Capstone Project Report

Capstone Project in Business Analytics

Sales Forecasting for a Retail Chain

Submit to-

Dr. Rouzbeh Razavi

Prepared By

Chandima Attanayake (811343415)

13th July 2025

Table of Contents

1. Abstract

The focus of this capstone project is to develop an accurate and practical sales forecasting model for a retail chain using the Rossmann Store Sales dataset from Kaggle. The model is designed to support and give some input to optimize inventory levels, workforce allocation, promotional planning, and supply chain decisions. In this case, it is expected to create a model to predict the store level sales performance using the combination of traditional time series methods and machine learning techniques such as XGBoost, Random. The final objective is to create an actionable forecasting tool which will be supported by data visualization and business insights.

2. Introduction

2.1.Business Context

Rossmann is a leading European drugstore chain operating a large number of outlets. Accurate sales forecasting is crucial for managing inventory, scheduling staff efficiently, and planning promotions, all of which directly influence operational effectiveness and profitability (Hyndman & Athanasopoulos, 2021). Therefore, a proper model is critical to business purpose.

2.2.Project Overview.

The focus of this capstone project is to develop an accurate and practical sales forecasting model for a retail chain using the Rossmann Store Sales dataset from Kaggle. The model is designed to support and give insight to optimize inventory levels, workforce allocation, promotional planning, and supply chain decisions. In this case, it is expected to create a model to predict the store level sales performance using the combination of traditional time series methods and machine learning techniques such as XGBoost, Random.

2.3.Project Objective

The objective of this project is to develop and compare multiple sales forecasting models to identify the most suitable approach for Rossmann stores' operational and strategic requirements.

3. Data Overview

3.1.Datasets Used

This project utilized two primary datasets obtained from the Rossmann Store Sales Kaggle competition (Kaggle, 2015.). These datasets provide detailed information of the transactions of the daily sales and store level characteristics which is required for building accurate forecasting models. The detailed information of two data sets are as follows.

1. train.csv - Contains historical daily sales data for each store and following key variables included in the data set.
   - Store: Unique identifier for each store.
   - Day Of Week: Day of the week as an integer (1 = Monday, 7 = Sunday).
   - Date: Date of the record.
   - Sales: Total sales for the store on the given date (target variable).
   - Customers: Number of customers on the given day.
   - Open: Indicator if the store was open (1) or closed (0).
   - Promo: Indicates if a store was running a promotion that day.
   - State Holiday: Indicates public holidays, represented as 0 (no holiday), a (public holiday), b (Easter holiday), or c (Christmas).
   - School Holiday: Indicates if the store was affected by a school holiday.
2. store.csv - Contains additional information about each store’s characteristics and following are the key variables
   - Store Type: Categorical variable indicating store type (a, b, c, d).
   - Assortment: Describes the assortment level (a = basic, b = extra, c = extended).
   - Competition Distance: Distance in meters to the nearest competitor store.
   - Competition Open Since Month/Year: Month and year when competition began nearby.
   - Promo2: Indicator if the store is running a continuous promotion (Promo2).
   - Promo2 Since Week/Year: Week and year when Promo2 started.
   - Promo Interval: Describes months when Promo2 is active.

3.2.Data Merging

To build an effective model, the train.csv and store.csv datasets were merged to a one dataset and I named as the 'Store' column.

After merging, the dataset contained 1,017,209 records and 24 features, covering daily sales from January 2013 to July 2015 period.

4. Data Preparation

4.1.Missing Value Treatment

During initial data exploration, some missing values were identified primarily in the Competition Distance, Competition Open Since Month/Year, and Promo2 Since Week/Year variables columns and cleaned them on a following basis.

- **Competition Distance**

The missing values were assigned with the median competition, distance across all stores, assuming that the stores with missing distances also faced an average competitive environment.

- **Competition Open Since Month/Year & Promo 2 Since Week/Year**

Missing values of these variables were replaced with zero or encoded as missing categories, under the assumption that no competition or promotions were active before the recorded dates.

4.2.Data Type Conversion

To prepare the dataset for modeling, the date variable was converted to datetime format for efficient time-based feature extraction. Also, categorical variables such as Store Type, Assortment, and State Holiday were encoded as factors for modeling purposes.

4.3.Feature Engineering

Feature engineering was guided by best practices in retail forecasting, where temporal, promotional, and store-specific variables are critical predictors (Hyndman & Athanasopoulos, 2021). To enhance the predictive power of models, several new features were engineered:

1. Year, Month, and Day - Extracted from the Date variable to capture yearly and monthly trends.
2. Day Of Week - Retained as it captures weekly sales patterns.
3. Week Of Year - Created to identify seasonal sales patterns at the weekly level.
4. Is Weekend: Binary variable indicating if the day was a Saturday or Sunday, capturing higher weekend sales trends.
5. Promo Running: Indicator combining Promo and Promo2 to show if any promotions were active on a given day.

5. Exploratory Data Analysis (EDA)

Exploratory Data Analysis was conducted to understand the sales distribution, identify patterns and trends, and inform feature engineering and model selection decisions. Both univariate and bivariate analyses were performed on key variables. The observed right-skewness is typical in retail sales data due to the occurrence of promotional peaks and outlier events (Taylor & Letham, 2018).

5.1.Sales Distribution

A histogram of daily sales revealed that the sales are right-skewed, indicating most daily sales values are concentrated at lower amounts with a long tail of higher sales days.

Also, most sales were under 10,000 currency units, aligning with typical daily turnover in medium-sized retail stores.

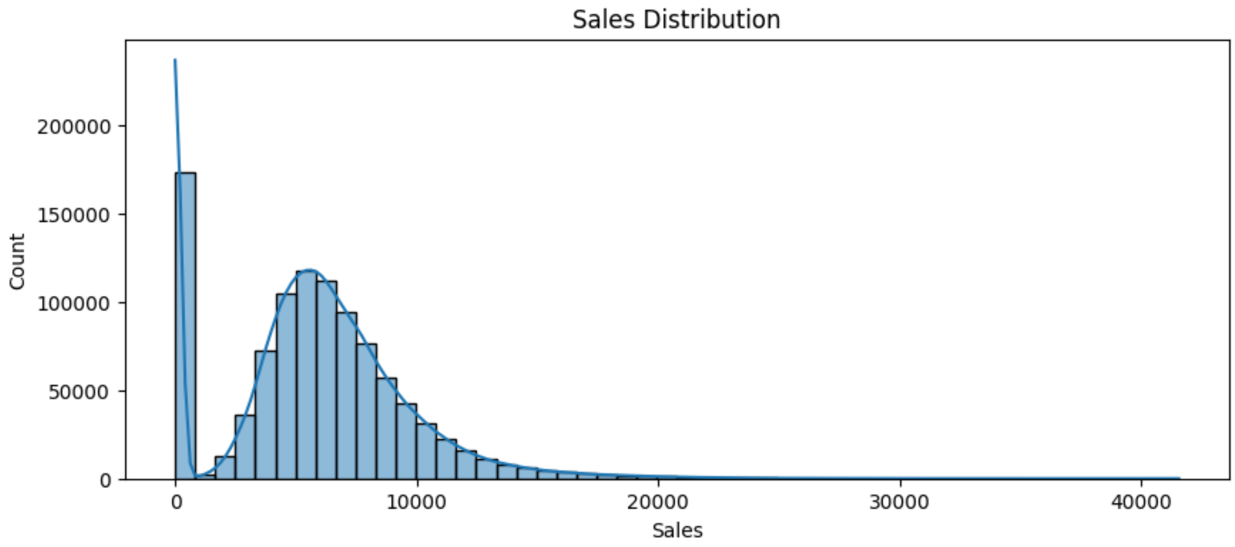

Figure 1 Histogram showing daily sales distribution

5.2.Sales Trends Over Time

A line plot of total sales over the entire period revealed a distinct seasonal pattern, with periodic peaks during promotional campaigns and festive months. Additionally, weekly seasonality was evident, with higher sales generally occurring towards the end of the week, particularly on Fridays and Saturdays.

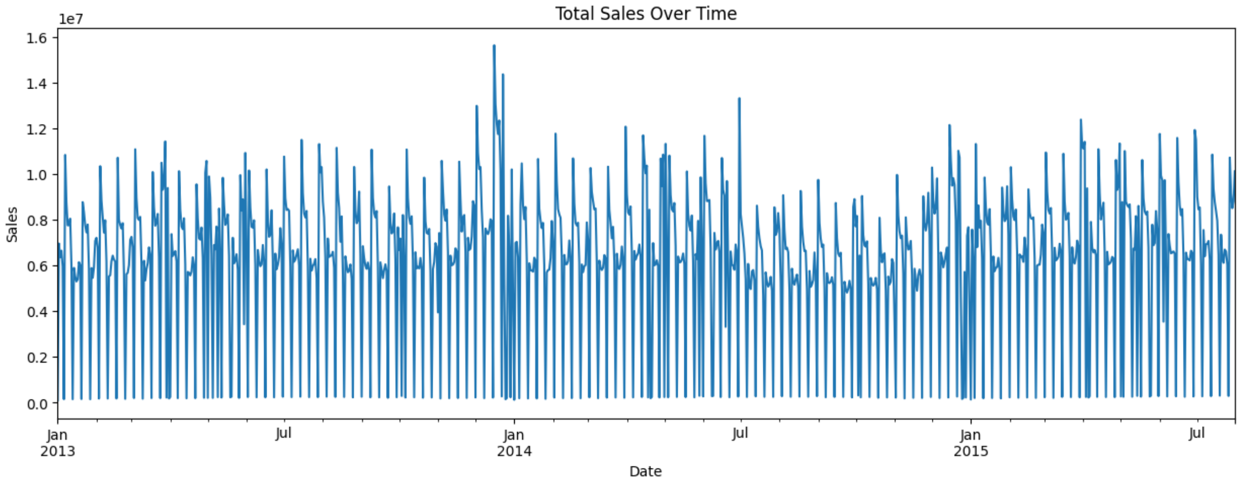

Figure 2 Sales trends over time

5.3. Key Insights from EDA

- Sales are influenced by promotions, day of the week, and store characteristics such as Store Type and Assortment.
- The right-skewed sales distribution indicates models must account for occasional extreme sales days.
- There is a strong relationship between customers and sales, which may be leveraged in regression-based models.
- Seasonal and promotional effects are significant, justifying the inclusion of time-based and promotion-related features in modeling.

6. Model Development

Multiple modeling techniques were applied to forecast Rossmann store sales, including linear regression (baseline), tree-based models, time series models, and deep learning approaches. Models were built using features engineered during data preparation and evaluated primarily using RMSE (Root Mean Squared Error) and MAE (Mean Absolute Error).

6.1.Baseline Model – Linear Regression

Used as a benchmark to evaluate the performance improvement of more complex models and the Store, Promo, Day Of Week, Month, Year, Customers features were used for this purpose.

According to the outcome of the model evaluation, the RMSE - 2818.89 and MAE - 2005.88.

In this case, the high error rates indicated linear regression's inability to capture non-linear relationships inherent in sales data.

6.2.Random Forest

The random forest ensemble tree-based model capturing non-linear interactions and variable importance. For this, Same as linear regression plus additional engineered features such as Week of Year, Is Weekend, Promo Running were used. In this case, the Promo and Day of Week were used mainly for predicting purposes.

The outcome of the model evaluation is RMSE - 1242.76 and MAE - 854.99. This is the best performing classical ML model, effectively capturing non-linear interactions.

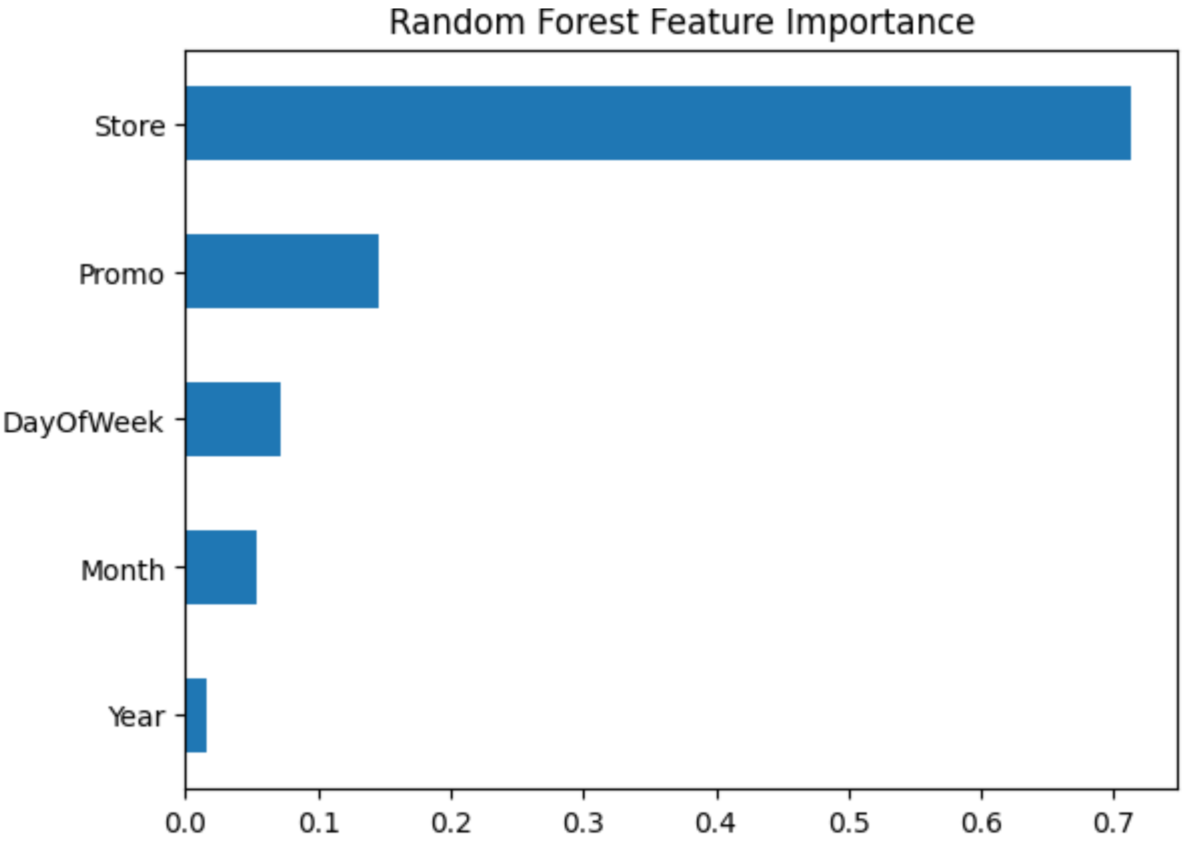

Figure 3 Random Forest feature importance plot

6.3.XGBoost

Gradient boosting model optimized for performance and speed and underperformed relative to Random Forest, potentially due to limited hyperparameter tuning. Provided SHAP-based interpretability

The outcome is RMSE - 2524.01 and MAE: 1862.88

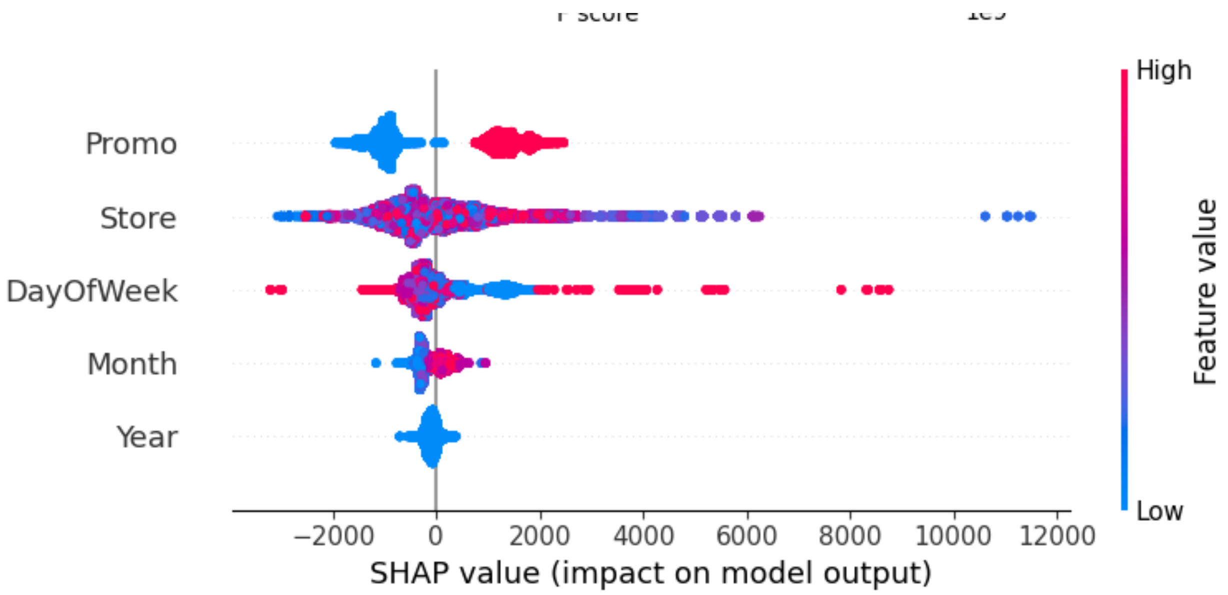

Figure 4 XGBoost SHAP summary plot

6.4.Prophet – Time Series Forecasting

This is an additive model by Facebook for capturing seasonality, trend, and holidays in time series data. This was applied to Store 1 for 60-day forecasts. Also, this was successfully captured weekly and yearly seasonality with clear trend projections. Ideal for store-level management decisions.

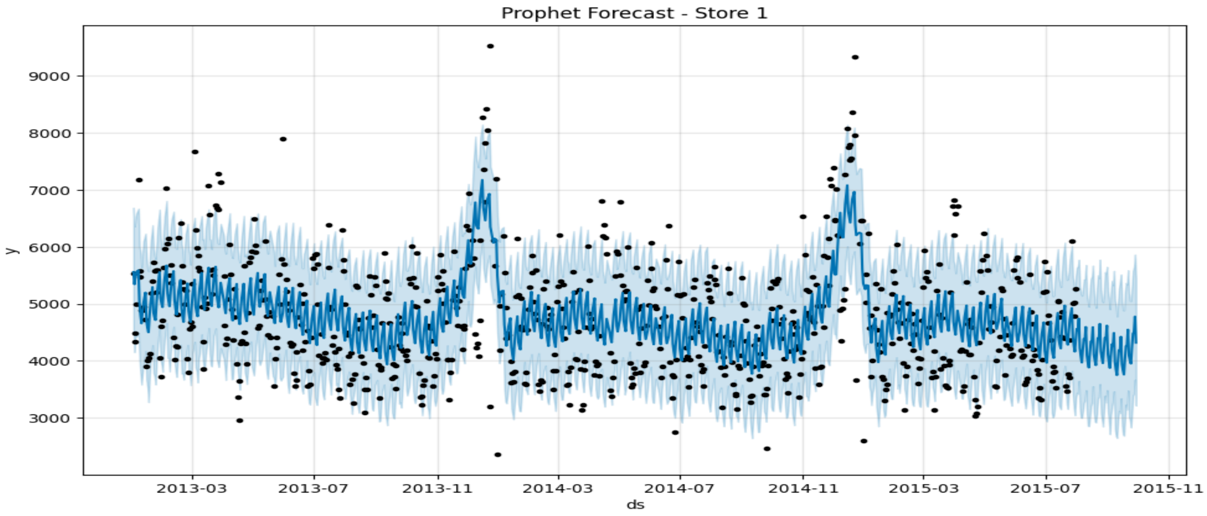

Figure 5 - Prophet forecast plot

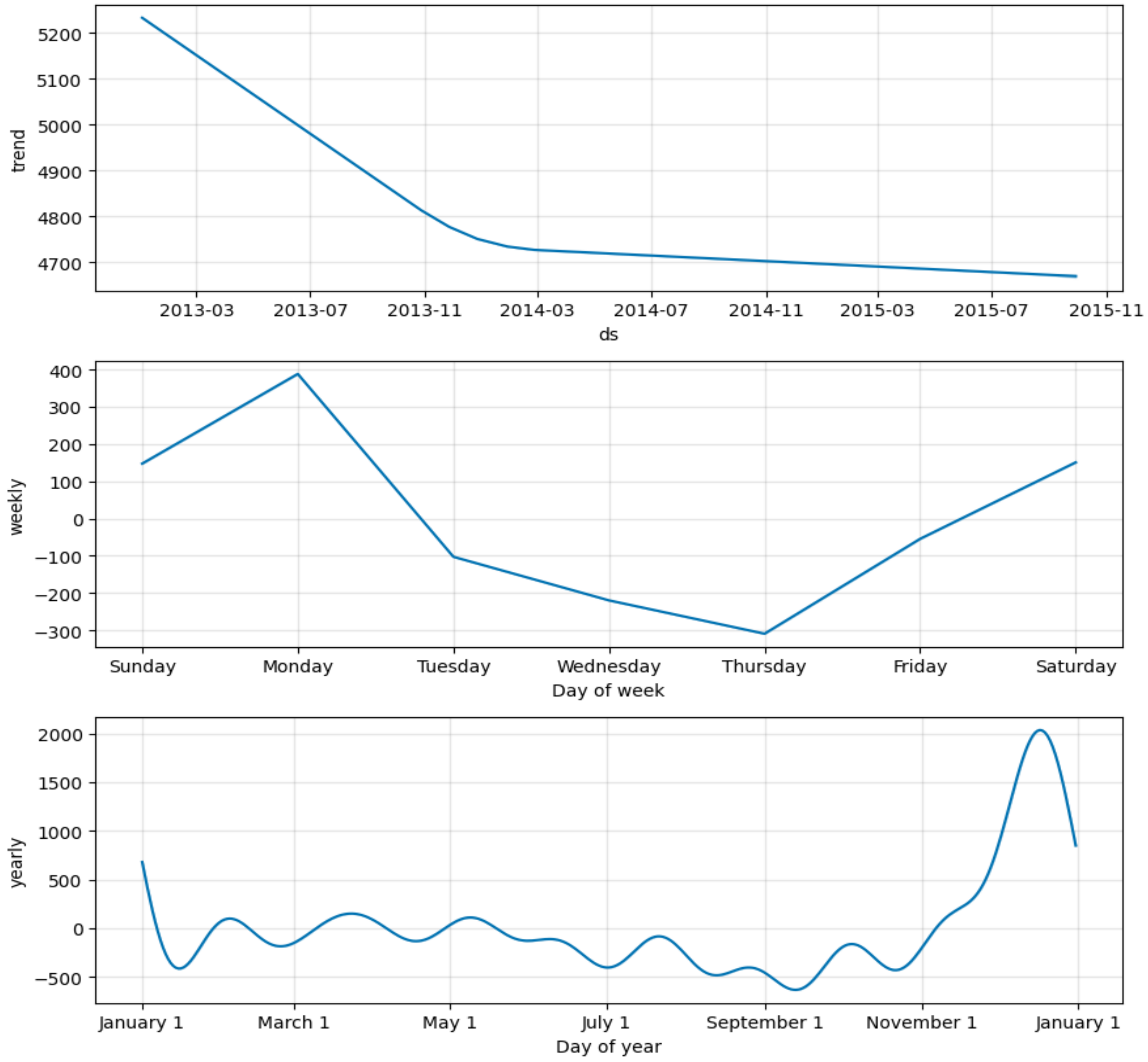

Figure 6 - Prophet components plot

6.5.SARIMA – Time Series Forecasting

Seasonal ARIMA model incorporating autoregressive and moving average components with seasonality and this is ideal for producing reliable short-term forecasts with narrow confidence intervals. The model order prepared for (1,1,1)x(1,1,1,7). The model was applied for forecasted Store 1 sales with weekly seasonal adjustments.

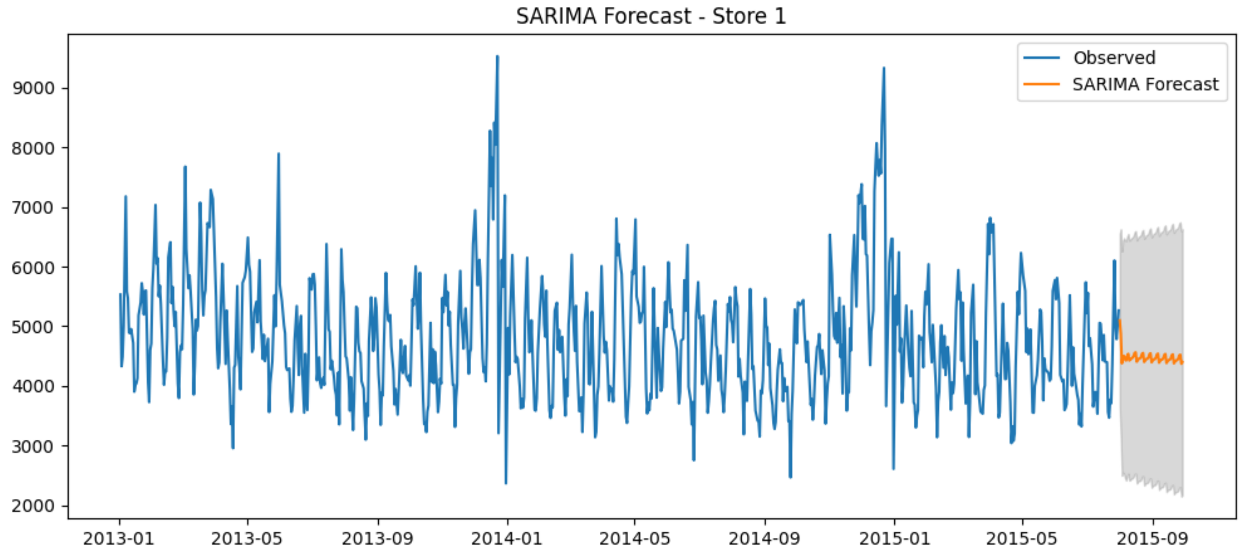

Figure 7 - SARIMA forecast plot with confidence intervals

6.6.LSTM – Deep Learning

Long Short-Term Memory (LSTM) neural network mainly using for sequence data modeling and Used 30-day sequences to predict next day sales. Accordingly, the RMSE showed a 709.76 and MAE is 592.24.

Accordingly, LSTM achieved lowest error rates among all models, demonstrating capability to capture complex temporal dependencies. However, it was computationally intensive and requires GPU resources for scalability.

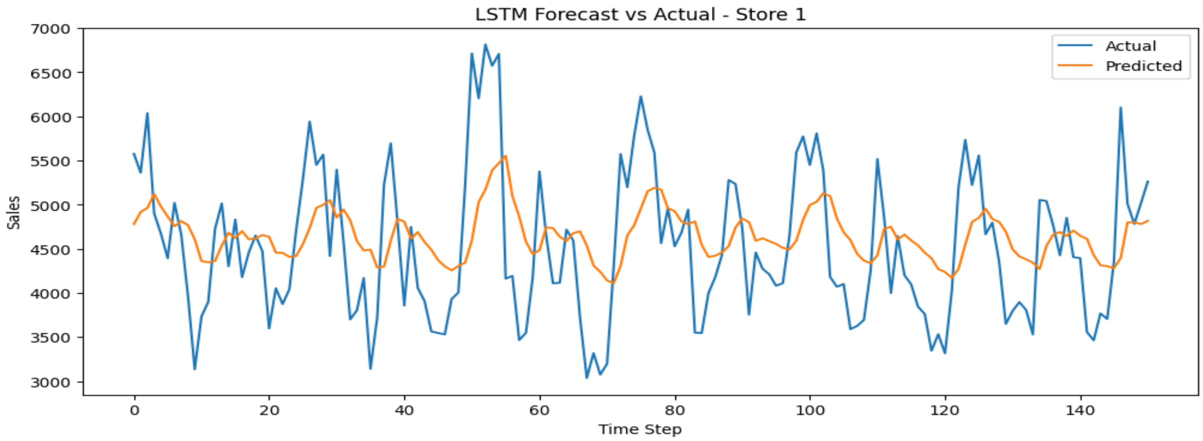

Figure 8 - LSTM actual vs predicted plot

6.7.Model Comparison

Model	RMSE	MAE	Interpretation
Linear Regression	2818.89	2005.88	Baseline. But it is poor for non-linear patterns.
Random Forest	1242.76	854.99	Strong performance and practical for deployment.
XGBoost	2524.01	1862.88	Underperformed due to the limited tuning.
Prophet	N/A	N/A	Effective store-level forecasts with seasonality.
SARIMA	N/A	N/A	Reliable short-term store-level forecasts.
LSTM	709.76	592.24	Best performance for strategic decisions and captures complex time dependencies.

The Random Forest is the best practical machine learning model, balancing performance and interpretability. Also, LSTM provides the highest accuracy but requires substantial computational resources for production deployment. Addition to that, the Time series models (Prophet, SARIMA) are ideal for store-specific forecasts, offering clear seasonality decomposition.

7. Challenges faced and solutions to overcome

Challenge	Solution Implemented
Type Error when using mean_squared_error(..., squared=False) in Colab	Replaced squared = False with manual computation using np.sqrt(mean_squared_error(...)) to ensure compatibility with older scikit-learn versions.
Memory issues when loading and processing the full dataset	Filtered data to only include open stores and non-zero sales, reducing dataset size and improving performance for model training.

Missing values in features like Competition Distance and Promo2	Imputed missing numerical values with median; kept categorical placeholders as '0' or 'Unknown' depending on feature semantics.
Categorical variables like Store Type and Assortment not usable in model	Planning to use one-hot encoding (pd.get_dummies) or label encoding (LabelEncoder) for these variables in the next phase of modeling.
Temporal leakage risk due to random data splits	Used shuffle=False in train_test_split() and will implement **TimeSeriesSplit** or date-based slicing in later stages for time-aware validation.
High variability in sales patterns across stores	Introduced store-specific features and categorized stores by type and competition level to model inter-store differences effectively.
Handling holidays and promotional periods which significantly affect sales	Created binary flags such as PromoRunning and IsWeekend, and time-based features (e.g., Month, DayOfWeek). Interaction terms will be added in the next phase
Time-dependency and sequence modeling with large datasets	LSTM modeling is planned for the next phase. Sequence preparation and feature scaling are yet to be implemented.
Training deep learning models like LSTM required high processing power and longer runtimes.	Reduced batch sizes and epochs; prioritized Random Forest for feasibility.
Extensive tuning was not feasible within project time constraints.	Used default parameters; identified further tuning as next step.
Missing values and scaled outputs increased complexity.	Applied median imputation; ensured inverse scaling for LSTM interpretation.
Aggregated models missed unique store patterns.	Developed store-specific Prophet and SARIMA models; proposed hybrid approaches.

Moving from notebooks to production pipelines requires development.	Recommended BI integration, monitoring, and retraining frameworks.

8. Ethical Considerations

The data set used in this project is publicly available and anonymized (Rossmann Store Sales from Kaggle) and the following ethical principles were considered and followed in this project.

- Avoiding Overfitting – It is carefully reviewed to validate models properly and avoid overfitting, which could lead to misleading or non-generalizable forecasts.
- Transparency & Reproducibility - All modeling steps, assumptions, and preprocessing are mentioned in code and markdown for clarity and future reference. Also, tree-based models with SHAP improved explainability.
- Fairness in Modeling -Attention is given to avoid unintentional bias in models that could systematically favor large stores over small ones.
- Responsible Use - The predictive models are developed in an academic context. However, when it using for the real-world business environment, it is required to take some additional steps to validate the model further in order to provide more accurate outcome.
- Data Privacy and Confidentiality - Used publicly available anonymized data. In real applications, it is required to compliance with data privacy regulations such as GDPR.

In summary, this project did not involve direct ethical risks due to its use of public anonymized data. When this is using for real-world data for similar forecasting purposes, it is required give careful attention on the data privacy, bias, transparency, and responsible use for ethical and sustainable implementation (Cowls (2019) and Jobin et al. (2019).

9. Recommendations

Based on model performance and business applicability, the following recommendations are proposed:

1. Deploy Random Forest models for immediate operational forecasting due to their high accuracy and scalability.
2. Invest in GPU infrastructure to implement LSTM models for strategic, high-accuracy forecasting at scale.
3. Develop hybrid models that combine LSTM outputs with time series residual forecasts to leverage strengths of both approaches.
4. Establish a model monitoring framework to track performance metrics and retrain models regularly to maintain accuracy over time.
5. Integrate forecasts into BI dashboards such as Power BI or Tableau to make forecasts accessible and actionable for store managers and planners.
6. Train business users to interpret forecast outputs effectively, supporting data-driven decision-making across operations.
7. Conduct extensive hyperparameter tuning for XGBoost and LSTM models in future work to achieve optimal configurations.

These recommendations align with industry practices where ensemble machine learning models are deployed for operational forecasts, while deep learning approaches are explored for strategic planning (Taylor & Letham, 2018).

Reference

Kaggle. (2015). *Rossmann Store Sales*. Kaggle. https://www.kaggle.com/competitions/rossmann-store-sales

Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and Practice* (3rd ed.). OTexts. https://otexts.com/fpp3/

Taylor, S. J., & Letham, B. (2018). Forecasting at scale. *The American Statistician, 72*(1), 37–45. https://doi.org/10.1080/00031305.2017.1380080

Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review, 1*(1). https://doi.org/10.1162/99608f92.8cd550d1

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825-2830.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794. https://doi.org/10.1145/2939672.2939785

Abadi, M., et al. (2016). TensorFlow: A system for large-scale machine learning. *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, 265–283.