University of New Haven
TAGLIATELA COLLEGE OF ENGINEERING
Electrical & Computer Engineering and Computer Science

Electrical & Computer Engineering & Computer Science (ECECS)

TECHNICAL REPORT

SEMESTER 2

CONTENTS

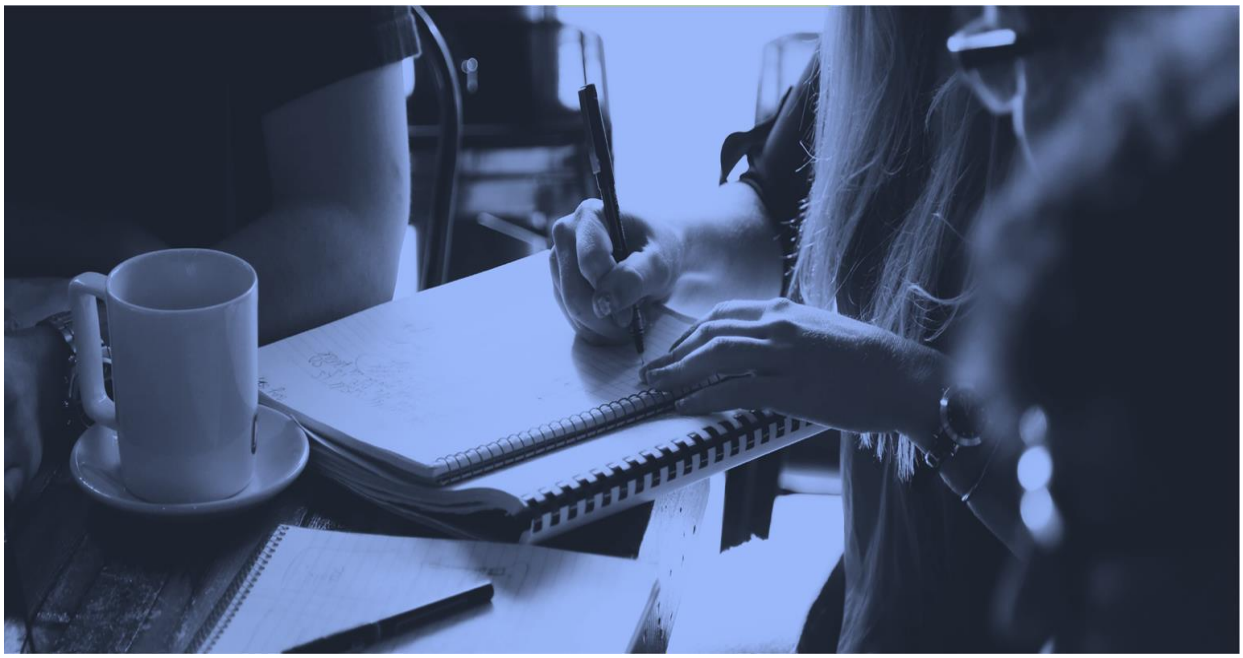
Executive Summary

Businesses and customers alike must keep an eye on dynamic price adjustments across e-commerce products in today's fast-paced digital retail environment. A real-time and batch ETL pipeline is implemented in this project, which gathers product pricing data, processes it effectively, uses machine learning to forecast future price reductions, and uses interactive dashboards to display actionable insights.

A mix of cloud-native services and technologies are used in the pipeline's design:

- Data Ingestion Layer: Using APIs, a Dockerized producer retrieves real-time product data and streams it as raw CSV files into an AWS S3 bucket.
- Data Processing (ETL Layer): The raw data is cleaned, converted, and then put back into S3 under an organized cleaned folder using AWS Glue Crawler and Glue ETL Jobs.
- Query and Analytics Layer: Amazon Athena uses SQL to directly query the cleaned S3 data.
- Power BI Desktop can see and analyze real-time data without the need for manual exports thanks to the Athena ODBC Connector.
- Machine Learning Layer: Using past trends, a RandomForestClassifier model was trained to forecast the possibility of a price decline.
- AWS EC2-hosted Streamlit was used to deploy the model, allowing for real-time prediction, risk classification, and suggested actions (Buy Now or Wait).
- Visualization Layer: KPIs such as average price drop, category-wise drop distribution, risk trends, and comparisons of anticipated and actual drops are available through Power BI dashboards that are live-connected to Athena.

Team Members



Questions?

Contact:

Team Members:

Stuti Bimali	sbima1@unh.newhaven.edu
Priyanka Singh	psing21@unh.newhaven.edu
Lakshman Rajith	lrong1@unh.newhaven.edu

TITLE

Real-Time and Batch ETL Pipeline for E-Commerce Price Drop

Highlights of Project:

This project delivers a fully functional, real-time and batch ETL pipeline integrated with machine learning and business intelligence, specifically targeting price drop prediction in e-commerce environments. It shows the convergence of cloud-native technologies, automated data processing, and intelligent analytics, and highlights the ability to scale and adapt solutions in a production-ready environment.

At the core of the solution is a Dockerized Python producer, designed to simulate a real-world data ingestion environment by streaming live product data—such as ID, title, category, current and historical prices—directly from an external API into Amazon S3 in real time. This real-time ingestion mechanism supports scalability and modularity, serving as the foundation of the data pipeline.

The ETL layer is implemented using AWS Glue Crawlers and Glue ETL Jobs, which automate schema detection, validate raw input, clean the data, and apply necessary transformations. These include calculating the percentage price drop, removing null values, and labeling entries as "drop" or "no drop." The transformed data is then stored

in an organized format within S3, under a separate /cleaned/ folder, making it ready for downstream analytics.

To enable efficient querying and analytics, the project leverages Amazon Athena, a serverless, pay-per-query service that allows direct SQL-based interaction with the structured data in S3. This eliminates the need for deploying and managing traditional database infrastructure. By using the Athena ODBC Connector, the system integrates seamlessly with Power BI Desktop, allowing dynamic and real-time visualization of business-critical metrics.

On the machine learning side, a Random Forest Classifier model was trained using historical product data to predict the likelihood of a future price drop. Features such as old prices, price trend, and product category were used in training. The trained model is deployed using Streamlit, a lightweight Python web app framework, and hosted on an AWS EC2 instance. The app enables users to input product information and receive immediate predictions, including confidence scores, drop amounts, risk levels, and actionable recommendations such as "Buy Now" or "Wait."

The Power BI dashboards provide live visual insights into KPIs such as average price drops, number of drops by category, historical price movement trends, and total product counts. Filters and slicers enhance interactivity, enabling granular exploration of trends by date, category, or drop type. This visualization transforms raw data into easily interpretable business intelligence for both technical and non-technical stakeholders.

This project stands out by integrating all critical components of modern data architecture—data engineering, machine learning, cloud computing, and visualization—into one cohesive system. It emphasizes production-readiness, fault tolerance, and end-to-end automation, ensuring minimum manual intervention. Furthermore, it demonstrates how businesses can gain real-time decision-making capability and leverage intelligent forecasting to enhance their competitive edge in fast-moving e-commerce markets.

Submitted on: 30th April 2025

Abstract

The implementation of a real-time and batch ETL pipeline for e-commerce price drop monitoring and analysis is the primary focus of this work. Live product data was collected via API using a Dockerized Python producer and ingested directly into AWS S3 storage. Schema detection, data cleaning, and transformation were automated through AWS Glue Crawler and ETL Jobs, resulting in structured datasets stored back into S3.

Amazon Athena provided serverless SQL querying over the cleaned data, while an Athena ODBC connector enabled seamless integration with Power BI Desktop for real-time visualization and analytics. Separately, a machine learning model based on Random Forest was trained to predict price drops and deployed through a Streamlit application hosted on an AWS EC2 instance.

The overall architecture combines real-time ingestion, batch processing, SQL querying, machine learning deployment, and dynamic dashboarding into a scalable and modular solution. This system enables efficient delivery of actionable insights regarding product price behaviors, supporting improved decision-making for businesses and consumers.

Introduction

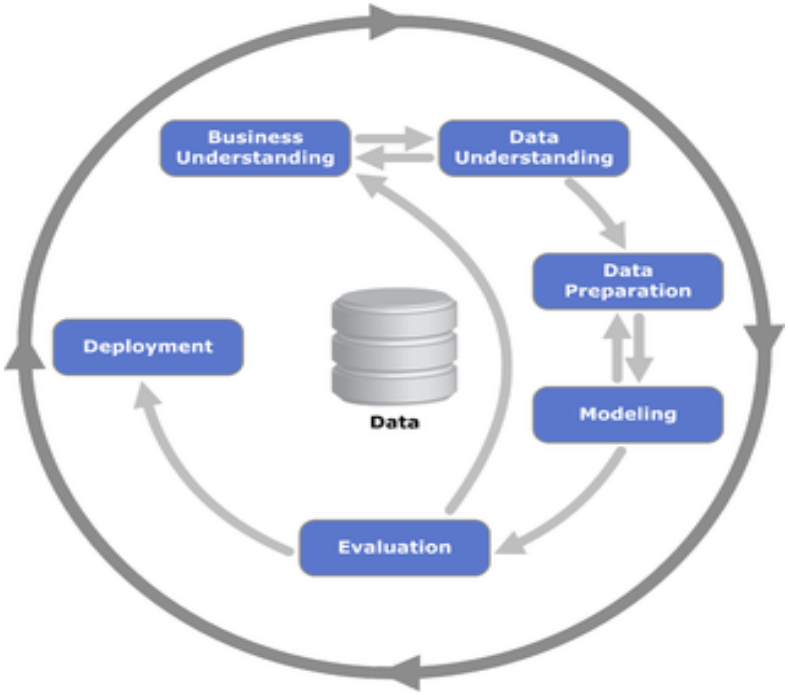
Today's e-commerce market is defined by extremely dynamic pricing methods, with product prices regularly changing in response to seasonal patterns, competition, and customer demand. Both customers looking for better discounts and organizations looking to maximize income now depend on tracking these price fluctuations and forecasting future adjustments. Organizations may make data-driven decisions and maintain their competitiveness in the market by having timely access to real-time insights into product pricing behavior.

To address this need, a real-time and batch ETL pipeline was developed using cloud-native technologies. A Dockerized Python producer was created to fetch live product data from public APIs and stream it into AWS S3 storage. AWS Glue Crawlers and ETL Jobs were employed to automate schema detection, clean the incoming raw data, and transform it into a structured format. Amazon Athena enabled serverless SQL querying over the cleaned datasets, allowing efficient retrieval of meaningful insights without the need for complex infrastructure management.

In parallel, a machine learning model based on the Random Forest Classifier was trained to predict the probability of future price drops. The model was deployed using a Streamlit application hosted on an AWS EC2 instance, offering real-time predictive capabilities. Power BI Desktop was connected via an Athena ODBC connector to create dynamic dashboards showcasing price drop trends, category-wise distribution, and historical pricing behavior. The overall architecture combines real-time ingestion, batch processing, predictive analytics, and interactive visualization into a modular, scalable, and production-ready solution aimed at enhancing decision-making for businesses and consumers.

Methodology

This project follows the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology, which provides a structured approach for implementing data science projects. Each phase is described below in the context of our project:



1. Real-Time Data Ingestion

A Dockerized Python producer was developed to continuously fetch live product data from a public API. The producer collected product attributes such as ID, title, category, current price, and old price. The data was periodically saved as CSV files and streamed directly into an AWS S3 bucket under the /price-stream/ directory, ensuring scalable and reliable cloud storage of real-time incoming data.

2. Schema Detection and Data Transformation

AWS Glue Crawlers were configured to automatically scan the raw data in S3 and create schema definitions inside the AWS Glue Data Catalog. An AWS Glue ETL Job was then developed to clean and transform the ingested data by removing null entries, calculating the percentage drop in prices, and labeling products based on whether a price drop

occurred. The cleaned datasets were subsequently stored back into S3 under a structured /cleaned/ folder.

3. Serverless Querying and Integration

Amazon Athena was used to run SQL queries directly over the cleaned datasets stored in S3. This serverless querying approach eliminated the need for setting up or managing traditional databases. An Athena ODBC connector was configured to enable Power BI Desktop to connect to Athena seamlessly, allowing the creation of real-time business intelligence dashboards and reports without manual data exports.

4. Machine Learning Model Deployment

A machine learning model based on the Random Forest Classifier algorithm was trained separately to predict the probability of future price drops. This model was deployed using a lightweight Streamlit web application, hosted on an AWS EC2 instance, providing users with real-time predictions through a simple and interactive web interface.

5. Visualization and Dashboarding

Power BI Desktop was connected through the Athena ODBC connector to visualize key business insights. Dashboards were built to display KPIs such as average price drops, category-wise price drop distributions, historical trends, and detailed product tables. Slicers and filters were added to enhance user interactivity and allow flexible exploration of the dataset.

6. Overall Architecture

The entire pipeline was designed to be modular, scalable, and production ready. By integrating real-time ingestion, batch ETL processing, machine learning deployment, and dynamic visualization, the project successfully demonstrated a cloud-native end-to-end solution for actionable e-commerce price drop analysis.

Results

The project implementation involved building an end-to-end real-time and batch ETL pipeline integrated with price drop analytics and visualization. Initially, a **Docker-based producer** was set up to continuously fetch real-time e-commerce product data via API and push it into an **Amazon S3** bucket under a designated "price-stream" folder. **AWS Glue Crawlers** were used to automatically detect schema changes from the raw data and catalog it into AWS Glue Data Catalog. Subsequently, an **AWS Glue ETL job** was designed and executed to clean, transform, and store the processed data into another S3 folder called "cleaned".

The cleaned dataset was then made queryable via **Amazon Athena**, enabling SQL-based access to the latest batch of product price records. To provide business insights, **Power BI Desktop** was connected to Athena using an **ODBC Connector**. This enabled the creation of live dashboards displaying critical metrics like total number of products, average price drop percentage, top dropping product categories, and drop trends over time. Users could filter dashboards based on product category or price drop status, facilitating interactive exploration of the market dynamics.

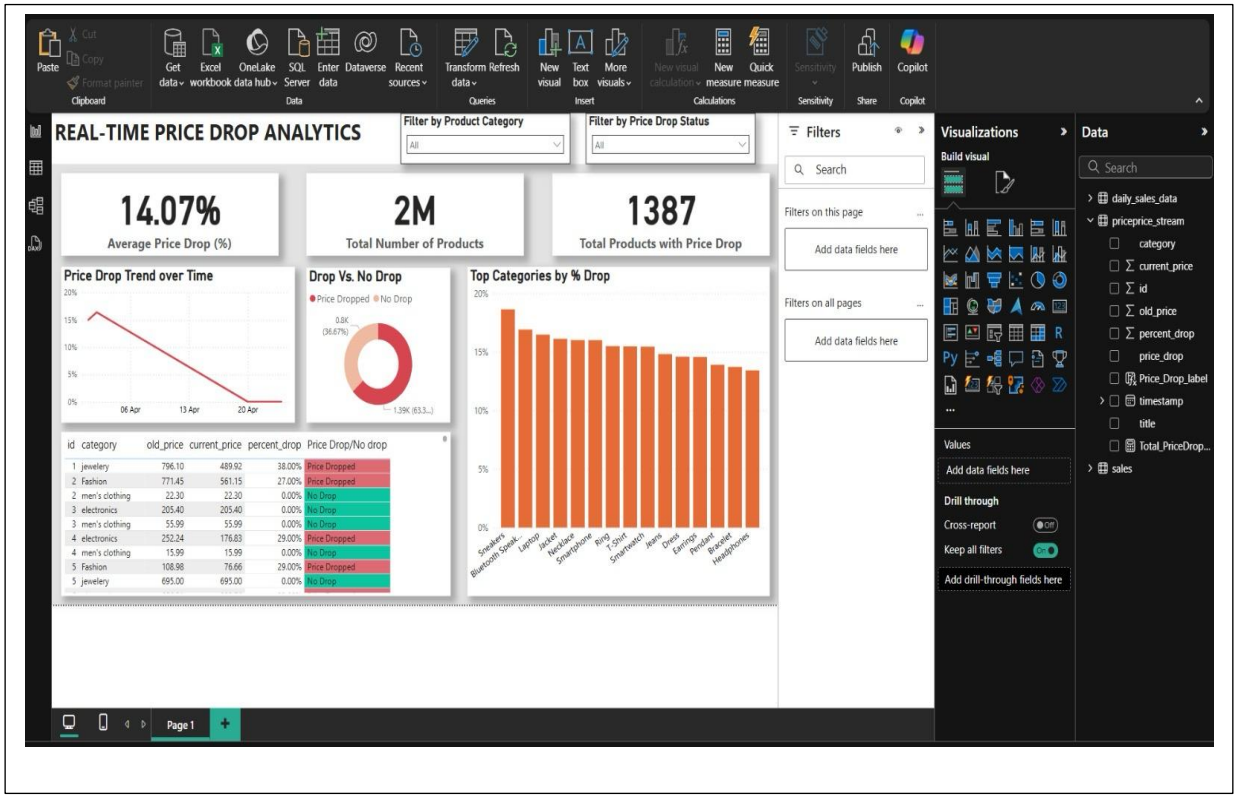
Parallelly, a **Streamlit app** was deployed on an **AWS EC2** instance to allow real-time product-level price drop prediction and risk evaluation. The app enabled users to select a product and instantly view details such as predicted price drop occurrence, drop amount, drop percentage, confidence score, risk level, and actionable recommendations.

The final results demonstrated the effectiveness of the architecture in continuously handling incoming product data, performing batch cleaning, supporting real-time queries, and delivering both high-level analytics through Power BI and detailed predictions via the deployed ML app. The setup ensures dynamic decision-making capability for monitoring product price trends and detecting drops in real time.

1. Power BI Dashboard Results

The Power BI dashboard built using ODBC connection to Athena provides live visualization of price drop analytics. The dashboard includes:

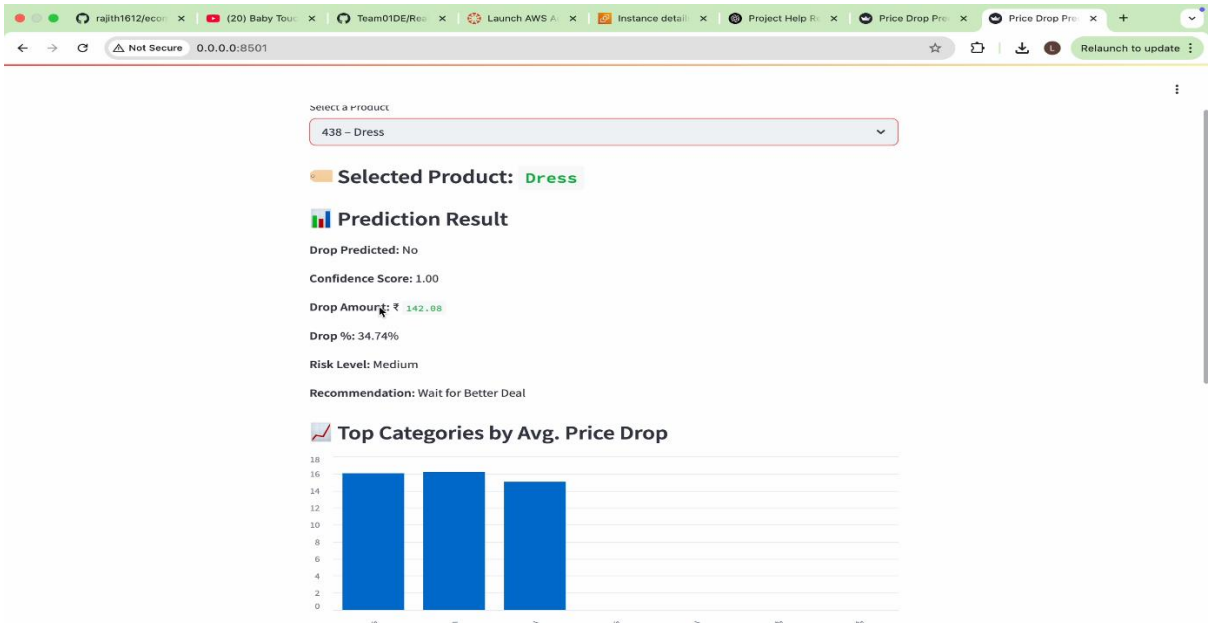
- **Average Price Drop % across all products**
- **Total number of products analyzed**
- **Total number of products that experienced a price drop**
- **Price Drop Trends over Time**
- **Top Categories by Percentage Drop**
- **Drop vs No Drop Split**



2. Streamlit ML Prediction App Results

The ML prediction web app deployed using Streamlit and hosted on EC2 enables users to:

- **Select a product based on ID and name**
- **Get real-time prediction**: whether a price drop is predicted
- **View additional details** like Confidence Score, Drop Amount, Risk Level, and Recommendation.



Discussion

The project successfully built a real-time batch ETL pipeline that streams e-commerce product data using Docker into AWS S3, processes it with Glue ETL jobs, queries it via Athena, and visualizes insights live in Power BI through ODBC. Additionally, an EC2-hosted Streamlit app provides real-time predictions on price drops and risk levels. The integration of ingestion, transformation, querying, visualization, and ML prediction offers a full-stack, scalable system for real-time product analytics. Despite minor challenges like permission setups, the overall system performs efficiently and achieves the project's objectives.

Conclusion

The development of a real-time batch ETL pipeline for price drop prediction successfully combined multiple AWS services, Docker, machine learning, and Power BI into a cohesive and functional system. The pipeline efficiently handled live data ingestion, automated ETL processing, and real-time visualization, while the EC2-hosted Streamlit app enabled interactive ML-based insights. This project demonstrated how modern cloud architectures can deliver scalable, real-time analytics solutions. It also provided valuable hands-on experience in integrating data engineering, machine learning, and business intelligence tools to solve real-world problems effectively.

References

1. **Amazon Web Services (AWS) Documentation. AWS Glue, Amazon S3, Amazon Athena, and EC2 documentation.**
   Link: https://docs.aws.amazon.com/

2. **Streamlit Documentation. Streamlit: Turn data scripts into shareable web apps.**
   Link: https://docs.streamlit.io/

3. **Power BI Documentation. Visualize your data with Power BI and connect to Amazon Athena using ODBC.**
   Link: https://learn.microsoft.com/en-us/power-bi/

4. **Docker Documentation. Developing and deploying containers for reproducible data ingestion.**
   Link: https://docs.docker.com/

5. **Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32.**
   Link: https://doi.org/10.1023/A:1010933404324

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