Walmart Sales Analysis for Retail Industry With Machine Learning

1. Introduction

Sale forecasting is a crucial process for businesses, especially in the retail industry, as it enables companies to make informed decisions about inventory management, marketing strategies, and overall business planning. For Walmart, one of the largest retail corporations, accurately predicting future sales is essential for optimizing supply chain and customer satisfaction. This project focuses on leveraging historical sales data from 45 Walmart stores to forecast sales and understand the impact of major holidays such as Christmas, Thanksgiving, Super Bowl, and Labor Day on sales trends.

By employing machine learning algorithms such as Random Forest, Decision Tree, XGBoost, and ARIMA, this project aims to identify patterns in sales data and provide Walmart with valuable insights into sales performance around major holidays. The models will be integrated into a Flask application and deployed on IBM Cloud, enabling real-time sales forecasts and supporting Walmart in making data-driven decisions to enhance its operations and maximize profitability.

1.1. Project Overview

The sales forecasting project using Walmart data focuses on predicting future sales and understanding the impact of major holidays (Christmas, Thanksgiving, Super Bowl, and Labor Day) on store sales. It involves data preparation, feature engineering, and model selection using algorithms such as Random Forest, Decision Tree, XGBoost, and ARIMA. After training and validating the models, the best-performing model will be integrated into a Flask application and deployed on IBM Cloud for real-time sales forecasting. The project aims to provide Walmart with insights into sales trends and recommendations to optimize sales strategies around holidays, ultimately supporting better decision-making and maximizing profitability.

1.2. Objectives

The objective of the project is to develop a comprehensive sales analysis tool for Walmart, a leading retail corporation, leveraging machine learning techniques. By analyzing past sales data, store information, and promotional markdown events, the project aims to forecast future sales accurately. Specifically, the project aims to determine the impact of holidays, including Christmas, Thanksgiving, Super Bowl, and Labor Day, on store sales. The ultimate goal is to equip Walmart with a robust forecasting model that aids in informed decision-making, enhances performance predictions, and optimizes resource allocation.

2. Project Initialization and Planning Phase

The "Project Initialization and Planning Phase" marks the project's outset, defining goals, scope, and stakeholders. This crucial phase establishes project parameters, identifies key team members, allocates resources, and outlines a realistic timeline.

It also involves risk assessment and mitigation planning. Successful initiation sets the foundation for a well-organized and efficiently executed machine learning project, ensuring clarity, alignment, and proactive measures for potential challenges.

2.1. Define Problem Statement

The problem at hand is the development of an accurate sales forecasting model for Walmart, a well-known retail company that owns a hypermarket chain. Using historical sales data, store data and promotional events, the goal is to forecast future sales and estimate the impact of holidays, including Christmas, Thanksgiving, the Super Bowl and Labor Day, on store sales. The dataset provided contains data from 45 trades every week and the task involves using machine learning algorithms such as Random Forest, Decision Tree, XGBoost and ARIMA to analyze and model the data. In addition, the project includes integrating the developed model into a Flask application and deploying it on IBM Cloud for accessibility.

Walmart Sales Analysis Problem Statement Report: Click Here

2.2. Project Proposal (Proposed Solution)

The objective of the project is to develop a comprehensive sales analysis tool for Walmart, a leading retail corporation, leveraging machine learning techniques. By analyzing past sales data, store information, and promotional markdown events, the project aims to forecast future sales accurately. Specifically, the project aims to determine the impact of holidays, including Christmas, Thanksgiving, Super Bowl, and Labor Day, on store sales. The ultimate goal is to equip Walmart with a robust forecasting model that aids in informed decision-making, enhances performance predictions, and optimizes resource allocation.

Walmart Sales Analysis Project Proposal Report: Click Here

2.3. Initial Project Planning

Initial Project Planning involves outlining key objectives, defining scope, and identifying stakeholders for a loan approval system. It encompasses setting timelines, allocating resources, and determining the overall project strategy. During this phase, the team establishes a clear understanding of the dataset, formulates goals for analysis, and plans the workflow for data processing. Effective initial planning lays the foundation for a systematic and well-executed project, ensuring successful outcomes.

Walmart Sales Analysis Project Planning Report: Click Here

3. Data Collection and Preprocessing Phase

The Data Collection and Preprocessing Phase involves executing a plan to gather relevant Walmart sales Analysis data from Kaggle, ensuring data quality through verification and addressing missing values. Preprocessing tasks include cleaning, encoding, and organizing the dataset for subsequent exploratory analysis and machine learning model development.

3.1. Data Collection Plan, Raw Data Sources Identified, Data QualityReport

The dataset for "Walmart Sales Analysis for retail industry with machine learning" is sourced from Kaggle. The provided sample data represents a subset of the collected information, encompassing variables such as store information, department, size, weekly sales, temperature, holidays, markdowns. Data quality is ensured through thorough verification, addressing missing values, and maintaining adherence to ethical guidelines, establishing a reliable foundation for predictive modeling.

Walmart Sales Analysis Data Collection Report: Click Here

3.2. Data Quality Report

The Data Quality Report will summarize data quality issues from the selected source, including severity levels and resolution plans. It will aid in systematically identifying and rectifying data discrepancies. Missing values, handling missing values, dealing with outliers. Other issues like warnings, etc.

Walmart Sales Analysis Data Quality Report: Click Here

3.3. Data Exploration and Preprocessing

Data Exploration involves analyzing the Walmart sales dataset to understand patterns, distributions, and outliers. Preprocessing includes handling missing values, scaling, and encoding categorical variables. These crucial steps enhance data quality, ensuring the reliability and effectiveness of subsequent analyses in the sales prediction.

Walmart Sales Analysis Data Exploration and Preprocessing Report: Click Here

4. Model Development Phase

The Model Development Phase entails crafting a predictive model for sales prediction. It encompasses strategic feature selection, evaluating and selecting models (Random Forest, Decision Tree, XGB, Arima model), initiating training with code, and rigorously validating and assessing model performance for informed decision-making in the lending process.

4.1. Feature Selection Report

The Feature Selection Report outlines the rationale behind choosing specific features (e.g., Store, Department, isHoliday, Month, Year, Size) for the sales prediction model. It evaluates relevance, importance, and impact on predictive accuracy, ensuring the inclusion of key factors influencing the model's ability to discern credible loan applicants.

Walmart Sales Analysis Feature Selection Report: Click Here

4.2. Model Selection Report

The Model Selection Report details the rationale behind choosing Random Forest, Decision Tree, Arima, and XGB models for sales prediction. It considers each model's strengths in handling complex relationships, interpretability, adaptability, and overall predictive performance, ensuring an informed choice aligned with project objectives.

Walmart Sales Analysis Model Selection Report: Click Here

4.3. Initial Model Training Code, Model Validation and EvaluationReport

The Initial Model Training Code employs selected algorithms on the Walmart sales Analysis dataset, setting the foundation for predictive modeling. The subsequent Model Validation and Evaluation Report rigorously assesses model performance, employing metrics like accuracy and precision to ensure reliability and effectiveness in predicting loan outcomes.

Walmart Sales Analysis Model Development Phase Template: Click Here

5. Model Optimization and Tuning Phase

The Model Optimization and Tuning Phase involves refining machine learning models for peak performance. It includes optimized model code, fine-tuning hyperparameters, comparing performance metrics, and justifying the final model selection for enhanced predictive accuracy and efficiency.

5.1. Hyperparameter Tuning Documentation

The Random Forest model was selected for its superior performance, exhibiting high accuracy during hyperparameter tuning and after cross validation. Its ability to handle complex relationships, minimize overfitting, and optimize predictive accuracy aligns with project objectives, justifying its selection as the final model.

5.2. Performance Metrics Comparison Report

The Performance Metrics Comparison Report contrasts the baseline and optimized metrics for various models, specifically highlighting the enhanced performance of the Random Forest model. This assessment provides a clear understanding of the refined predictive capabilities achieved through hyperparameter tuning. Random Forest has the maximum accuracy. With minimum loss of information as compared to other models. The RMSE and MSE score is low for the random forest model making it a optimal model for sales prediction.

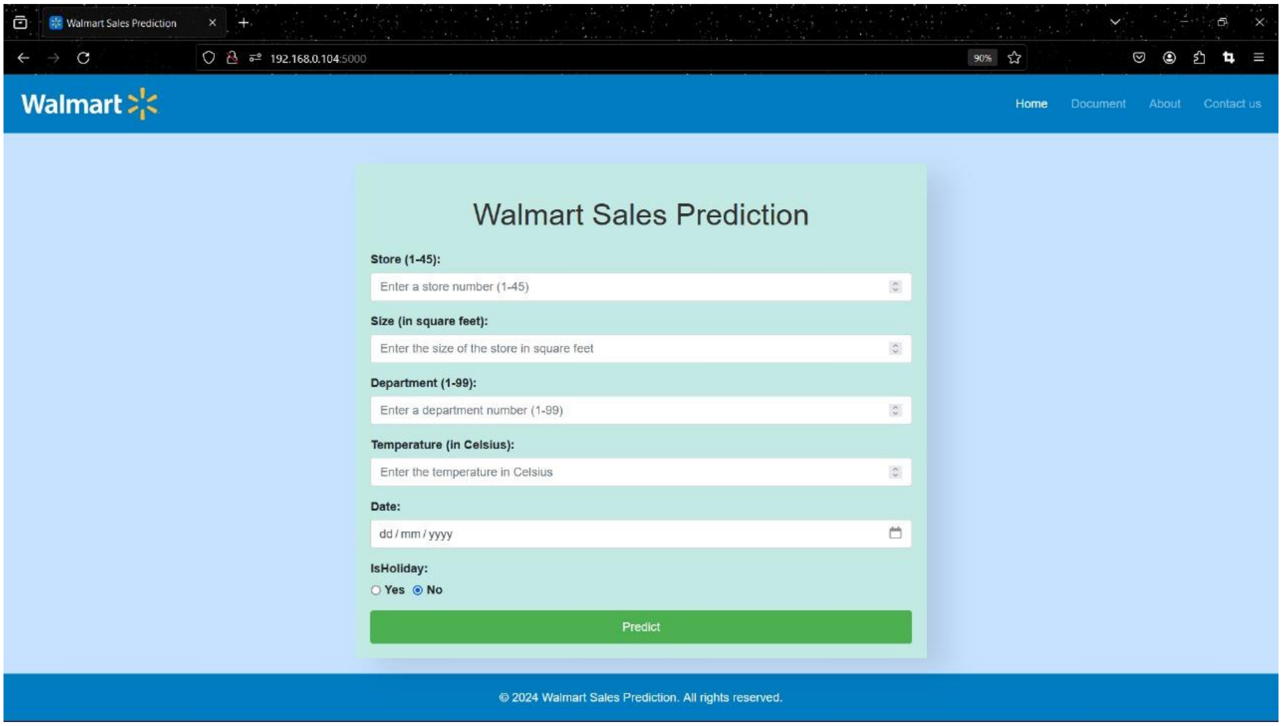
5.3. Final Model Selection Justification

The Final Model Selection Justification articulates the rationale for choosing Random Forest as the ultimate model. Its exceptional accuracy, ability to handle complexity, overcome the problem of overfitting and successful hyperparameter tuning align with project objectives, ensuring optimal sales predictions.

Walmart Sales Analysis Model Optimization and Tuning Phase Report: Click Here

6. Results

6.1 Output Screenshots



Walmart Sales Prediction

Store (1-45):
Enter a store number (1-45)

Size (in square feet):
Enter the size of the store in square feet

Department (1-99):
Enter a department number (1-99)

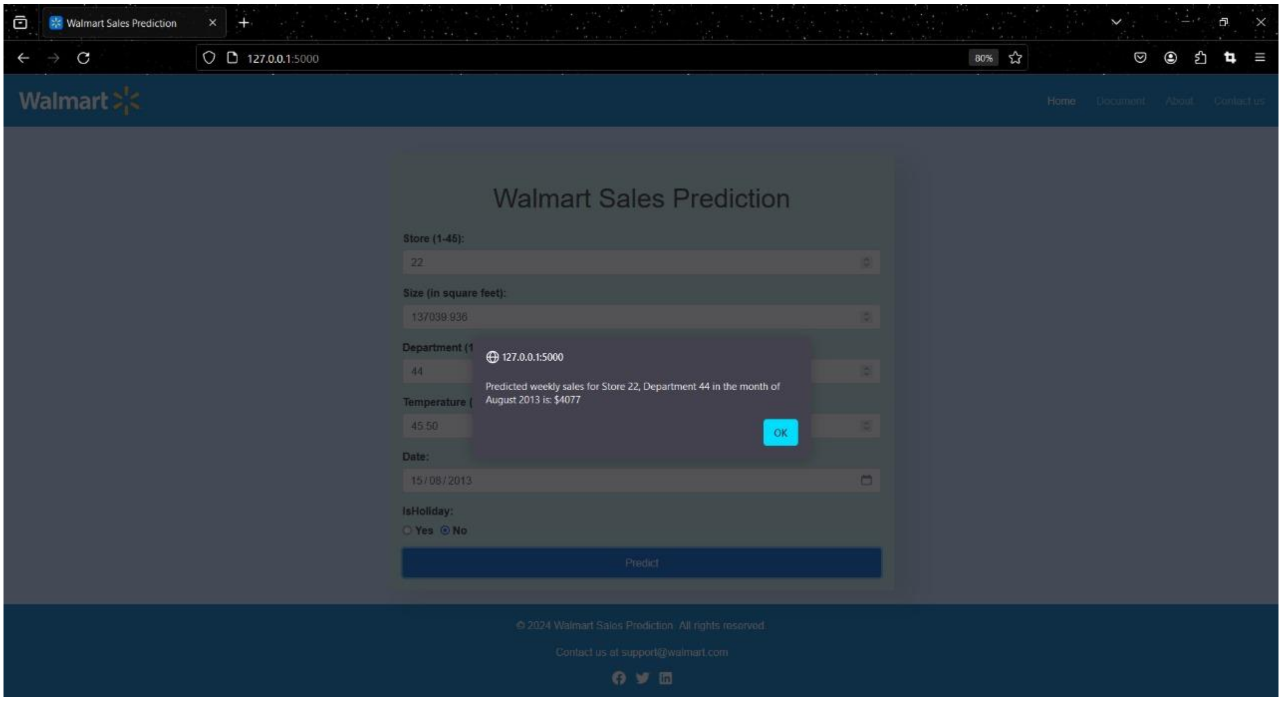
Temperature (in Celsius):
Enter the temperature in Celsius

Date:
dd / mm / yyyy

IsHoliday:
○ Yes ● No

Predict

© 2024 Walmart Sales Prediction. All rights reserved.



Walmart Sales Prediction

Store (1-45):
22

Size (in square feet):
137039.936

Department (1
44

Temperature (
45.50

Date:
15 / 08 / 2013

IsHoliday:
○ Yes ● No

Predict

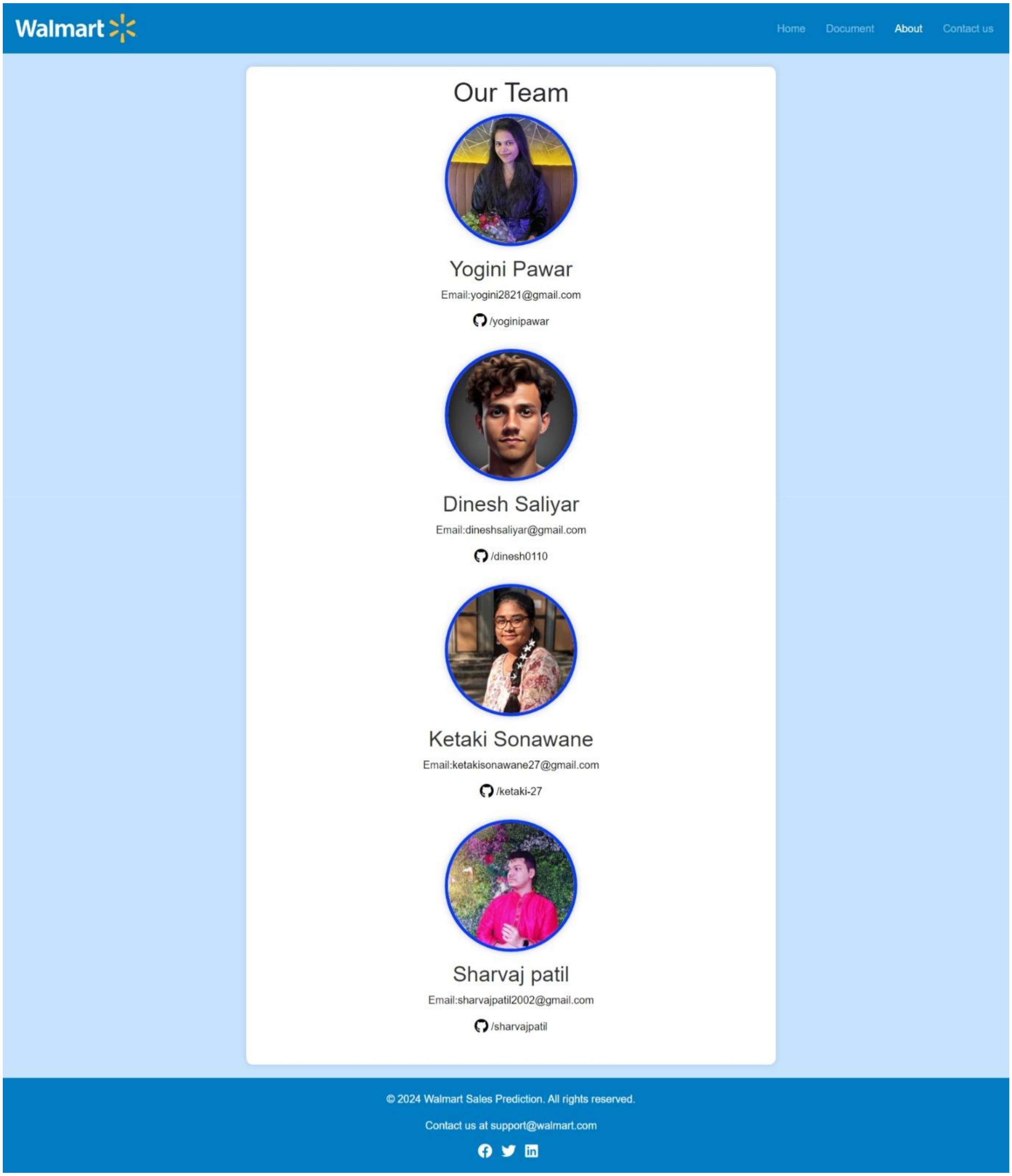
127.0.0.1:5000

Predicted weekly sales for Store 22, Department 44 in the month of August 2013 is: $4077

OK

© 2024 Walmart Sales Prediction. All rights reserved.

Contact us at support@walmart.com



Our Team

Yogini Pawar

Email:yogini2821@gmail.com

/yoginipawar

Dinesh Saliyar

Email:dineshsaliyar@gmail.com

/dinesh0110

Ketaki Sonawane

Email:ketakisonawane27@gmail.com

/ketaki-27

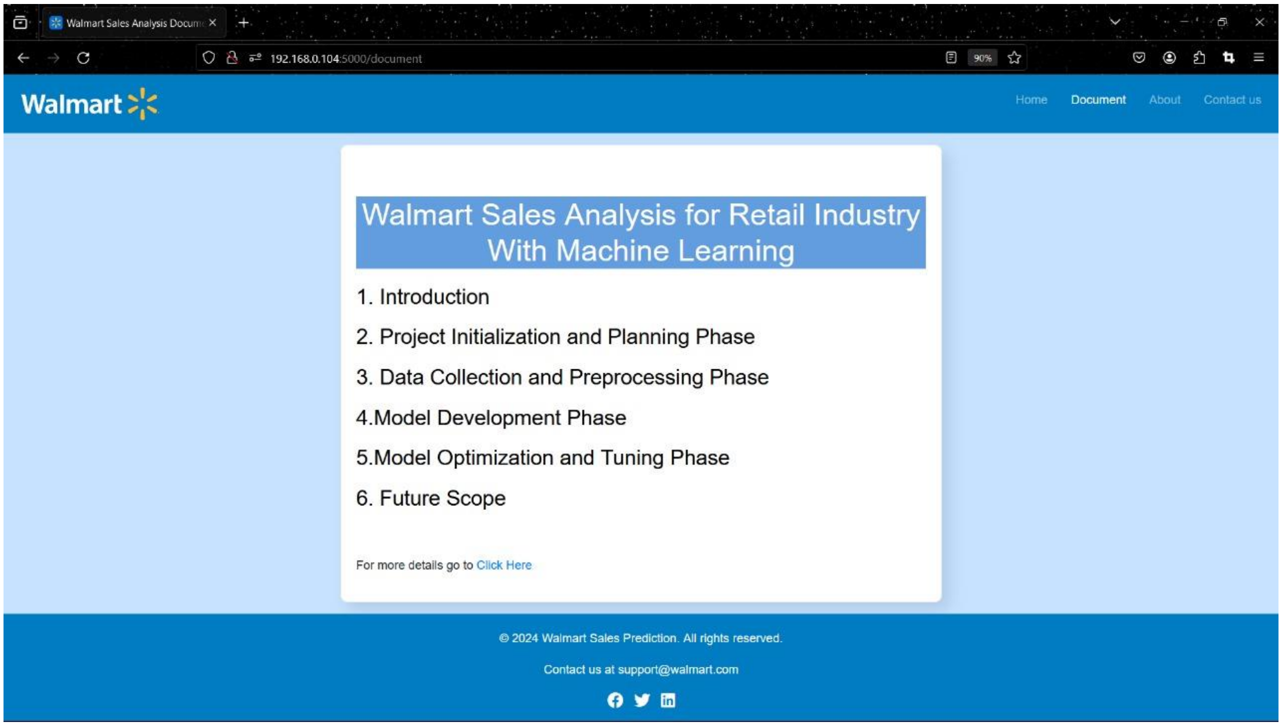
Sharvaj patil

Email:sharvajpatil2002@gmail.com

/sharvajpatil

© 2024 Walmart Sales Prediction. All rights reserved.

Contact us at support@walmart.com



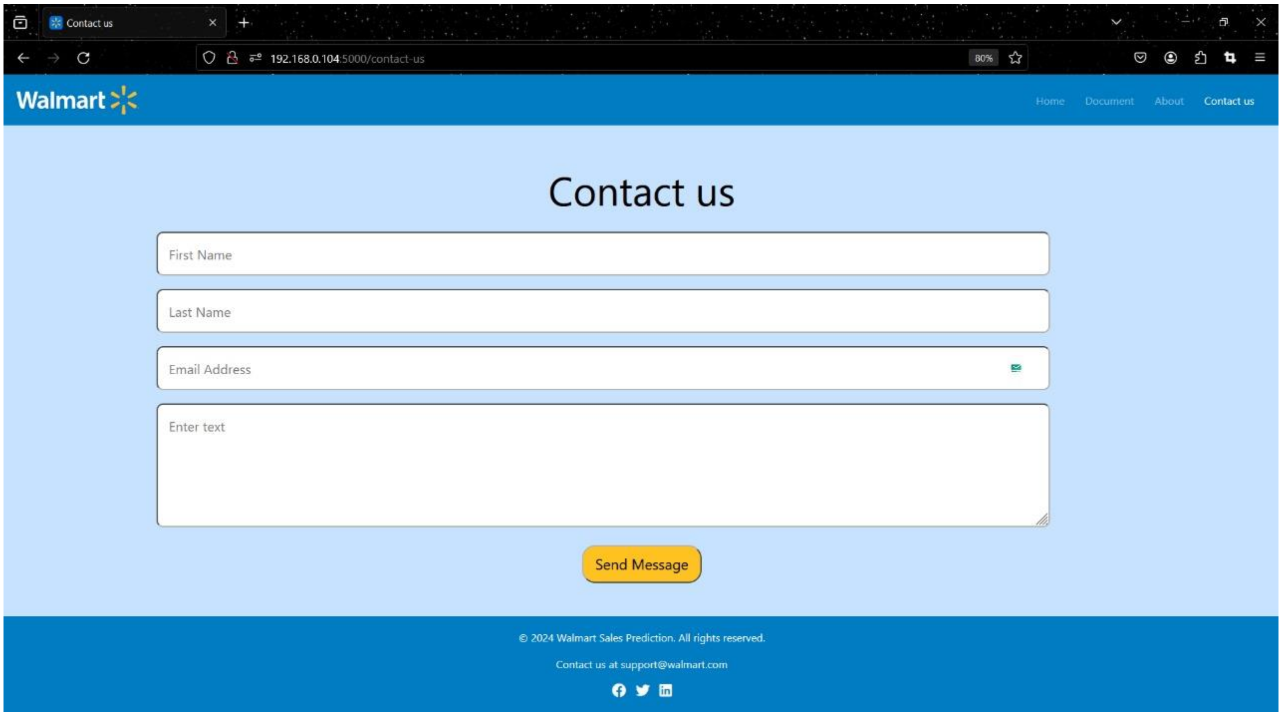
Walmart Sales Analysis for Retail Industry With Machine Learning

1. Introduction
2. Project Initialization and Planning Phase
3. Data Collection and Preprocessing Phase
4.Model Development Phase
5.Model Optimization and Tuning Phase
6. Future Scope

For more details go to Click Here

© 2024 Walmart Sales Prediction. All rights reserved.

Contact us at support@walmart.com



Contact us

First Name

Last Name

Email Address

Enter text

Send Message

© 2024 Walmart Sales Prediction. All rights reserved.

Contact us at support@walmart.com

7. Advantages and Disadvantages

Advantages

1. Accurate forecasting allows businesses to make informed decisions about inventory, staffing and promotions.
2. Machine learning models can continuously learn and adapt to changing trends.

Disadvantages

1. Model performance may degrade over time if the data distribution changes.
2. Requires expertise in data science and machine learning for implementation and maintenance.
3. The integration and deployment of models (e.g., with Flask and IBM Cloud) can be complex and time-consuming.

8. Conclusion

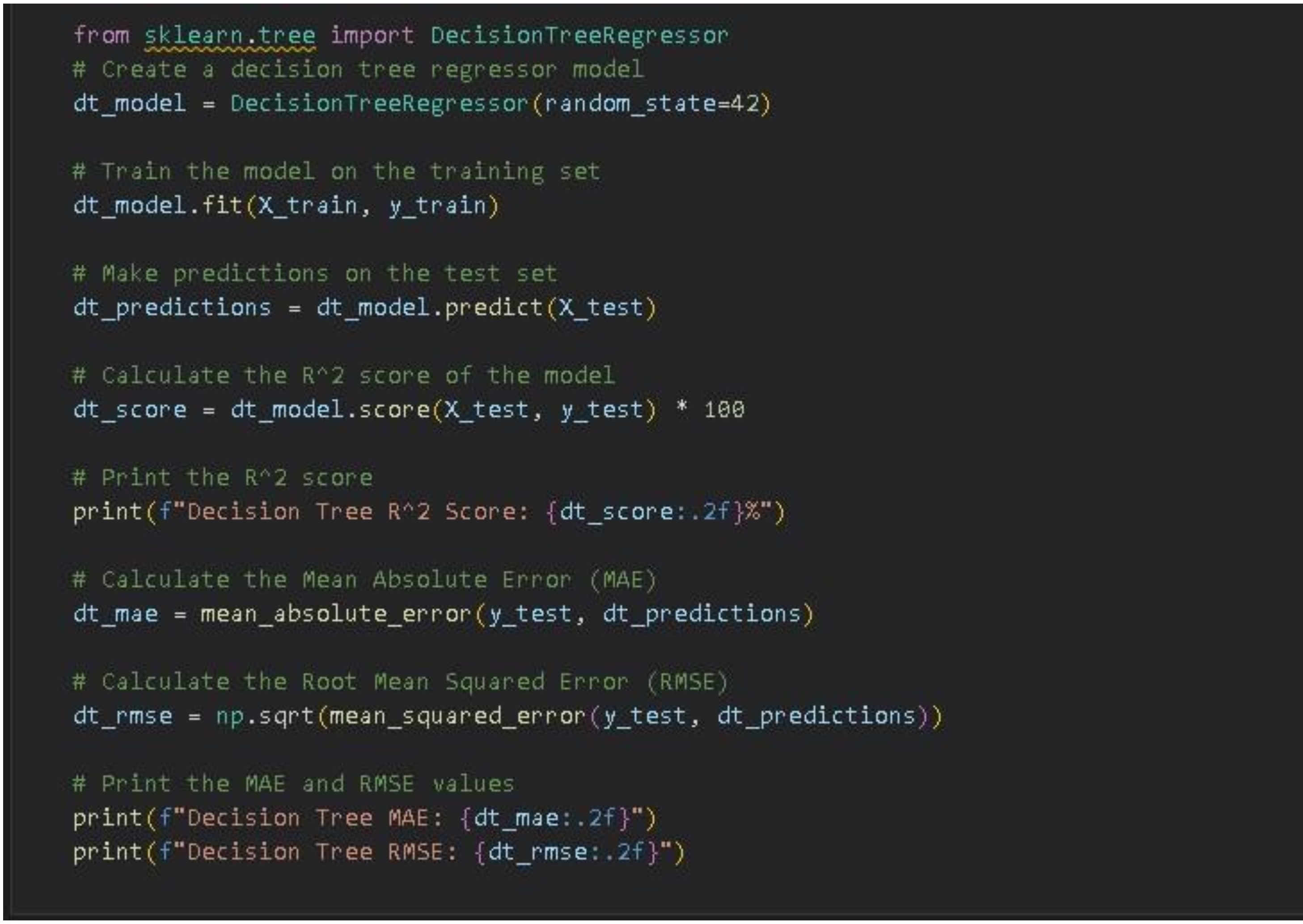
The conclusion of a sales forecasting project for Walmart, focusing on the impact of holidays on sales, includes identifying the best-performing model such as Random Forest, Decision Tree, XGBoost, or ARIMA based on evaluation metrics. The project quantifies the impact of holidays like Christmas, Thanksgiving, Super Bowl, and Labor Day on sales, enabling Walmart to plan promotions and inventory effectively. Business insights into customer behavior and sales trends inform strategic decision-making. The successful integration and deployment of the model via a Flask application and IBM Cloud allow for real-time access to forecasts. Ongoing monitoring and maintenance ensure the model's performance remains robust over time, providing Walmart with accurate forecasts to optimize business performance around holidays.

9. Future Scope

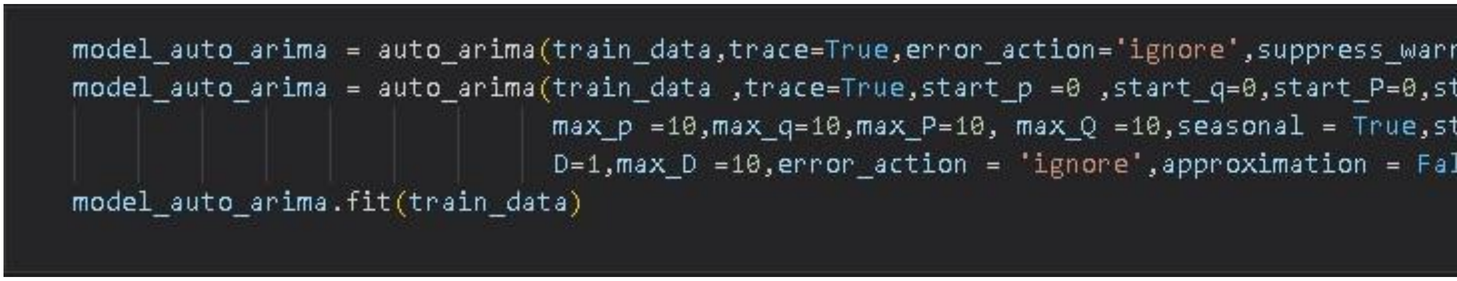
The future scope of the sales forecasting project using Walmart data includes incorporating additional data sources such as economic indicators, weather patterns, and regional events for a comprehensive understanding of sales trends. Advanced machine learning models like deep learning and ensemble methods can be explored to improve forecasting accuracy. Real-time data processing and forecasting will allow Walmart to quickly respond to market changes. Hyperparameter tuning and model optimization can further enhance accuracy and robustness. Geographical and store-level analysis can offer targeted insights for specific regions or stores. Seasonal and long-term trend analysis can guide strategic planning, while improved user experience and visualization will aid decision-making. Integrating the forecasting system with existing business processes such as inventory management and supply chain planning can lead to efficient operations. Additionally, exploring causal relationships between sales, promotions, and holidays can provide deeper insights into sales fluctuations.

10. Appendix

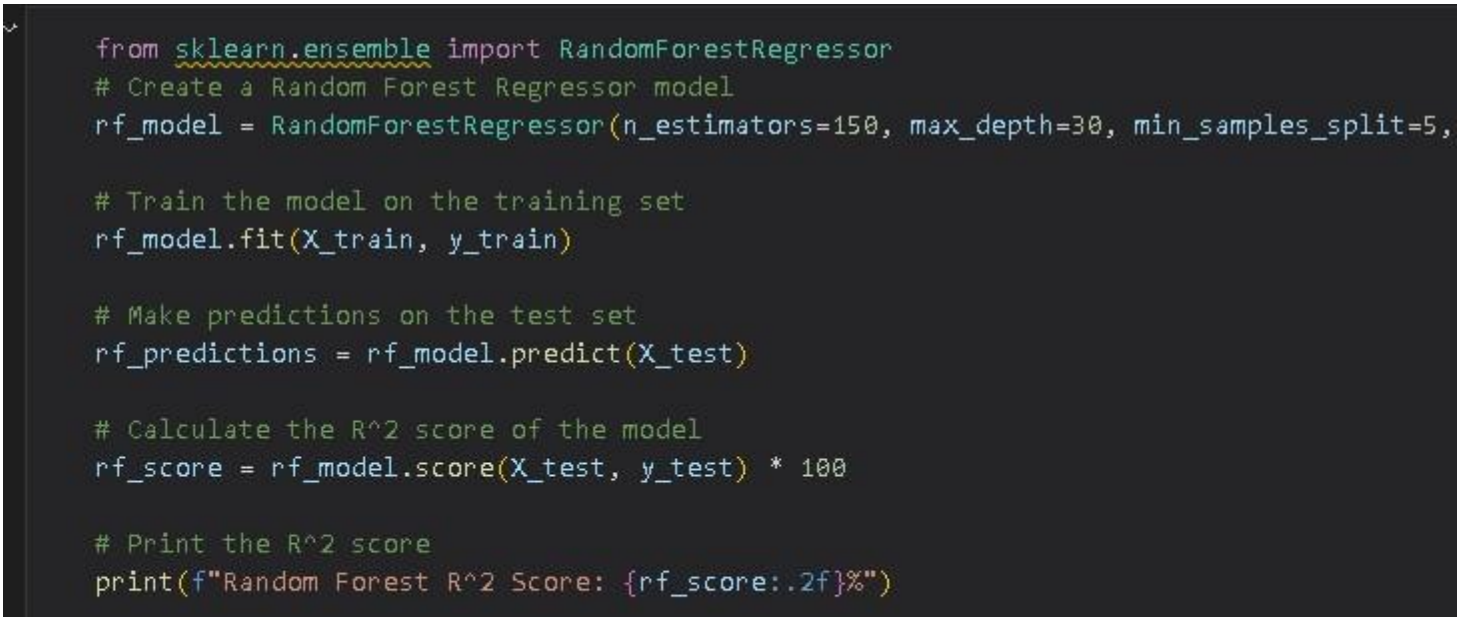
10.1. Source Code



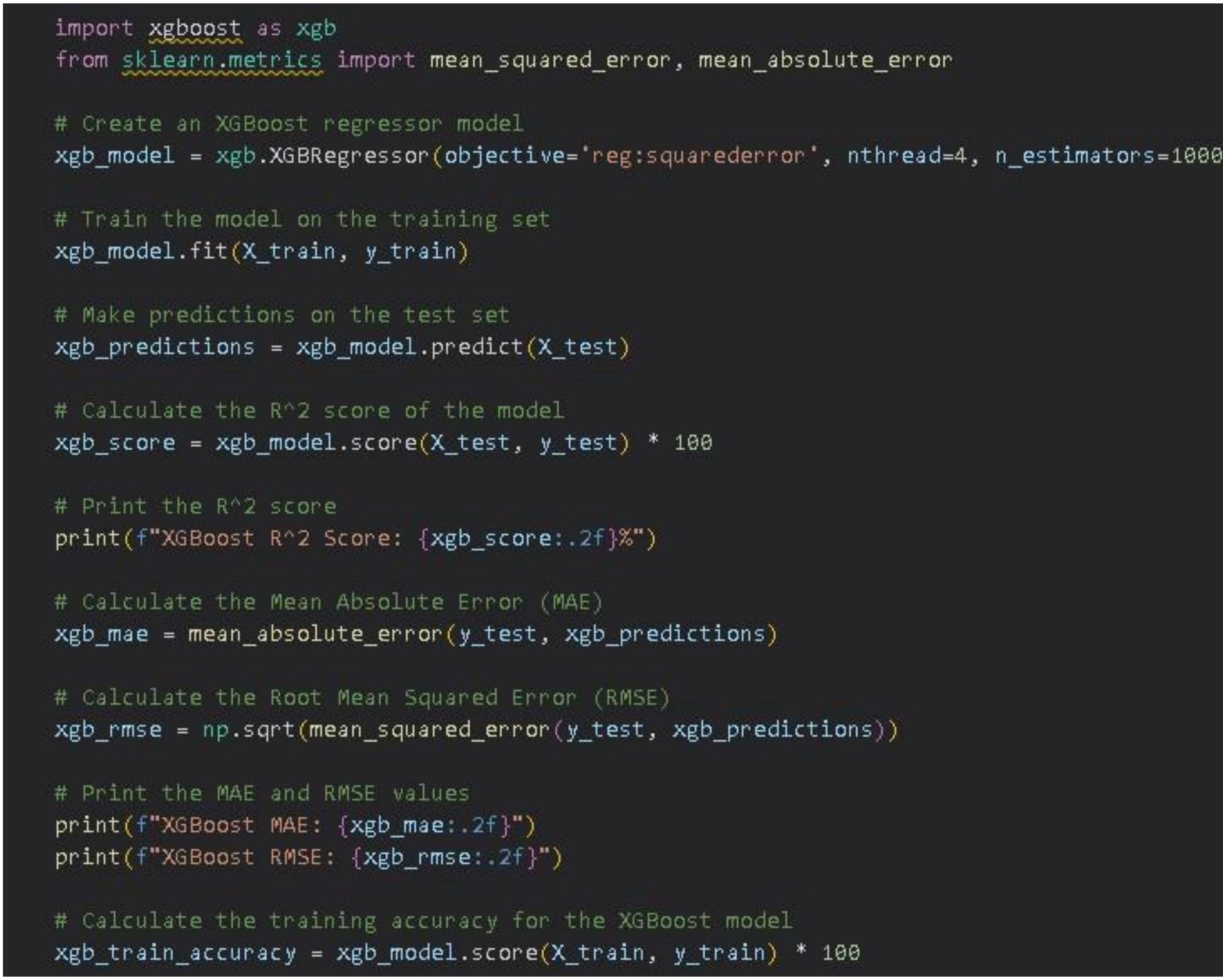
```
from sklearn.tree import DecisionTreeRegressor
# Create a decision tree regressor model
dt_model = DecisionTreeRegressor(random_state=42)

# Train the model on the training set
dt_model.fit(X_train, y_train)

# Make predictions on the test set
dt_predictions = dt_model.predict(X_test)

# Calculate the R^2 score of the model
dt_score = dt_model.score(X_test, y_test) * 100

# Print the R^2 score
print(f"Decision Tree R^2 Score: {dt_score:.2f}%")

# Calculate the Mean Absolute Error (MAE)
dt_mae = mean_absolute_error(y_test, dt_predictions)

# Calculate the Root Mean Squared Error (RMSE)
dt_rmse = np.sqrt(mean_squared_error(y_test, dt_predictions))

# Print the MAE and RMSE values
print(f"Decision Tree MAE: {dt_mae:.2f}")
print(f"Decision Tree RMSE: {dt_rmse:.2f}")
```



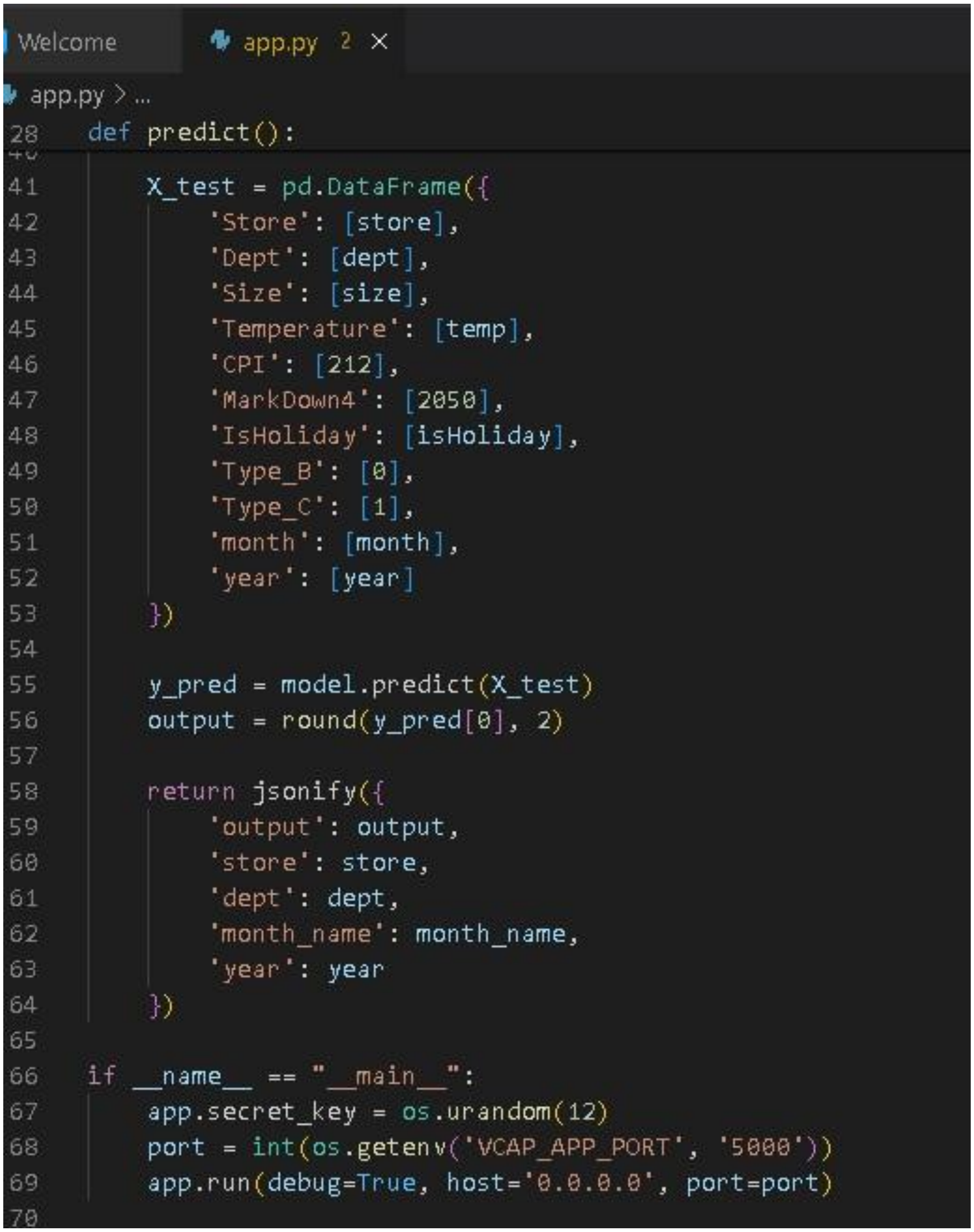
```
model_auto_arima = auto_arima(train_data,trace=True,error_action='ignore',suppress_warn
model_auto_arima = auto_arima(train_data ,trace=True,start_p =0 ,start_q=0,start_P=0,st
                              max_p =10,max_q=10,max_P=10, max_Q =10,seasonal = True,st
                              D=1,max_D =10,error_action = 'ignore',approximation = Fal
model_auto_arima.fit(train_data)
```



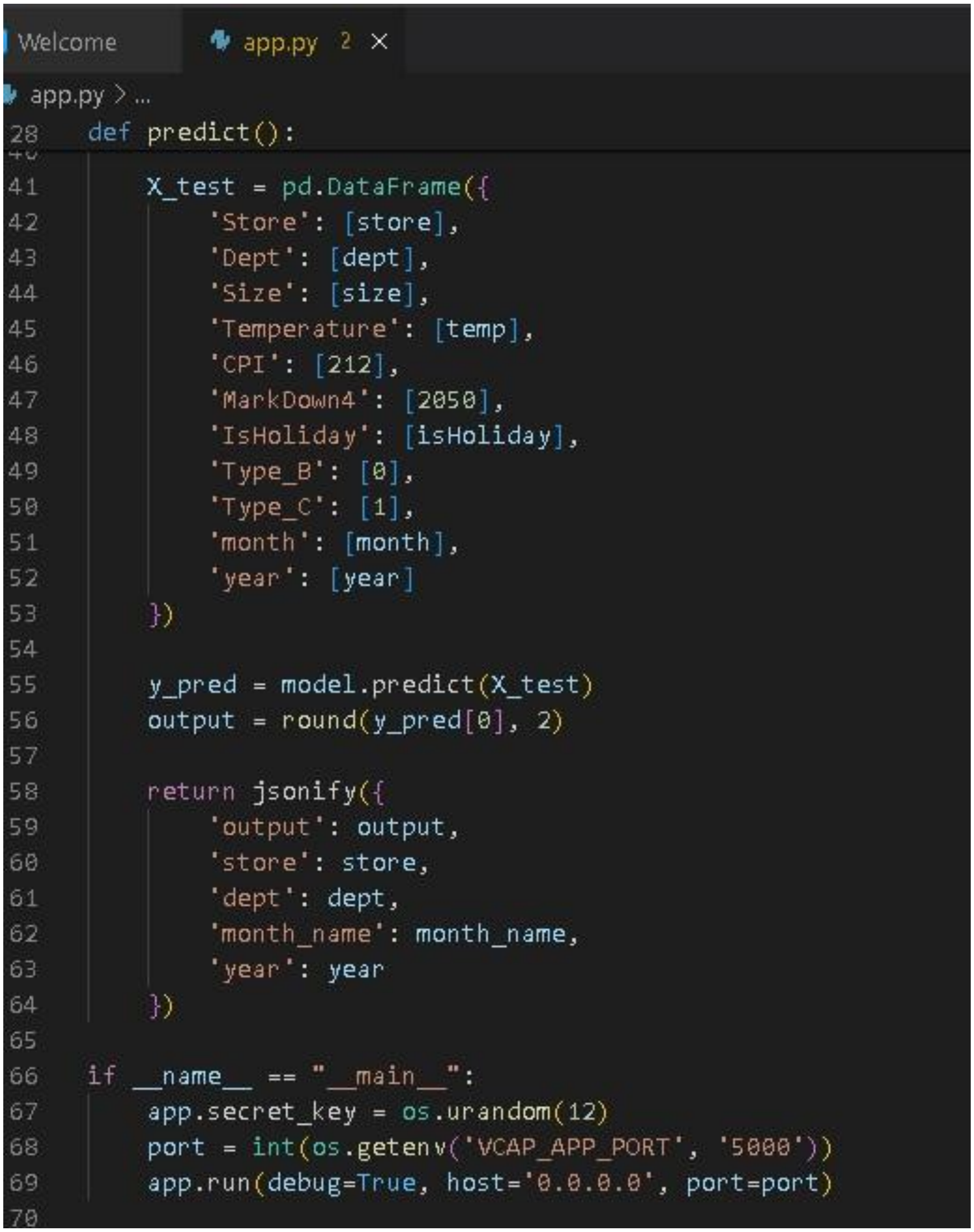
```
from sklearn.ensemble import RandomForestRegressor
# Create a Random Forest Regressor model
rf_model = RandomForestRegressor(n_estimators=150, max_depth=30, min_samples_split=5,

# Train the model on the training set
rf_model.fit(X_train, y_train)

# Make predictions on the test set
rf_predictions = rf_model.predict(X_test)

# Calculate the R^2 score of the model
rf_score = rf_model.score(X_test, y_test) * 100

# Print the R^2 score
print(f"Random Forest R^2 Score: {rf_score:.2f}%")
```



```
import xgboost as xgb
from sklearn.metrics import mean_squared_error, mean_absolute_error

# Create an XGBoost regressor model
xgb_model = xgb.XGBRegressor(objective='reg:squarederror', nthread=4, n_estimators=1000

# Train the model on the training set
xgb_model.fit(X_train, y_train)

# Make predictions on the test set
xgb_predictions = xgb_model.predict(X_test)

# Calculate the R^2 score of the model
xgb_score = xgb_model.score(X_test, y_test) * 100

# Print the R^2 score
print(f"XGBoost R^2 Score: {xgb_score:.2f}%")

# Calculate the Mean Absolute Error (MAE)
xgb_mae = mean_absolute_error(y_test, xgb_predictions)

# Calculate the Root Mean Squared Error (RMSE)
xgb_rmse = np.sqrt(mean_squared_error(y_test, xgb_predictions))

# Print the MAE and RMSE values
print(f"XGBoost MAE: {xgb_mae:.2f}")
print(f"XGBoost RMSE: {xgb_rmse:.2f}")

# Calculate the training accuracy for the XGBoost model
xgb_train_accuracy = xgb_model.score(X_train, y_train) * 100
```



```
def predict():
    X_test = pd.DataFrame({
        'Store': [store],
        'Dept': [dept],
        'Size': [size],
        'Temperature': [temp],
        'CPI': [212],
        'MarkDown4': [2050],
        'IsHoliday': [isHoliday],
        'Type_B': [0],
        'Type_C': [1],
        'month': [month],
        'year': [year]
    })

    y_pred = model.predict(X_test)
    output = round(y_pred[0], 2)

    return jsonify({
        'output': output,
        'store': store,
        'dept': dept,
        'month_name': month_name,
        'year': year
    })

if __name__ == "__main__":
    app.secret_key = os.urandom(12)
    port = int(os.getenv('VCAP_APP_PORT', '5000'))
    app.run(debug=True, host='0.0.0.0', port=port)

```



```
def predict():
    X_test = pd.DataFrame({
        'Store': [store],
        'Dept': [dept],
        'Size': [size],
        'Temperature': [temp],
        'CPI': [212],
        'MarkDown4': [2050],
        'IsHoliday': [isHoliday],
        'Type_B': [0],
        'Type_C': [1],
        'month': [month],
        'year': [year]
    })

    y_pred = model.predict(X_test)
    output = round(y_pred[0], 2)

    return jsonify({
        'output': output,
        'store': store,
        'dept': dept,
        'month_name': month_name,
        'year': year
    })

if __name__ == "__main__":
    app.secret_key = os.urandom(12)
    port = int(os.getenv('VCAP_APP_PORT', '5000'))
    app.run(debug=True, host='0.0.0.0', port=port)
```

10.2. GitHub & Project Demo Link

Project demo link

github link