

WALMART SALES ANALYSIS WITH XGBOOST

- Submitted to Prof. Jawahar Panchal

Presented by:

- JENITA SHIRLEY DEVADOSS SAMRAJ
- VINITHA INAGANTI
- KUNDETI NAGA SRI ADITYA
- LAKSHMA REDDY BAIRI
- PRAPUL KUMAR PODILI

Overview

Utilize advanced analytics to optimize retail sales by leveraging historical data, customer information, and product details. Employ the XG Boost algorithm for predictive modeling, incorporating feature engineering to enhance performance. Through interpretability techniques, analyze factors influencing sales and provide actionable recommendations for store managers. Aim to empower retailers with informed decision-making tools for sustainable sales growth.

What we seek to Address

Develop accurate XG Boost models for predicting future store sales, aiding store managers in effective planning. Uncover insights from data regarding factors influencing store sales, such as customer behavior and product preferences. Optimize sales strategies by identifying improvements in pricing, inventory management, and marketing tactics. Enable retail businesses to make data-driven decisions for enhanced store sales performance.

Project Methodology

- **Data Preparation:**

Collect and clean historical sales, customer, and product data, addressing missing values and encoding categorical variables.

- **EDA:**

Analyze data for sales-affecting patterns. Visualize insights and conduct statistical analysis.

- **Modeling:**

Develop an optimized XG Boost model, assessing performance with metrics like MAE, MSE, and RMSE through cross-validation.

- **Interpretation:**

Interpret model results using feature importance and SHAP values, deriving actionable insights.

- **Reporting:**

Generate clear reports and dashboards. Collaborate with store managers for implementation, refining strategies based on real-world feedback and model performance.

Data Acquisition

Dataset: Obtained from a Kaggle competition by Walmart ([Link](#)), it includes historic weekly sales data for 45 Walmart stores, with department-wide details.

- Testing Data: 'test.csv' is used only for predicting values with the lowest WMAE score, lacking the target variable 'Weekly Sales.'
- Data Division: 'train.csv' is split into training and validation sets, covering weekly sales from 2010-02-05 to 2012-11-01, totaling 421,570 rows in training and 115,064 in testing.
- Store Information: 'stores.csv' provides detailed info on the type and size of 45 stores.
- Impact Analysis: Aims to predict department-wide weekly sales, assessing the influence of factors like temperature, fuel prices, holidays, markdowns, unemployment rate, and consumer price indexes using 'features.csv.'
- Submission Record: 'sampleSubmission.csv' records results of the most accurate model for creating a Power BI dashboard, aligned with 'stores' and 'features' datasets.

Text Data Processing

Combine separate data frames into one for unified analysis. Sales data covers 2010-2012 weekly. Enhance analysis by splitting the date column for year, month, and week details, enabling a more granular exploration of temporal sales patterns.

```
```{r}
library(dplyr)
library(lubridate)

mergeddf <- traindf %>%
 left_join(storesdf, by = "Store") %>%
 left_join(featuresdf, by = c("Store", "Date"))

testingdf_merged <- testingdf %>%
 left_join(storesdf, by = "Store") %>%
 left_join(featuresdf, by = c("Store", "Date"))

splitdf_date <- function(df) {
 df$Date <- as.Date(df$Date)
 df$Year <- year(df$Date)
 df$Month <- month(df$Date)
 df$Day <- day(df$Date)
 df$WeekOfYear <- week(df$Date)
 return(df)
}

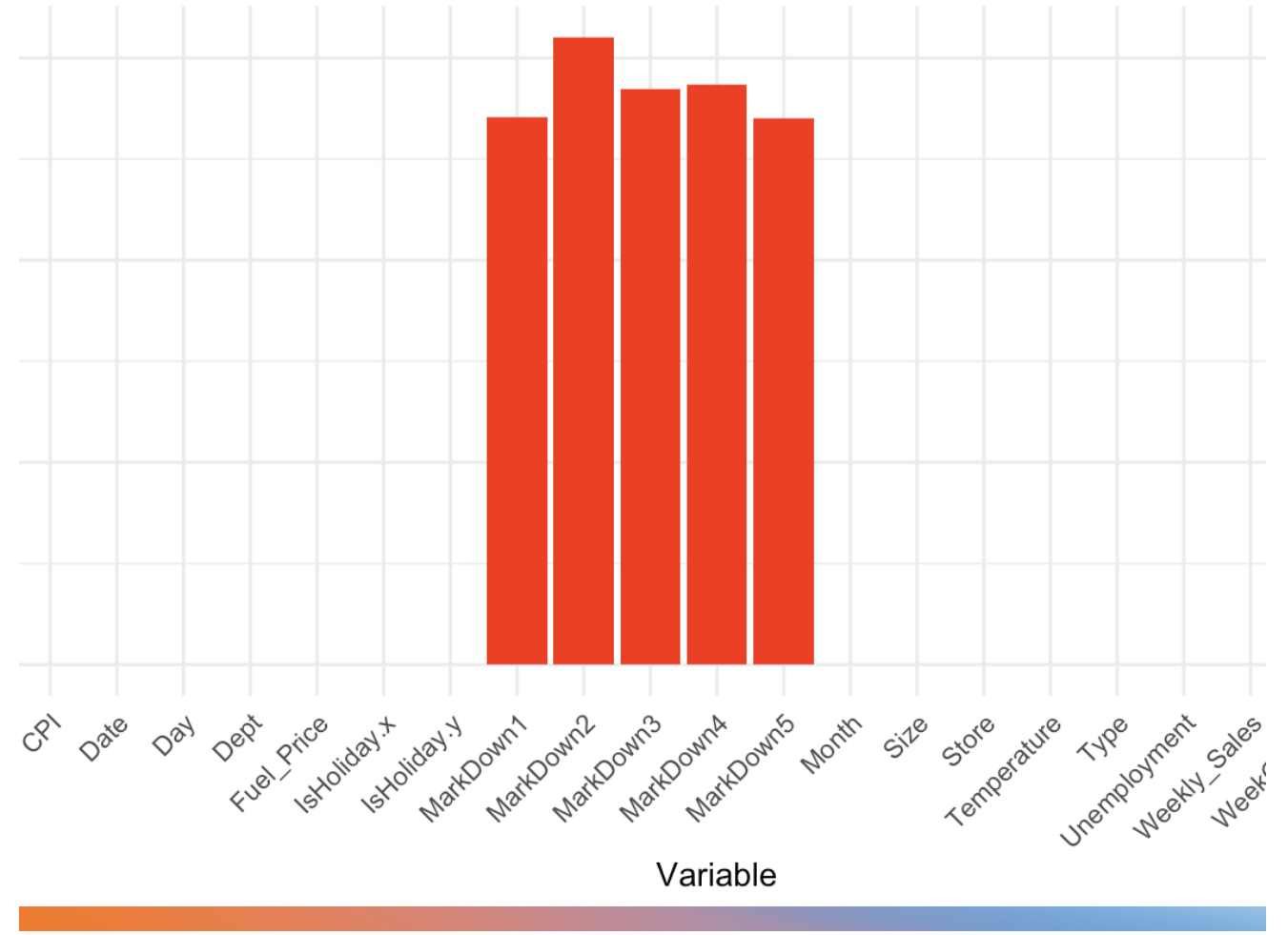
mergeddf <- splitdf_date(mergeddf)
testingdf_merged <- splitdf_date(testingdf_merged)
mergeddf
testingdf_merged

```
```

Missing Values

- All columns in the dataset are complete, except for Markdown 1 to 5, which have over 250,000 missing values each. These columns represent store-specific promotional activities initiated after November 2011, implemented inconsistently. The substantial NaN values in these columns are expected.
- Exploratory data analysis (EDA) will scrutinize the relationship between Markdown columns and weekly sales. This analysis guides decisions on managing missing values and determines the relevance and significance of Markdown data in the context of overall sales patterns.

Missing Values

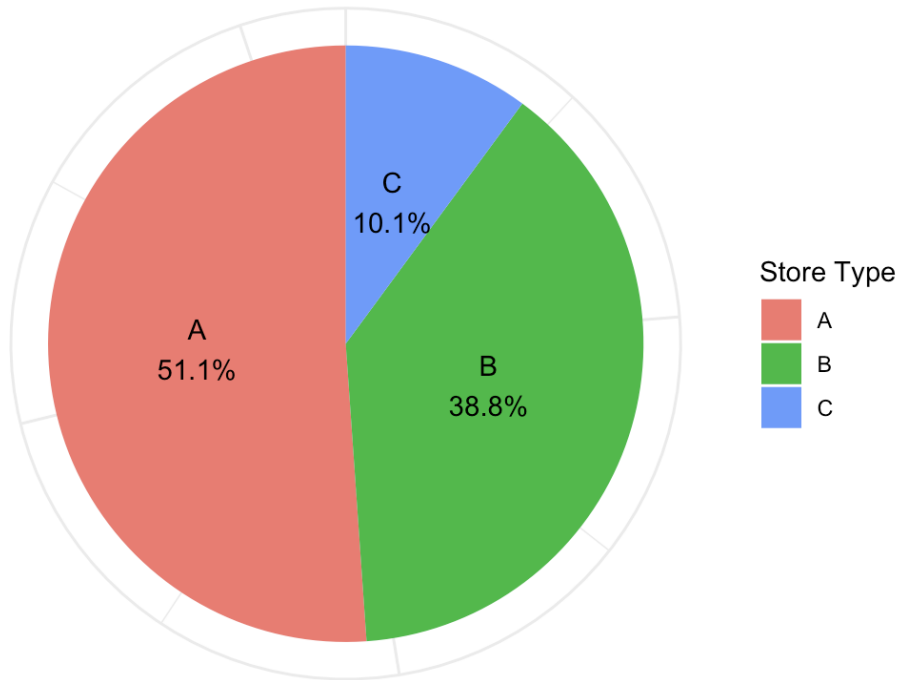


Exploratory Data Analysis

An in-depth exploratory data analysis (EDA) is vital for understanding the dataset and enhancing predictive models. Utilizing tools in R we can analyze and explore main dataset characteristics. Visualizations generated through ggplot2 offers insights into weekly sales patterns, holiday impacts, regional variations, and relationships with factors like CPI, fuel price, temperature, and unemployment. These visualizations inform potential modeling approaches in subsequent project stages.

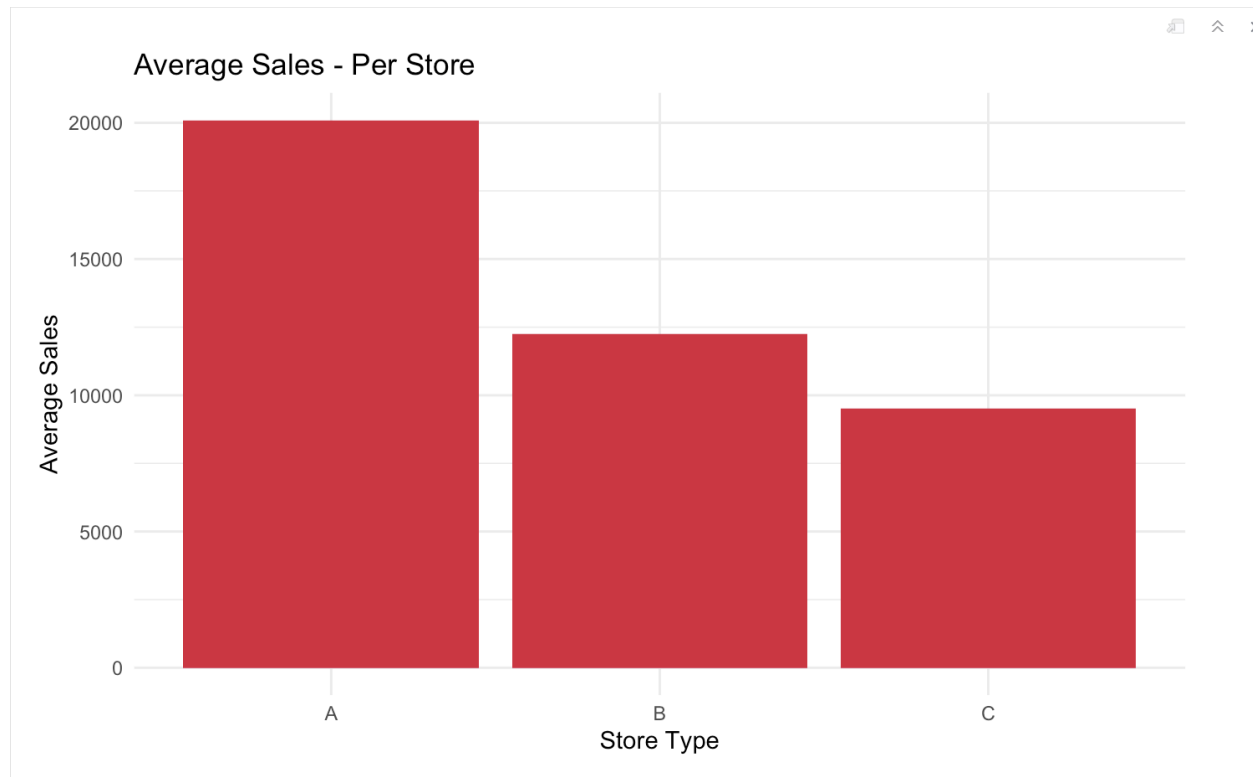
Types of Stores and their Popularity

Popularity of Store Types



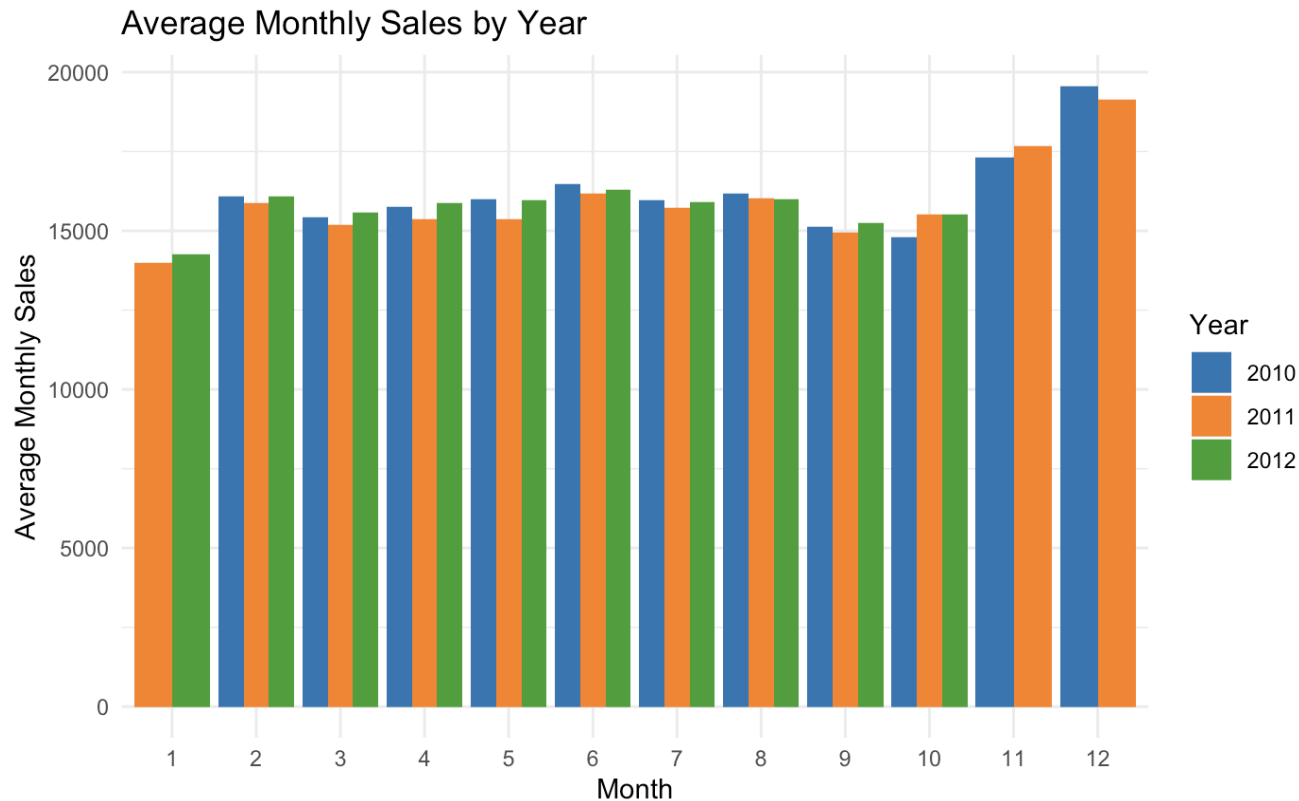
Type A stores surpass B and C in popularity. Analysis indicates Type A stores as notably favored, considering factors like customer footfall or sales figures. In-depth exploration of each store type's characteristics and performance is warranted for a comprehensive understanding of the observed popularity gap.

Average Sales - Store Type



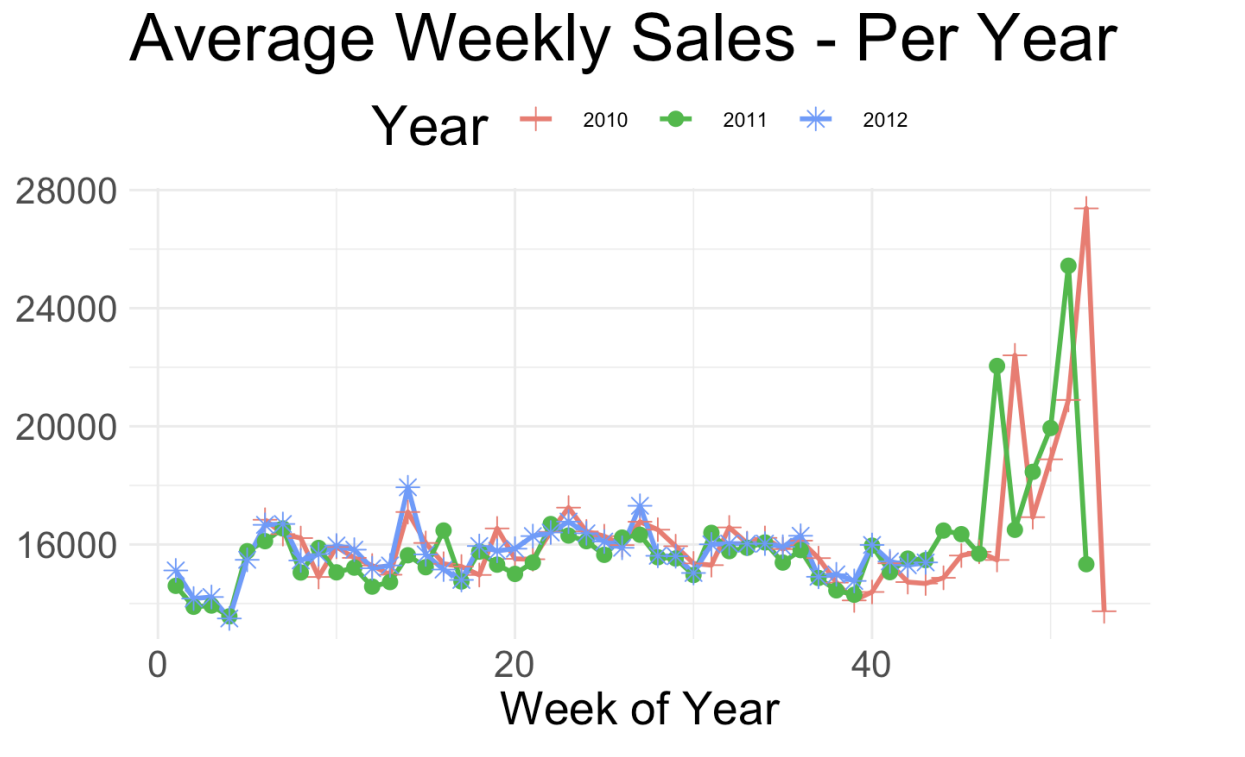
Type A stores excel in sales compared to Type B and Type C. This performance superiority may be attributed to factors like location, product assortment, and promotional strategies. A detailed investigation into these specific aspects can yield valuable insights for strategic decision-making and optimizing overall retail performance.

Average Monthly Sales - Per Year



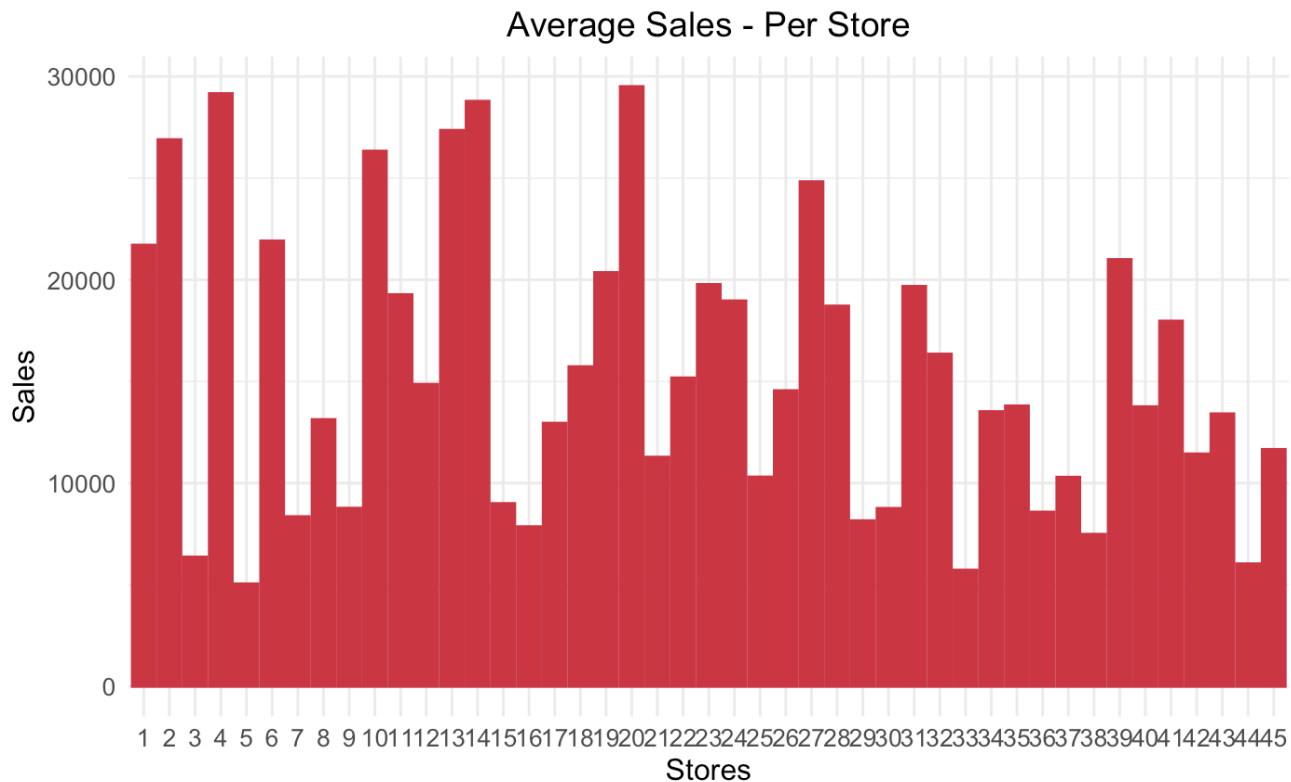
January consistently reports the lowest sales in 2011 and 2012, with unavailable weekly data for January 2010. Weekly sales stabilize around 15,000 from February to October, indicating a consistent market demand. Sales peak in November and December for 2010 and 2011, but 2012 data is absent. The analysis underscores distinct monthly sales trends, emphasizing January's consistently lower sales and peaks during the holiday months. Limited 2012 data hinders conclusive insights, warranting further exploration into factors influencing peak months.

Average Weekly Sales - Per Year



- Weekly sales data for 2010 and 2011 consistently highlights peak sales during the weeks encompassing Thanksgiving and the week just before Christmas.
- In 2012, a notable deviation is observed, with week number 14 standing out as the period with the highest sales. This is distinct from the established trend tied to recognized holidays or special events.
- The deviation in 2012 suggests a unique sales trend not directly linked to holidays. Further analysis is required to understand the factors contributing to this anomaly, such as market dynamics, promotional activities, or external factors.
- Investigating the specifics of week 14 in 2012 may uncover insights contributing to a nuanced understanding of the sales dynamics during that period.

Average Store Sales



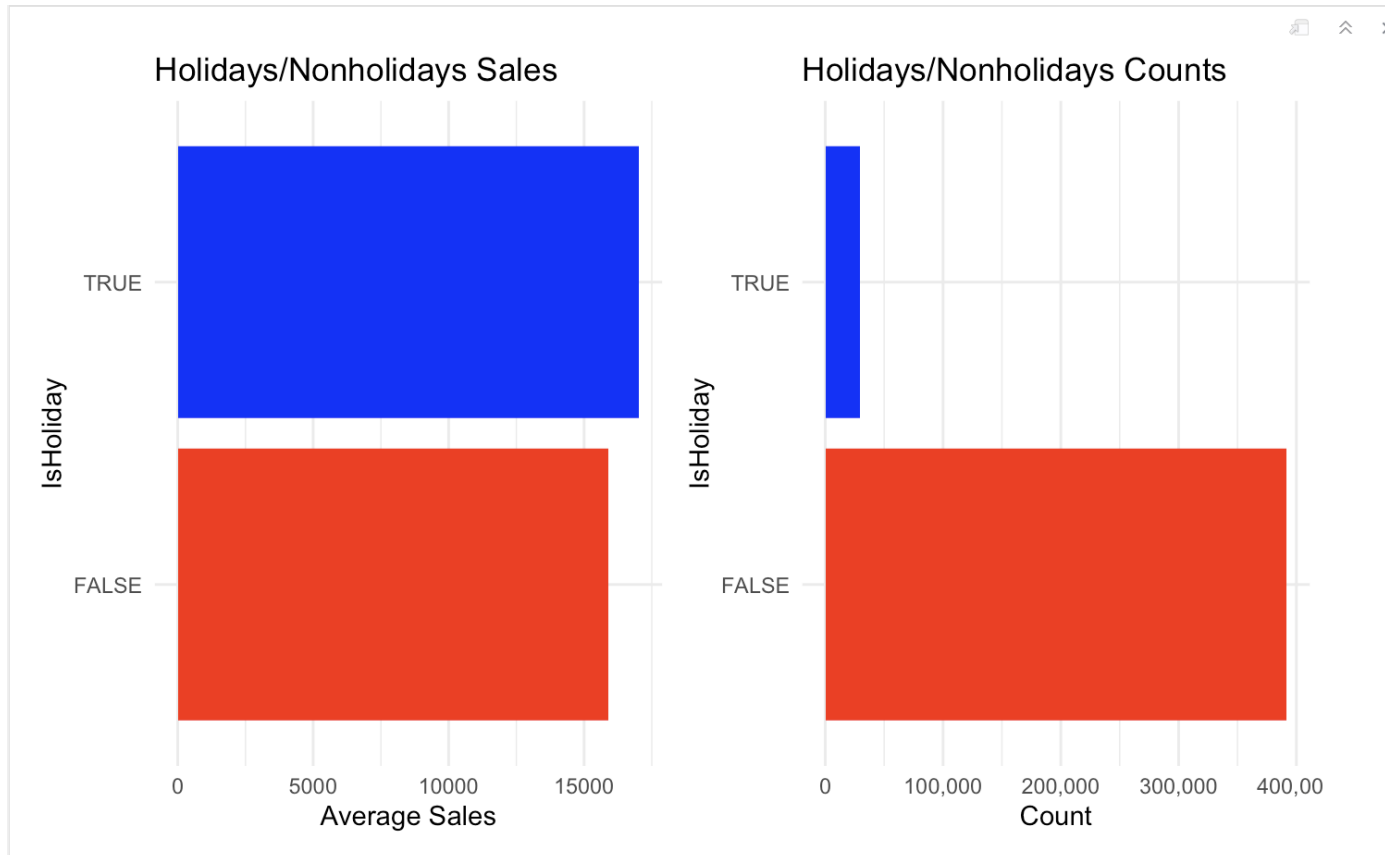
- Considerable variations in sales performance exist among the 45 stores.
- The magnitude of these variations depends on the store category and the specific week within the given year.
- Different store categories and specific weeks play significant roles in influencing sales outcomes, emphasizing the need for tailored strategies based on unique store characteristics and temporal context.

Average Store Sales - Year Wise

- A consistent overall trend in store sales is observed across the three years, primarily influenced by store type and size.
- Stores 2, 4, 13, 14, and 20 consistently stand out with the highest sales figures throughout all three years.
- The classification and physical dimensions of stores play a crucial role in determining sales performance, and understanding the distinguishing factors of the top-performing stores could provide valuable insights for optimizing sales strategies and improving overall store performance.

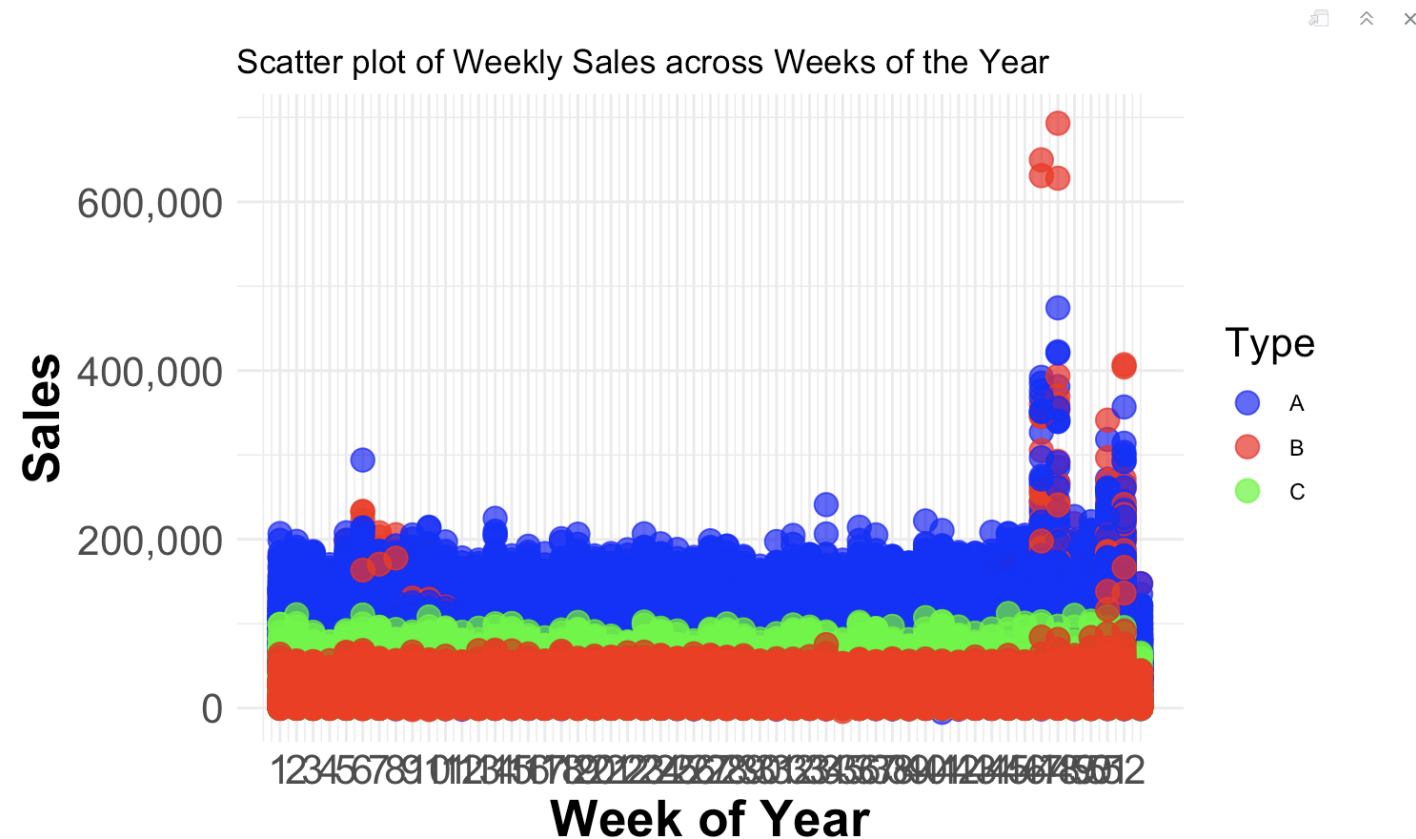


Holidays Vs Nonholidays Sales



- Holiday weeks constitute only 7 percent of the total weeks in the dataset.
- Despite the low percentage, average sales during holiday weeks are higher compared to non-holiday weeks.
- This finding emphasizes the importance of considering holiday weeks in sales analysis, as they disproportionately contribute to overall sales performance. Further investigation into factors influencing consumer spending during holidays could provide insights for targeted marketing strategies and resource allocation during peak sales periods.

Relationship: Week of Year vs Sales



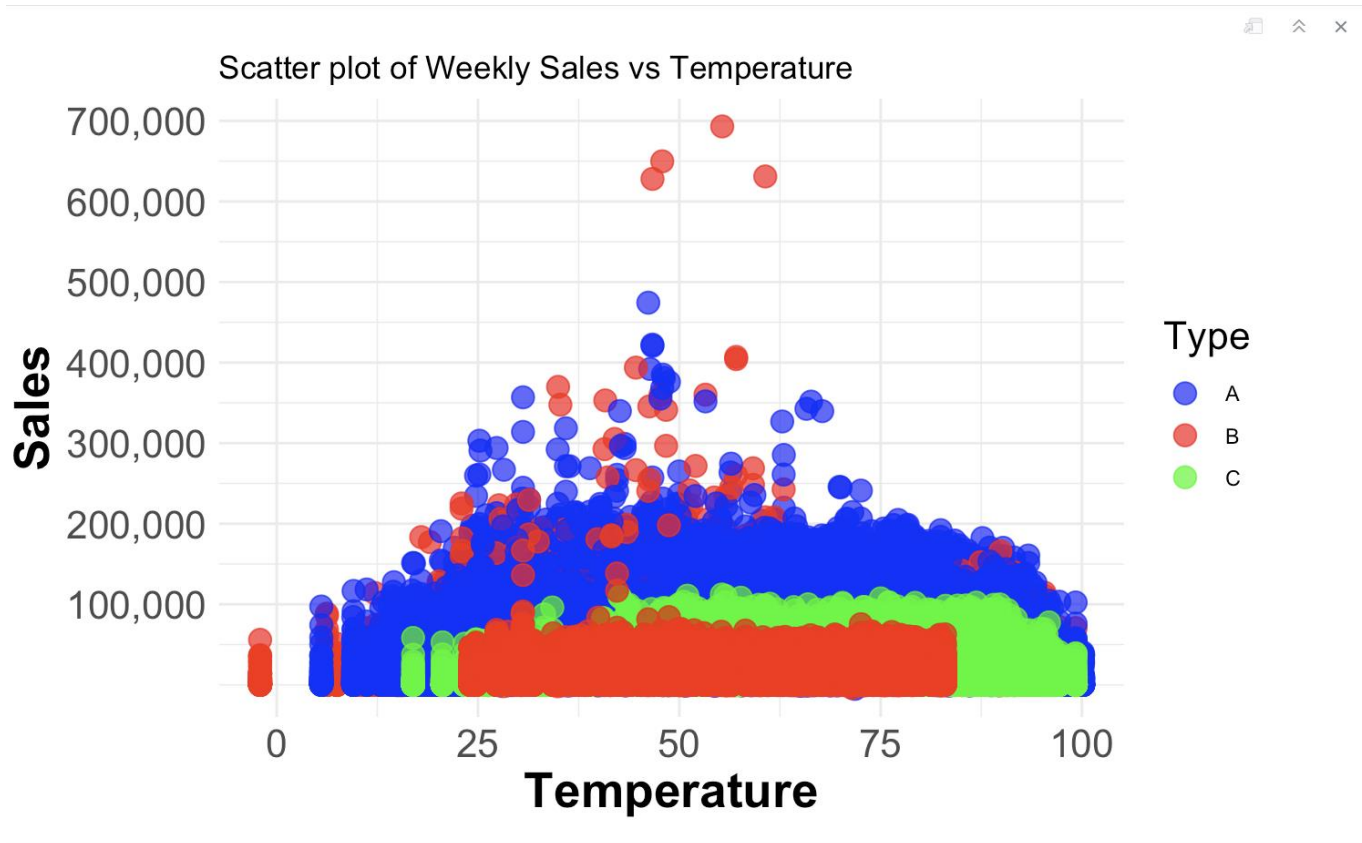
- A modest correlation is evident, indicating a slight relationship between variables.
- Weekly sales show a gradual increase towards the end of the year, suggesting a nuanced trend in the data.
- The subtle uptick in sales during the latter part of the year may be influenced by factors like holiday shopping, year-end promotions, or seasonal trends. Understanding this relationship is crucial for adapting strategies, recognizing increased consumer activity, and refining marketing approaches for optimized sales outcomes.

Relationship: Size of Store vs Sales



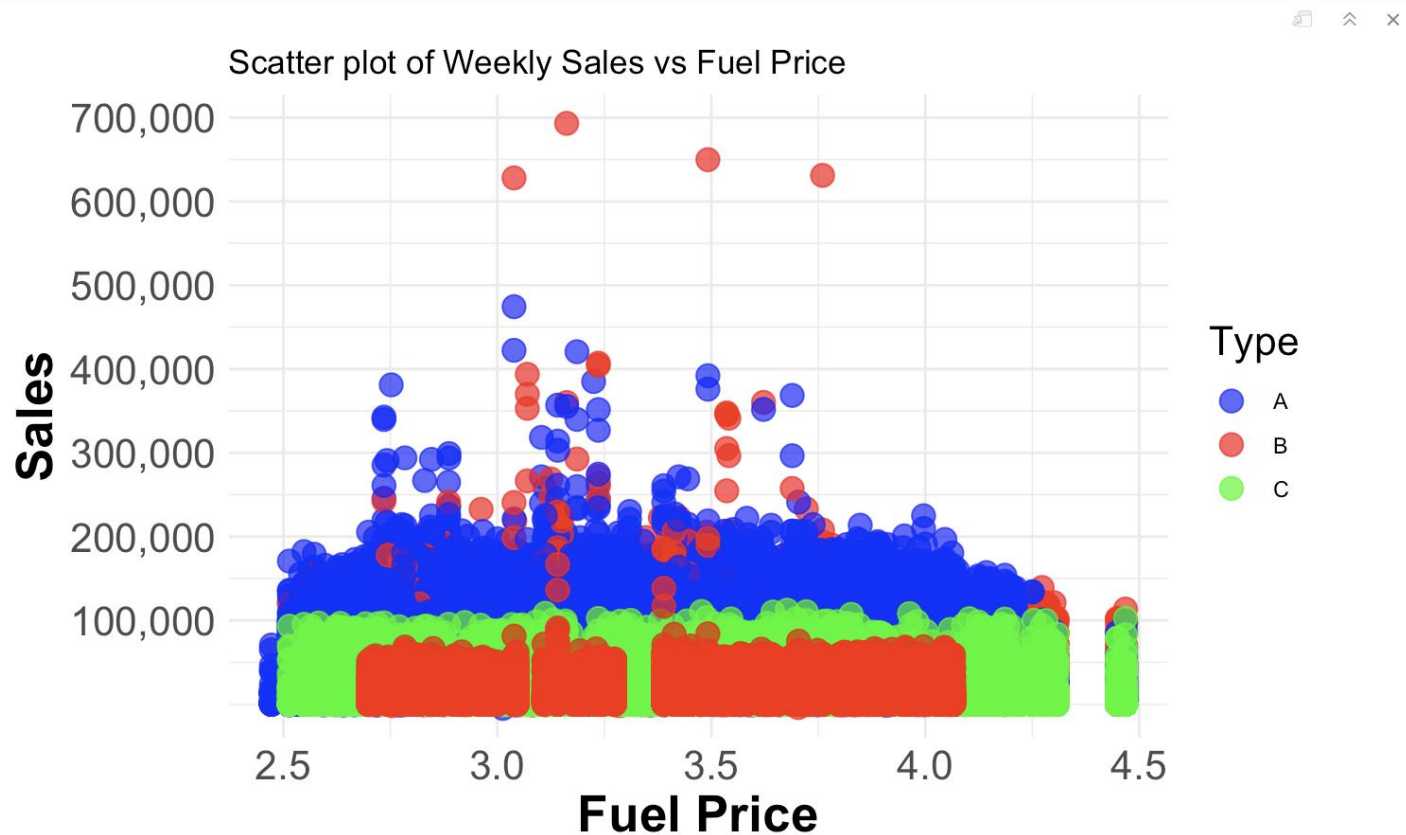
- A discernible linear relationship is evident between the size of the store and weekly sales.
- Generally, there is a tendency for sales to increase with an expansion in store size, though exceptions exist, indicating that other factors may influence sales outcomes in certain cases. Understanding these exceptions is crucial for tailoring effective business strategies.

Relationship: Temperature vs Sales



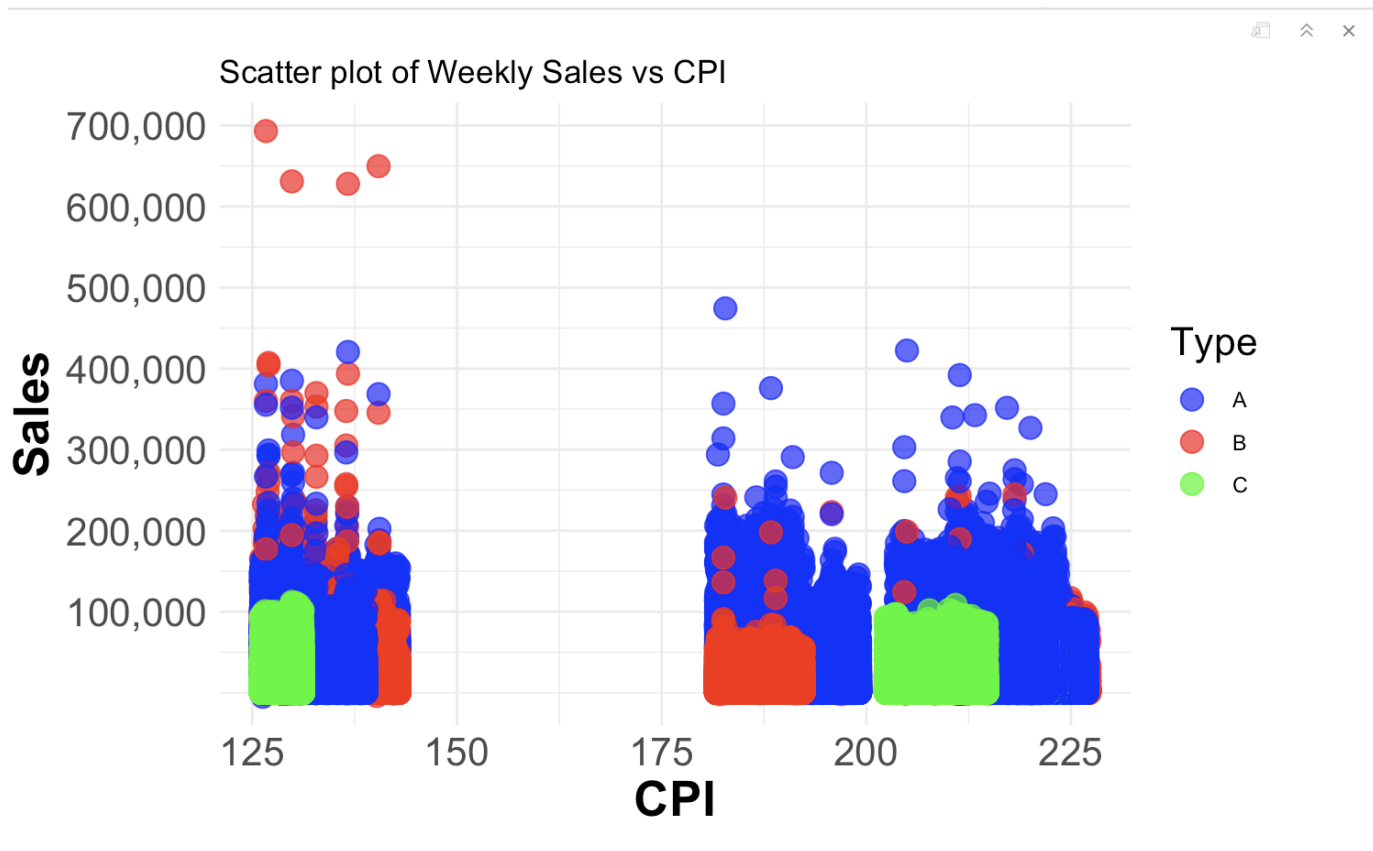
- There is no discernible relationship between the temperature in the region and the weekly sales of the stores.
- Although a slight dip in sales is observed during periods of both low and very high temperatures, the lack of a clear and consistent relationship suggests that other factors, such as consumer behavior, local events, or marketing strategies, might play a more influential role in determining weekly sales.

Relationship: Fuel Price vs Sales



- There is no apparent and distinct relationship between fuel prices and sales.
- Fluctuations in fuel prices do not consistently correlate with changes in sales, suggesting that other factors like consumer spending habits, economic conditions, or marketing strategies may have a more prominent influence on sales outcomes. Understanding this nuanced interplay is crucial for businesses to adapt strategies effectively in response to changing economic variables.

Relationship: CPI vs Sales



- The data reveals three distinct clusters, indicating clear grouping within the dataset.
- Despite discernible clustering, there is no evident and consistent correlation between the Consumer Price Index (CPI) and weekly sales. This suggests that other factors, possibly internal or external to the dataset, may influence observed patterns, emphasizing the need for a more comprehensive analysis to understand the nuanced relationship between the CPI and weekly sales.

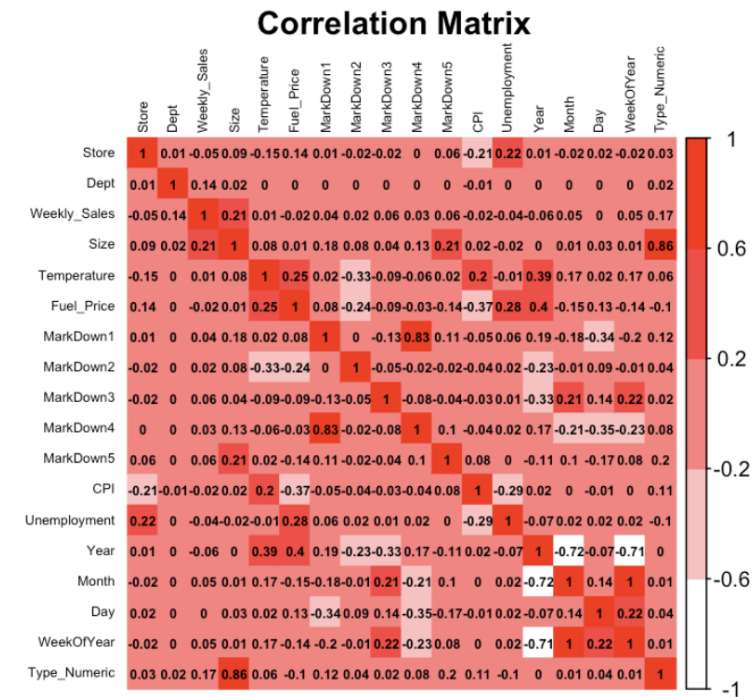
Relationship: Unemployment vs Sales



- The Unemployment rate appears to have no discernible impact on weekly sales.
- Fluctuations in the Unemployment rate do not consistently correlate with changes in weekly sales, suggesting that other factors like consumer confidence, economic conditions, or industry-specific dynamics may be more influential in determining sales outcomes. Understanding this interplay is crucial for businesses to develop responsive strategies in the face of changing market conditions.


Correlation Matrix

- Features such as "Department," "Store size," and "Type" exhibit a moderate correlation with weekly sales, indicating their reasonable influence on sales figures.
- Columns related to "Markdown1" through "Markdown5" and economic indicators like "Temperature," "Fuel price," "CPI," and "Unemployment" show weak correlation with weekly sales and will be excluded from further analysis. The "IsHoliday" column, indicating sales during holiday weeks, will be retained, and "Month" and "Day" columns will be excluded, as their temporal information is already captured in the "WeekOfYear" column. These decisions aim to focus on the most relevant variables for understanding and predicting weekly sales dynamics.



Data Preparation for Model Training

- Columns with weak relationships with the target variable, identified through EDA and correlation analysis, will be dropped to retain only the most relevant features for model training.
- The data preparation involves creating separate dataframes for input features and the target variable, scaling the input features to a standardized range, splitting the dataset into training and validation sets, and defining a performance measurement function to assess the model's effectiveness consistently. These steps ensure a structured and informed approach to preparing data for subsequent model training, enhancing generalization and predictive capabilities.



The screenshot displays a workspace with five data frames and an R console. The data frames are:

- data.frame 421570 x 9
- data.frame 115064 x 8
- data.frame 421570 x 9
- data.frame 115064 x 8
- R Console (Loading required package: lme4)

Below the data frames, a description of the first data frame is shown:

Description: df [421,570 x 9]

| Store | Dept | Weekly_Sales | IsHoliday.x | Size | IsHoliday.y | Year | WeekOfYear | Type_Numeric |
|-------|-------|--------------|-------------|--------|-------------|-------|------------|--------------|
| <int> | <int> | <dbl> | <lgl> | <int> | <lgl> | <dbl> | <dbl> | <dbl> |
| 1 | 1 | 24924.50 | FALSE | 151315 | FALSE | 2010 | 6 | 3 |
| 1 | 1 | 46039.49 | TRUE | 151315 | TRUE | 2010 | 7 | 3 |
| 1 | 1 | 41595.55 | FALSE | 151315 | FALSE | 2010 | 8 | 3 |
| 1 | 1 | 19403.54 | FALSE | 151315 | FALSE | 2010 | 9 | 3 |
| 1 | 1 | 21827.90 | FALSE | 151315 | FALSE | 2010 | 10 | 3 |
| 1 | 1 | 21043.39 | FALSE | 151315 | FALSE | 2010 | 11 | 3 |
| 1 | 1 | 22136.64 | FALSE | 151315 | FALSE | 2010 | 12 | 3 |
| 1 | 1 | 26229.21 | FALSE | 151315 | FALSE | 2010 | 13 | 3 |
| 1 | 1 | 57258.43 | FALSE | 151315 | FALSE | 2010 | 14 | 3 |
| 1 | 1 | 42960.91 | FALSE | 151315 | FALSE | 2010 | 15 | 3 |

Navigation: 1-10 of 421,570 rows Previous 1 2 3 4 5 6 ... 100 Next

Machine Learning

This study will evaluate the performance of various machine learning models, including:

- **Linear Regression:** Establishes a linear relationship between input features and the target variable, suitable for regression tasks.
- **Ridge Regression:** Introduces regularization to address multicollinearity issues and enhance model robustness, similar to linear regression.
- **Decision Tree:** A tree-like model making decisions based on feature splits, versatile for regression and classification tasks.
- **Random Forest:** An ensemble method combining multiple decision trees to improve accuracy and reduce overfitting.
- **Gradient Boosting Machine:** An ensemble technique building sequential weak learners, correcting errors made by previous models and achieving high predictive accuracy.

a. Linear Regression

Utilizing linear regression for Walmart sales analysis, we aim to model and understand the relationship between various factors (e.g., promotions, holidays) and sales performance. The method provides insights into the impact of these variables, aiding strategic decision-making for optimizing sales forecasts and resource allocation.

```
```{r}
Load required libraries
library(caret)
library(Metrics)

Create and train the model
model <- lm(train_targets ~ ., data = train_inputs)

Generate predictions on training data
train_preds <- predict(model, newdata = train_inputs)

Compute WMAE on training data
train_wmae <- weighted.mean(abs(train_preds - train_targets), weights = train_inputs$IsHoliday)
cat('The WMAE loss for the training set is', train_wmae, '\n')

Generate predictions on validation data
val_preds <- predict(model, newdata = val_inputs)

Compute WMAE on validation data
val_wmae <- weighted.mean(abs(val_preds - val_targets), weights = val_inputs$IsHoliday)
cat('The WMAE loss for the validation set is', val_wmae, '\n')
```
```

```
The WMAE loss for the training set is 14573.56
```

```
The WMAE loss for the validation set is 14568.17
```

b. Ridge Regression

Ridge Regression enhances Walmart sales analysis by introducing regularization to mitigate multicollinearity, improving model robustness. Its application results in refined predictions, addressing potential overfitting challenges and optimizing the accuracy of the sales forecasting model.

```
```{r}
Load required libraries
library(glmnet)

Create and train the Ridge model
model_ridge <- glmnet(as.matrix(train_inputs), train_targets, alpha = 0, lambda = 1)

Generate predictions on training data
train_preds <- predict(model_ridge, newx = as.matrix(train_inputs), s = 1)

Compute WMAE on training data
train_wmae <- weighted.mean(abs(train_preds - train_targets), weights = train_inputs$IsHoliday)
cat('The WMAE loss for the training set is', train_wmae, '\n')

Generate predictions on validation data
val_preds <- predict(model_ridge, newx = as.matrix(val_inputs), s = 1)

Compute WMAE on validation data
val_wmae <- weighted.mean(abs(val_preds - val_targets), weights = val_inputs$IsHoliday)
cat('The WMAE loss for the validation set is', val_wmae, '\n')

```
```

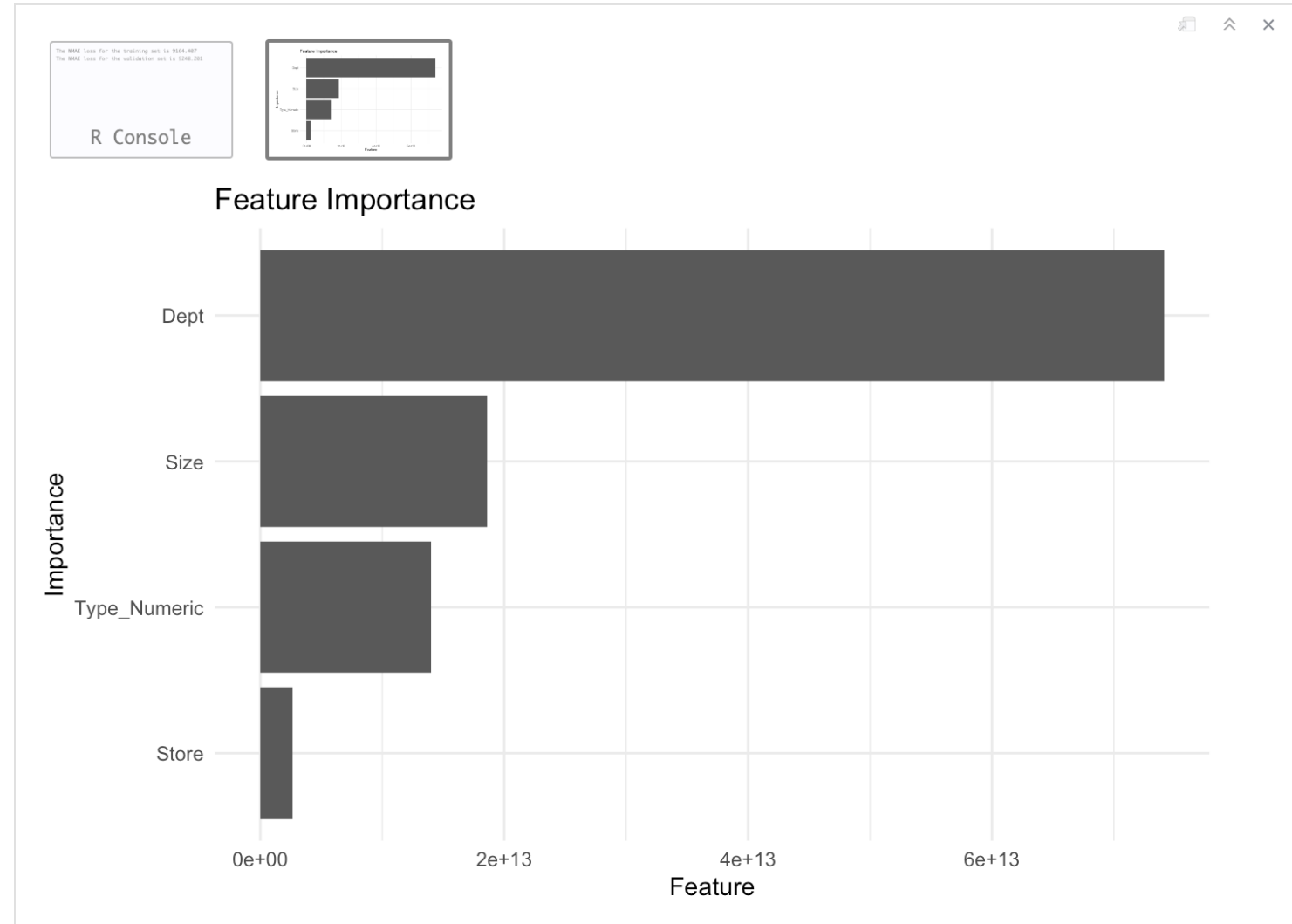
The WMAE loss for the training set is 14573.4

The WMAE loss for the validation set is 14568.01

c. Decision Tree

Utilizing a Decision Tree model for Walmart sales analysis provides a clear, interpretable structure for understanding sales patterns. Its ability to capture non-linear relationships and feature importance can offer valuable insights into factors influencing sales performance. Consider leveraging Decision Trees to enhance decision-making in Walmart's sales strategy.

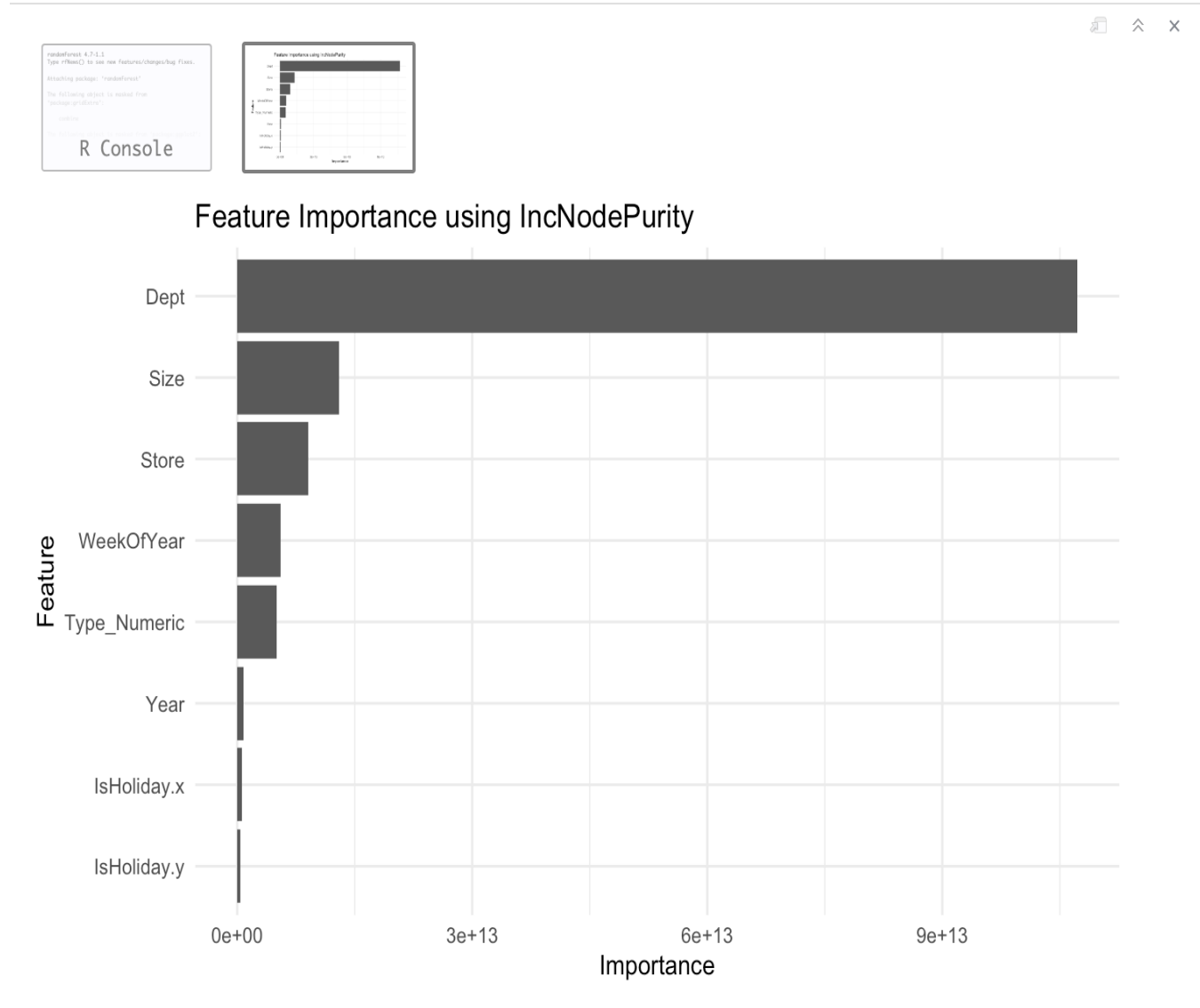
- Decision Tree: Strong fit to training data (WMAE: 9164.407).
- Concerns: Higher WMAE on validation set (9248.201) indicates potential overfitting.
- Recommendations: Explore model complexity adjustments (e.g., pruning) and compare with alternative models for enhanced performance.



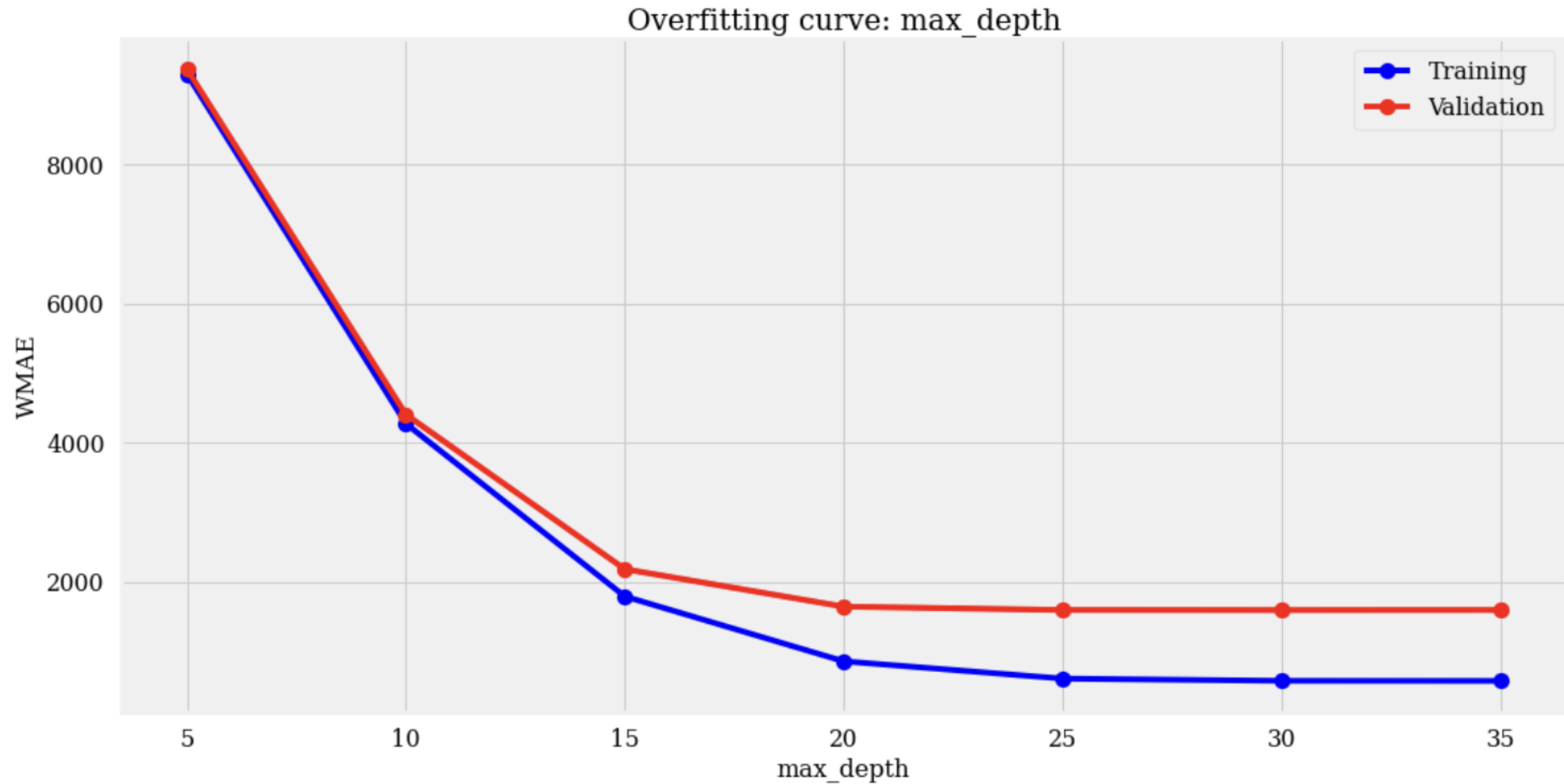
d. Random Forest

Utilizing Random Forest for Walmart sales analysis yields robust predictive performance, leveraging an ensemble of decision trees. The model excels in capturing complex relationships, providing valuable insights into sales patterns and contributing to accurate forecasting for informed business decisions.

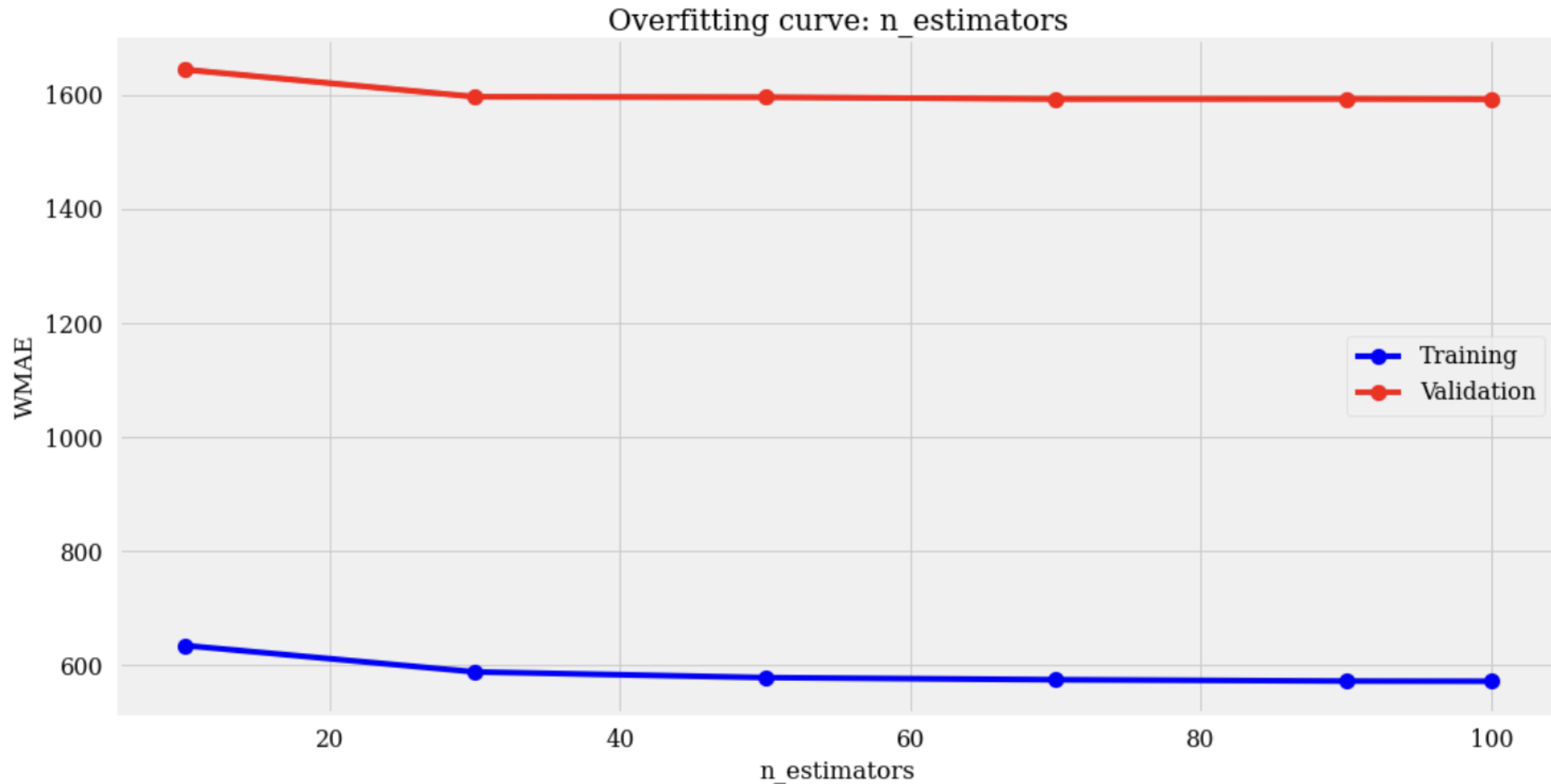
- Random Forest model excels in training (WMAE: 2238.897) with slight increase in validation (2713.409), indicating improved generalization.
- Explore hyperparameter tuning for potential enhancements in performance.



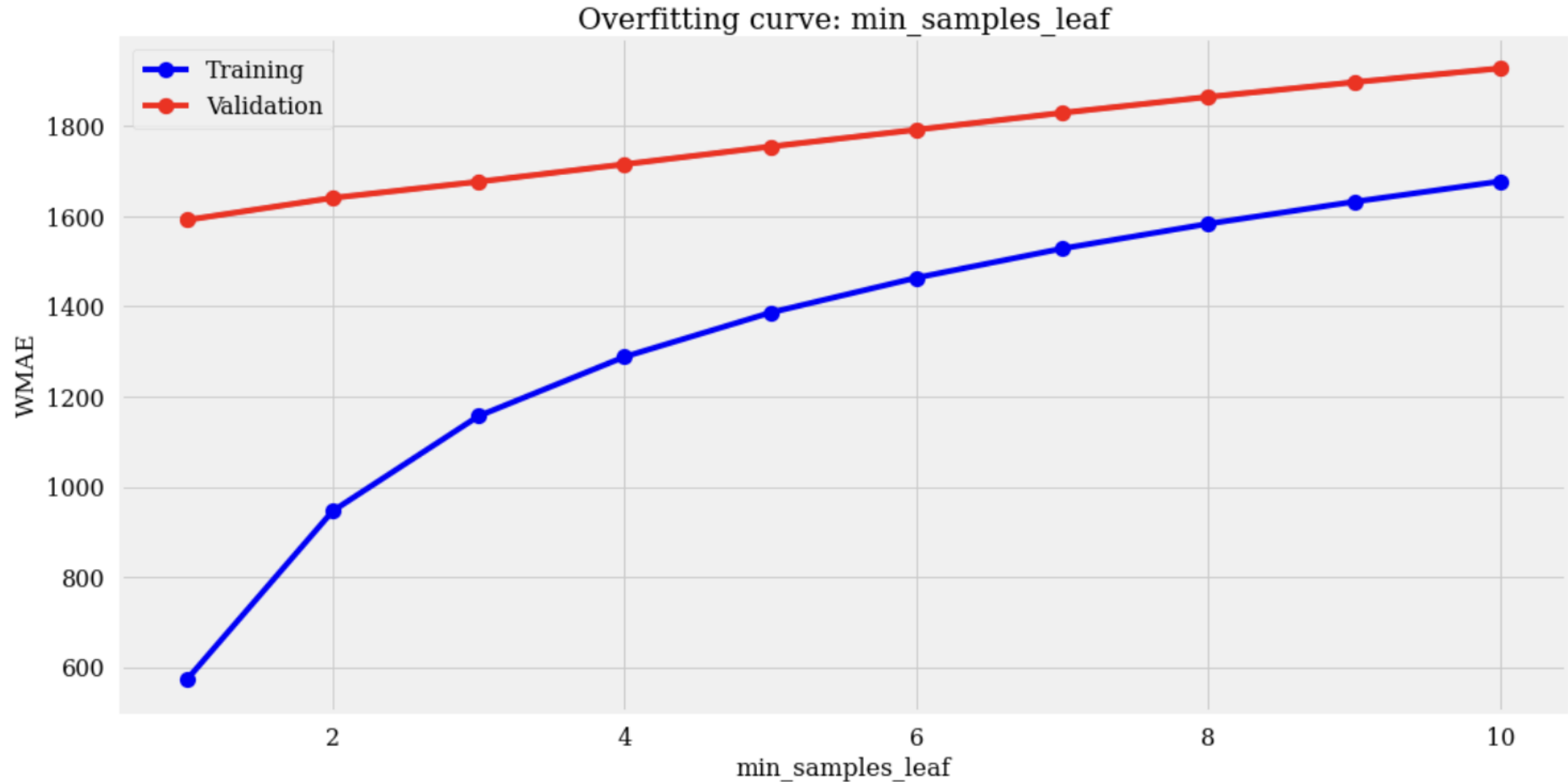
Overfitting Curve - max_depth:



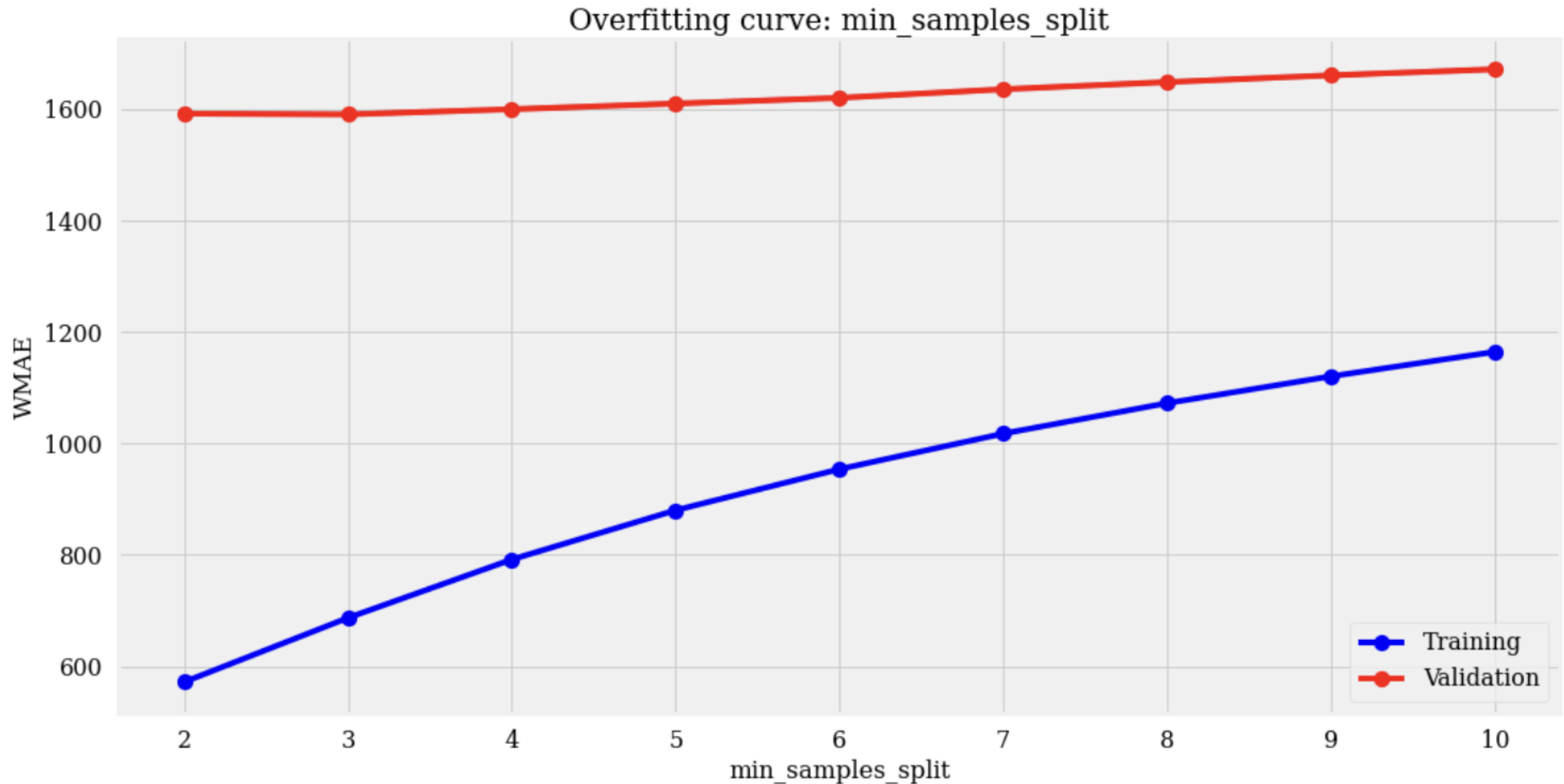
Overfitting Curve - n_estimators:



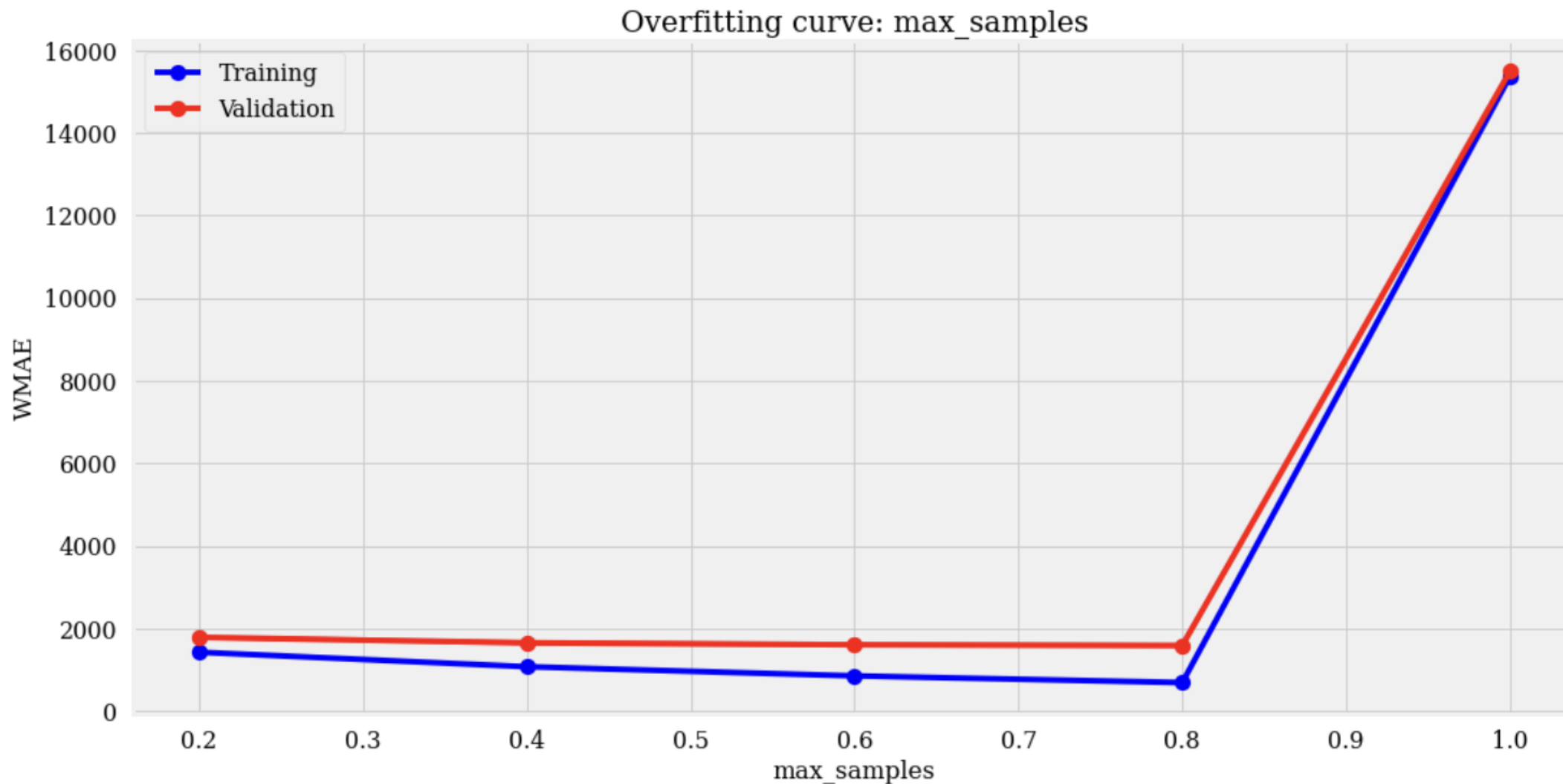
Overfitting Curve - min_samples_leaf:



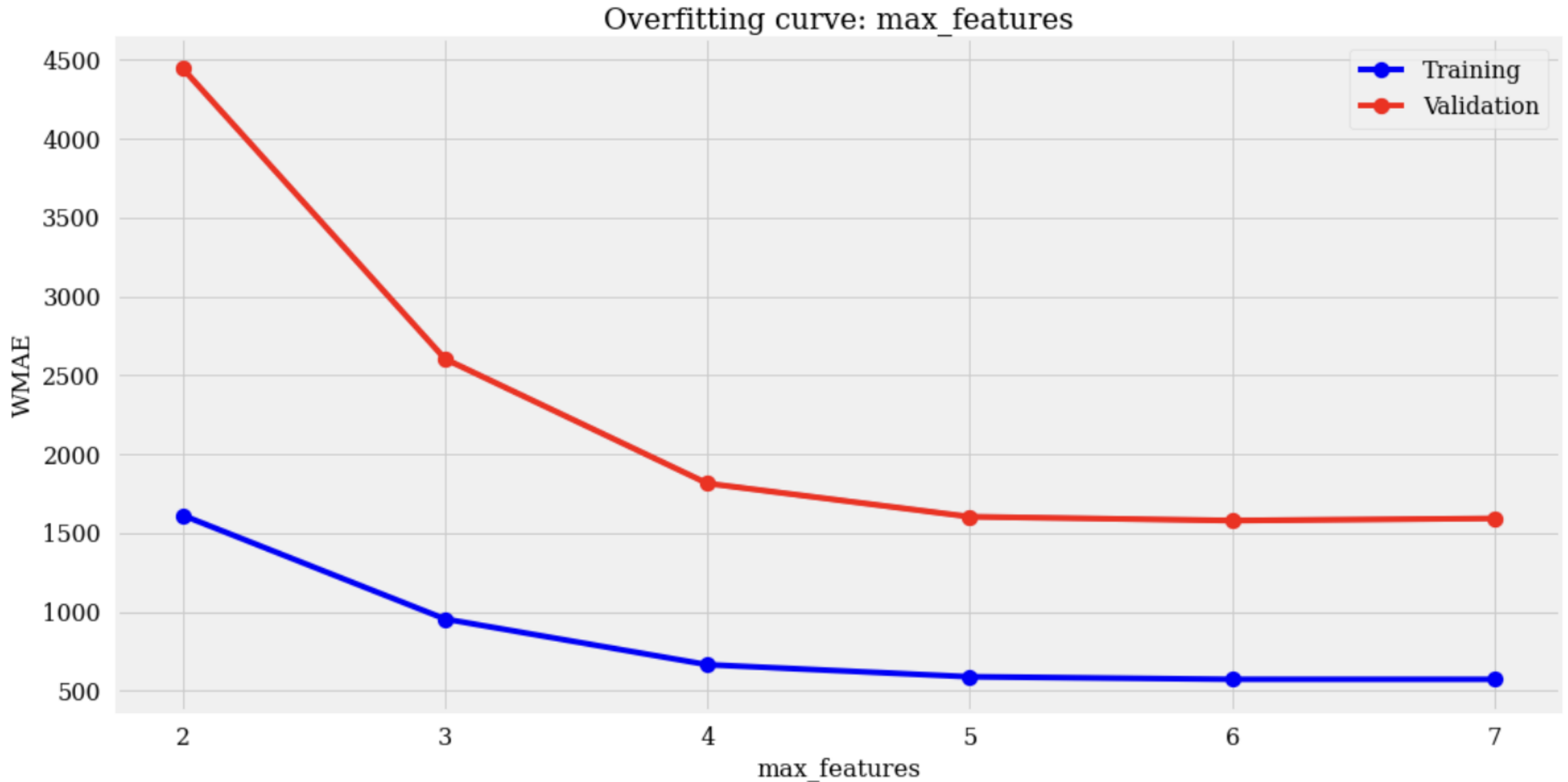
Overfitting Curve - min_samples_split:



Overfitting Curve - min_samples_samples:



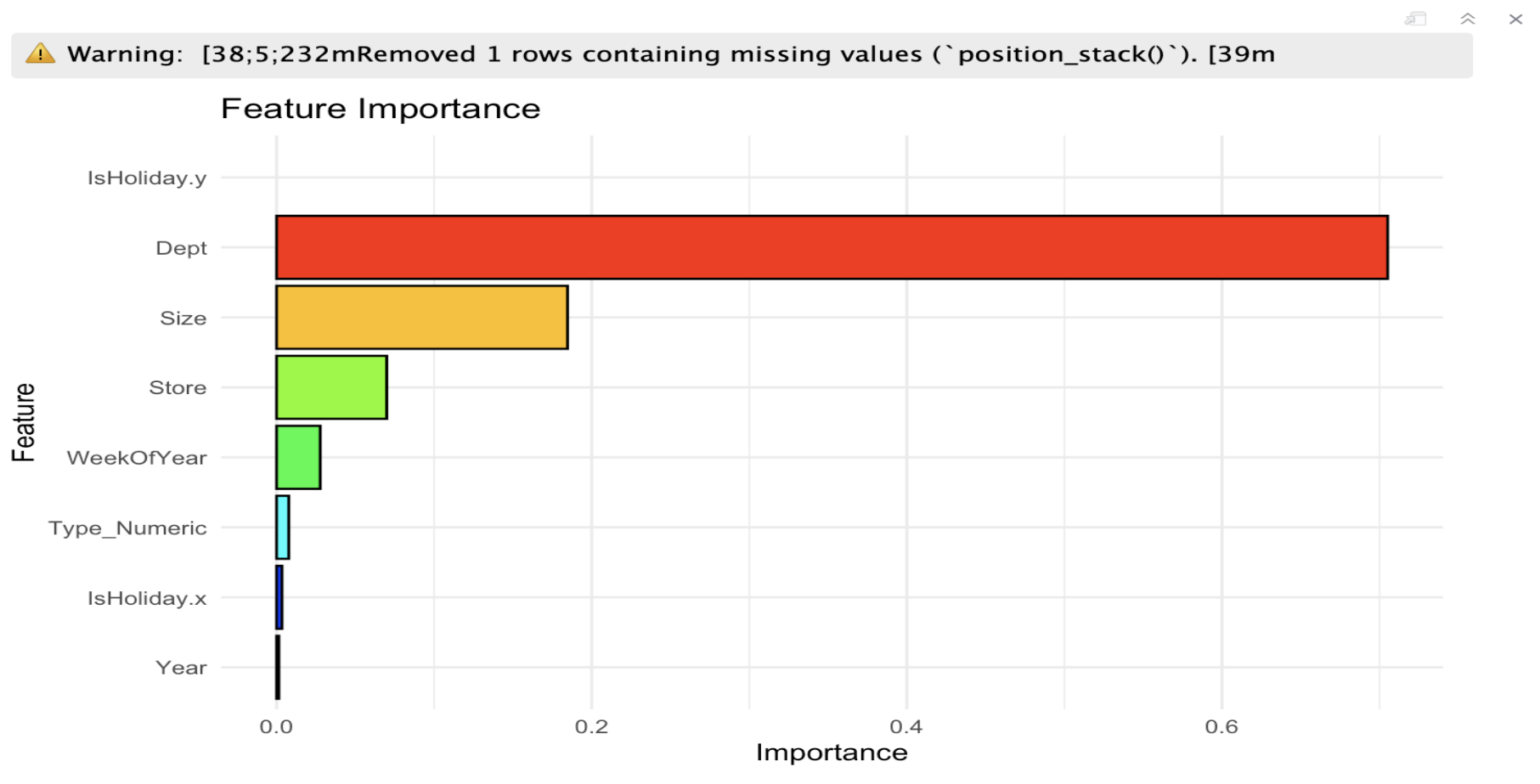
Overfitting Curve - max_features:



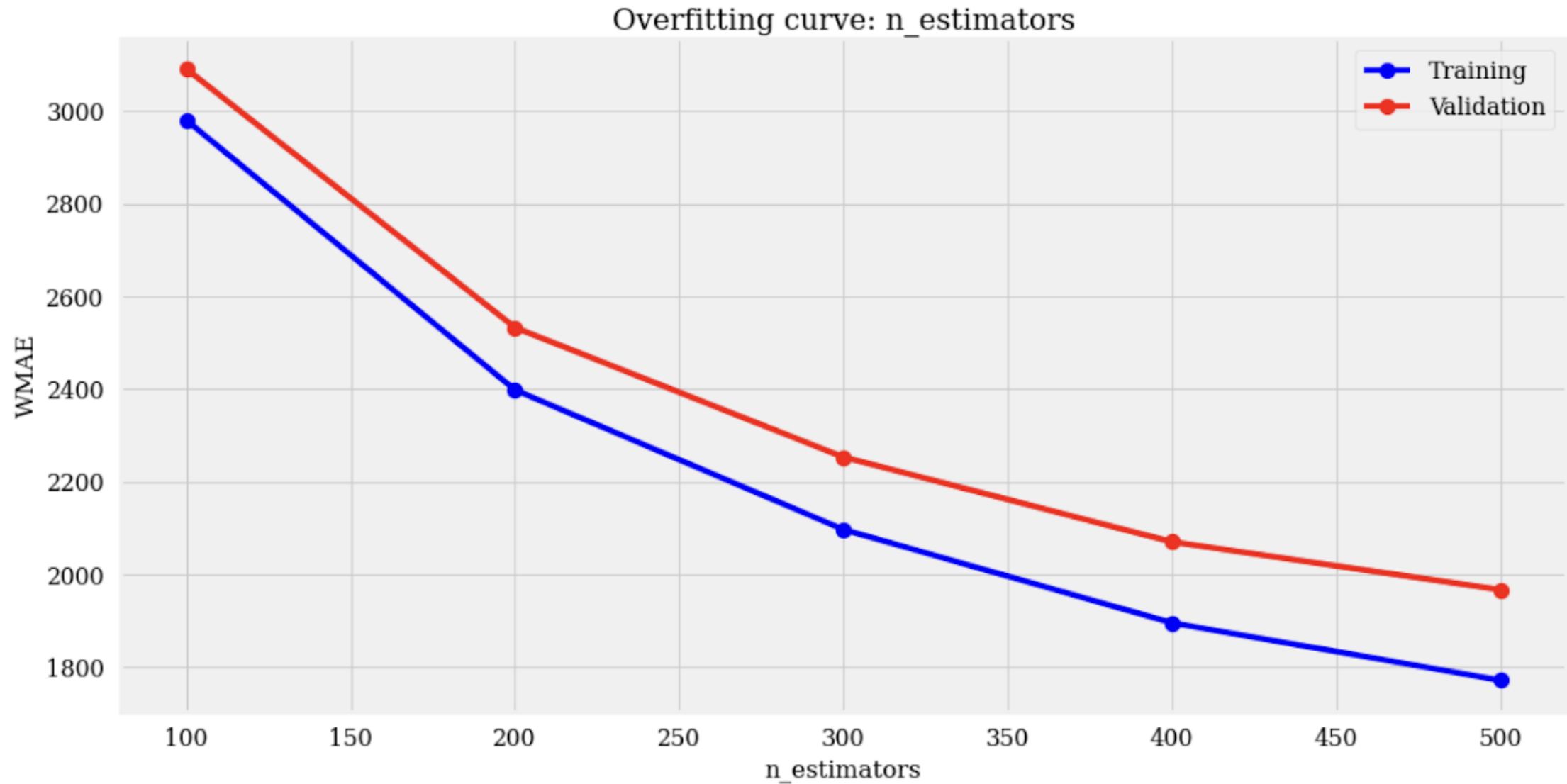
Gradient Boosting

- Utilizing Gradient Boosting for Walmart sales analysis yields superior predictive accuracy through ensemble learning. The model sequentially builds weak learners, refining predictions and capturing complex patterns, enhancing forecasting precision. Its adaptability makes it a powerful tool for optimizing sales predictions.

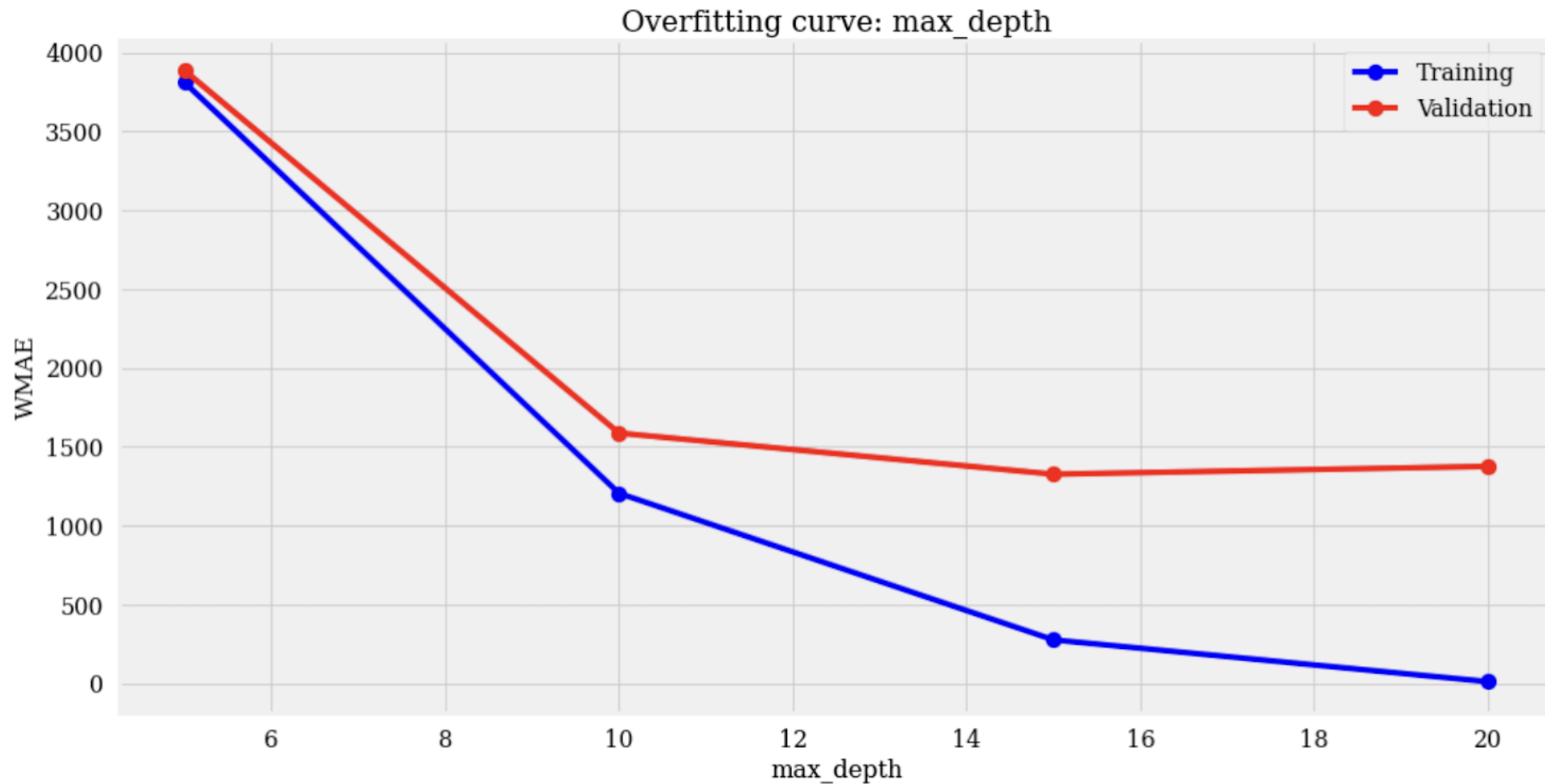
Feature Importance



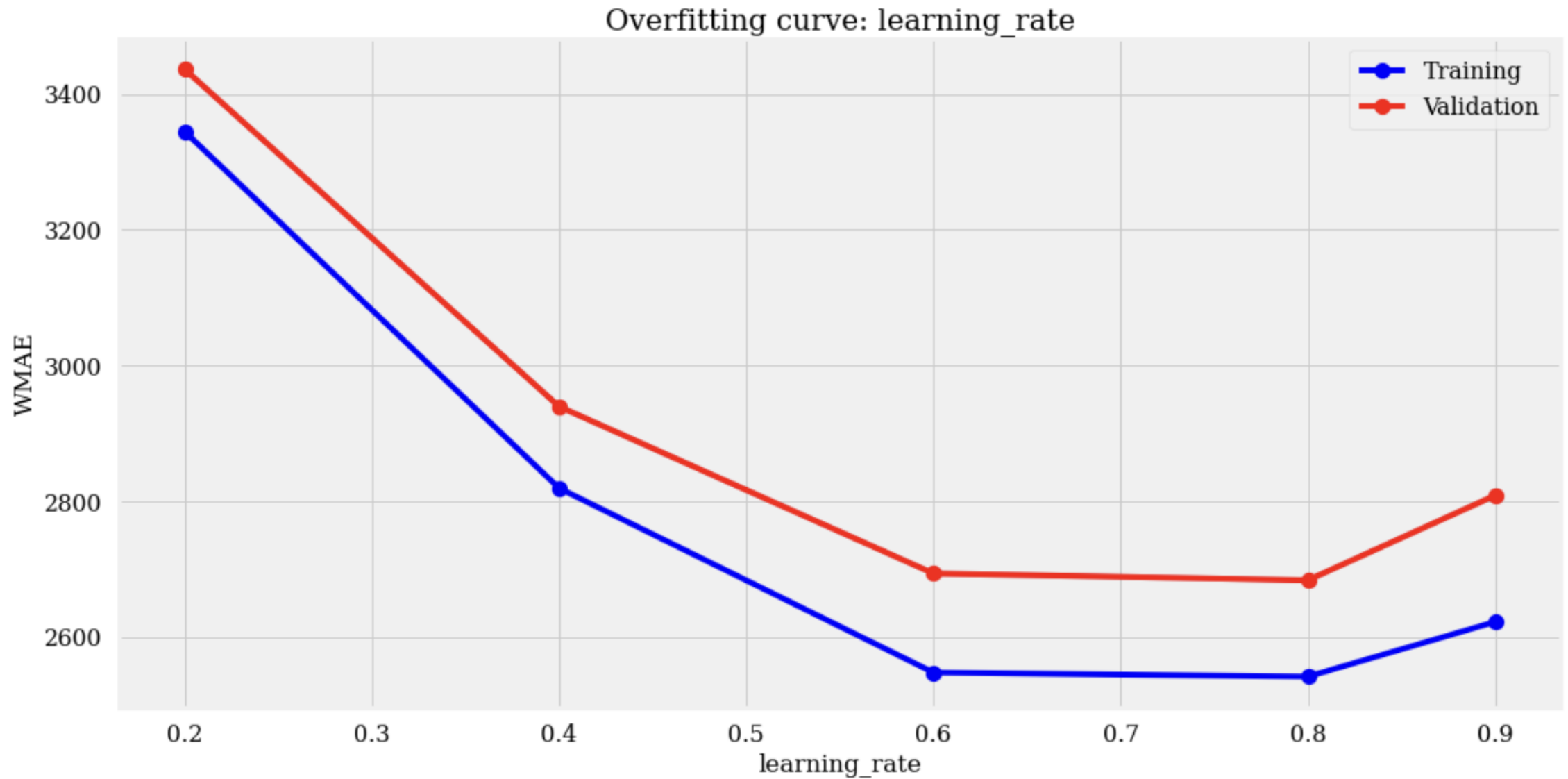
Tuning - Overfitting Curve : n_estimators:



Tuning - Overfitting Curve : max_depth:



Tuning - Overfitting Curve : learning_rate:



Models Comparison

- Type 'A' stores are more popular and outperform 'B' and 'C' types in terms of size and average weekly sales.
- Weekly sales are significantly influenced by the week of the year, with holiday weeks experiencing higher sales.
- Store size is a crucial factor, as larger stores contribute substantially to overall sales.
- Sales vary across departments, emphasizing the importance of considering departmental dynamics.
- The Gradient Boosting Machine with tuned hyperparameters is identified as the optimal model for predicting future sales, demonstrating superior performance compared to other models.
- In conclusion, the analysis highlights key factors impacting weekly sales and identifies an effective predictive model for future sales predictions.