LOVELY PROFESSIONAL UNIVERSITY

Report

made on

Continuous Assignment 2

INT315

CLUSTER COMPUTING

Shubham Raj
Reg. No. – 12109415
Roll No.- RK21AKA07

Student Declaration

To Whom It May Concern

I, Shubham, a student of B. Tech under the CSE/IT Discipline at Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Shubham Raj
Reg. No. – 12109415
Roll No.- RK21AKA07

Acknowledgement

I would like to express my sincere gratitude to the following individuals and institutions for their invaluable support and guidance throughout this project:

- Professor **Saqib UL Sabha**, Sir: For their unwavering support, encouragement, and expertise in the subject matter, which greatly contributed to the depth and quality of this project.
- **Lovely Professional University**: For providing unwavering support and encouragement, and for the opportunity to have worked on this project.
- Friends: For their invaluable feedback and support, which helped me to develop and refine my ideas.
- Family: For their essential support and encouragement throughout the completion of this project.

Shubham Raj
Reg. No. – 12109415
Roll No.- RK21AKA07

Purchase Completion Prediction Using Apache Spark

1. Introduction

- Understanding customer behavior is vital for enhancing e-commerce platforms.
- Predicting purchase completion is a critical aspect.
- Traditional analytics approaches are often insufficient.
- Machine learning (ML) offers a data-driven solution for predicting user intent.
- Project investigates ML algorithms to predict purchase completion using session-level data.
- Aims to identify effective approaches for improving conversion rates.
- Lays groundwork for intelligent recommendation and retargeting systems.

2. Objective of the Project

- Leverage machine learning to enhance purchase completion prediction in e-commerce.
    - **2.1 Evaluating ML Models:** Assess performance of supervised ML algorithms (logistic regression, decision trees, random forests, gradient boosting) in predicting purchase completion.
    - **2.2 Feature Analysis:** Identify key behavioral and session-level features influencing purchase decisions (time spent, pages viewed, device type, traffic source).
    - **2.3 Model Optimization:** Improve predictive performance through feature engineering, handling class imbalance, and hyperparameter tuning.
    - **2.4 Practical Implementation:** Propose strategies for integrating predictive models into real-time systems and personalized marketing to improve conversion rates.
- Aims to address limitations of traditional methods and enhance proactive management of cardiovascular health.

1.2. Description of the Project

- Applies machine learning to predict customer purchase completion based on session behavior.
    - **Data Utilization:** Features include session time, pages visited, referral source, and cart abandonment rate. Target variable is binary (purchase completed or not).
    - **Algorithm Selection:**
        - Decision Trees: Chosen for interpretability and handling of categorical/numerical features.
        - Indexing & Vectorization: Used for transforming categorical data into numerical representations.
    - **Performance Metrics:**
        - Accuracy: Overall correctness of the model.
        - Confusion Matrix: Provides insight into true positives, true negatives, false positives, and false negatives.
    - **Implementation:**

Shubham Raj
Reg. No. – 12109415
Roll No.- RK21AKA07

- Apache Spark (PySpark): For scalable data processing and ML pipeline creation.
- Databricks: Unified platform for executing project in a collaborative cloud environment.
- Matplotlib & Seaborn: For generating visualizations (pie charts, bar graphs, scatter plots).
- Pandas: Used for data conversion and manipulation for visualization.

1.3. Scope of the Project

- Outlined in short-term and long-term objectives to improve conversion prediction and guide future enhancements.

Short-Term Scope

1. Build and evaluate ML models that predict purchase completion with high accuracy.
2. Identify key behavioral indicators (session time, referral source, cart abandonment rate).
3. Visualize user behavior patterns using graphs and plots.

Long-Term Scope

1. Expand dataset with additional behavioral metrics (mouse movement, time on page, product interactions).
2. Integrate real-time prediction models for personalized offers/reminders.
3. Employ advanced models (ensemble techniques, neural networks) to improve accuracy.
4. Explore user segmentation and personalization strategies to optimize sales funnel.

3. Software Used

1. **Apache Spark (PySpark)**
   - Purpose: Distributed data processing and machine learning.
   - Components:
     - SparkSession: Initialize and manage Spark environment.
     - spark.read.csv(): Read CSV files with schema inference.
     - DataFrame APIs: Data manipulation (select, groupBy, filter, withColumn).
     - MLlib: Model training and feature engineering.
2. **PySpark MLlib**
   - Feature Engineering:
     - StringIndexer: Converts categorical variables into indexed numeric form.
     - VectorAssembler: Combines input columns into a single features vector.
   - Modeling:
     - DecisionTreeClassifier: Supervised learning algorithm for classification, provides interpretable models and feature importance.
3. **Databricks**
   - Purpose: Unified analytics platform for data engineering, data science, and ML.
   - Components:

- Databricks Notebooks: Interactive environment for running Spark code, visualizing results, and documenting findings.
- Cluster Management: Simplifies Spark cluster management by provisioning and scaling resources.
- Optimized Spark Runtime: Accelerates large-scale data processing tasks.
- Integration with MLflow: Tools for managing the ML lifecycle (tracking, packaging, managing models).

4. About The Datasets

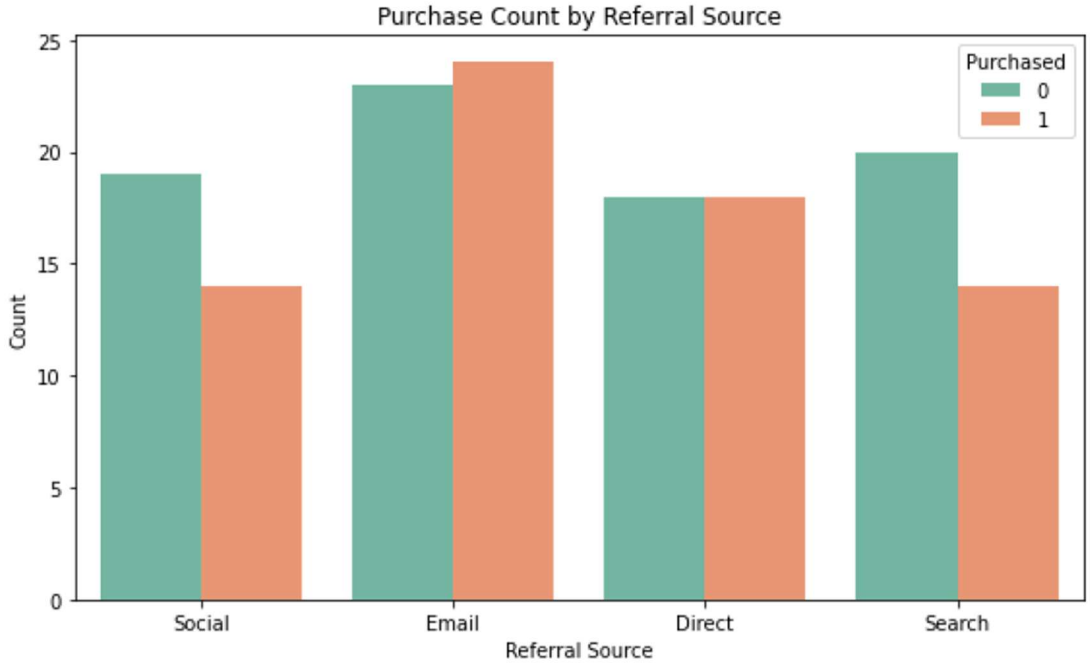
- Used to analyze and predict customer purchase completion based on session behavior.
- Comprises behavioral metrics and categorical features reflecting customer journey and source of arrival.

Column Name	Data Type	Description
SessionTime	Float	Duration (in minutes) a user spent in a session. Indicates engagement level.
PagesVisited	Integer	Number of pages the user viewed during their session.
ReferralSource	Categorical	Channel through which the user landed on the website. Values include: **Search, Social, Email, Direct**.
CartAbdonRate	Float	A score (0 to 1) indicating how likely a user is to abandon the cart. Higher

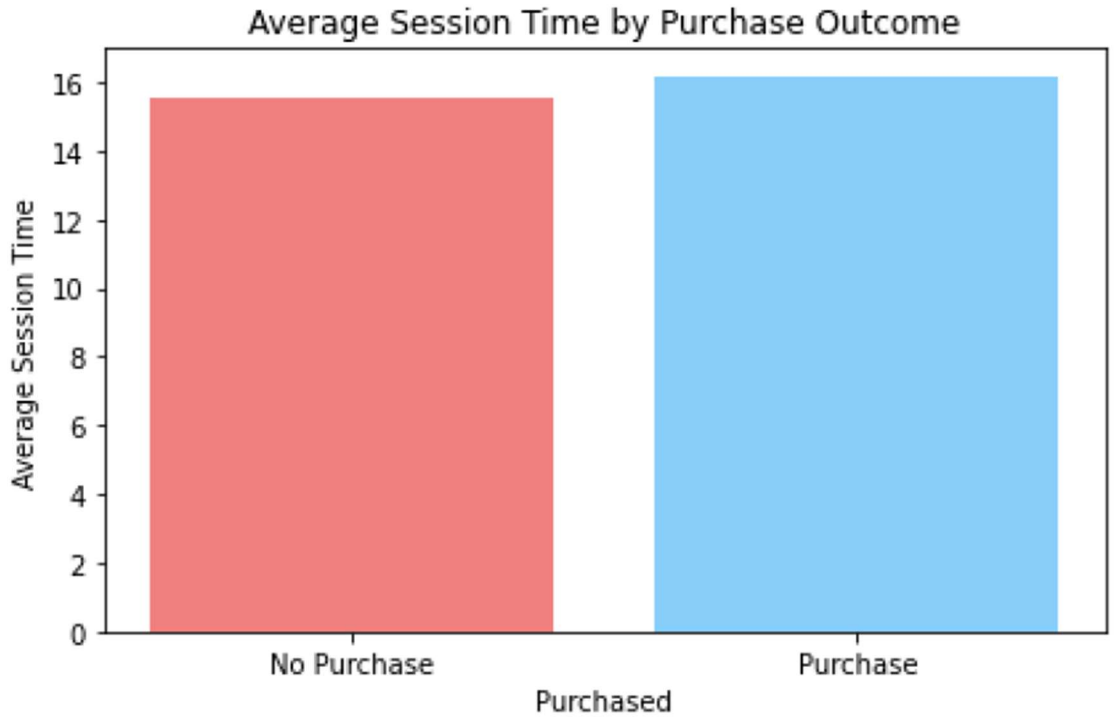
Column Name	Data Type	Description
Purchased	Binary	values mean higher likelihood of abandonment. Target variable – **1** if the user made a purchase, **0** otherwise.

Observations & Insights

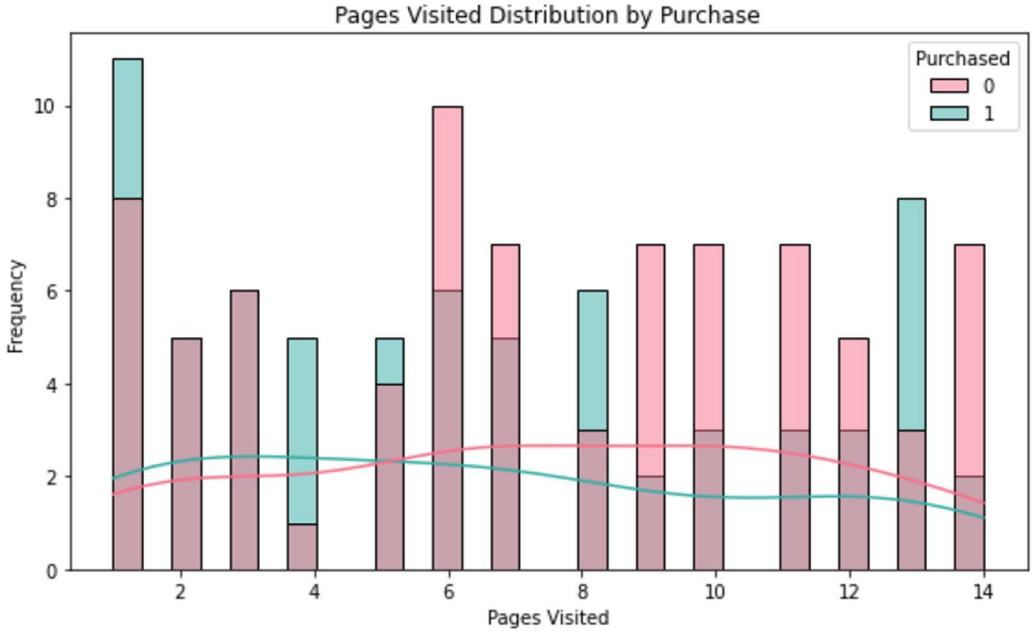
I.



Shubham Raj
Reg. No. – 12109415
Roll No.- RK21AKA07



III.



Shubham Raj
Reg. No. – 12109415
Roll No.- RK21AKA07

Bibliography:

Databricks and Apache Spark Resources

1. Databricks Documentation

- URL: https://docs.databricks.com/
- Description: Offers how-to guides and reference materials for data analysts, data scientists, and data engineers using the Databricks Data Intelligence Platform. It covers Spark cluster setup, notebook usage, MLflow integration, and optimized PySpark runtime.

2. Apache Spark MLlib Guide

- URL: https://spark.apache.org/docs/latest/ml-guide.html
- Description: Information about Spark MLlib, including transformers, estimators, and classification algorithms like the Decision Tree Classifier.

3. PySpark API Documentation

- URL: https://spark.apache.org/docs/latest/api/python/
- Description: Details the API syntax for PySpark DataFrames, StringIndexer, VectorAssembler, and model evaluation.

4. Pandas Documentation

- URL: https://pandas.pydata.org/docs/
- Description: User guide for Pandas, useful for converting Spark DataFrames to Pandas DataFrames for visualization and analysis.

5. Matplotlib and Seaborn Documentation

- Matplotlib URL: https://matplotlib.org/stable/contents.html
- Seaborn URL: https://seaborn.pydata.org/
- Description: Guides for creating visualizations like pie charts, bar charts, and box plots to visualize prediction outcomes and feature distributions.

Citations:

1. https://docs.databricks.com
2. https://docs.databricks.com/

GitHub -: https://github.com/the-shubham-raj/Purchase-Completion-Prediction-Using-Apache-Spark.git

Shubham Raj
Reg. No. – 12109415
Roll No.- RK21AKA07

Shubham Raj
Reg. No. – 12109415
Roll No.- RK21AKA07