



amazon ELECTRONICS REVIEWS CLASSIFICATION USING APACHE SPARK

APPLIED BIG DATA & VISUALISATION

TEAM MEMBERS : ASHWIN, ANURAG , N BALA, NITISH , AND KISHORE



TABLE OF CONTENT

01 THE IMPORTANCE
OF DATA PRIVACY
AND COMPLIANCE

02 UNDERSTANDING
DATA PRIVACY
REGULATIONS

03 IMPLEMENTING
DATA PROTECTION
STRATEGIES

04 MANAGING
DATA BREACHES

05 DATA PRIVACY
COMPLIANCE
TOOLS

06 BUILDING A
CULTURE OF
DATA PRIVACY



INTRODUCTION & DATASET OVERVIEW



- Big Data Challenges:
 - Volume, Velocity, Variety, Veracity
- Dataset:
 - Amazon Electronics Reviews from Stanford SNAP (contains both structured & unstructured data)
- Tools & Environment:
 - Apache Spark, Databricks Community Edition, SparkML



DATA INGESTION, CLEANING & EDA

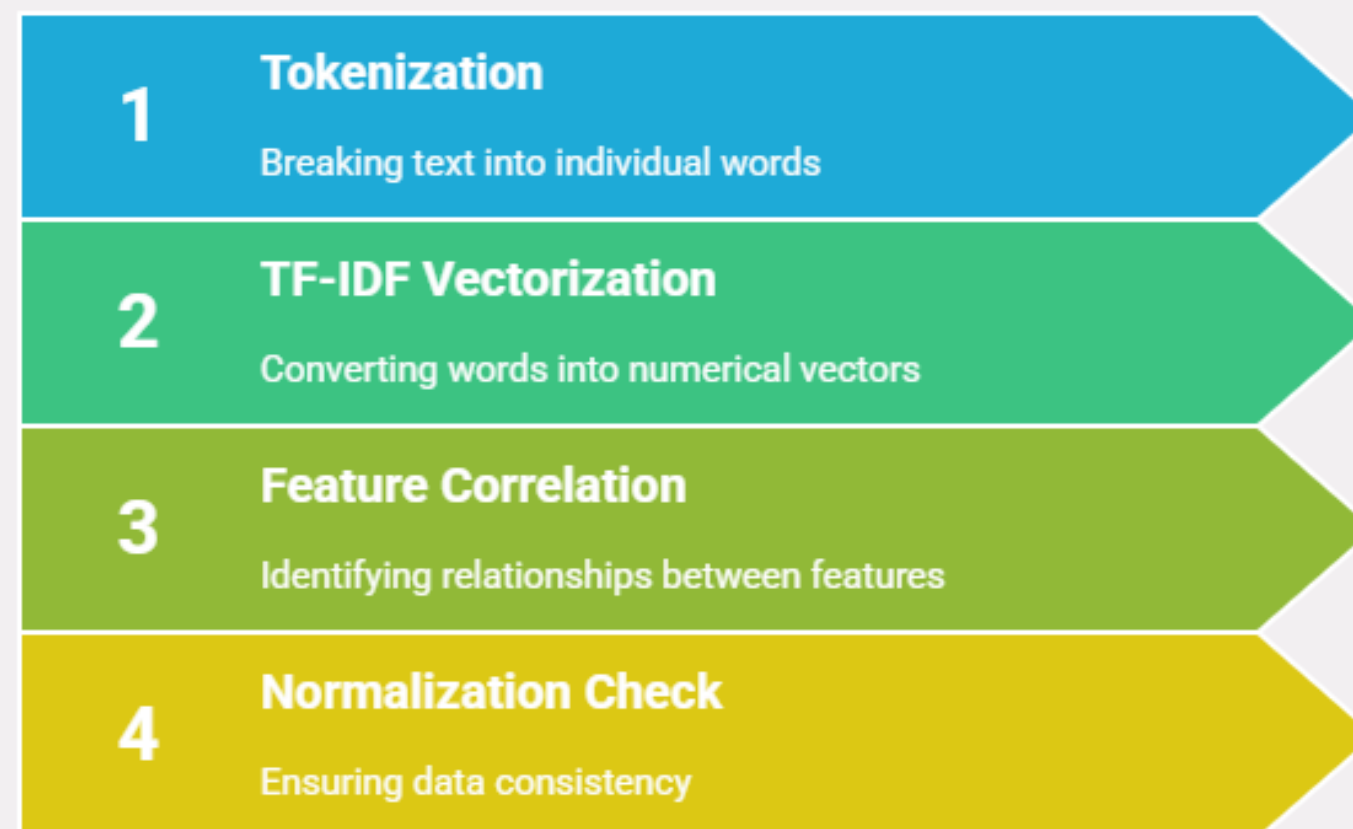


Data Analysis Process Overview



FEATURE ENGINEERING & ANALYSIS

Review Data Processing Funnel



MACHINE LEARNING IMPLEMENTATION

Binary classification: Define positive reviews (≥ 4 stars) and negative reviews (< 4 stars)



APPROACH

- Train/Test Split: 80/20
- Model: Classification- Random Forest
- (scalable with TF-IDF features)

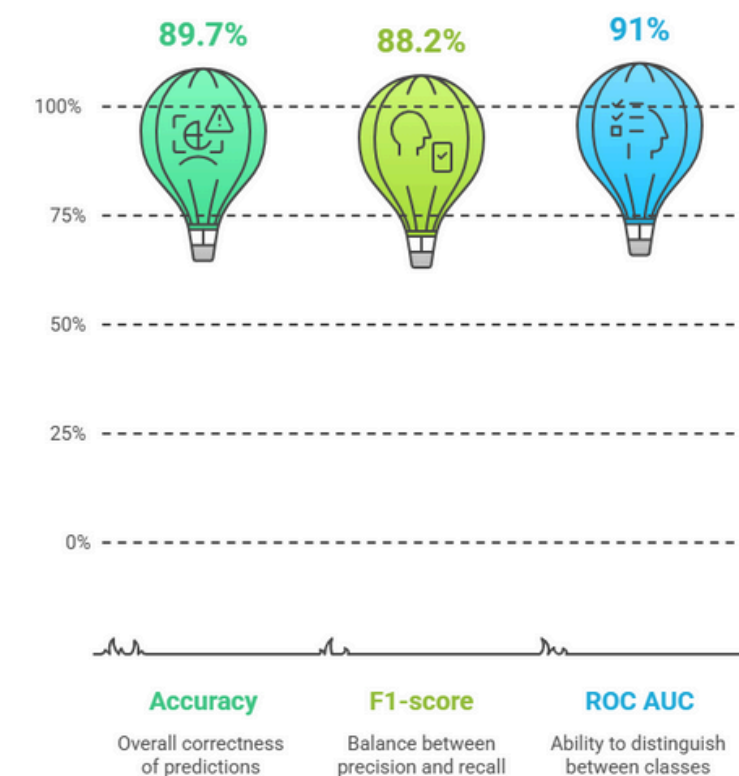


PERFORMANCE METRICS

- Accuracy: 80.29%



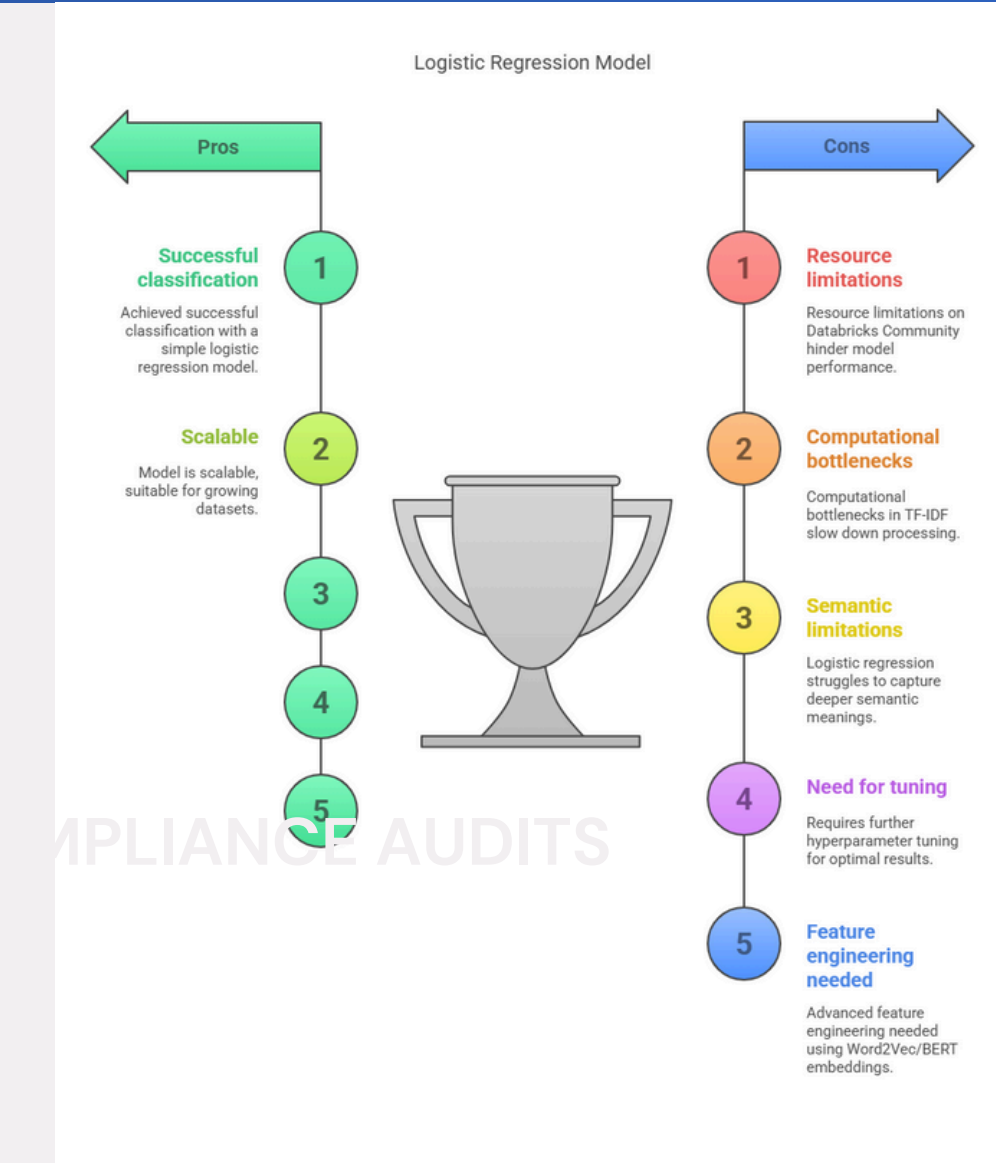
Performance Metrics of Logistic Regression Model





CONCLUSION

- Achievements:
 - Successful classification with a simple and scalable logistic regression model
- Challenges Faced:
 - Resource limitations on Databricks Community
 - Computational bottlenecks in TF-IDF; inability of logistic regression to capture deeper semantics
- Future Directions:
 - Further hyperparameter tuning via cross-validation
 - Advanced feature engineering using Word2Vec/BERT embeddings
 - Experiment with alternative models (Random Forest, Gradient Boosting) for improved performance



MONITORING AND AUDITING FOR COMPLIANCE

Regular monitoring and auditing are essential to maintain compliance with data privacy regulations and to identify potential areas of improvement.

- ▶ CONTINUOUS MONITORING
- ▶ PERIODIC AUDITS
- ▶ COMPLIANCE REPORTING
- ▶ THIRD-PARTY ASSESSMENTS





THANK YOU

FOR TAKING THE TIME TO LEARN ABOUT OUR
DATA PRIVACY AND COMPLIANCE SERVICES

