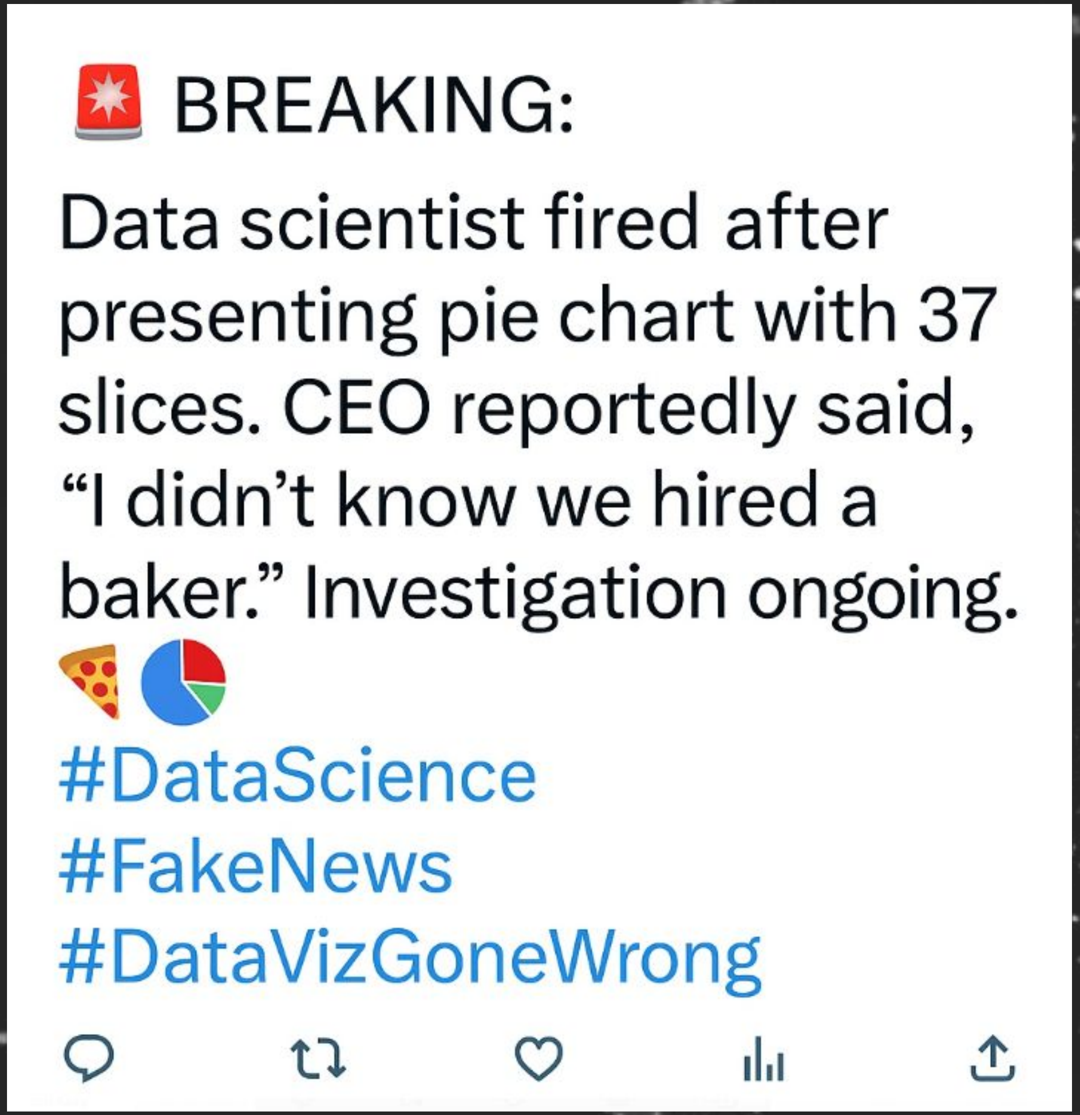
LIVE ●

M.I.S

FAKE NEWS DETECTION

Using NLP

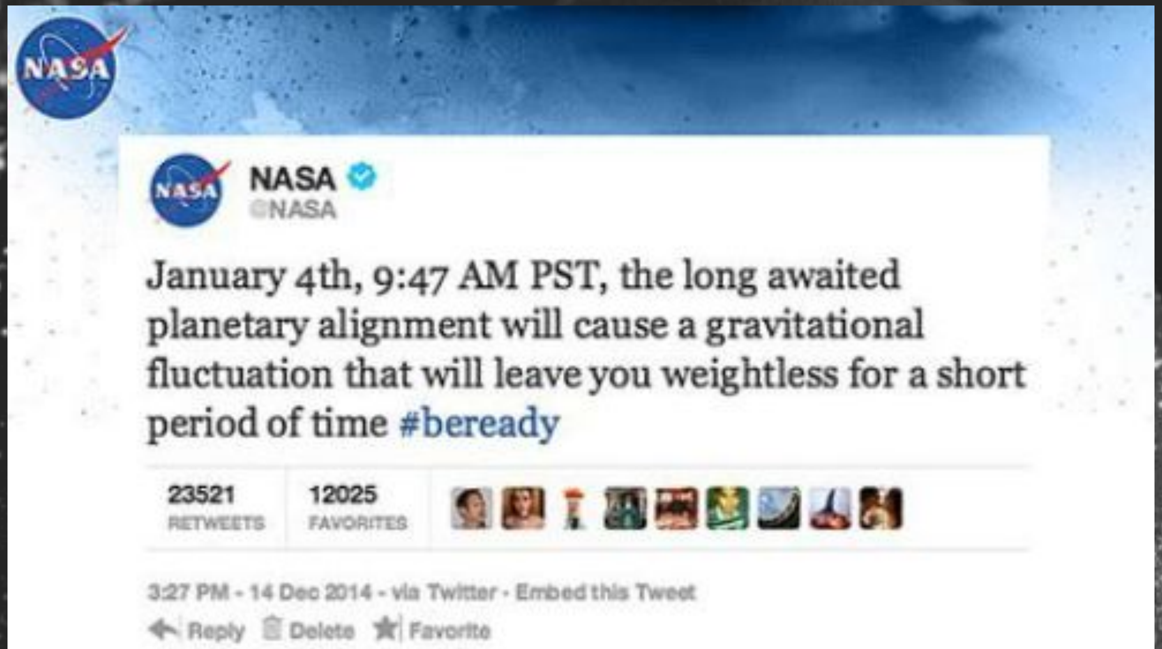
LIVE ●



🚨 BREAKING:

Data scientist fired after presenting pie chart with 37 slices. CEO reportedly said, “I didn’t know we hired a baker.” Investigation ongoing. 🍕

#DataScience
#FakeNews
#DataVizGoneWrong

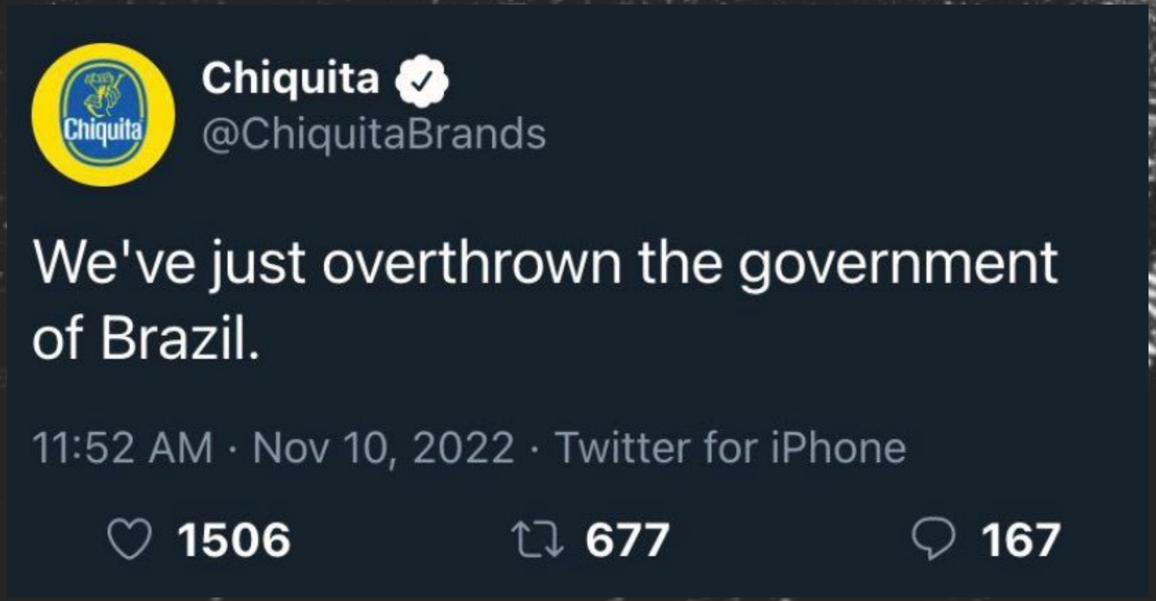


NASA ✓
@NASA

January 4th, 9:47 AM PST, the long awaited planetary alignment will cause a gravitational fluctuation that will leave you weightless for a short period of time #beready

23521 RETWEETS 12025 FAVORITES

3:27 PM - 14 Dec 2014 - via Twitter - Embed this Tweet



Chiquita ✓
@ChiquitaBrands

We've just overthrown the government of Brazil.

11:52 AM · Nov 10, 2022 · Twitter for iPhone

1506 677 167

STOCKS 100% ▼ STOCKS 100% ▼ STOCKS 100% ▲ STOCKS 100% ▼ STOCKS 100% ▲

Asmita Sonavane

Tanvi Takavane

Srushti Shah

Ruchi Jha

Meet The Team ☺️

STOCKS 100% ▼ STOCKS 100% ▼ STOCKS 100% ▲ STOCKS 100% ▼ STOCKS 100% ▲

Business Objective

The Problem: Rise of Misinformation

- ✔ AI-generated articles make it harder to detect fabricated content
- ✔ Existing detection tools rely on basic keyword matching—insufficient for today's complex threats

Our Solution: Intelligent Stance-Based Detection

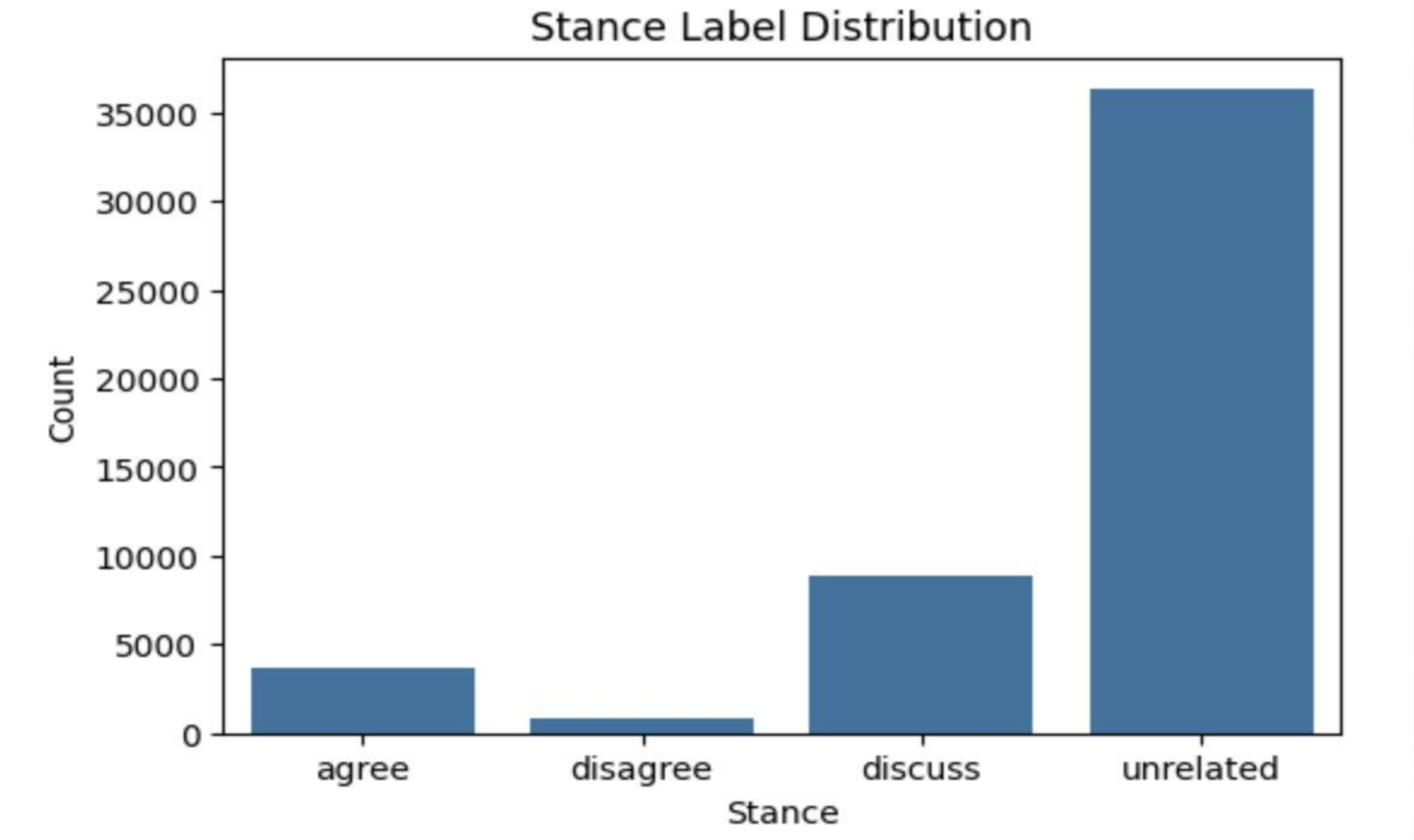
- ✔ Leverage **FNC-1 dataset** to classify the stance between headlines and article bodies (agree, disagree, discuss, unrelated)
- ✔ Use TF-IDF, word embeddings, and semantic similarity to capture deeper linguistic patterns
- ✔ Build a robust ML pipeline with interpretability and high accuracy

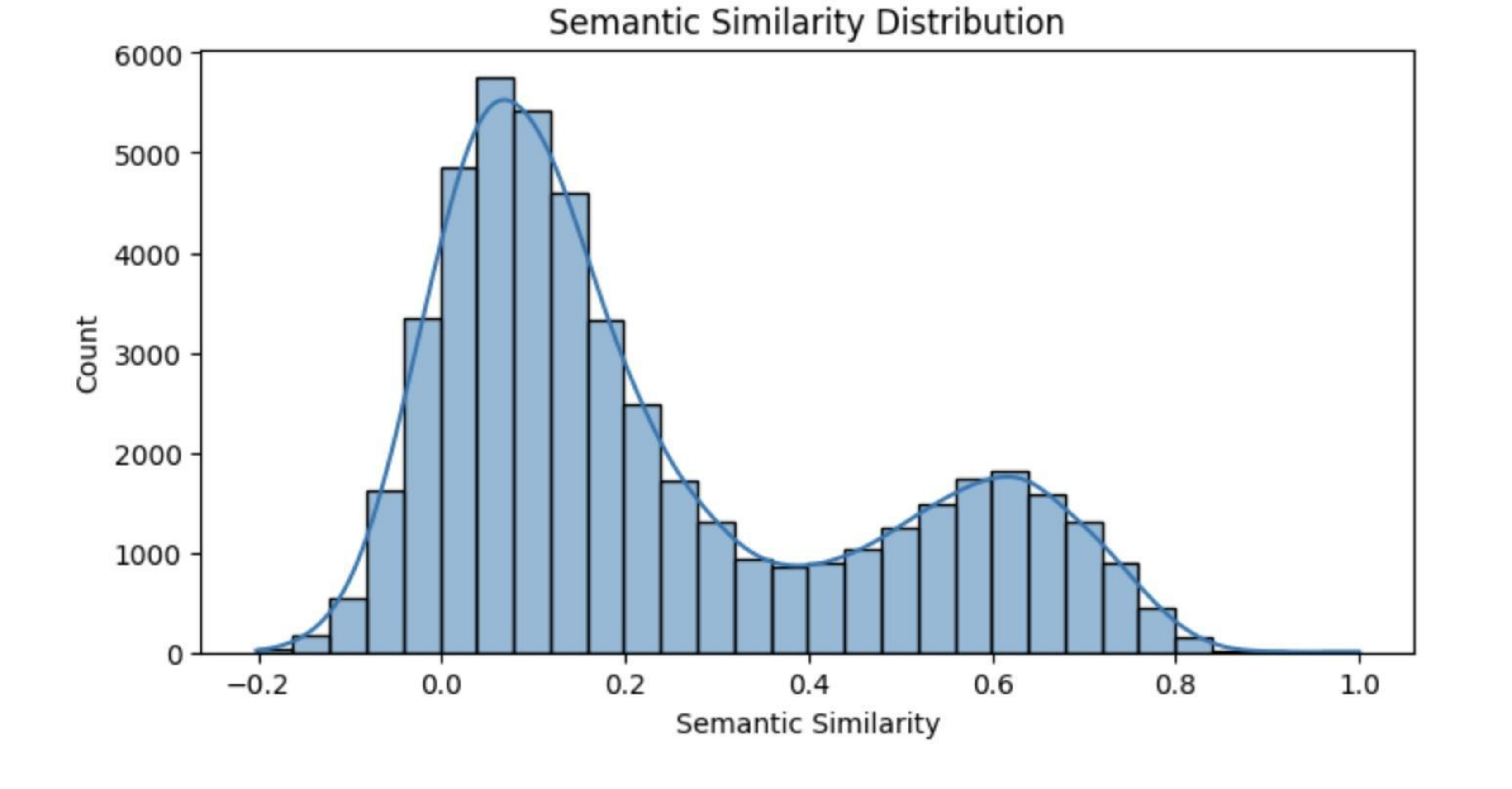
Impact

- ✔ Automate fact-checking systems to quickly flag suspicious articles.
- ✔ Shape regulatory policies by understanding the patterns and prevalence of fake news.

STOCKS 100% ▼ STOCKS 100% ▼ STOCKS 100% ▲ STOCKS 100% ▼ STOCKS 100% ▲

Semantic Similarity Distribution

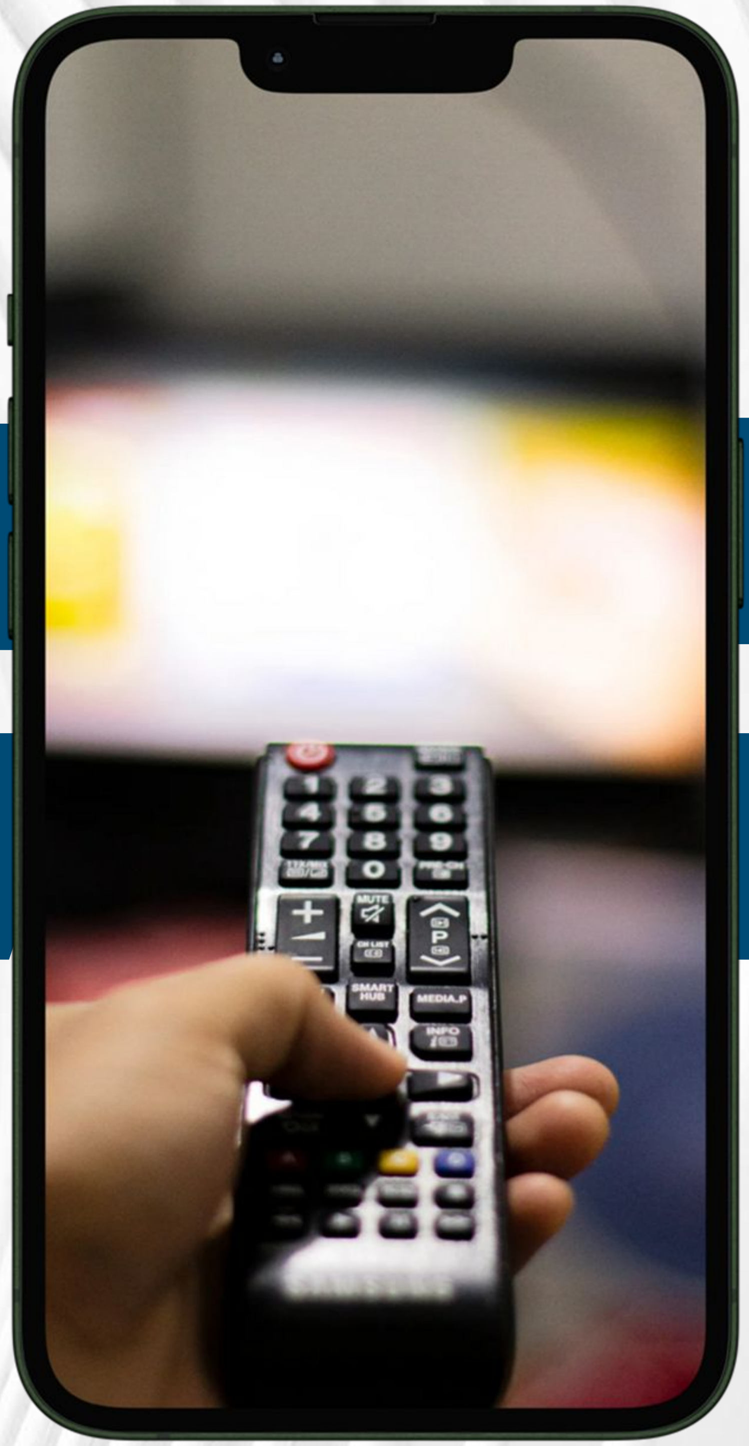




- ***Shows an imbalance in class distribution.***

- ***Suggests semantic similarity is an effective feature to differentiate between related and unrelated pairs-critical for stance distribution.***
- ***Bimodal distribution-peaks show headline-body pairs are either weakly or strongly related,with fewer in-between.***

Smart Feature Engineering & Key Takeaways



TF-IDF, N-gram Overlap, Word Overlap

surface-level lexical similarity

Cosine Similarity

semantic closeness between headline and article

GloVe Embeddings + Sentence Transformers

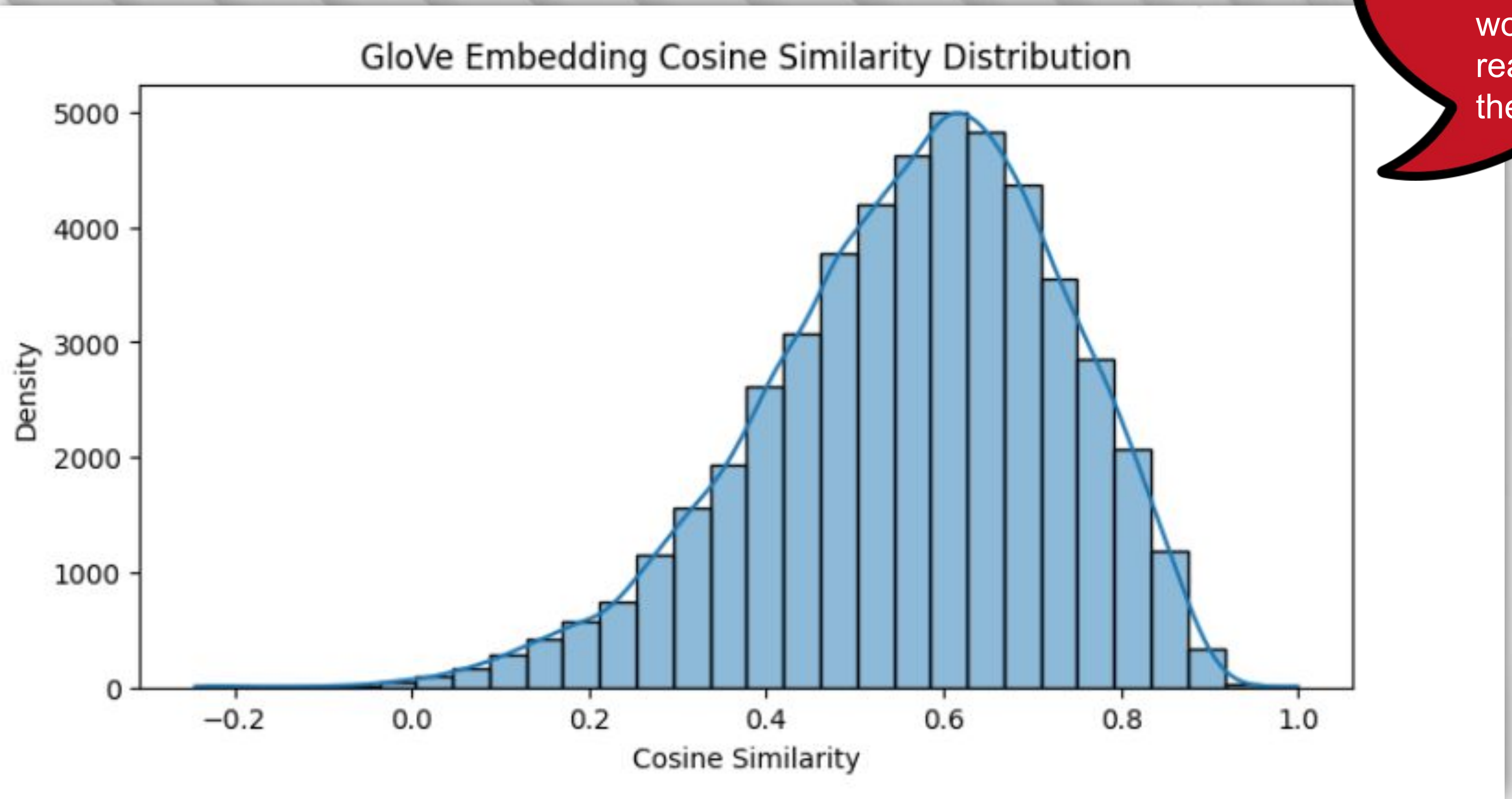
context-rich and vector-based similarity

Manual Features

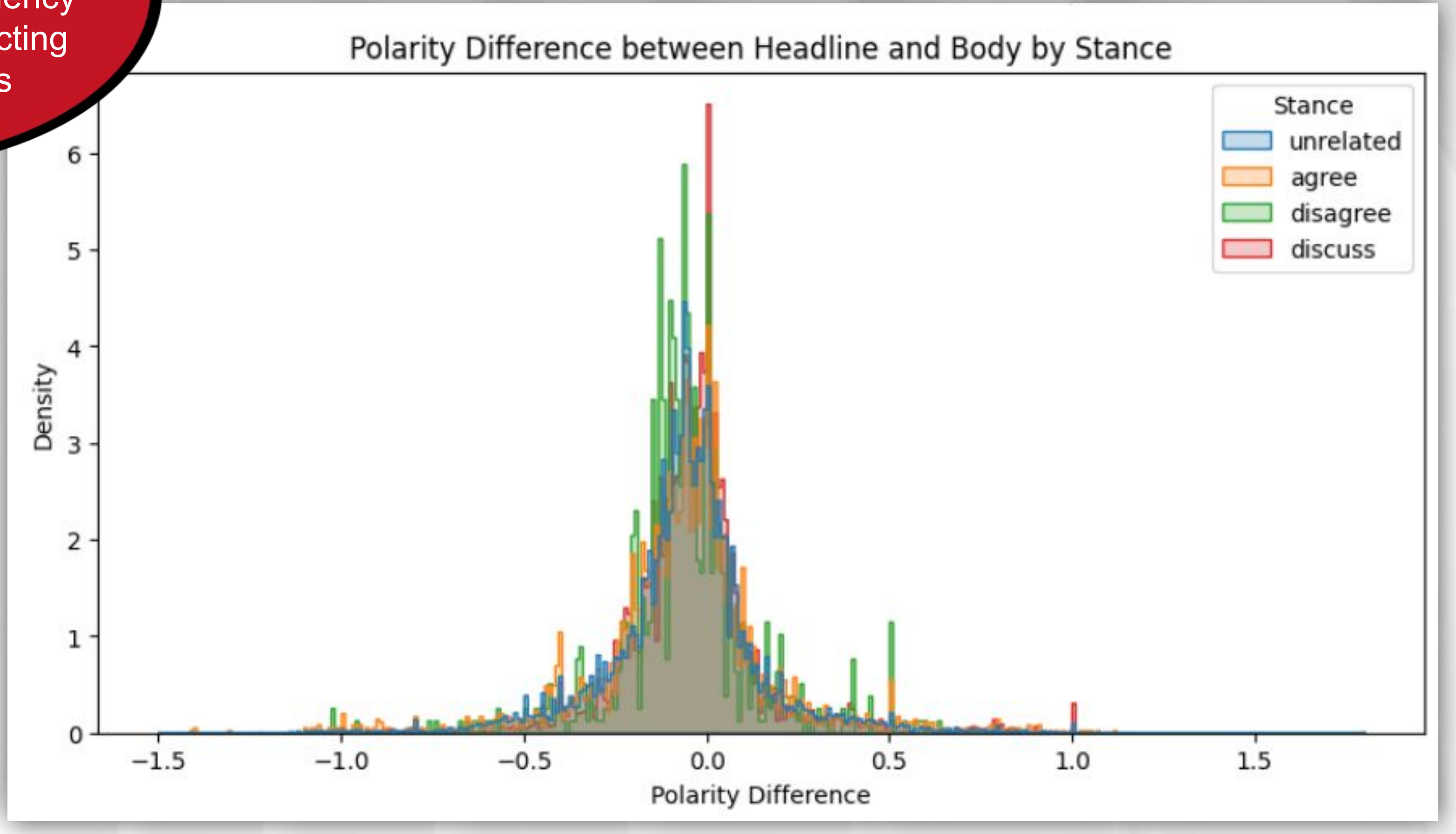
Refuting word presence (e.g., “hoax”, “fake”, “false”)

STOCKS 100% ▼ STOCKS 100% ▼ STOCKS 100% ▲ STOCKS 100% ▼ STOCKS 100% ▲

Exploratory Insights: Semantics & Sentiment Patterns in Fake News



Sensitive areas like **politics**, **international affairs**, and **health** dominate high-frequency word usage — reflecting real-world fake news themes.



- ***Cosine similarity from GloVe embeddings reveals stance alignment.***

✔ Headline-body pairs with **high cosine similarity** are more likely to be labeled *"agree"* or *"discuss"*.

✔ Pairs with **low similarity** tend to reflect *"disagree"* or *"unrelated"* stances.

✔ This confirms that **semantic closeness** is a strong predictor of stance alignment.

- ***Polarity gaps highlight emotional framing in disagreement stances***

✔ **Disagree** stances show a **wider polarity gap**, indicating emotional tension or contrast between headline and body.

✔ **Agree** and discuss stances tend to have **closer or more neutral polarity**.

Model Fitting

- Logistic Regression
- Random Forest
- XG Boost
- KNN Classifier
- Light GBM

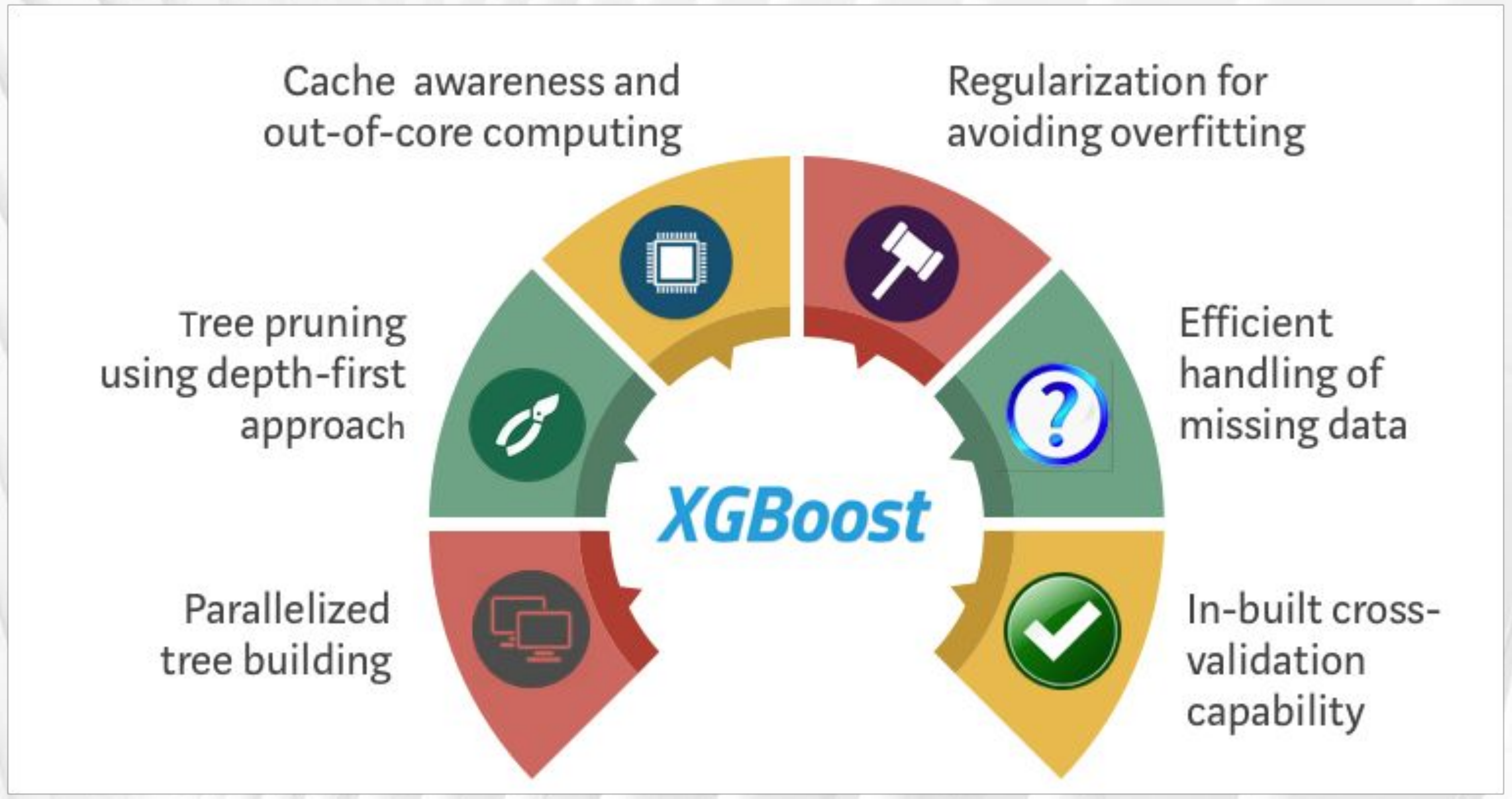
STOCKS 100% ▼ STOCKS 100% ▼ STOCKS 100% ▲ STOCKS 100% ▼ STOCKS 100% ▲

What is XG Boost?

- **XGBoost** is an optimized gradient boosting algorithm.

- It builds decision trees sequentially to improve model accuracy.

- It's fast, handles missing data, and reduces overfitting with regularization.

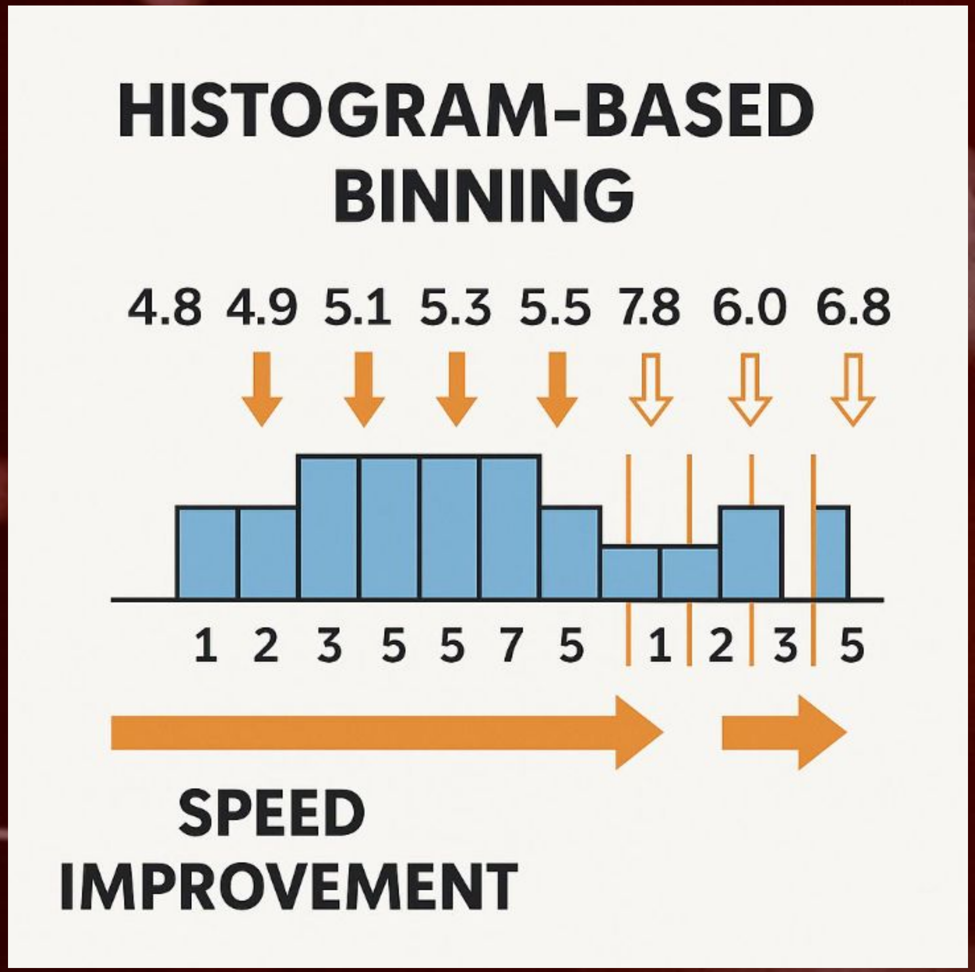


Cache awareness and out-of-core computing
Regularization for avoiding overfitting
Tree pruning using depth-first approach
Efficient handling of missing data
XGBoost
Parallelized tree building
In-built cross-validation capability

M.I.S

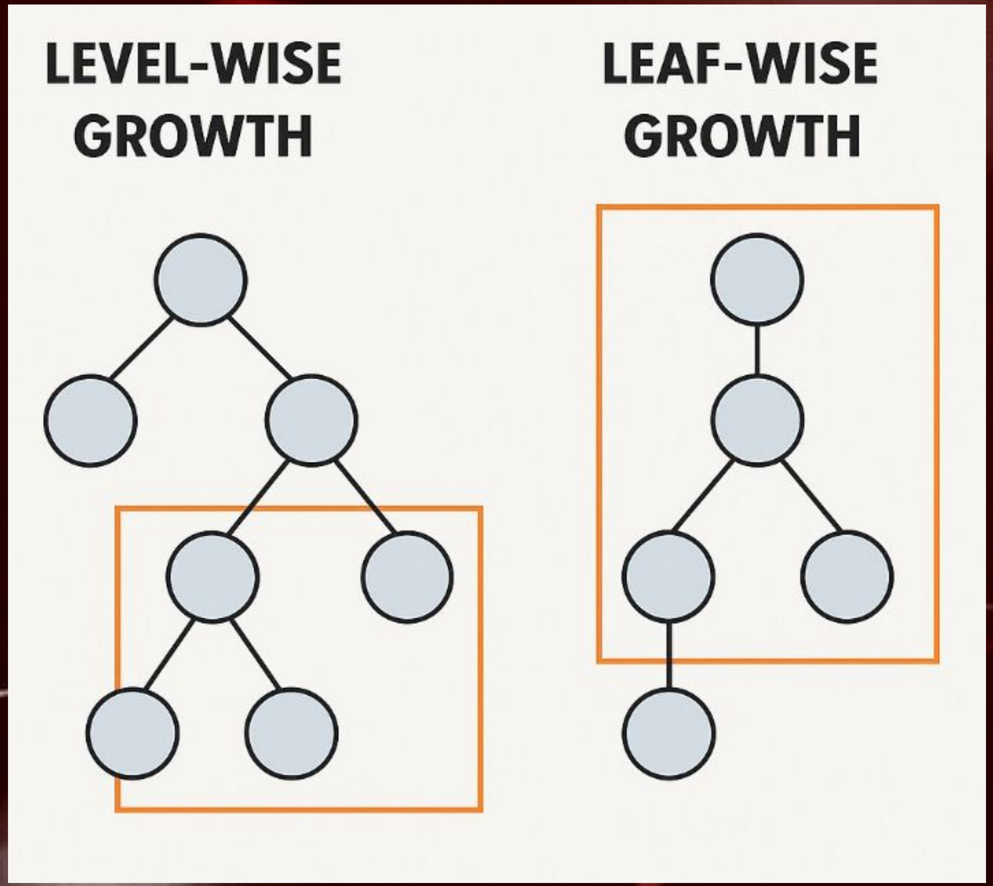
Why LightGBM?

A HIGH-PERFORMANCE IMPLEMENTATION OF GRADIENT-BOOSTED DECISION TREES (GBDT) BY MICROSOFT.
GROWS TREES LEAF-WISE (BEST-FIRST) RATHER THAN LEVEL-WISE FOR FASTER ACCURACY GAINS



HISTOGRAM-BASED BINNING

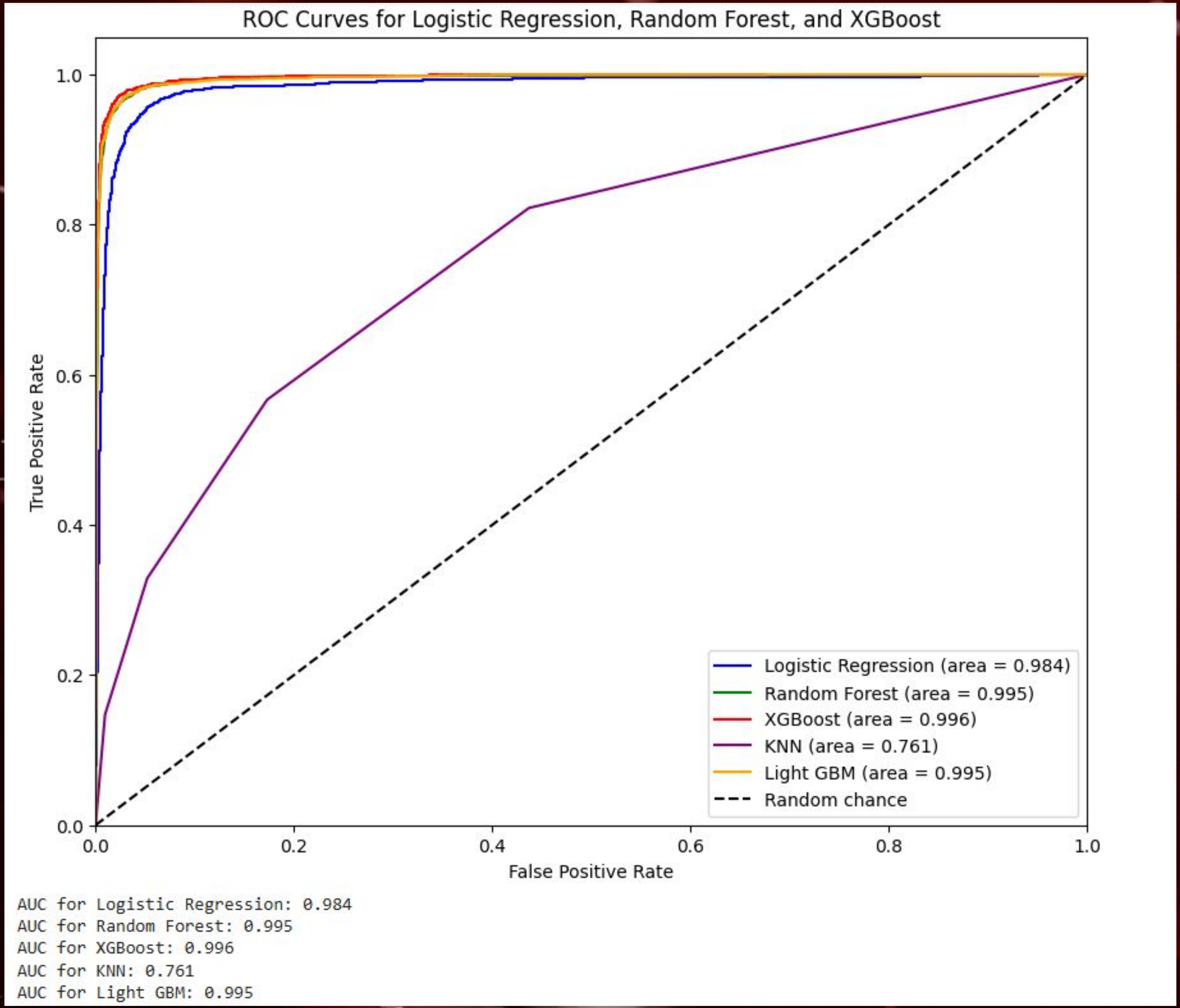
SPEED IMPROVEMENT



STOCKS 100% ▼ STOCKS 100% ▼ STOCKS 100% ▲ STOCKS 100% ▼ STOCKS 100% ▲

M.I.S

Performance Curve

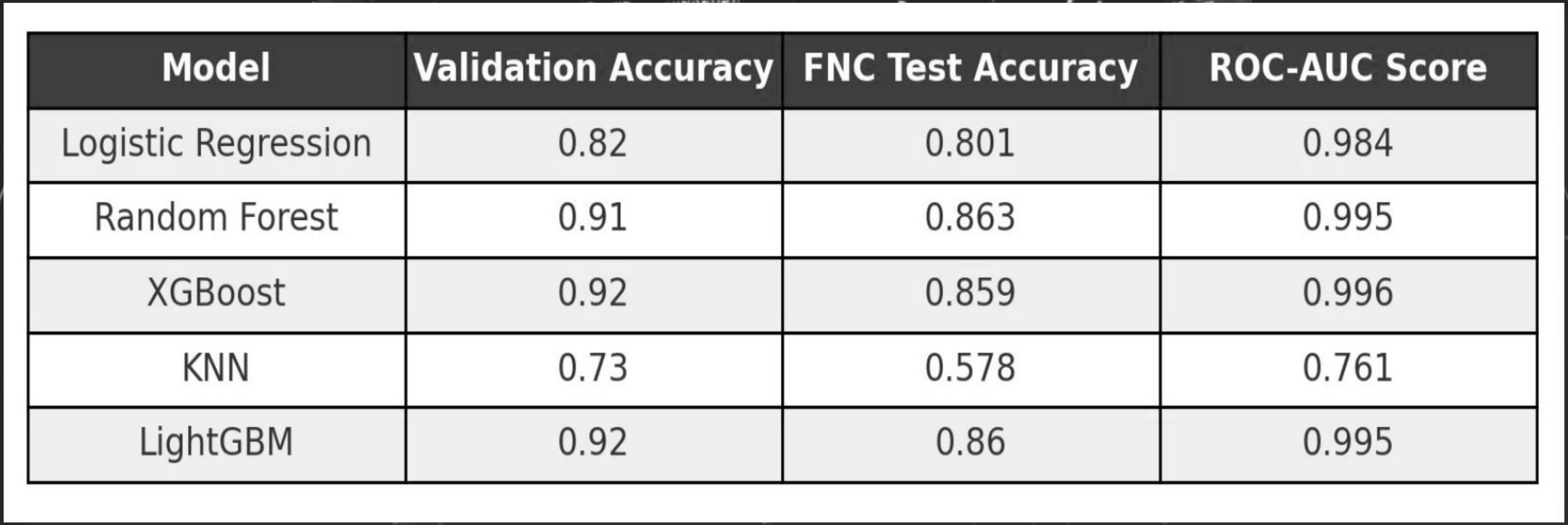


```
AUC for Logistic Regression: 0.984
AUC for Random Forest: 0.995
AUC for XGBoost: 0.996
AUC for KNN: 0.761
AUC for Light GBM: 0.995
```

STOCKS 100% ▼ STOCKS 100% ▼ STOCKS 100% ▲ STOCKS 100% ▼ STOCKS 100% ▲

Performance Metrics

LIVE ●



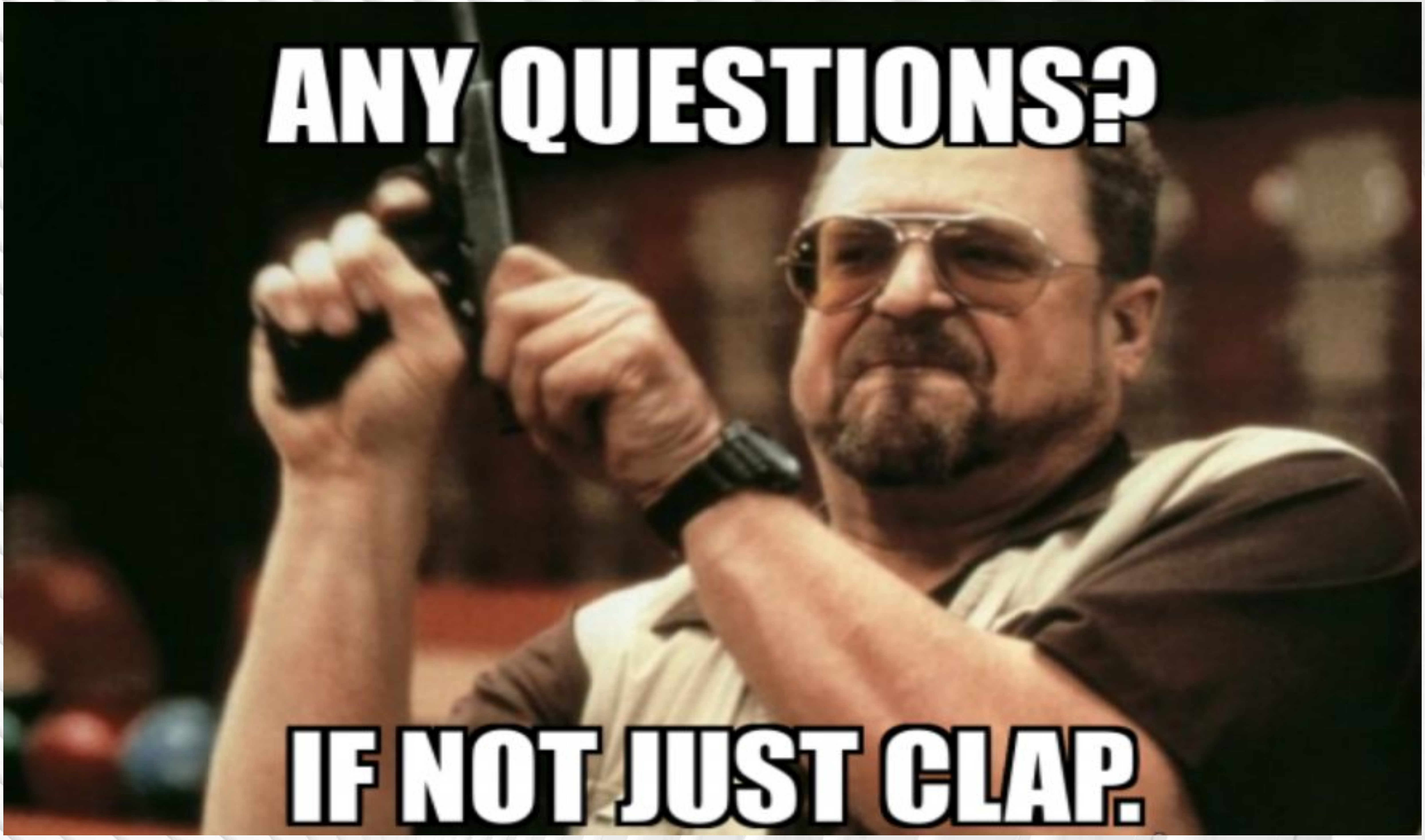
Model	Validation Accuracy	FNC Test Accuracy	ROC-AUC Score
Logistic Regression	0.82	0.801	0.984
Random Forest	0.91	0.863	0.995
XGBoost	0.92	0.859	0.996
KNN	0.73	0.578	0.761
LightGBM	0.92	0.86	0.995

STOCKS 100% ▼ STOCKS 100% ▼ STOCKS 100% ▲ STOCKS 100% ▼ STOCKS 100% ▲

Future Scope

- End-to-End Fact-Checking Integration
- Imbalanced Data Strategies
- Ensemble & Hybrid Architectures
- Richer & Multimodal Signals
- Online Learning & Drift Detection

THANK YOU!! ☺



STOCKS 100% ▼ STOCKS 100% ▼ STOCKS 100% ▲ STOCKS 100% ▼ STOCKS 100% ▲