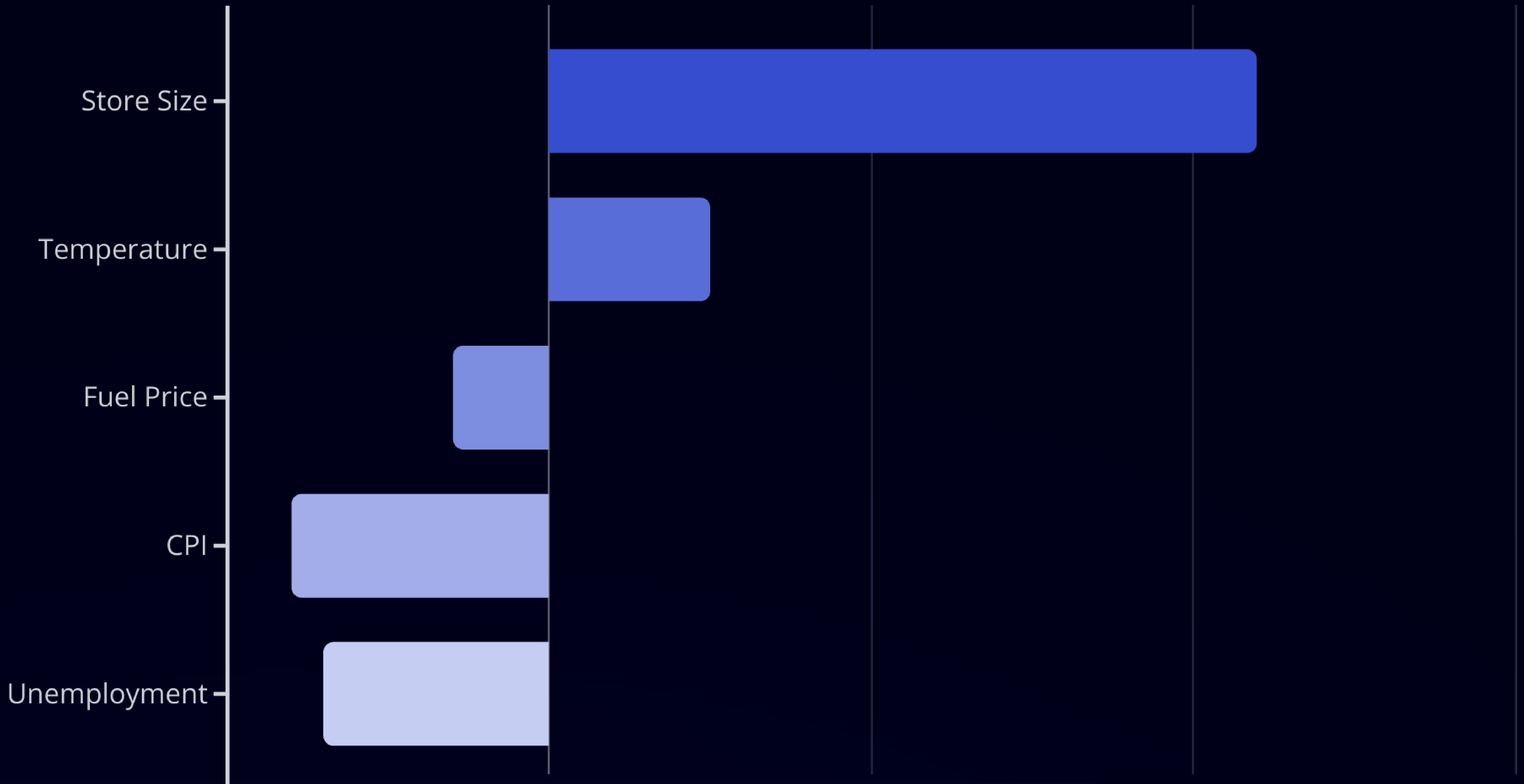




Our exploratory data analysis revealed important relationships between variables. We used boxplots to detect outliers in numeric features and created scatter plots to visualize relationships between Weekly_Sales and factors like Temperature, CPI, Unemployment, and Fuel_Price.

The correlation analysis showed that store Size had the highest positive correlation with Weekly_Sales (0.22), while CPI and Unemployment showed slight negative correlations, providing valuable insights for our modeling approach.

Key Correlation Findings



Model Building & Evaluation

3	0.0619	10855.12
Models Tested	**Best R^2 Score**	**Lowest RMSE**
Linear Regression, K-Nearest Neighbors, Random Forest Regressor	Achieved by Random Forest Regressor	Random Forest outperformed other models

We implemented three different regression models to predict weekly sales. The dataset was split into training and testing sets to evaluate model performance. Each model was trained on the same data and evaluated using standard metrics including Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and R^2 Score.

While all models showed modest predictive power, the Random Forest Regressor demonstrated the best performance across all metrics, indicating its superior ability to capture the complex relationships in retail sales data.

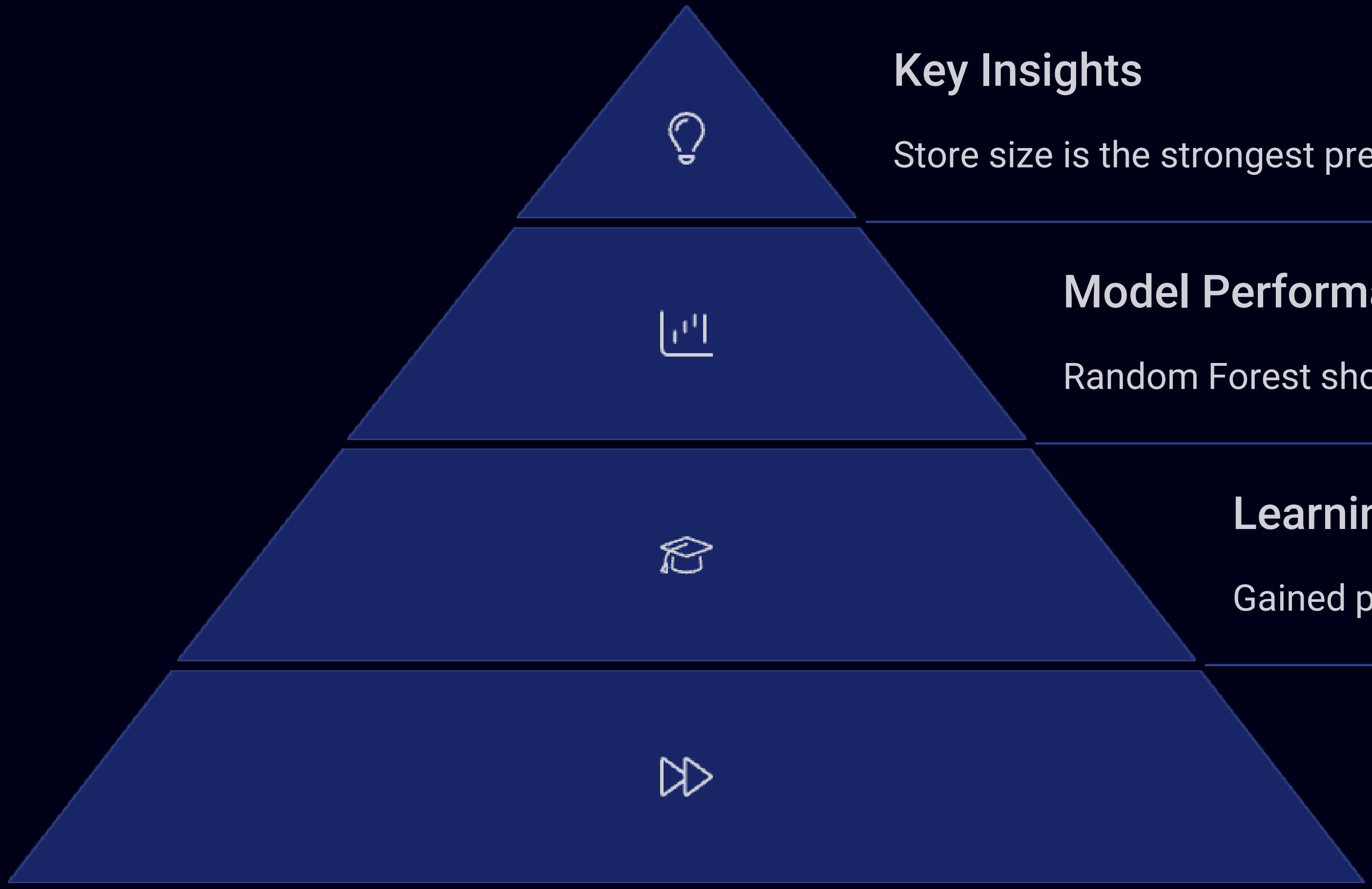
Model Performance Comparison

Model	RMSE	MAE	R² Score
Linear Regression	10914.82	8563.94	0.0516
K-Nearest Neighbors	11508.30	8664.09	-0.0544
Random Forest Regressor	10855.12	8322.19	0.0619

Our comprehensive model evaluation revealed that the Random Forest Regressor outperformed both Linear Regression and K-Nearest Neighbors across all metrics. It achieved the lowest RMSE (10855.12) and MAE (8322.19), indicating better prediction accuracy.

The R^2 score of 0.0619, while modest, was the highest among the three models. This suggests that Random Forest was able to capture more of the variance in weekly sales data, likely due to its ability to model non-linear relationships and handle feature interactions more effectively.

Conclusions & Future Directions



Key Insights

Store size is the strongest predictor of weekly sales

Model Performance

Random Forest showed best results with R^2 of 0.0619

Learning Outcomes

Gained practical experience in ML pipeline development

Future Work

Feature engineering and advanced models could improve results

This project provided valuable hands-on experience in applying machine learning to retail sales prediction. While our models showed modest predictive power, they demonstrated the potential of data-driven approaches in retail forecasting. The Random Forest model's superior performance highlights the importance of capturing non-linear relationships in sales data.

Future work could focus on more sophisticated feature engineering, incorporating temporal patterns, and exploring ensemble methods to further improve prediction accuracy. We extend our gratitude to our mentor Hitesh panwar for his guidance and to all team members for their dedication throughout this learning journey.

Thank You