Data Science Canvas

Project: Online News Popularity

Team: Rajat Chaudhary, Srinivas Shavukapu Kattegummula, Sonu Goyal, Yuvaraj G

Problem Statement

Business Case & Value Added
Media companies and content creators struggle to predict which articles will gain traction online.

Accurate predictions help optimize content strategy, improve engagement and increase ad revenue by focusing on high-impact articles.

Data Landscape
Online News Popularity - UCI Machine Learning Repository

Model Selection

Regression models
Linear regression, logistic regression, Random Forest, XGBoost
KNN

Classification models
XGBoostClassifier
LightGBM
RandomForest
Adaboost
Naive Bayes
Stacked Ensemble

Model Requirements
Models must handle multivariate, tabular data with both numerical and categorical features. Should be robust to skewed distributions and class imbalance.

Software & Libraries
Python, pandas, numpy, scikit-learn, XGBoost, LightGBM, matplotlib, seaborn, SHAP and LIME, streamlit, TensorFlow and PyTorch

Skills
Python skills for data cleaning, feature engineering, and EDA; ML expertise for model building, tuning, and evaluation. Deployment and interpretability using Streamlit/Flask, SHAP/LIME, plus Git for collaboration.

Execution & Evaluation

Model Evaluation
Use MAE, MSE, RMSE for regression; accuracy, precision, F1-score, macro-averaged metrics for classification. Monitor model performance, especially on minority classes (Trending, Viral).

Data Storytelling
Present results with clear visuals (charts, tables). Focus on actionable insights for editors (e.g., which features drive virality). Communicate limitations and next steps.

Data Collection & Preparation

Data Selection & Cleansing
Use all relevant features except identifiers (e.g., URL). Remove or correct outliers (e.g., invalid ratios, negative keyword shares). One-hot encode categorical, standardize numerical.

Data Collection
Dataset is pre-collected from UCI; no additional data required. Ensure data meets requirements: sufficient size, relevant features, no missing values.

Data Integration
All data is from a single source (UCI dataset); no integration needed.

Explorative Data Analysis
Check for skewness, outliers, and feature distributions. Use univariate, bivariate, and multivariate analysis to understand relationships. Identify that popularity is driven by complex, non-linear interactions.