DS5110 Iteration 04

Katherine Barney and Nick Sheft

November 2025

1 Dataset Description

1.1 Datasets:

- https://www.kaggle.com/datasets/atomicd/retail-store-inventory-and-demand-forecasting
- https://www.kaggle.com/datasets/anirudhchauhan/retail-store-inventory-forecasting-dataset

1.2 About the Datasets:

These datasets both contain retail store inventory and demand forecasting data, which makes them very relevant for our project goals. The datasets contain the following columns:

- Date: Date of the record.
- Store ID: Unique identifier for the store.
- Product ID: Unique identifier for the product.
- Category: Product category.
- Region: Geographical region of the store.
- Inventory Level: Units available in stock.
- Units Sold: Units sold on that day.
- Units Ordered: Units ordered for restocking.
- Price: Product price.
- Discount: Discount applied, if any.
- Weather Condition: Weather on the day of the record.
- Promotion: 1 if there was a promotion, 0 otherwise.

- Competitor Pricing: Price of a similar product from a competitor.
- Seasonality: Season (e.g., Winter, Spring).
- Epidemic: 1 if an epidemic occurred, 0 otherwise.
- Demand: Daily estimated demand for the product

Both datasets were obtained from Kaggle.

1.3 Why this Dataset was Selected:

These datasets were selected due to their historical usage in retail demand forecasting, thus making them optimal for our needs. Additionally, because we are focusing on seasonal demand, these datasets were ideal due to their extensive date ranges and their identical data formatting and column arrangements.

2 Tools and Methodologies

2.1 Tools, Models, Frameworks, and Techniques

- Data Storage / Processing: SQL Server, PySpark (Databricks)
- Analysis and Visualization: Pandas, Plotly, Seaborn, Tableau
- Feature Engineering: NumPy, Pandas, Scikit-learn
- Forecasting Models: LinReg, Neural Networks (TensorFlow/PyTorch), XGBoost, ARIMA
- Collaboration and Version Control: GitHub, Jupyter Notebooks / VS Code

These tools will support our objectives as both team members have experience with a range of these tools inside and outside of class, thus making them ideal for the purposes and needs of this project. A majority of these packages are robust and well-defined, which makes them key resources for completion of our project.

3 Preliminary Timeline

3.1 Weekly Timeline

This weekly timeline can be found in our GitHUb repository in the format of an Excel time tracker file. Milestones for data preparation, model development, evaluation, visualization, and final reporting are included.

4 Team Member Contributions

By collaboratively learning new skills, techniques, models, etc. we can ensure that all members are comfortable with the technology. Each member is aware of their specific role. As of Iteration 04, both teammates have contributed equally through a joint effort in completing the first three tasks as outlined in the weekly timeline file.

As of now, all tasks have been completed by team members in parallel. As project tasks become more complex or time-consuming, roles may evolve to support this through more individualized work as outlined in the weekly timeline file.

5 Progress:

So far, the proper datasets have been acquired, the project scope has been defined, data cleaning and pre-processing have been handled, and feature engineering has been implemented. Ongoing work includes exploratory analysis, visualizations, identification of demand patterns, and the collection of correlational data.

6 Next Steps:

Next steps include the following:

- Data Splitting for Testing and Training
- Baseline Model Development
- Feature Selection
- Predictive Model Engineering
- Model Evaluation
- Uncertainty Quantification
- Scenario Testing
- Reports and Suggestions Based on Findings