Flipkart Sentiment Analysis Prediction and MLOps Integration

Overview:

This report provides insights into Sentiment Analysis Prediction for Flipkart reviews, emphasizing the integration of MLOps to enhance machine learning functionality and improve the user experience.

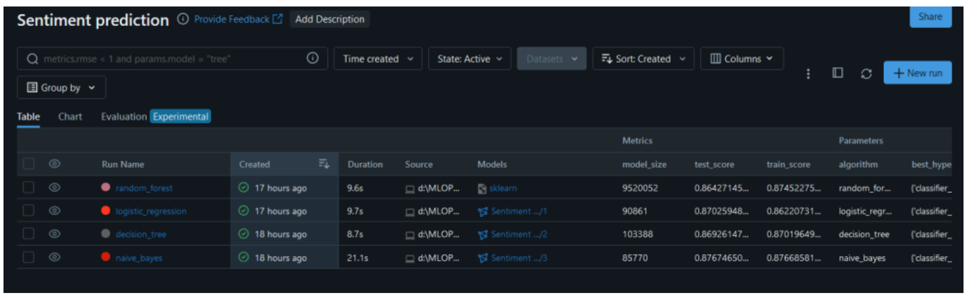
Agenda:

- Introduction
- Experimentation Overview
- Run Analysis
- Metadata Tracking (Tags, Parameters, Metrics)
- Artifacts Management (Output Files)
- MLFlow Features Overview
- Integration Commands

Introduction

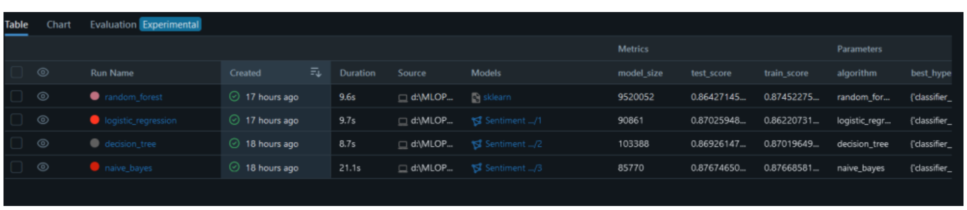
In today's digital era, online reviews play a crucial role in influencing purchasing decisions. Understanding customer sentiment expressed in these reviews is vital for businesses to improve their products and services. This report explores the application of sentiment analysis techniques to Flipkart reviews and demonstrates how MLOps integration can optimize the machine learning pipeline.

Experimentation Overview:



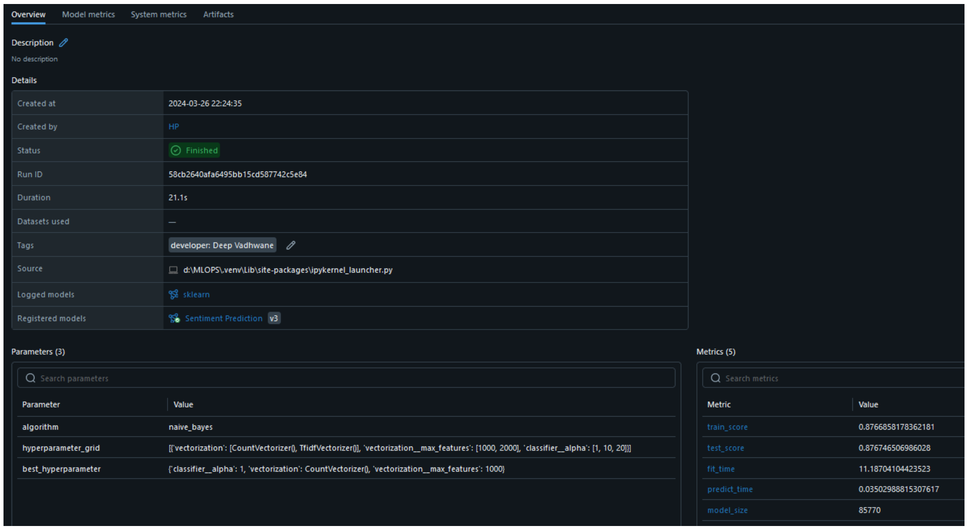
I conducted experiments using various machine learning algorithms, including Naive Bayes, Decision Trees, Random Forest, and Logistic Regression, to predict sentiment from Flipkart reviews. Each algorithm was trained, tested, and evaluated using performance metrics such as accuracy, precision, recall, and F1-score.

Run Analysis



The analysis revealed insights into the performance of different models, highlighting their strengths and weaknesses in sentiment prediction. By comparing performance metrics across multiple runs, we identified the most effective algorithms for sentiment analysis on Flipkart reviews.

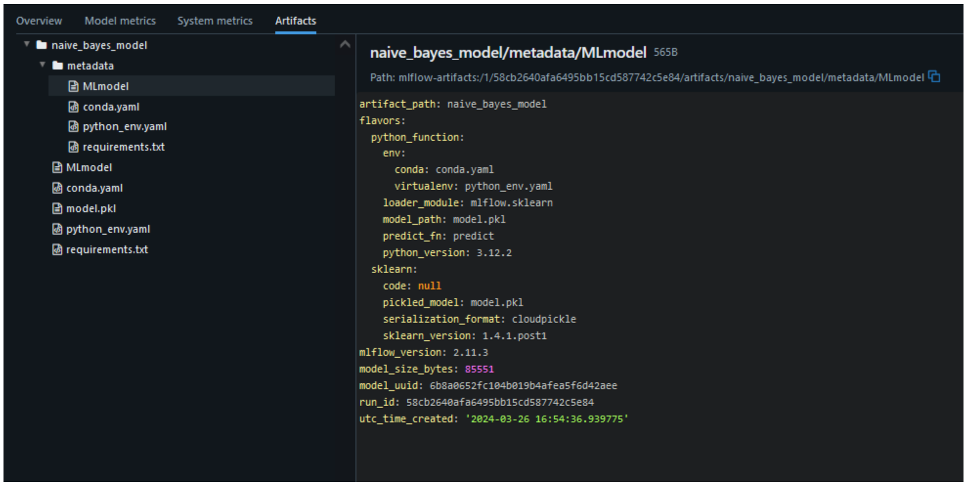
Metadata Tracking



MLFlow's metadata tracking feature allowed us to capture essential information about each experiment run, including tags, parameters, and performance metrics. This metadata provided valuable insights into the experimental setup and facilitated reproducibility and collaboration among team members.

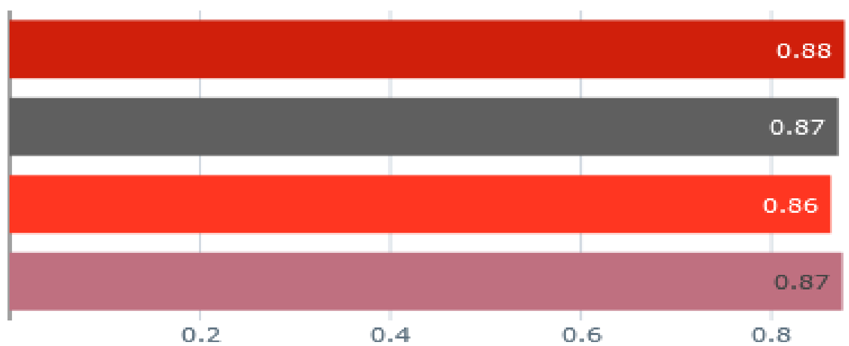
Artifacts Management

MLFlow's artifact management capabilities enabled us to organize and store output files generated during experiment runs, such as model checkpoints, evaluation reports, and visualizations. These artifacts were instrumental in model evaluation, versioning, and deployment processes.



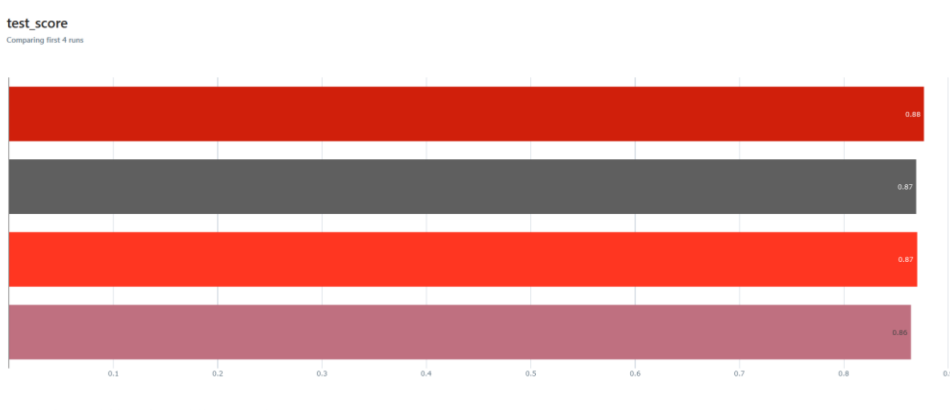
Visulization

Train Score



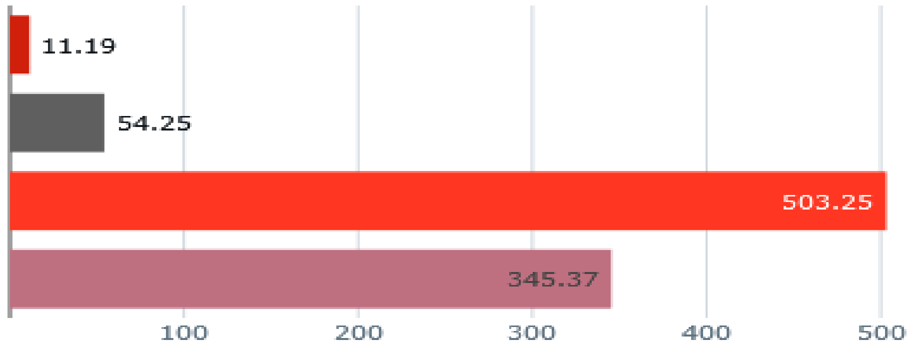
Train Score: - Consistent Training Scores Across Models Showcase Robust Learning Dynamics

Test Score



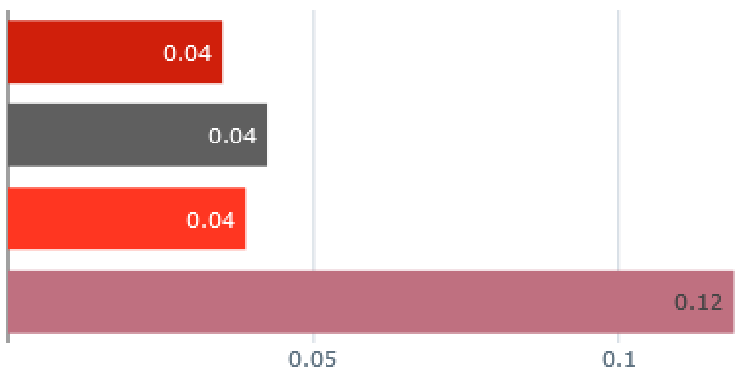
Test score: - Consistent Test Scores Across Models Showcase Robust Learning Dynamics

Fit time



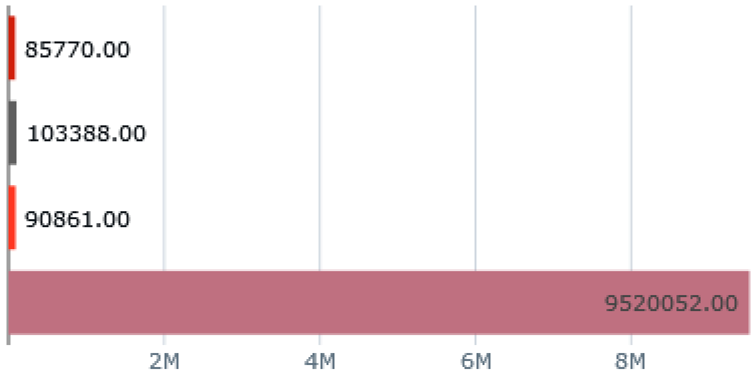
Efficiency in Model Fitting: Naive Bayes Demonstrates Faster Training Time Compared to Other Models

Predict time



Prediction Efficiency: Linear and Naive Bayes Models Outperform Random Forest and Decision Tree in Training Time

Model Size

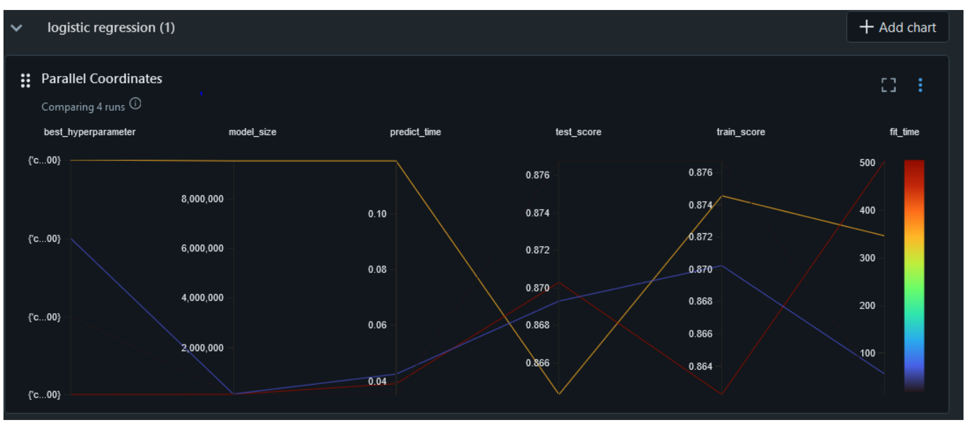


- Random forest has 9520052,
- logistic regression size = 90861
- Decision tree =103388
- naive bayes = 85770

Among these, Naive Bayes exhibits the smallest model size, followed by Logistic Regression, Decision Tree, and Random Forest. This indicates that Naive Bayes requires the least amount of memory storage, making it more efficient in terms of computational resources compared to the other algorithms

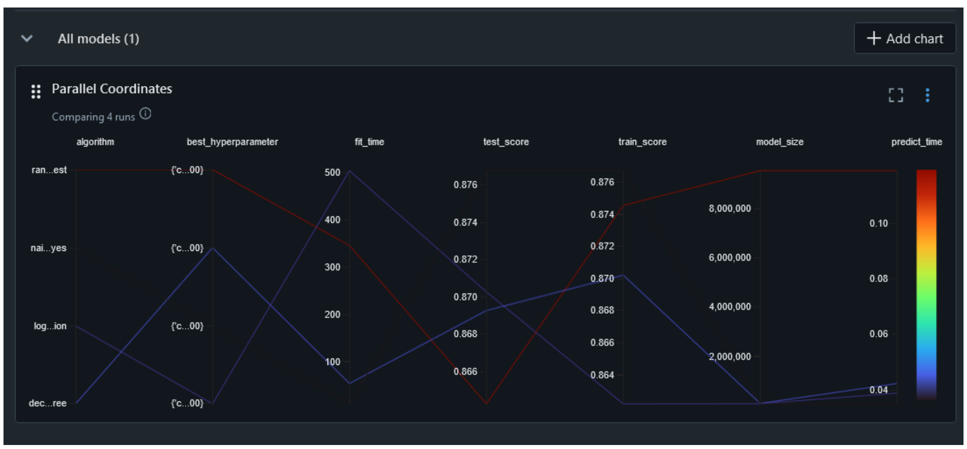
Hyper-parameter Plots

Logistic Regression Parallel Coordinates



Decision Tree Parallel Coordinates

All models



Model Registry:

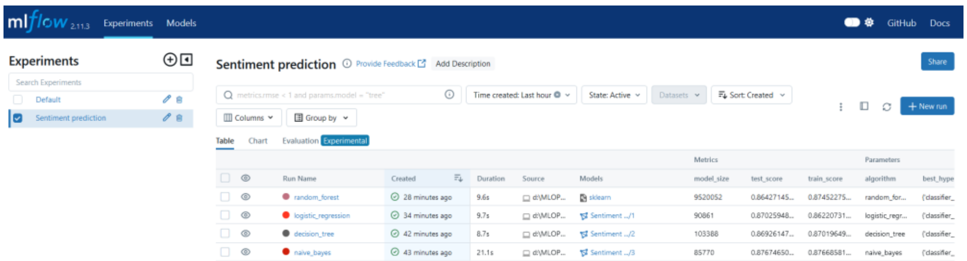
- The Model Registry is a central platform for managing and versioning machine learning models and metadata.
- It facilitates collaboration among data scientists and machine learning engineers throughout the model lifecycle.

Key Features:

- **Model Registration:** Allows seamless registration of models with associated metadata.
- **Model Versioning:** Provides version control capabilities for tracking changes to models.
- **Stage Transitions:** Enables smooth transitions between different stages of the model lifecycle.
- **Intra Team Collaboration:** Promotes collaboration and sharing of models within the team.

Model Versioning

- Versions are tagged as **Archived, Staged,** or **Production** to denote different stages in the model lifecycle.
- Archived versions are no longer in active use, while staged versions are ready for deployment pending final validation.
- Production versions are actively serving users in live environments and undergo strict release procedure



Conclusion

Upon thorough analysis of the metadata, it becomes evident that Naive Bayes is the optimal choice for our sentiment analysis prediction task. This conclusion is drawn from a comprehensive assessment of multiple metrics, including time complexity, training and test scores, and prediction time. Among the models evaluated, Naive Bayes consistently demonstrates superior performance, showcasing lower time complexity for model fitting, competitive training and test scores, and notably faster prediction times compared to other models such as Random Forest, Decision Tree, and Logistic Regression. Therefore, based on these insights, Naive Bayes emerges as the most suitable model for our sentiment analysis prediction, promising enhanced efficiency and effectiveness in our machine learning workflow