Sales Forecasting & Time Series Analysis

Rossmann Store Sales — Kaggle Dataset

SARIMAX • Random Forest • XGBoost

Report generated: February 24, 2026

Project Type: Analytics + Machine Learning

1,115	~1M	94	644.67
Stores	Training Rows	Features	Best MAE (XGBoost)

1. Project Overview

This project builds a complete end-to-end sales forecasting pipeline on the Rossmann Store Sales dataset from Kaggle. The goal is to predict daily sales for 1,115 stores across Germany using historical sales data combined with store metadata, promotional campaigns, and holiday indicators.

The pipeline consists of five stages:

- **EDA** — Exploratory analysis of trends, seasonality, promotions and store characteristics.
- **Feature Engineering** — 94 features: lags, rolling statistics, Fourier terms, cyclical encodings.
- **SARIMAX Modeling** — Statistical model with exogenous features fitted on Store 1.
- **ML Modeling** — Random Forest and XGBoost trained on all 1,115 stores simultaneously.
- **Model Comparison** — Side-by-side evaluation using MAE, RMSE, and MAPE.

Dataset Summary

File	Rows	Columns	Description
train.csv	1,017,209	9	Historical daily sales per store
store.csv	1,115	10	Store metadata (type, assortment, competition)
test.csv	41,088	8	Stores & dates to forecast

2. Exploratory Data Analysis — Key Insights

2.1 Target Variable (Sales)

- Distribution is heavily **right-skewed** — log1p transform applied for modeling.
- Strong positive correlation with Customers (Pearson r ≈ 0.82).

2.2 Temporal Patterns

- **Monday** records the highest average sales; **Saturday** the lowest.
- **December** is the peak month — Christmas surge clearly visible.
- Strong weekly and annual seasonality confirmed via rolling mean and Fourier analysis.

2.3 Promotions

- Promo active on ~40% of training days.
- Mann-Whitney U test: Promo **significantly increases sales (p < 0.001)**.
- Confirmed as the single most important categorical predictor across all models.

2.4 Store Type & Assortment

- Store **Type b** achieves the highest average daily sales.
- Assortment **Type b (Extra)** outperforms Extended and Basic.

2.5 Competition

- CompetitionDistance shows a very weak linear correlation with Sales.
- Effect is non-linear — log1p(CompetitionDistance) is more informative.

---

3. Feature Engineering

A total of **94 features** were engineered from raw data. All lag and rolling features were computed per-store with a one-day shift to prevent data leakage.

Category	# Features	Key Features
Lag Features	14	Sales & log(Sales) at lags 1,2,3,7,14,21,28 days
Rolling Statistics	14	7/14/28-day mean, std, max, min + EWM span=7,28
Fourier Terms	12	3 sin/cos pairs for weekly (7) and annual (365) periods
Date / Calendar	15	Year, Month, Day, WeekOfYear, IsMonday, IsDecember…
Cyclical Encoding	4	Month_sin/cos, WeekOfYear_sin/cos
Promo Features	7	Promo, IsPromo2Active, WeeksSincePromo2, interactions
Holiday Features	8	StateHoliday_enc, IsChristmas, BeforeHoliday, AfterHoliday
Competition	4	log1p(CompDist), CompOpen_Months, HasCompetitor
Categoricals	12	StoreType one-hot, Assortment ordinal, DayOfWeek one-hot
Store Aggregates	6	Per-store mean, median, std, max, min sales, promo rate
Missing Flags	1	CompDist_WasNull

Train / Validation Split

A strict **time-based 80/20 split** — never random — to replicate real-world forecasting.

- Train : 675,958 rows (up to ~November 2014)
- Validation : 168,380 rows (~November 2014 onward)

---

4. SARIMAX Model (Improved Statistical Baseline)

4.1 Why SARIMAX over plain ARIMA

Plain ARIMA is purely univariate — it only sees past Sales values and completely ignores external signals like Promo or holidays. SARIMAX (Seasonal ARIMA with eXogenous regressors) solves this by allowing external features to be passed directly into the model as exogenous regressors. This produced a significant improvement over the original ARIMA baseline.

	ARIMA(2,1,2)	SARIMAX(2,1,2)+exog
Exogenous features	None	16 features
Captures Promo effect	No	Yes
Captures holidays	No	Yes

Seasonality	Via AR/MA terms	Via Fourier sin/cos terms
Expected MAE	~695	489.65 (actual)

4.2 Exogenous Features Used (16 total)

- **Promo signals:** Promo, Promo_x_Monday, Promo_x_SchoolHol
- **Holiday signals:** SchoolHoliday, IsPublicHoliday, IsChristmas, BeforeHoliday, AfterHoliday
- **Calendar flags:** IsMonday, IsSaturday, IsMonthEnd, IsQ4
- **Fourier seasonality:** fourier_weekly_sin/cos_7_1, fourier_annual_sin/cos_365_1

4.3 Methodology

- **Stationarity:** ADF test applied — first differencing used (d=1).
- **Order:** (p=2, d=1, q=2) confirmed via ACF / PACF plots.
- **Fit:** SARIMAX trained on Store 1 train split with train_exog.
- **Forecast:** model.forecast(steps=n, exog=val_exog) — val exog required.
- **Scope:** Univariate per store — Store 1 used as representative example.

4.4 SARIMAX Limitations

- Still per-store — not scalable to all 1,115 stores without automation.
- ML models still outperform SARIMAX because they use all 94 features including lags.

---

5. ML Models — Random Forest & XGBoost

Both ML models were trained on all 1,115 stores simultaneously using the full 94-feature set. Target was **log1p(Sales)**; predictions converted back with expm1 for evaluation in original sales scale.

5.1 Random Forest

- n_estimators=200, max_depth=10, min_samples_leaf=10, max_features=sqrt.
- Parallel training with n_jobs=-1.
- Captures non-linear patterns and feature interactions.

5.2 XGBoost

- n_estimators=500, learning_rate=0.05, max_depth=6.
- Regularisation: reg_alpha=0.1, reg_lambda=1.0 to prevent overfitting.
- Subsampling: subsample=0.8, colsample_bytree=0.8.
- Validation set used for early stopping monitoring.

5.3 Results

Model	Scope	MAE	RMSE	MAPE%
Random Forest	All 1,115 stores	736.83	1,098.15	—
XGBoost	**All 1,115 stores**	**644.67**	**933.08**	**—**

5.4 Feature Importance (XGBoost) — Top Predictors

- **Lag features** (Sales_lag_1, Sales_lag_7) — strongest signal, captures autocorrelation.
- **Rolling statistics** (roll_mean_7, roll_mean_28, ewm_7) — local trend context.
- **Promo** — statistically significant sales driver (confirmed in EDA).
- **DayOfWeek** dummies — weekly seasonality pattern.
- **Store aggregates** (Store_SalesMean) — per-store baseline.
- **Fourier terms** — annual seasonality encoding.

---

6. Final Model Comparison

SARIMAX and ML models serve different purposes. SARIMAX provides an interpretable statistical baseline for a single store with explicit exogenous regressors, while XGBoost provides the best predictive accuracy across all 1,115 stores using the full feature set.

Model	Type	Scope	MAE	RMSE	Features	Winner
SARIMAX(2,1,2)+exog	Statistical	Store 1	489.65	653.29	16 exog	
Random Forest	ML Ensemble	All 1,115	736.83	1,098.15	94	
XGBoost	**ML Boosting**	**All 1,115**	**644.67**	**933.08**	**94**	**BEST**

Why XGBoost Wins

- Uses all 94 features — Promo, holidays, lags — that SARIMAX cannot scale to use globally.
- Single global model across all 1,115 stores learns cross-store patterns.
- Gradient boosting handles non-linear competition and assortment interactions.
- Regularisation prevents overfitting on the large training set.

SARIMAX vs plain ARIMA

- SARIMAX dramatically outperforms plain ARIMA by incorporating 16 exogenous regressors.
- Promo alone explains a large fraction of the variance that plain ARIMA cannot capture.
- Fourier terms in SARIMAX explicitly model weekly and annual seasonality.

Error Pattern Analysis

- **December** has the highest MAE across all models — Christmas spikes are hard to predict.
- **Monday** shows elevated error due to large swing from Sunday closures.
- XGBoost outperforms Random Forest on the majority of the 1,115 stores.

---

7. Conclusions & Future Work

Conclusions

- Complete pipeline from raw data → 94 features → three trained models successfully built.

- XGBoost achieves MAE 644.67 / RMSE 933.08 — significantly better than Random Forest and naive baseline.
- SARIMAX+exog achieves MAE 489.65 / RMSE 653.29 on Store 1 — strong result for a statistical model.
- Lag and rolling features are the strongest predictors, confirming strong autocorrelation.
- Promo is the most important categorical predictor — consistent across EDA, SARIMAX, and ML.
- SARIMAX with exogenous features substantially improves over plain univariate ARIMA.

Future Work

- **LightGBM / CatBoost** — likely to match or exceed XGBoost with less tuning.
- **Multi-store SARIMAX** — automate SARIMAX fitting across all 1,115 stores in parallel.
- **LSTM / Transformer** — deep learning sequence models for further improvement.
- **Hyperparameter tuning** — Optuna for automated XGBoost / RF optimisation.
- **Store clustering** — group similar stores and train one model per cluster.
- **Production deployment** — Streamlit app serving XGBoost forecasts interactively.

---

8. Project Artifacts

Artifact	Location	Description
01_eda.ipynb	notebooks/	Exploratory data analysis
02_feature_engineering.ipynb	notebooks/	94-feature engineering pipeline
03_arima_model.ipynb	notebooks/	SARIMAX with exogenous features — Store 1
04_ml_models.ipynb	notebooks/	Random Forest & XGBoost training
05_model_comparison.ipynb	notebooks/	Final comparison & error analysis
data_preprocessing.py	src/	Data cleaning & merging pipeline
feature_engineering.py	src/	Feature engineering (94 features)
arima_model.py	src/	SARIMAX pipeline with exogenous features
arima_model.pkl	models/	Saved SARIMAXResultsWrapper (Store 1)
random_forest.pkl	models/	Trained Random Forest regressor
xgboost.pkl	**models/**	**Trained XGBoost regressor (BEST MODEL)**
featured_sales_data.csv	data/processed/	Full 94-feature dataset
train_featured.csv	data/processed/	Training split (675,958 rows)
val_featured.csv	data/processed/	Validation split (168,380 rows)
feature_list.csv	data/processed/	List of 94 feature names
final_comparison.csv	reports/	All model MAE/RMSE/MAPE results
final_report.pdf	reports/	This document
