DS5110 Iteration 2

Nick Sheft and Katherine Barney

October 26th 2025

1 Project Kickoff

1.1 Goals

- Forecast product demand across different retail categories using historical sales and external seasonal data.
- Identify key drivers of demand fluctuations (e.g., holidays, promotions, weather, regional factors).
- Develop predictive models to improve forecasting accuracy.
- Quantify uncertainty and provide actionable insights for inventory and supply-chain decisions.

1.2 Scope

The data should include historical sales, holidays, store metadata, weather, region-level data, etc. modeled with time series forecasting. It should utilize big data tools like PySpark with SQL integration.

1.3 Deliverables and Phases

1. Problem Definition and Data Collection: Gather historical retail datasets and define forecast targets
2. Data Cleaning and Feature Engineering: Handle missing data, aggregate time-series, engineer seasonal and holiday features
3. Model Development: Train baseline models
4. Model Evaluation and Optimization: Compare models, fine-tune hyperparameters, interpret feature importances
5. Deployment and Visualization: Present forecasts and insights
6. Final Report and Presentation: Summarize findings, business implications, and limitations

1.4 Capabilities

Both group members have knowledge in this domain but need further education on big data technologies like Spark.

1.5 Datasets

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

2 Team Discussions

2.1 Core Skills

Each member has core background competencies in areas like mathematics, coding, machine learning, database design, data visualization, and model building that are pivotal for this assignment. The current missing skills surround specific technologies or methods like using PySpark and specific time series forecasting models. Both team members have skills coding in Python and SQL, which are the primary languages that will be used for the project.

3 Skills and Tools Assessment

3.1 External Resources

Individauls like professors/TAs can assist in building knowledge where needed. Additionally, there are resources online e.g., GeeksForGeeks, Khan Academy, W3 Schools, etc. which are useful.

3.2 Tools

- 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

3.3 Team Roles

By collaboratively learning new:

- skills
- techniques
- models
- approaches/methods
- and more!

We can ensure that all members are comfortable with the technology. Each member is aware of their specific role.

4 Initial Setup

We will be using a GitHub containing interactive python notebooks for most of the model development. Every group member has access to the GitHub repository.