Predicting Future Rossmann Store Sales

Kamil Pacana, Julian Savini, and Ved Rajesh

Northeastern University

Introduction/Abstract

This project explores whether simple machine learning models that we learned in a undergraduate machine learning course can be used to accurately predict future retail sales. Our work examines the Rossmann store dataset and implements two ML models, Multi-Layer Perceptron (MLP) and SARIMA.

Motivation and Data

Being able to accurately forecast sales give companies an upper hand in the business world. Accurately forecasting future sales allows businesses to make informed decisions on inventory management, staffing, and budgeting. It can greatly increase customer satisfaction by ensuring products are available when needed and reduces costs by preventing excess inventory. .

During the COVID-19 pandemic, there was a shortage of toilet paper across the country. Could a machine learning model have predicted customer behavior, enabling stores to prepare more strategically?

We used the Rossmann Store Sales dataset. The set included:

- Sales data for 1,000 stores from Jan 2013 to July 2015
- Daily records include sales, open/closed status, and store type
- External data includes store metadata (e.g., competition distance, promo start dates)

For our project, we narrowed down the scope to focus solely on one store. Furthermore, we aggregated daily sales into weekly sales to reduce noise for our SARIMA model.

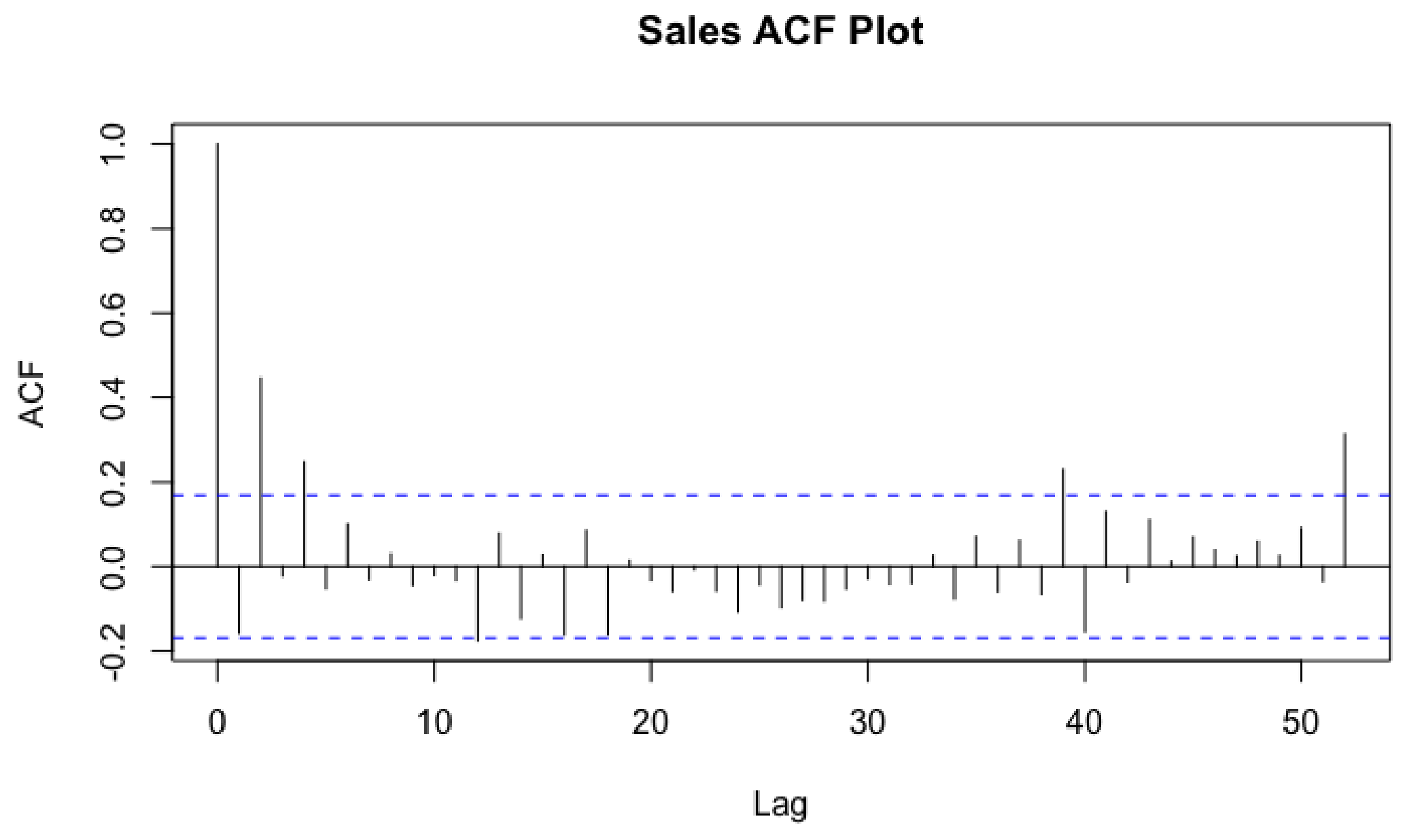


Fig. 1: Autocorrelation Function (ACF) plot of weekly sales.

During our exploratory data analysis, we ran an autocorrelation function (ACF) plot on the sales data. One of the most important takeaways was that the correlation in sales wasn't significant. This suggested that a model relying on recent sales would likely under perform, so we knew it would be very important to incorporate other measures into our models.

Methodology

MLP Model: We processed the Rossmann dataset by:

- Adding lag features
- Creating rolling means and standard deviations for 7, 14, and 28 days
- Generating exponentially weighted features with multiple alpha values (0.1 to 0.5)
- Adding weekend indicators and cyclical encoding for day/month

Our MLP used:

- One hidden layer with 80 nodes and the ReLU activation
- Linear output layer (for regression), L2 regularization, learning rate decay

SARIMA Model: We used the SARIMA$(p, d, q)(P, D, Q)(s)$ model where one cycle, s, is equal to 52 weeks. The (p, d, q) parameters focus on trends across the entire time series while the (P, D, Q) parameters capture yearly cycles. R's forecast library identified the following optimal parameters: $(p, d, q) = (2, 1, 0)$ and $(P, D, Q) = (0, 1, 0)$

Results

MLP Daily Sales Results:

- MAE: 260.72
- RMSE: 320.18
- R-squared: 0.9704

SARIMA Weekly Sales Results:

- MAE: 1,703.30
- RMSE: 2,156.71
- R-squared: 0.6631

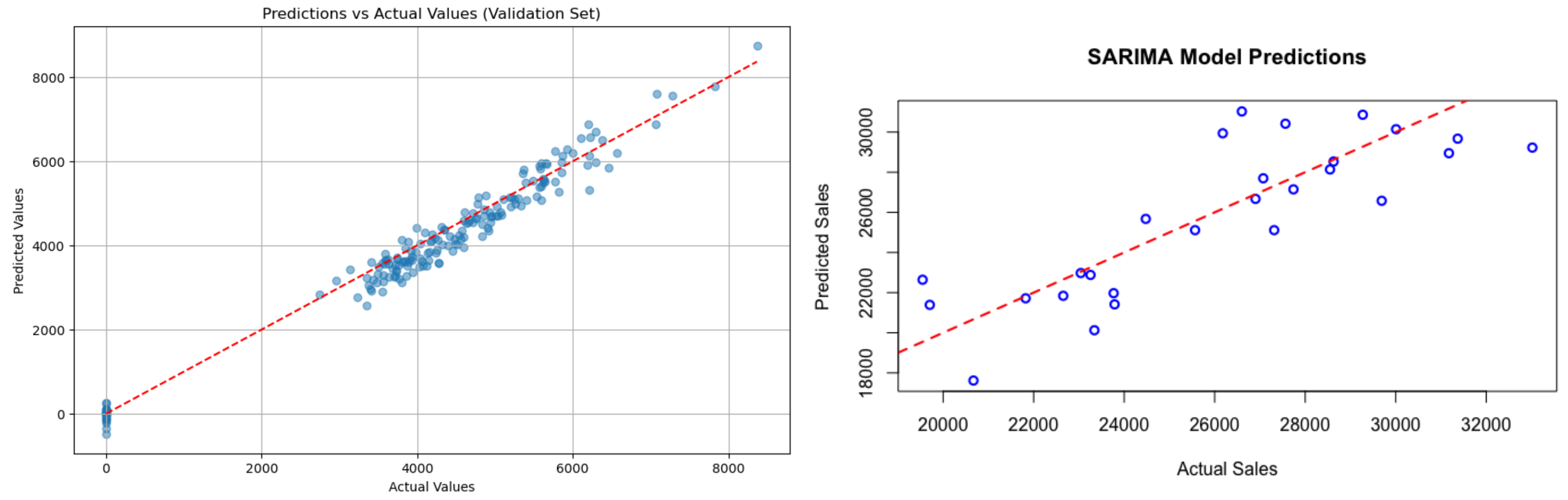


Fig. 2: Predicted weekly sales compared to actual sales for our MLP model and predicted

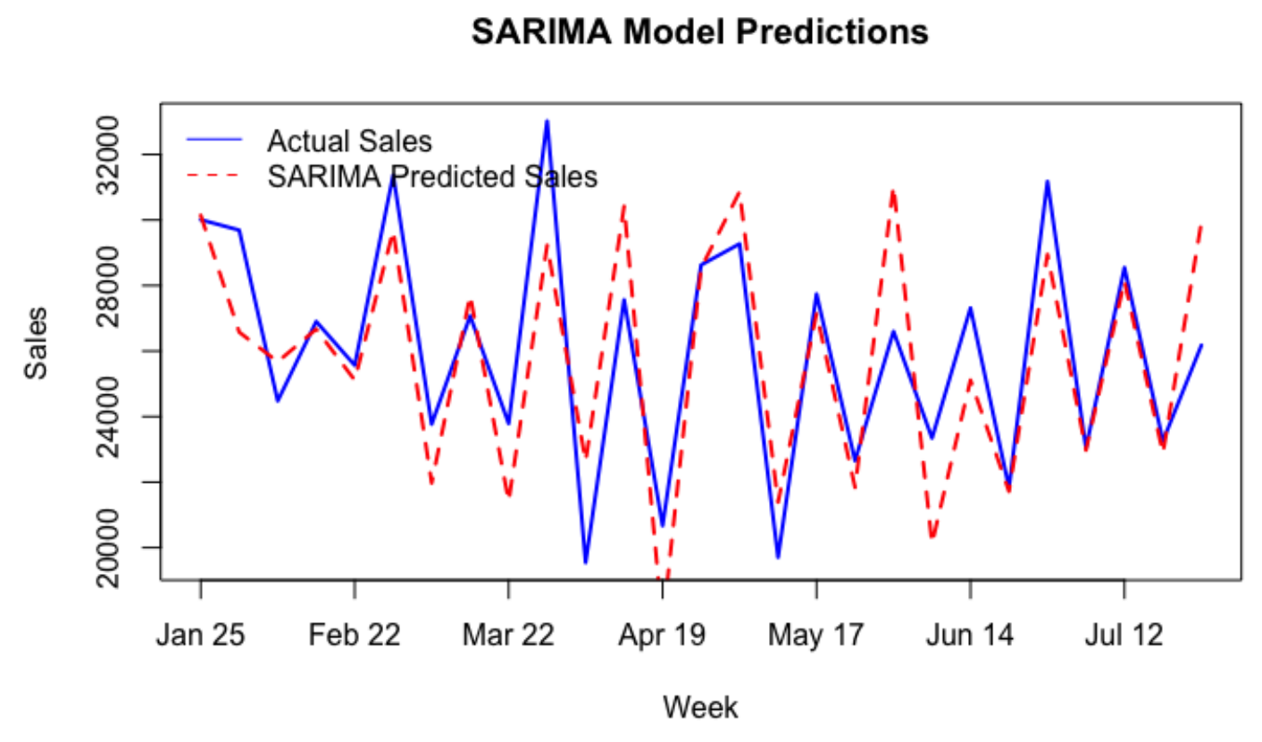


Fig. 3: Our SARIMA model's predictions compared to actual sales for 2015

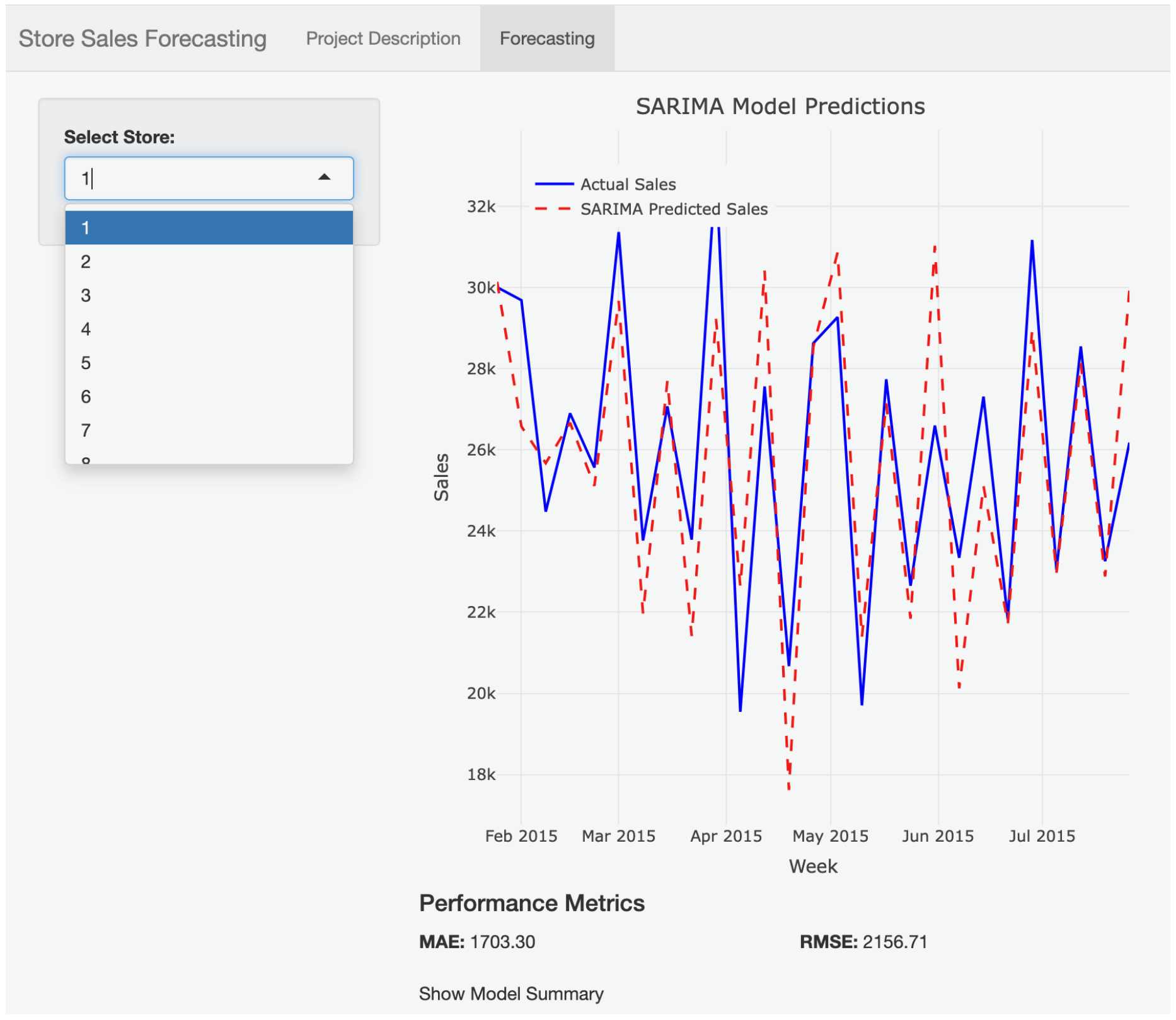
Both models delivered very impressive results. It's important to note that the results of the two models shouldn't be compared because the MLP model is using daily sales data while the SARIMA model is using weekly sales data.

Discussion/Future Work/Web App

For the SARIMA model, we chose to aggregate the daily sales data into weekly sales data. We did this because the SARIMA model isn't able to account for multiple cycles, and sales data tends to have weekly and yearly cycles. For our model, we chose to focus on the yearly cycle by aggregating the data into weekly data. However, our MLP model greatly benefits from having more data points, so we decided to use the daily sales data for that model.

One major takeaway from our models is that seasonality plays a very significant role in predicting sales data. Our SARIMA model relied very heavily on seasonality and was able to predict future sales well. Furthermore, one of the factors our MLP model used was cyclical patterns. We believe this factor significantly contributed to the success of our model. We also believe there are several features that could potentially improve both models such as weather and on-going promotions.

While our SARIMA model performed well on Store 1's data, it's unclear how well it will perform on other stores with different patterns. To address this question, we are building a web app that allows a user to select any Rossmann store and run the model on it. This will help us understand the accuracy of the model across several stores.



Select Store: 1

Performance Metrics

MAE: 1703.30

RMSE: 2156.71

Fig. 4: Screenshot of the app

Future work includes expanding the MLP model to include data from multiple stores and adding more features such as the weather and on-going promotions.

References

Qureshi, N. U. H., Javed, S., Javed, K., Naqvi, S. M. R., Raza, A., Saeed, Z. (2024). Demand forecasting in supply chain management for Rossmann stores using weather enhanced deep learning model. IEEE Access, 12, 145570-145581. https://doi.org/10.1109/ACCESS.2024.3472499