From Clicks To Conversions

Under Professor:

Amit Patel

Team Members:

Sai Ram Purimetla
Anvesh Varma Dantuluri
Pavan Kishore Ramavath

Objective:

- To analyze and understand customer behavior in an e-commerce environment by processing large-scale retail data using PySpark, with the aim of uncovering patterns in user actions—such as product views, cart additions, and purchases—across different times and dates. The project seeks to generate actionable insights through scalable big data techniques and data visualizations, ultimately supporting data-driven strategies to improve retail conversions.

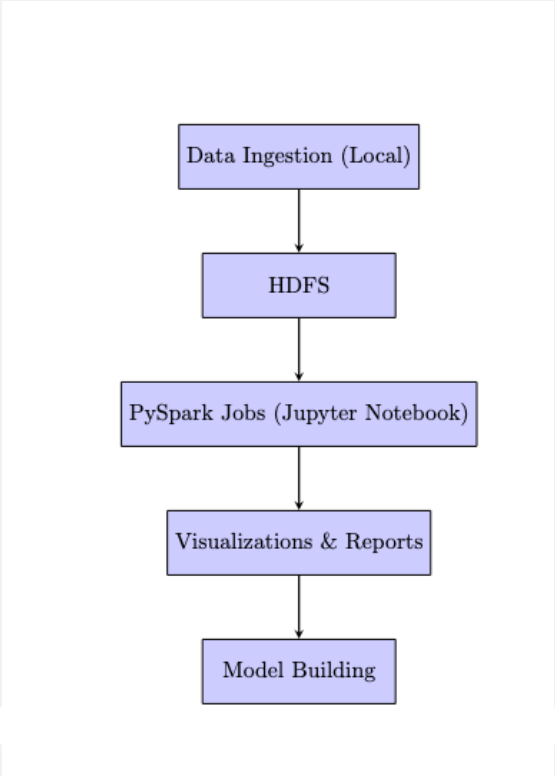
Project Overview

1. **Analyze category & brand performance**: Evaluate which product categories and brands drive the most engagement and sales.
2. **Understand cart abandonment**: Identify where and why customers abandon their carts during the shopping process.
3. **Purchase trends**:Track and visualize how purchases fluctuate across different times and days.
4. **Predict purchases**:Use data-driven models to forecast future buying behavior based on historical patterns.
5. **Drive business actions**:Translate analytical insights into actionable strategies to improve conversions and revenue.

TOOLS AND DATASET

- **DATASET:**
  - 3 million rows (CSV format)
  - https://www.kaggle.com/datasets/mkechinov/ecommerce-behavior-data-from-multi-category-store?select=2019-Nov.csv
- **TOOLS:**
  - Spark
  - Hive
  - Python
  - Dataproc
  - Jupyter Notebook

Work Flow

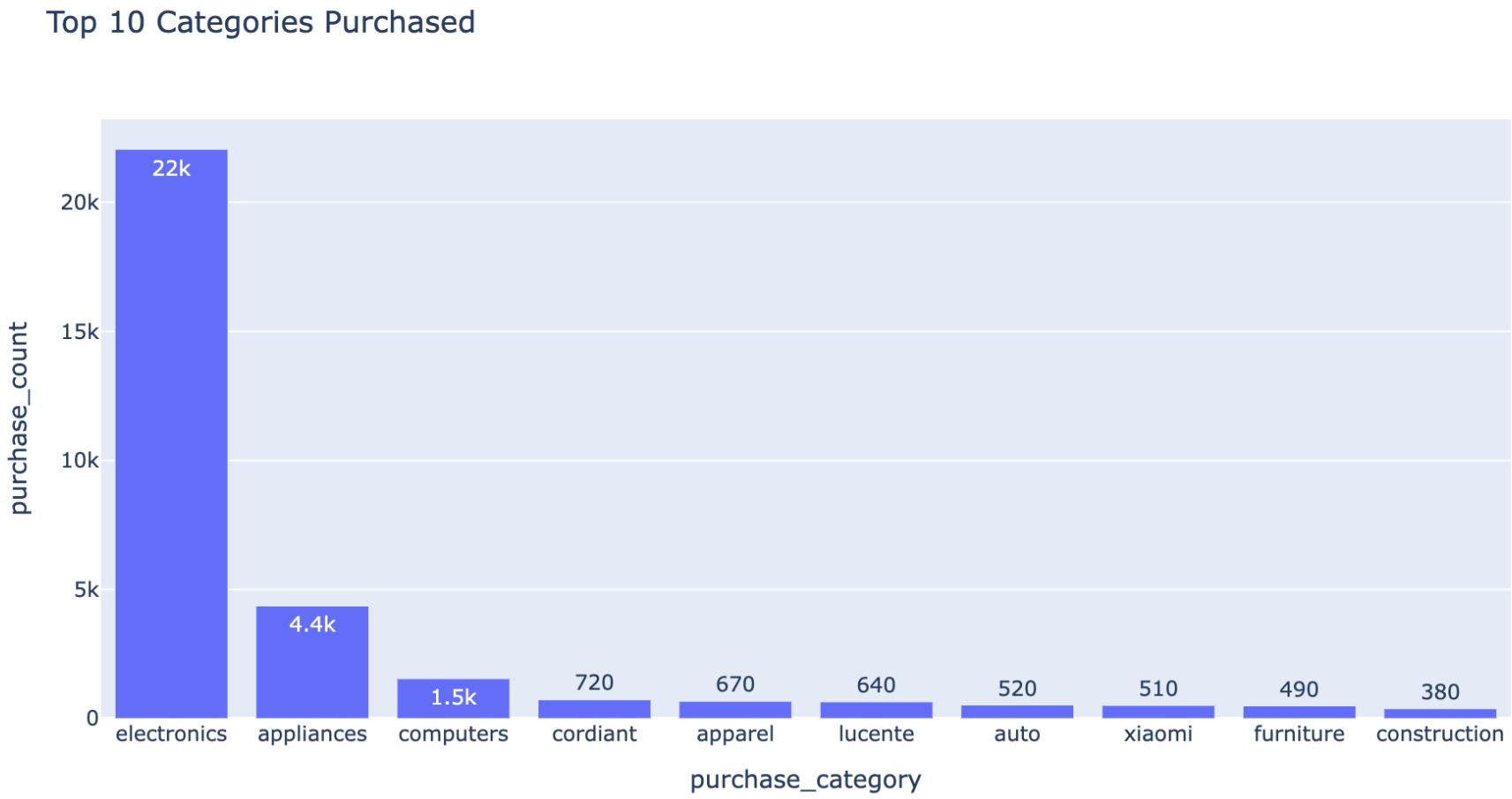

Data Ingestion (Local)
HDFS
PySpark Jobs (Jupyter Notebook)
Visualizations & Reports
Model Building

Exploratory Data Analysis (EDA)

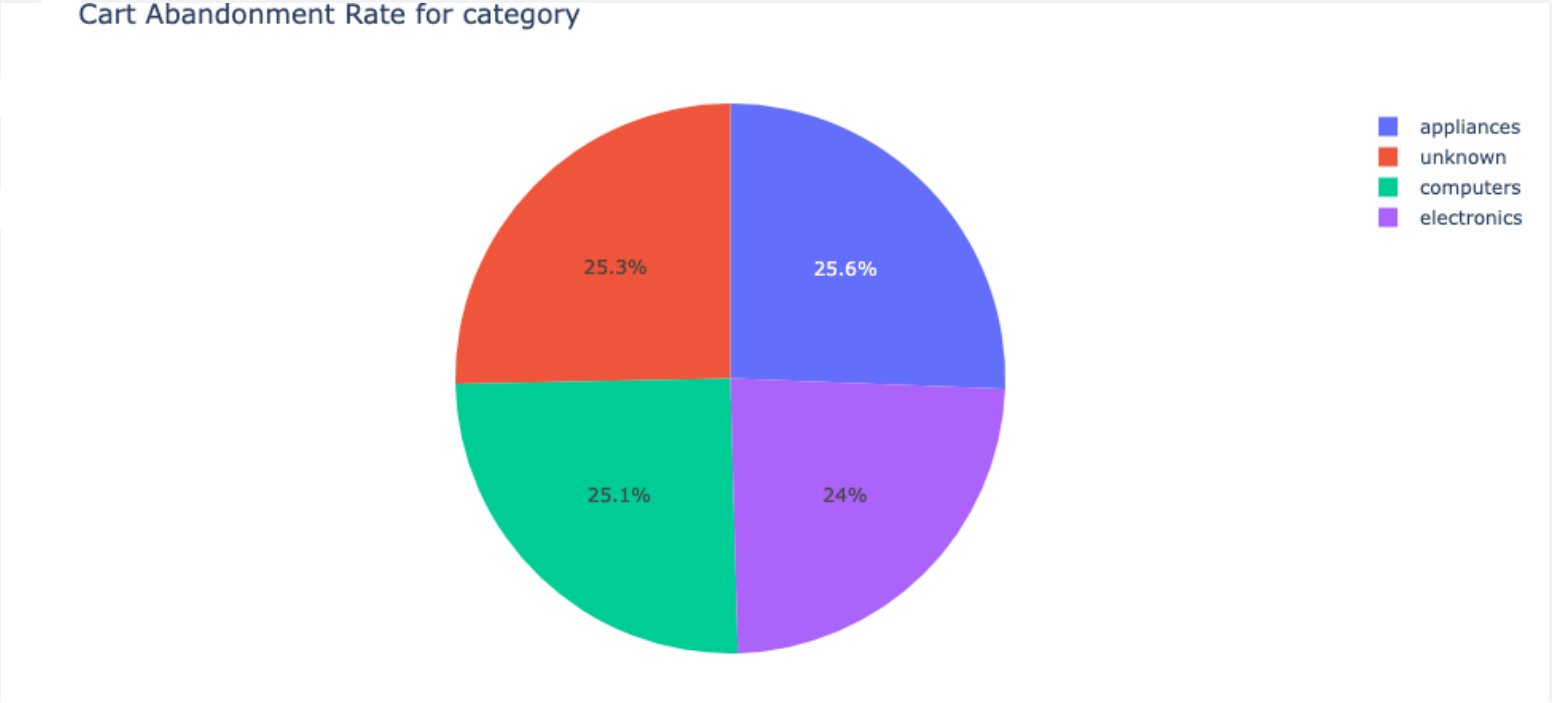
Exploratory Data Analysis (EDA) is mainly performed on the following three topics to uncover insights, identify patterns, and guide further analysis:

- Part 1 – Category and Product Performance
- Part 2 – Cart Behavior and Abandonment
- Part 3 – Temporal Purchase Trends

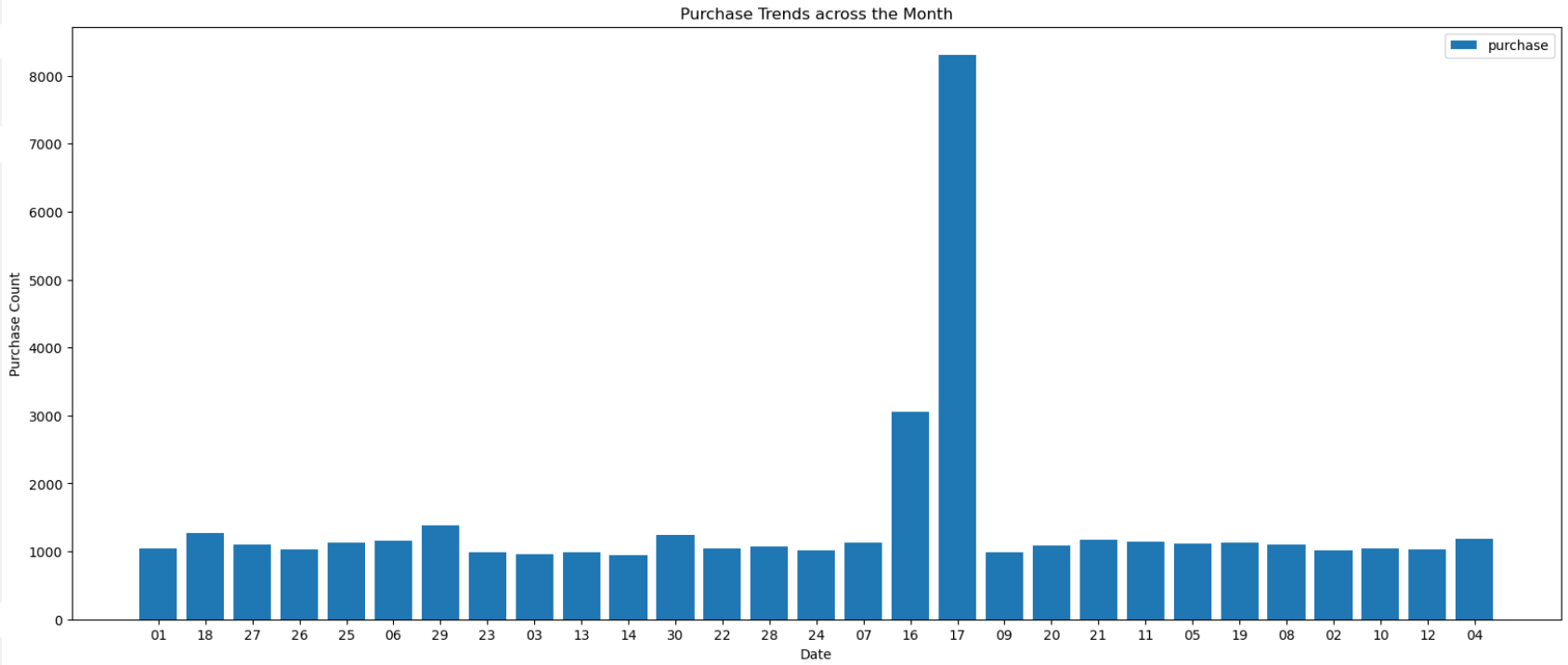
Part 1 – Category and Product Performance



Part 2 – Cart Behavior and Abandonment



Part 3 – Temporal Purchase Trends



Model Development

1. **Three tree-based machine learning models were developed and evaluated:**
    - **Random Forest**: An ensemble method that combines multiple decision trees to improve predictive accuracy and control overfitting.
    - **Decision Tree**: A single tree-based model that splits data based on feature values to make predictions.
    - **Gradient Boosted Trees**: An ensemble technique that builds trees sequentially, each correcting errors from the previous one.
2. **Training & Evaluation:**
    - Models were trained on a labeled dataset and evaluated using accuracy on both training and test sets.
    - The goal was to assess both predictive performance and generalization ability.

RESULTS

Model	Train Accuracy	Test Accuracy
Random Forest	0.7671	0.7645
Decision Tree	1.0000	1.0000
Gradient Boosted Trees	1.0000	1.0000

1. **Random Forest**
   - Moderate accuracy on both training and test sets.
   - Very small gap between training and test accuracy → **Good generalization**.
   - Likely a **realistically performing model** with **no overfitting**.
2. **Decision Tree**
   - Perfect accuracy on both train and test sets.
   - This is **suspicious** and may indicate **overfitting** or **data leakage**.
   - Should be examined further, especially if the dataset is large or complex.
3. **Gradient Boosted Trees**
   - Also shows perfect scores.
   - Similar concerns as Decision Tree: may suggest **overfitting** or **data leakage**.
   - Needs **validation of training/testing process**.

FINAL MODEL CONCLUSION

- **Random Forest** is the most trustworthy model in this comparison.

- **Decision Tree** and **Gradient Boosted Trees** require careful investigation to rule out data leakage or overly simplistic data

Technical Challenges

- HDFS integration
- Spark tuning
- Jupyter memory issues

Future Scope

- Integrate Apache Kafka for Real-Time Data Ingestion.
- Deploy a Real-Time Data Processing Pipeline.
- Develop a Scalable Real-Time Dashboard.
- Automate Pipeline Scaling and Monitoring.

Thank You