Sales Forecasting for Store Sales - Time Series Forecasting

Project Description:

The project aims to develop a predictive analytics model to forecast store sales for a retail chain using time series data. The primary goal is to accurately predict future sales trends to support data-driven decisions in inventory management, staffing, and marketing.

The dataset will combine multiple data sources, including:

- **train.csv** – daily sales for each store and product family.
- **stores.csv** – metadata such as store location, type, and cluster.
- **oil.csv** – daily oil prices that may influence economic activity.
- **holidays_events.csv** – records of national and regional holidays and special events.
- **transactions.csv** – the number of daily transactions per store.

After merging these datasets, the final data is expected to contain approximately **3,000,888 rows and 13 columns**.

The project will begin with extensive **data preprocessing and cleaning** to ensure data integrity and consistency. This includes handling missing values, removing irrelevant or redundant features, and encoding categorical variables for modeling. Following this, **Exploratory Data Analysis (EDA)** will be conducted to uncover key insights such as sales patterns, seasonal trends, store performance differences, and the influence of promotions and holidays on consumer behavior.

To enhance data quality and improve model accuracy, transformations such as **logarithmic scaling** and **outlier removal** will be applied to stabilize variance and reduce noise. The dataset will then be resampled at **monthly intervals** to capture broader sales trends and long-term seasonality.

Various forecasting models (e.g., SARIMAX and other time series approaches) will be trained and evaluated using **Mean Absolute Error (MAE)** and **Root Mean Squared Error (RMSE)** as performance metrics. Visualization will play a key role in assessing model performance through plots of actual vs. predicted values, forecasted sales with confidence intervals, and trend analysis.

Ultimately, this project will demonstrate a complete end-to-end workflow for building and evaluating a sales forecasting model — from data integration and exploration to model development and visualization — to provide actionable insights that support strategic business planning and operational efficiency.

Team Members:

- - Amr Ashraf (**Team Leader**)
- - Radwa Amr
- - Lourina Email
- - Mostafa Mohamed
- - Fares Mahmoud
- - Mostafa Hesham

Roles:

Member	Milestone 1	Milestone 2	Milestone 3	Milestone 4	Milestone 5
Amr Ashraf	1. Data Collection 2. Preprocessing	1. Time Series Visualizations 2. Feature Engineering	1. Model Selection 2. Model Training 3. Model Evaluation	1. Deploy Model Interface 2. Enable Real-Time Predictions 3. Version control with DVC	1. Presentation
Mostafa Mohamed	1. EDA 2. Preprocessing	1. Visualize Trends, Seasonality	1. Model Selection 2. Model Training 3. Model Evaluation	1. Version control with DVC 2. Monitor model performance 3. Create feedback loop	1. Presentation
Radwa Amr	1. Statistics, outliers, correlations 2. Preprocessing	1. Time Series Visualizations	1. Model Selection 2. Model Training 3. Model Evaluation	1. Track experiments with MLflow 2. Monitor model performance 3. Create feedback loop	1. Report
Lourina Emil	1. Encoding	1. interactive dashboards	1. Model Selection 2. Model Training 3. Model Evaluation	1. Deploy Model Interface 2. Enable Real-Time Predictions 3. Track experiments with MLflow	1. Presentation
Fares Mahmoud	1. Data Exploration Report	1. Analysis Report	1. Model Selection 2. Model Training 3. Model Evaluation 4. Forecasting Model Performance Report	1. Version control with DVC 2. MLOps Report	1. Report
Mostafa Hesham	1. Data Exploration Report	1. Interactive visualization	1. Model Selection 2. Model Training 3. Model Evaluation	1. Performance Reporting	1. Report

Team Leader: Amr Ashraf

Project Objectives:

The primary objective of this project is to develop a robust sales forecasting and optimization model that accurately predicts future retail sales based on historical time-series data. Leveraging advanced data science techniques, including data preprocessing, feature engineering, statistical modeling, and model optimization, the project aims to identify key sales patterns, capture seasonality, and quantify the impact of promotions and holidays on store performance.

The forecasting model will be trained and evaluated using metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) to ensure accuracy and reliability. Ultimately, this project seeks to deliver a data-driven forecasting solution that enhances decision-making in inventory management, marketing planning, and business strategy, while demonstrating a complete end-to-end workflow aligned with the IBM Data Science Track framework.

Stakeholder Analysis:

The success of the Sales Forecasting and Optimization project depends on understanding the interests and influence of key stakeholders involved in or affected by the project. The following table summarizes the main stakeholders, their roles, level of influence, and communication strategies.

Stakeholder	Role / Description	Power	Interest	Communication Strategy
Project Team (Data Scientists, Developers, Designers)	Responsible for data processing, model development, and deployment	High	High	Daily coordination via GitHub/Slack and weekly meetings
Team Leader (Amr Ashraf)	Oversees project progress, task distribution, and communication with supervisor	High	High	Weekly progress reports and milestone reviews
Academic Supervisor	Provides technical and academic guidance, reviews deliverables	High	High	Weekly review meetings and feedback reports
University / Evaluation Committee	Evaluates final submission and presentation	High	Medium	Final report submission and presentation demo
Store Managers /	End users of sales forecasts for decision-	Medium	High	Dashboard demo and

Stakeholder	Role / Description	Power	Interest	Communication Strategy
Retail Owners	making			monthly reports
IT / DevOps (if applicable)	Supports model deployment and monitoring	Medium	Medium	Collaboration during deployment and maintenance
Data Providers (Kaggle dataset)	Source of historical data	Low	Medium	Data documentation and update tracking

Summary:
High-power stakeholders such as the project leader, supervisor, and evaluation committee require regular updates and milestone reports. High-interest stakeholders such as store managers should be engaged through interactive dashboards and demos to ensure usability and satisfaction.

Tools & Technologies:

- - Python (Pandas, NumPy, Scikit-learn, Statsmodels, Prophet, XGBoost/LightGBM, TensorFlow/PyTorch if needed)
- - Jupyter Notebook / VS Code
- - Plotly / Matplotlib / Seaborn for visualizations
- - Dash or Power BI for dashboarding
- - Git & GitHub for version control

Milestones & Deadlines:

Milestone	Description	Status	Deadline
Milestone 1	Data Collection, Exploration, and Preprocessing	Completed	20/09/2025
Milestone 2	Data Analysis and Visualization	Completed	01/10/2025
Milestone 3	Forecasting Model Development and Optimization	In Progress	17/10/2025
Milestone 4	MLOps, Deployment, and Monitoring	Planned	01/11/2025
Milestone 5	Final Documentation and Presentation	Planned	15/11/2025

Key Performance Indicators (KPIs)

Data Quality & Feature Engineering:

- Percentage of missing values handled: 100%
- Data accuracy after preprocessing: 98%
- Dataset diversity: Balanced representation across stores/items.
- Identification of ≥ 3 major seasonal patterns per store/product.

Forecasting Model Performance:

- Model accuracy (e.g., RMSE/ MAE target): RMSE< 0.35 (target) / MAE as baseline
- At least 4 models tested (SARIMAX, Prophet, XGBoost, Arima).
- Overfitting gap ≤ 5% between training and validation sets.
- Model prediction speed (Latency): < 100 ms per request (batch predictions acceptable)
- Error rate (e.g., bias / large outliers): Minimize large error events; track per-store error

Model Monitoring & Deployment:

- Deploy forecasting API or dashboard with **≥ 99% uptime**.
- Achieve forecast response time ≤ 2 seconds per query.
- **Implement daily performance monitoring (track RMSE and MAE drift over time).**
- Trigger **automated retraining** if error metrics degrade by ≥ 2% from baseline.

Business Value & Impact:

- - Forecast accuracy improvement ≥ 20% compared to baseline.
- - Inventory imbalance reduction ≥ 10%.
- - Faster and data-driven restocking and promotion planning.
- - Stakeholder satisfaction ≥ 8/10 during final demo.

Reporting & Communication:

- - Full documentation and reproducible Jupyter workflow.
- - Interactive dashboards displaying actual vs predicted sales.
- - 100% milestone coverage (data, modeling, deployment, evaluation).

Stakeholder Engagement:

- - Achieve stakeholder satisfaction score ≥ 8/10 during the final demonstration and feedback survey.

Future Improvements

Future work will explore hybrid and deep learning models such as Temporal Fusion Transformers and enhanced feature selection methods to further improve forecasting precision and adaptability. Integration with live data streams and automated retraining pipelines will ensure the model remains accurate and scalable in production environments.

Contact & Notes:

Team Leader: Amr Ashraf
Email: amrashraf.official@gmail.com
GitHub: GitHub Profile