Stores Sales Forecasting

Done by:

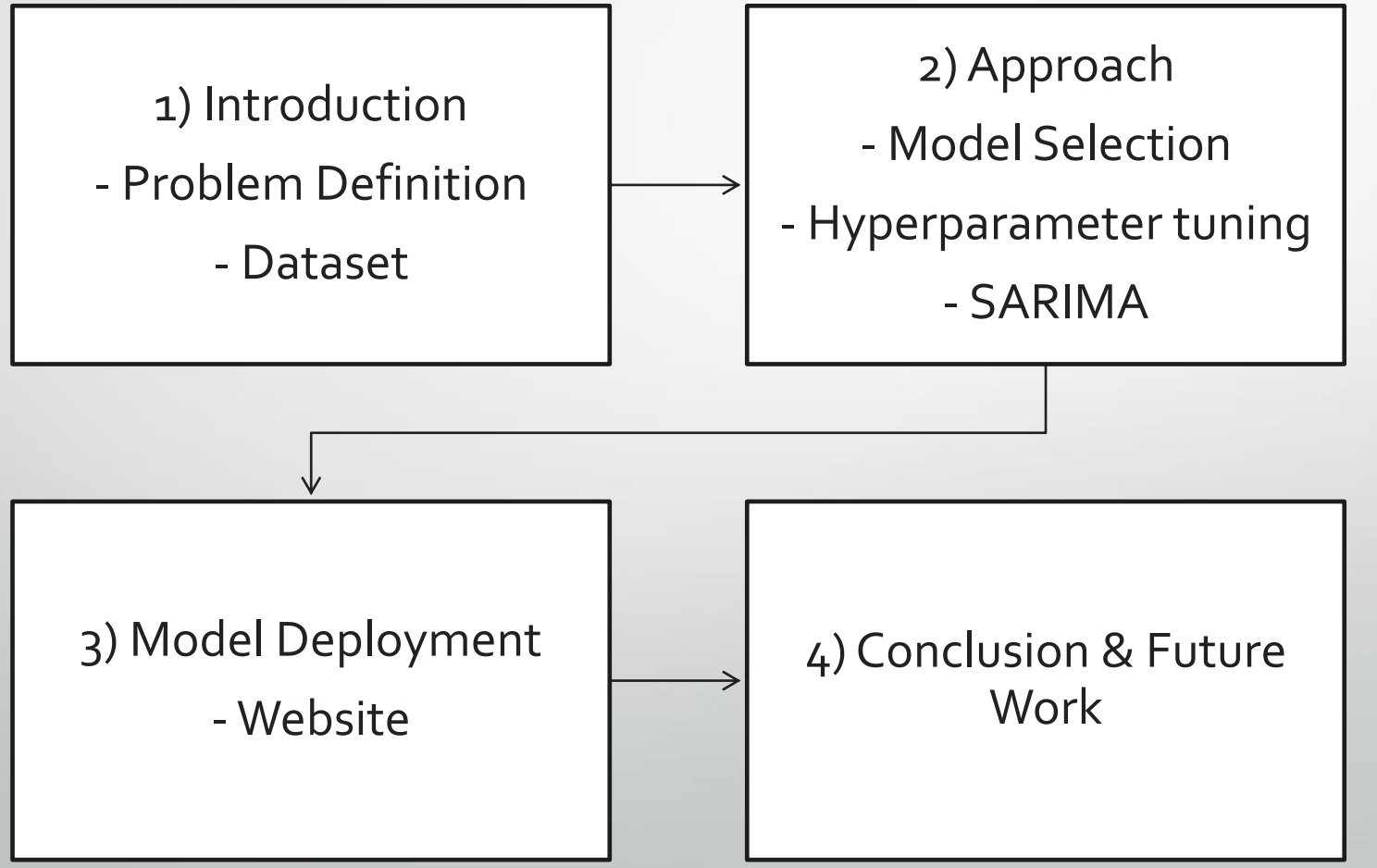
Hamza Naser

Diaa Alqadi

Dana Ghazal

Co-funded by the European Union

Outlines


1) Introduction
- Problem Definition
- Dataset
2) Approach
- Model Selection
- Hyperparameter tuning
- SARIMA
3) Model Deployment
- Website
4) Conclusion & Future Work

Introduction

Problem Definition

- **Project objective:** Enhance sales performance and drive growth in a competitive retail landscape.
- **Sales improvement goals:** Achieve a 10% increase in weekly sales over the next 3 months, through strategic expansion and improved customer engagement.
- **Machine learning forecasting:** Utilize machine learning algorithms to predict weekly sales using Walmart's historical sales data for 45 stores from February 2010 to October 2012.
- **Key benefits:** Accurate sales predictions, optimized inventory management, and data-driven decision-making for increased profitability.

Dataset

The final **merged** and **preprocessed** dataset includes various features.

	Store	Dept	Year	Month	Weekly_Sales	IsHoliday	Type	Size	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemployment
0	1	1	2010	2	24924.50	0	A	151315	5.727778	2.572	0.00	0.00	0.0	0.00	0.00	211.096358	8.106
1	1	2	2010	2	50605.27	0	A	151315	5.727778	2.572	0.00	0.00	0.0	0.00	0.00	211.096358	8.106
2	1	3	2010	2	13740.12	0	A	151315	5.727778	2.572	0.00	0.00	0.0	0.00	0.00	211.096358	8.106
3	1	4	2010	2	39954.04	0	A	151315	5.727778	2.572	0.00	0.00	0.0	0.00	0.00	211.096358	8.106
4	1	5	2010	2	32229.38	0	A	151315	5.727778	2.572	0.00	0.00	0.0	0.00	0.00	211.096358	8.106
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
421565	45	93	2012	10	2487.80	0	B	118221	14.916667	3.882	4018.91	58.08	100.0	211.94	858.33	192.308899	8.667
421566	45	94	2012	10	5203.31	0	B	118221	14.916667	3.882	4018.91	58.08	100.0	211.94	858.33	192.308899	8.667
421567	45	95	2012	10	56017.47	0	B	118221	14.916667	3.882	4018.91	58.08	100.0	211.94	858.33	192.308899	8.667
421568	45	97	2012	10	6817.48	0	B	118221	14.916667	3.882	4018.91	58.08	100.0	211.94	858.33	192.308899	8.667
421569	45	98	2012	10	1076.80	0	B	118221	14.916667	3.882	4018.91	58.08	100.0	211.94	858.33	192.308899	8.667

421570 rows × 17 columns

Approach

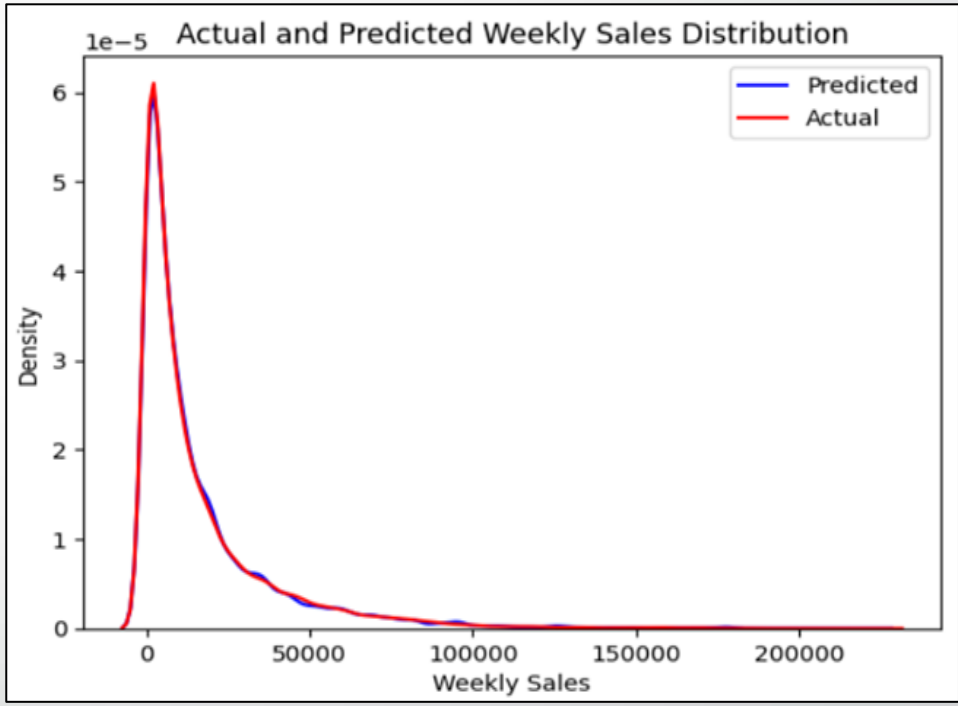
Model Selection

- Train test split: test split was determined based on last **7 months** of sales, 21% of the data.
- We prepared the data by performing one-hot encoding on categorical columns and scaling the numerical features.

```
Training time for LinearRegression(): 9.4810 seconds
Training time for Ridge(alpha=10): 1.8486 seconds
Training time for Lasso(alpha=10): 24.5039 seconds
Training time for DecisionTreeRegressor(): 25.5643 seconds
Training time for RandomForestRegressor(n_estimators=10): 161.5180 seconds

Model: LinearRegression
MAE: 7988.7424
RMSE: 12236.2510
R-squared: 0.6926

Model: Lasso
MAE: 7914.7157
RMSE: 12268.2787
R-squared: 0.6910

Model: RandomForest
MAE: 1974.3647
RMSE: 4123.2044
R-squared: 0.9651

Model: Ridge
MAE: 7979.8804
RMSE: 12231.8539
R-squared: 0.6928

Model: DecisionTree
MAE: 2533.2258
RMSE: 5404.2042
R-squared: 0.9400
```

Hyperparameter tuning

- We enhanced the Decision Tree Regressor's performance by tuning hyperparameters with **GridSearchCV** and converting categorical features to numeric using **OrdinalEncoder**.
- R-squared: **94.2%** in the test data, and visualized by KDE plot.




```
grid_search.best_params_
```

```
{'criterion': 'squared_error', 'max_depth': None, 'min_samples_split': 30}
```

Time Series Forecasting

- Time series forecasting is a technique for the prediction of events through a sequence of time. It predicts future events by analyzing the trends of the past, on the assumption that future trends will hold similar to historical trends.

ARIMA vs. SARIMA

ARIMA (Auto Regressive Integrated Moving Average)

- **Purpose:** ARIMA is used for time series forecasting, capturing non-seasonal patterns.
- **Components:** It consists of autoregressive (AR) and moving average (MA) components, along with differencing for stationarity.
- **Advantages:**

  - Effective for non-seasonal data.

  - Simplicity and ease of use.
- **Limitations:**

  - Not suitable for data with clear seasonality.

  - Assumes stationary data.

ARIMA vs. SARIMA

SARIMA (Seasonal Auto Regressive Integrated Moving Average)

- **Purpose:** SARIMA extends ARIMA to capture both non-seasonal and seasonal patterns in time series data.
- **Components:** SARIMA includes seasonal AR, seasonal differencing, and seasonal MA components, in addition to non-seasonal ones.
- **Advantages:**
  - Suitable for data with strong seasonal patterns.
  - Enhanced forecasting accuracy.
- **Limitations:**
  - More complex than ARIMA.
  - Requires identification of seasonal patterns.

Time Series Forecasting with SARIMA

- The ‘forecast’ function is a tool for time series forecasting, specifically tailored to predict weekly sales data for a designated store.

Time Series Forecasting with SARIMA

- **Key Steps:**

- Data is filtered to isolate information pertinent to the chosen store.
- Mean weekly sales are computed through date-based grouping.
- The dataset is partitioned into training and test sets.
- SARIMA model parameters are specified, and the model is trained on the training data.
- Forecasts are generated for a predefined future period.

Time Series Forecasting with SARIMA

- Forecasted values are refined based on the specified month and year.
- In cases with no forecasted values, observed data from the training set is employed.
- The mean weekly sales for the selected month and year are calculated and displayed.
- The R-squared score, indicating the model's goodness of fit, is presented.
- A visual representation illustrates the observed and forecasted sales data.

Time Series Forecasting with SARIMA

- **Bias Detection:**

Bias is evaluated by quantifying forecast errors and subjecting them to a hypothesis test for statistical significance.

A bias near zero signifies that the model's forecasts are, on average, accurate.

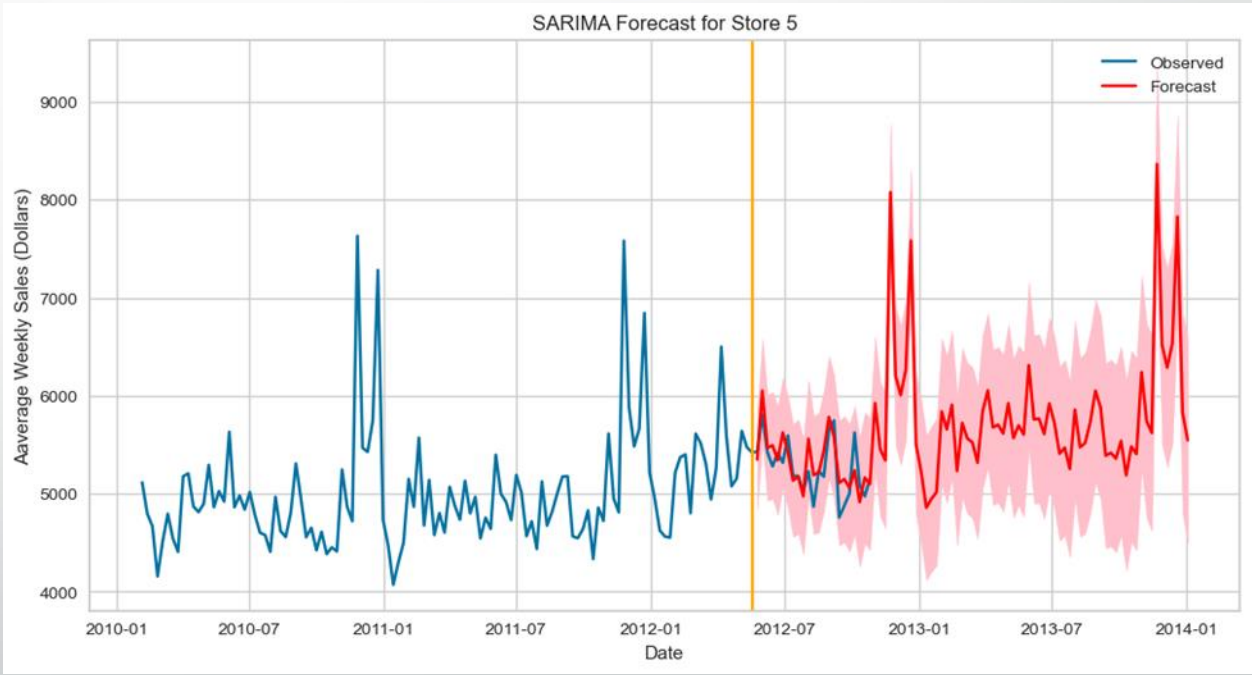
Running An Example

- Forecasting weekly sales for store 5 until the end of 2013 and finding the average weekly sales for January,2013:

The forecast does not have a significant bias.

The mean Weekly sales for month 1 and year 2013 is 5011.95$

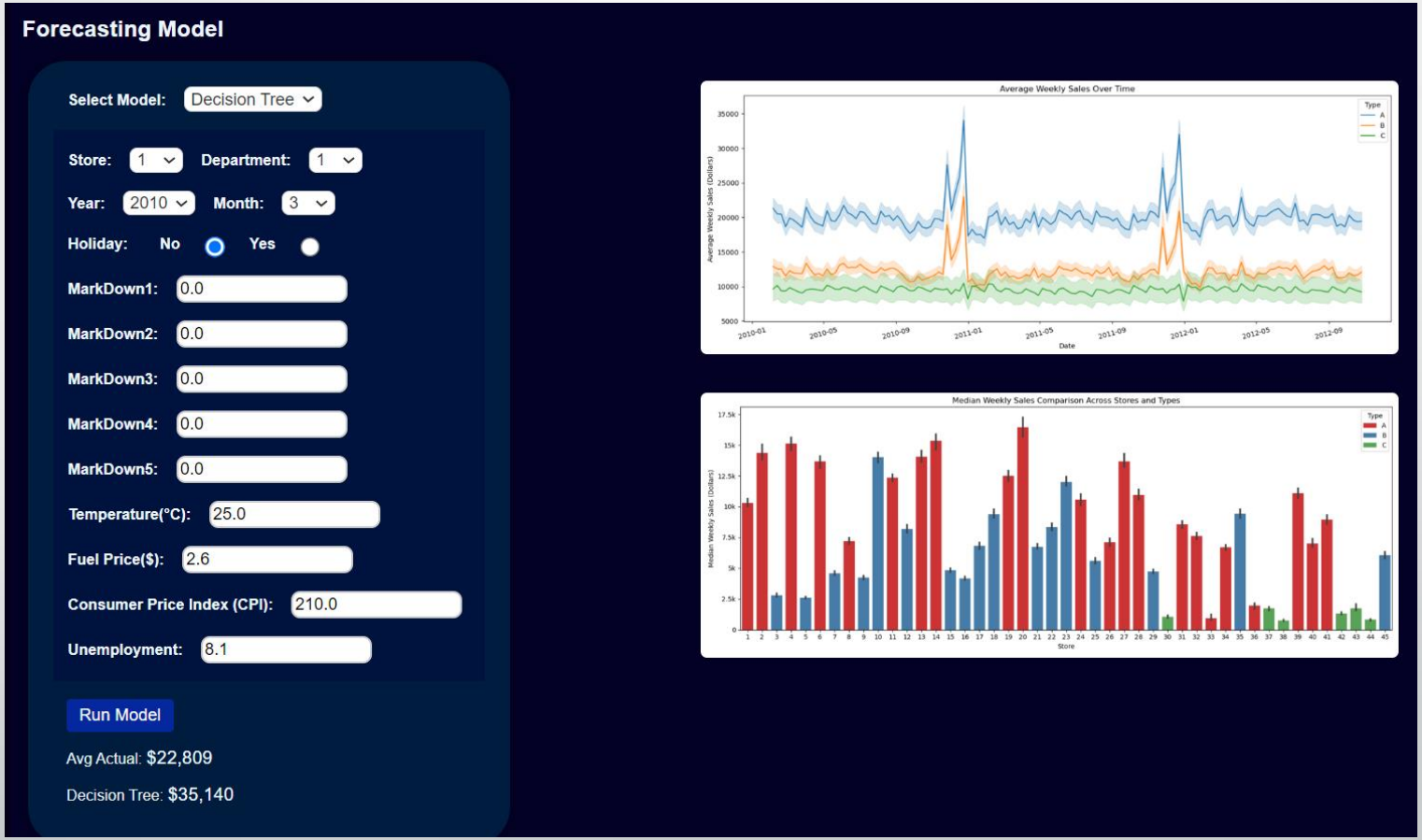
The R-Squared for Test set is equal to 0.456



Model Deployment

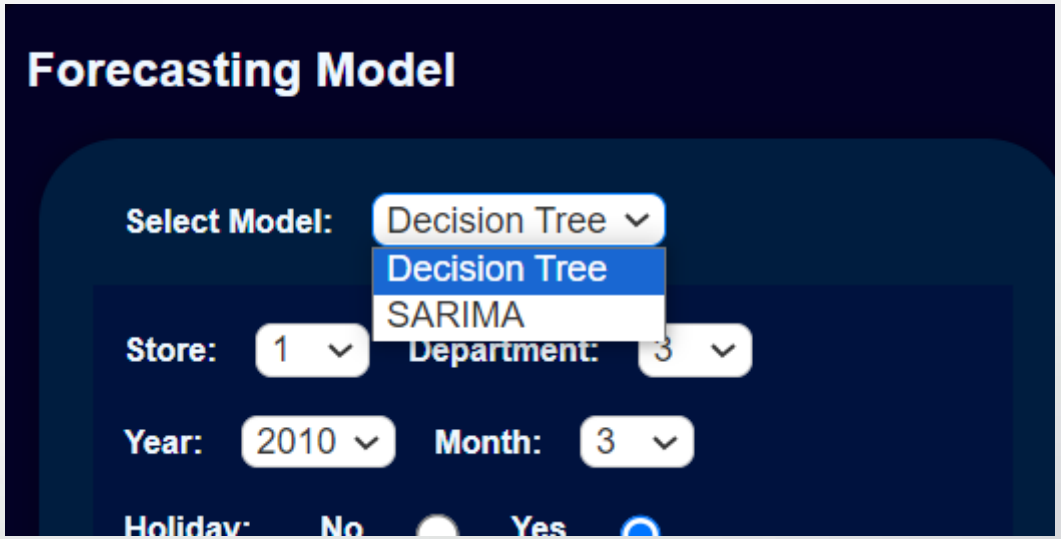
Web-app

- HTML, CSS, JS
- Python
- Flask app
- Cloud server (link)



Models

- Decision Tree
- SARIMA

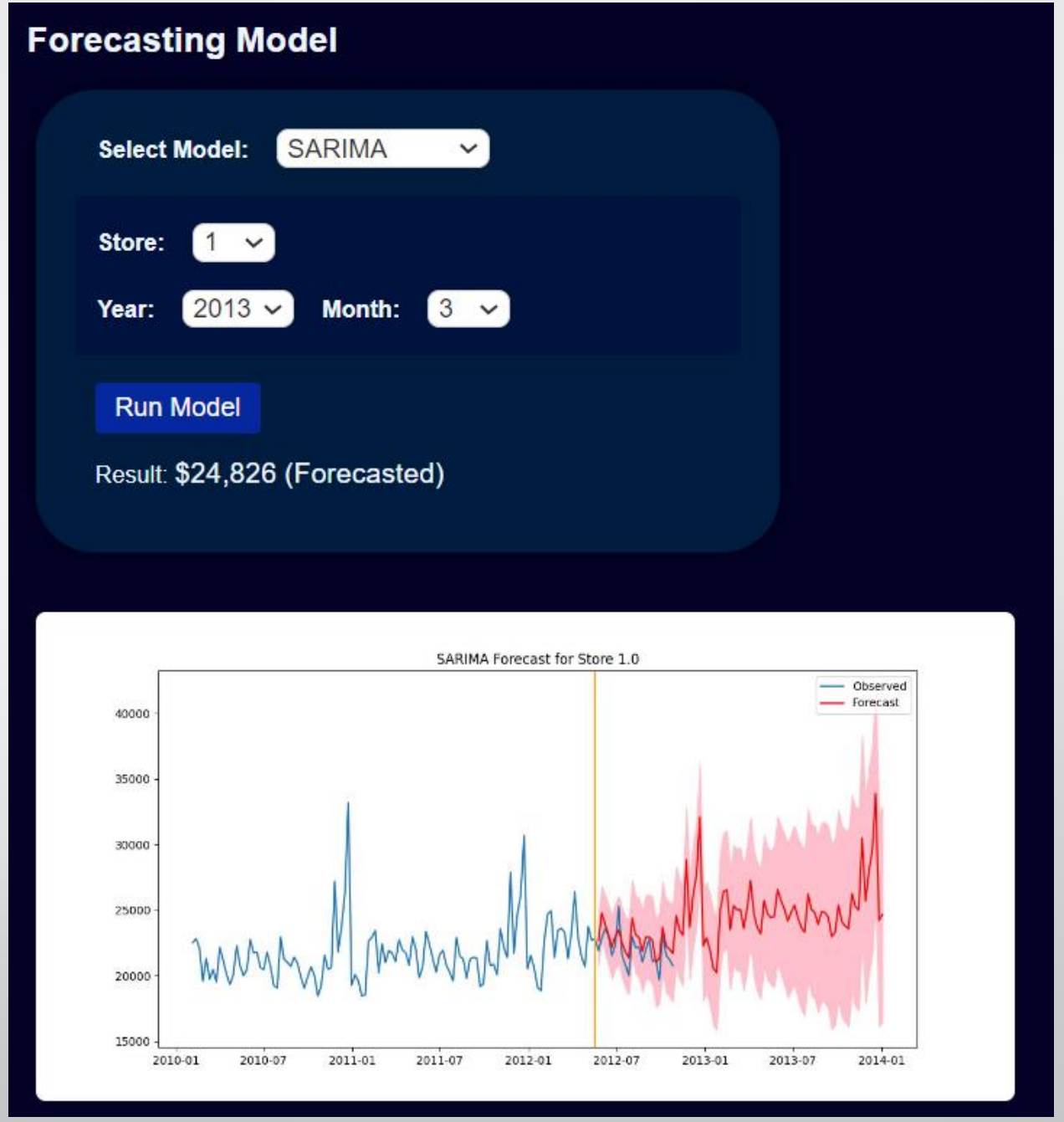


Decision Tree

- Machine Learning Model.
- Result affected by several features.

SARIMA

- Time Series Model.
- Results affected by date and sales pattern.



Challenges

- Web development Knowledge.
- Web Server not working on the cloud:
  - **Disk Space**
  - (Libraries and Frameworks) **Versions Mismatch**.

Conclusion & Future Work

Conclusion

- Project outcome (Website).
- Better Expansion decisions based on forecasted sales.
- Better Sales prediction based on Economics, Climate Change and Seasonality.

Future Improvement

- Consider to use other Forecasting model techniques (STL, Holt-winters, Exponential Smoothing,….).
- User Interface Enhancements.
- Apply the model in various domains.

Thank you