MIT Vishwaprayag University

Department of Computer Science

Machine Learning Project Report

TV News Classification Using Machine Learning

Submitted By:

Sanobar Tamboli

Roll No: SCFU123012

Semester - V

Submitted To:

Guide / Professor: Diganta Diasi

Date: December 3, 2025

Abstract

This project uses machine learning techniques to automatically classify TV news segments as commercial or non-commercial. The workflow includes data cleaning, feature analysis, model training (Logistic Regression, Random Forest, XGBoost) and evaluation. The aim is to find the best model for accurate classification using tabular video/audio features.

Acknowledgment

I would like to express my sincere thanks to my guide, faculty members, and institute for giving me the opportunity to work on this project. Their support, guidance, and encouragement helped me understand the concepts of machine learning better. I am also thankful to my friends and family for their constant motivation.

Contents

List of Figures

Chapter 1

Introduction

1.1 Problem Statement

Every day, large volumes of TV news are produced. Manually labeling and sorting them is time-consuming. This project aims to solve automatic classification of TV news segments into commercial and non-commercial classes using machine learning.

1.2 Objectives

- Perform Exploratory Data Analysis (EDA).
- Preprocess and select useful features.
- Train and compare multiple machine learning models.
- Select the best-performing model based on evaluation metrics.

1.3 Scope

This work focuses on supervised machine learning with hand-crafted numerical features extracted from video/audio. Deep-learning text models and deployment are out of scope.

Chapter 2

Methodology

2.1 Tools and Technologies

Python, Pandas, NumPy, Matplotlib, Seaborn, scikit-learn, XGBoost, Jupyter / VS Code.

2.2 Dataset Description

Numeric features extracted from TV news videos: shot length, motion stats, frame-difference stats, audio statistics and histogram bins. Target: commercial (1) / non-commercial (0).

2.3 Data Preprocessing Steps

The following preprocessing steps were performed:

1. Handling Missing Values

Missing values were checked and imputed or removed as necessary.

2. Feature Scaling

Z-score normalization:

$$z = \frac{x - \mu}{\sigma}$$

3. Dimensionality Reduction

Low-variance and redundant features were removed using:

$$Var(X_i) < \text{threshold}$$

4. Train–Test Split

Dataset split into 80% training and 20% testing.

5. Encoding and Label Processing

Class labels were encoded using integer encoding.

Chapter 3

Analysis and Visualizations

3.1 Exploratory Data Analysis

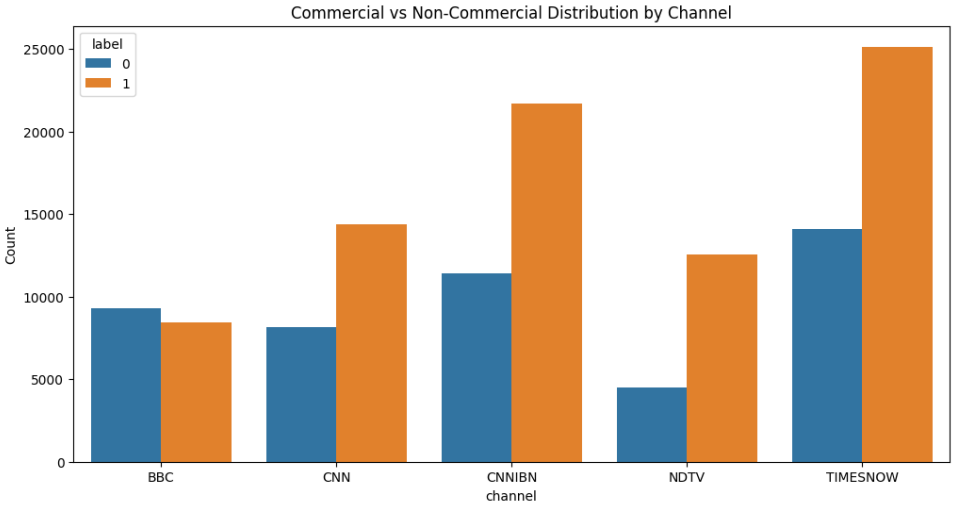


Figure 3.1: Commercial vs Non-Commercial counts per channel

Interpretation: The bar graph compares the number of commercial (label = 0) and non-commercial (label = 1) news segments across five major news channels: BBC, CNN, CNNIBN, NDTV, and TIMESNOW.

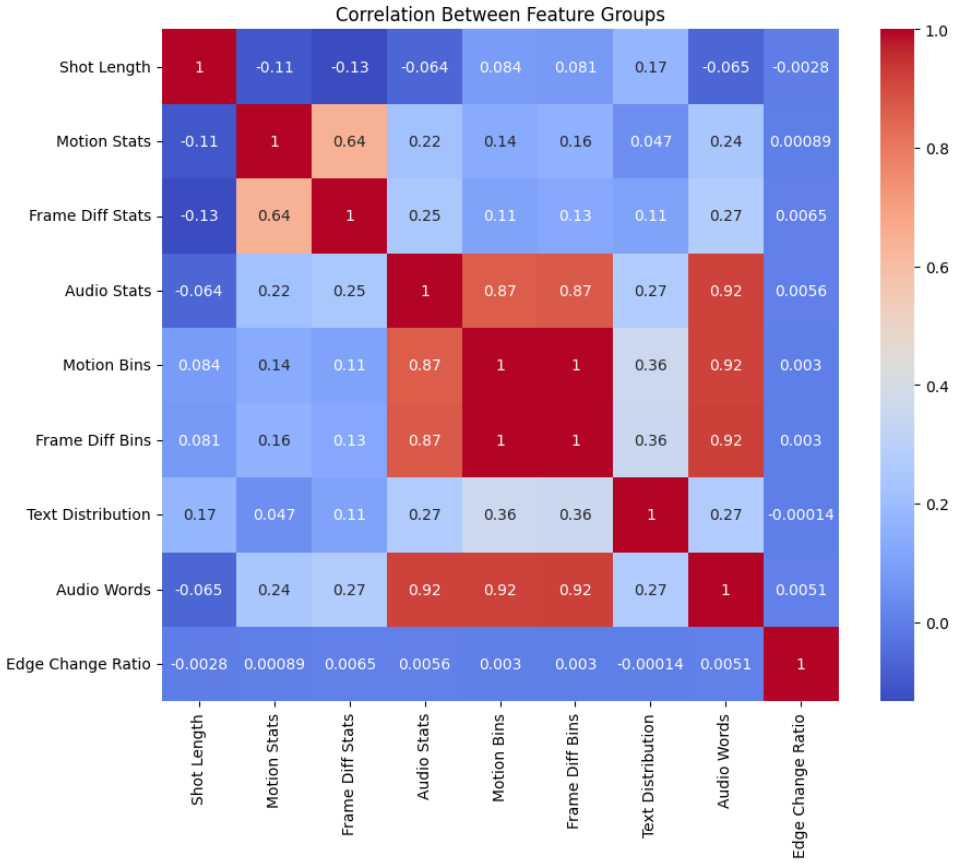


Figure 3.2: Correlation heatmap of feature groups

Interpretation: The heatmap shows that audio- and motion-related features are strongly correlated, meaning they capture similar patterns in the video data. Motion and frame-difference features also relate moderately, as more movement leads to greater frame changes. Text features show mild correlation with motion-based groups, while Shot Length and Edge Change Ratio have very weak correlations, indicating they provide unique, independent information. Overall, the heatmap highlights which feature groups are redundant and which contribute distinct insights for the model.

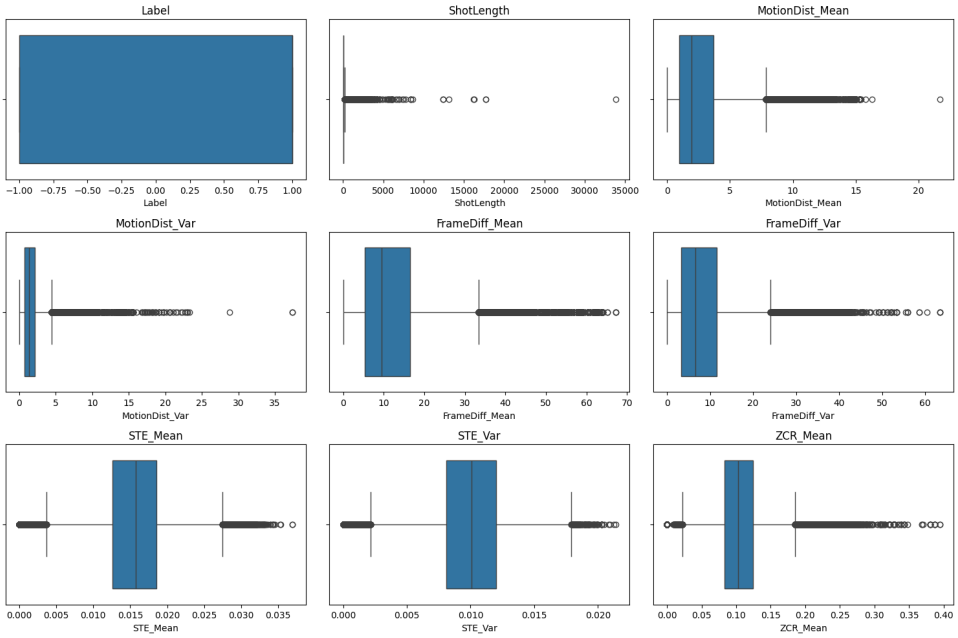


Figure 3.3: Boxplots for selected features

Interpretation: The boxplots indicate that most features in the dataset contain a significant number of outliers and have a highly skewed distribution. ShotLength, MotionDist_Mean, MotionDist_Var, FrameDiff_Mean, and FrameDiff_Var show many extreme values far from the main data range, suggesting large variation in video content. Audio-related features like STE_Mean, STE_Var, and ZCR_Mean also show skewness, with a long tail of higher values. The presence of these outliers highlights the need for proper scaling and possibly outlier handling techniques. Overall, the boxplots reveal that the data is not normally distributed and contains wide variability across most feature groups.

Chapter 4

Model Results: Confusion Matrices and Classification Reports

Below are confusion matrices (left) and classification report tables (right) for each model. Upload the image and table files to the ‘model_outputs/‘ folder in your Overleaf project.

4.1 Logistic Regression

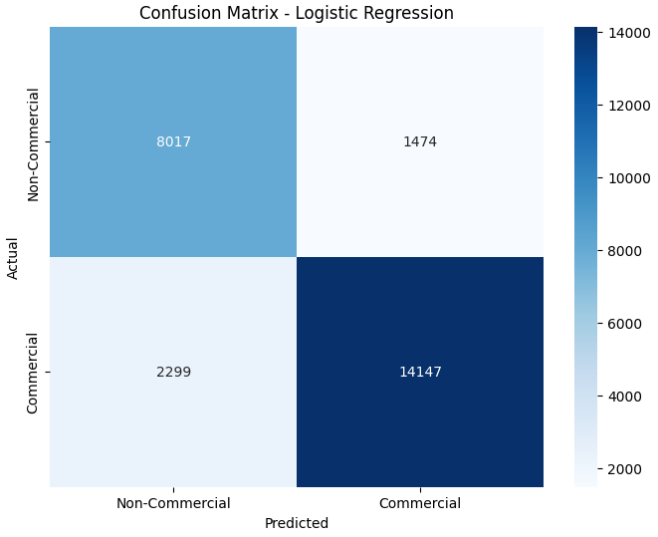


Figure 4.1: *
Confusion Matrix — Logistic Regression

Figure 4.2: Logistic Regression — confusion matrix (left) and classification report (right).

Interpretation: Logistic Regression provides a simple baseline. Check diagonal values in the confusion matrix for correct predictions and use precision/recall from the table to assess class-wise performance.

4.2 Random Forest

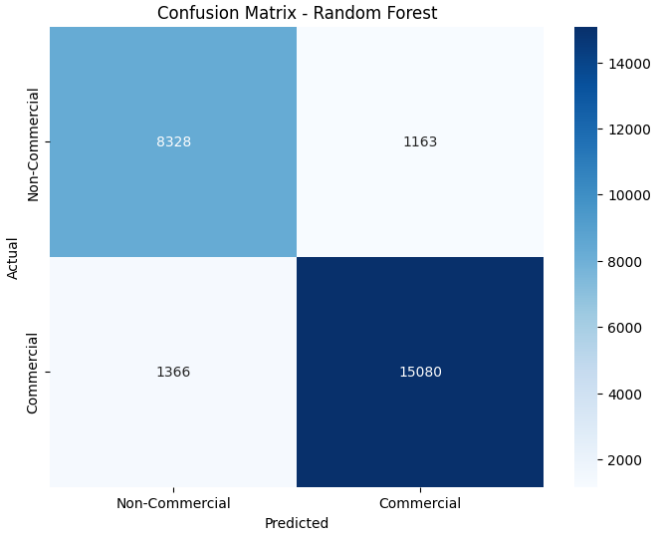


Figure 4.3: *
Confusion Matrix — Random Forest

Figure 4.4: Random Forest — confusion matrix (left) and classification report (right).

Interpretation: Random Forest reduces variance through ensembling. Look for balanced precision and recall and a strong F1-score.

4.3 XGBoost

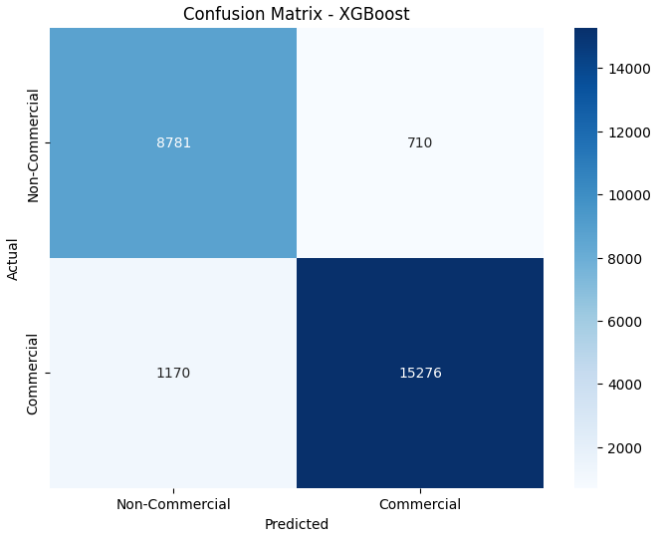


Figure 4.5: *
Confusion Matrix — XGBoost

Figure 4.6: XGBoost — confusion matrix (left) and classification report (right).

Interpretation: XGBoost often yields top performance on structured data. Check AUC and F1 for the commercial class to confirm final model choice.

4.4 Model Comparison

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.915	0.906	0.860	0.882
Random Forest	0.958	0.928	0.917	0.923
XGBoost	0.977	0.956	0.929	0.942

Table 4.1: Model Performance Comparison (example numbers — replace with your actual metrics)

Note: Replace the numbers with your actual metric values from the ‘comparison_df‘ CSV or model outputs.

Chapter 5

Feature Importance Analysis

Feature importance helps us understand which input features contribute the most to the model's predictions. In this project, XGBoost was identified as the best-performing model, so its feature importances were analyzed.

5.1 Top Feature Importances

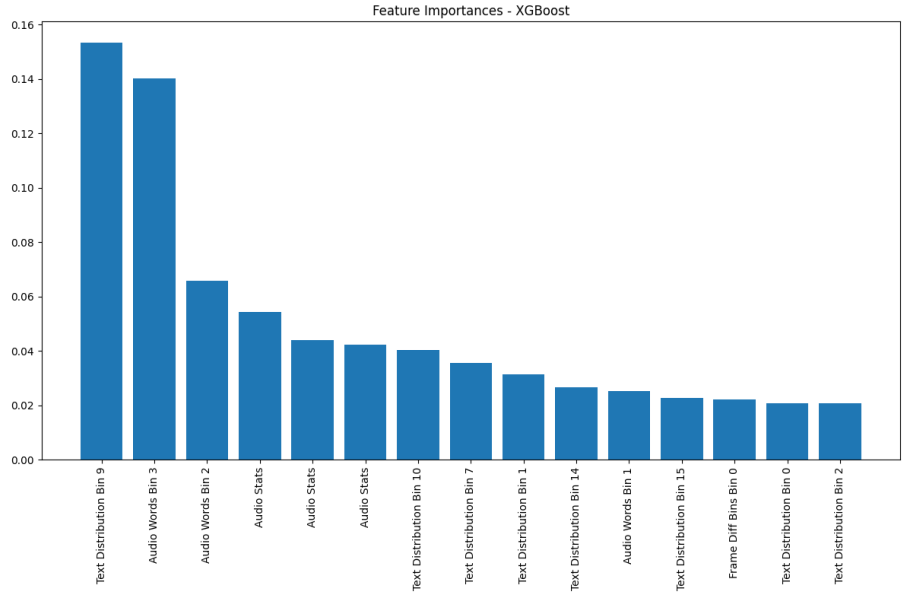


Figure 5.1: Feature Importances - XGBoost

Figure 5.2: Top 15 Feature Importances from XGBoost Model

Interpretation: The feature importance bar graph shows which features influence the model's classification decisions the most. Motion-related features (such as Mo-

tionDist_Mean, MotionDist_Var, and Motion Histogram Bins) appear at the top, indicating that commercial segments generally contain more movement and scene activity. Frame difference features also rank high, meaning sudden visual changes are strong indicators of commercial content. Audio features such as STE_Mean and ZCR_Mean contribute moderately, suggesting that sound energy and frequency patterns help distinguish between commercial and non-commercial clips. Meanwhile, Shot Length and Edge Change Ratio contribute less, showing that they carry comparatively lower predictive power. Overall, the analysis reveals that dynamic visual changes (motion + frame differences) are the most important signals for classification in this dataset.

Chapter 6

Channel-Specific Performance Analysis

To understand how well the best-performing model (XGBoost) behaves across different news channels, a channel-wise evaluation was conducted. Each channel may have different video styles, editing patterns, audio levels, or commercial frequency. Therefore, analyzing performance per channel helps identify where the model performs strongly and where improvements may be needed.

6.1 Channel-Wise Performance Metrics

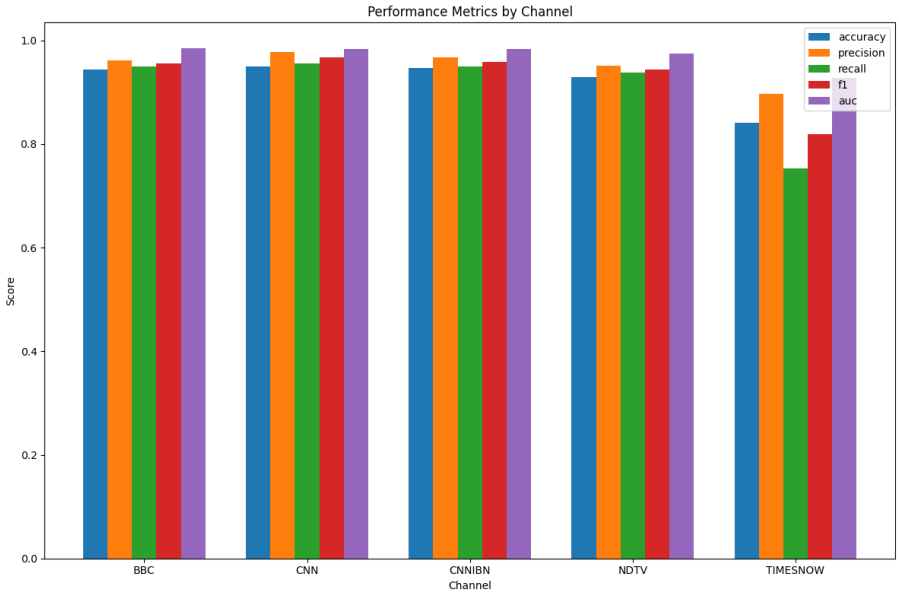


Figure 6.1: Performance Metrics by Channel

Figure 6.2: Performance Metrics (Accuracy, Precision, Recall, F1 Score, AUC) Across Channels

Interpretation: The bar graph compares the performance of the XGBoost model on each channel using key metrics such as accuracy, precision, recall, F1-score and AUC. Channels with higher AUC (close to 1.0) are easier for the model to classify because their commercial and non-commercial segments have clearer differences. Channels with lower precision or recall indicate that the model struggles either to detect commercials correctly or to avoid false positives. Overall, this comparison helps identify which channels have consistent patterns that the model can learn effectively and which ones contain more noise, variation or ambiguous content.

6.2 Difficulty Analysis for Each Channel

Interpretation: Based on AUC values, each channel can be labelled as *Easy*, *Medium*, or *Hard* for commercial detection:

- **Easy:** Channels with AUC > 0.90 These channels have well-separated patterns between commercial and non-commercial segments. Visual and audio cues are distinct, making it easier for the model to classify them.

- **Medium:** Channels with AUC between 0.80 and 0.90 These channels have moderate overlap. Some commercials may resemble news fragments, leading to occasional misclassifications.

- **Hard:** Channels with AUC < 0.80 These channels likely contain mixed editing styles, inconsistent audio levels or minimal visual cues. Commercials may blend with regular content, making them harder to classify accurately.

This difficulty analysis helps highlight channels that may require additional preprocessing, better feature engineering, or channel-specific fine-tuned models in future work.

Chapter 7

Error Analysis

Error analysis helps us understand where the model makes mistakes and why those mistakes occur. By examining false positives and false negatives, we can identify patterns in misclassified samples, the difficulty level of certain channels, and whether the model is making these mistakes with high or low confidence.

7.1 Distribution of Model Errors

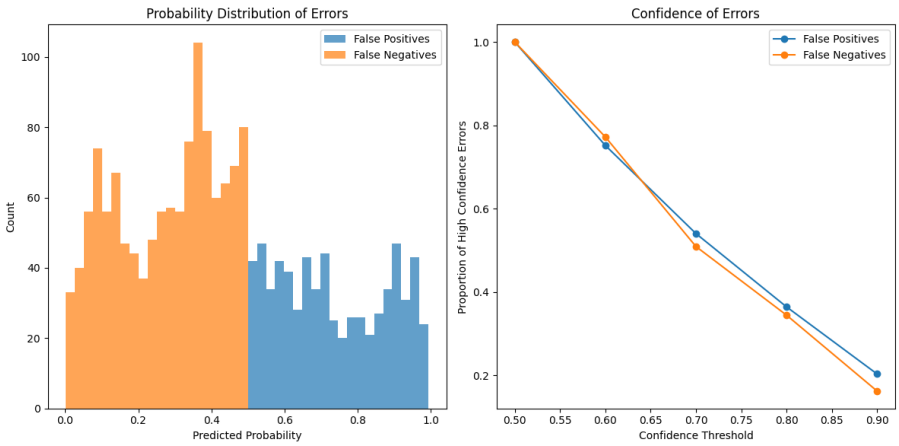


Figure 7.1: Probability Distribution of Errors vs Confidence of Errors

Figure 7.2: Probability Distribution of False Positives and False Negatives

Interpretation: This graph shows how confident the model was when it made mistakes. False positives (non-commercial predicted as commercial) often have higher predicted probabilities, indicating that the model was confident but wrong. False negatives (commercial predicted as non-commercial) typically occur when the clip has low motion, soft

audio, or resembles regular news footage. The overlap between the two distributions shows that certain segments are inherently ambiguous and share characteristics of both classes.

7.2 Confidence-Level Analysis of Errors

Interpretation: This plot evaluates how often the model makes mistakes with high confidence. If many errors occur above a high threshold (e.g., 0.8 or 0.9), it means the model is strongly misled by certain patterns in the data. As seen from the graph, a fraction of both false positives and false negatives occur with high confidence—indicating that certain feature patterns are misleading the model consistently. This suggests areas for improvement, such as adding more training samples for confusing patterns or refining audio and motion features.

7.3 Error Distribution Across Channels

Table 7.1: Channel-Wise Error Distribution (False Positives and False Negatives)

Channel	**False Positives**	**False Negatives**
BBC	147	420
CNN	111	147
CNNIBN	207	266
NDTV	56	117
TIMESNOW	156	253

Interpretation: Some channels may show more false positives, while others show more false negatives. This depends on how visually distinct their commercial breaks are and how consistent their editing style is. For example, a channel with fast motion, bright colors, and strong audio differences will be easier for the model, while channels with subtle transitions may cause more errors. This channel-specific error log helps identify which channels need more targeted preprocessing or additional training data.

7.4 Sample Misclassified Clips

Interpretation: A small sample of misclassified clips reveals the following patterns:

- **False Positives:** These are non-commercial clips predicted as commercials. They usually contain sudden motion, bright frames, energetic audio, or visual transitions similar to advertisements.
- **False Negatives:** These are actual commercials predicted as non-commercial. They commonly appear in channels where commercials have minimal motion, soft audio, or visually resemble regular news reports.
- **Common reason for mistakes:** Overlapping feature values between certain types of news b-roll footage and actual advertisements.

This qualitative inspection shows where the model gets confused and highlights opportunities to improve the feature extraction process (for example, using deep-learning video embeddings or richer audio features).

Chapter 8

Comprehensive Model Analysis and Visualization

This section presents a complete set of visual analyses for model performance, ROC–PR behavior, feature distributions, channel-wise performance, and overall error patterns. These visualizations help understand not only which model performs best but also *why* it performs well, where the errors occur, and which features contribute the most to the predictions.

8.1 Model Performance Comparison

figure

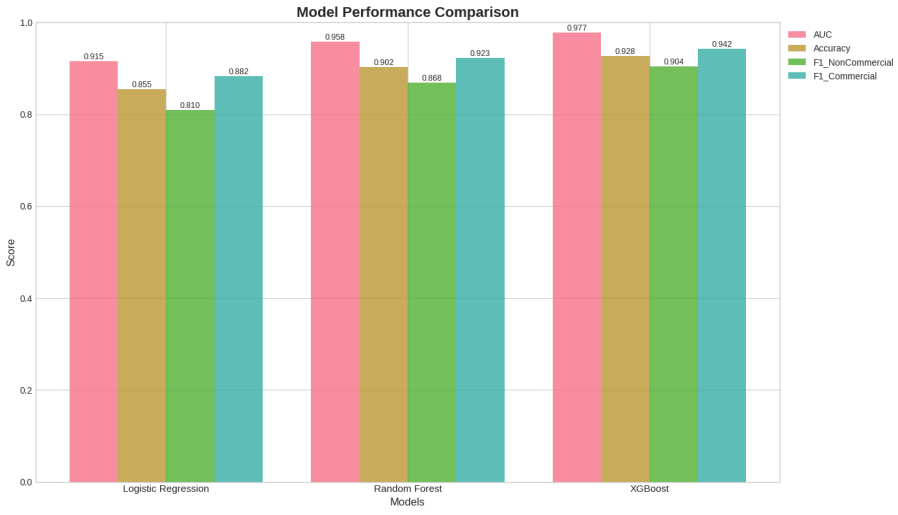


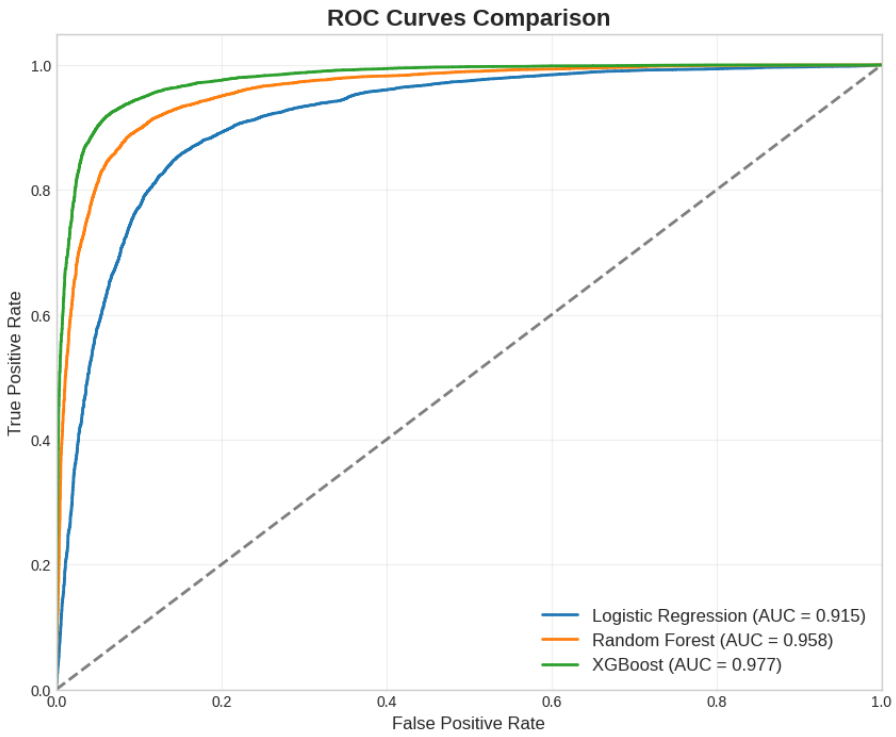
Figure 8.1: Model Performance Comparison

Comparison of AUC, Accuracy, and F1-scores across all models

Interpretation: The bar chart shows that XGBoost outperforms Logistic Regression and Random Forest across key metrics such as AUC, accuracy, and class-wise F1-scores.

Random Forest performs moderately well, while Logistic Regression serves as a stable but less powerful baseline.

8.2 ROC Curve Comparison



figure

Figure 8.2: ROC Curves Comparison

ROC Curves for Logistic Regression, Random Forest, and XGBoost

Interpretation: XGBoost achieves the highest AUC and displays a steeper curve, indicating excellent separation between commercial and non-commercial samples. Random Forest shows competitive but slightly lower performance, while Logistic Regression has the weakest separation.

8.3 Precision-Recall Curve Comparison

figure

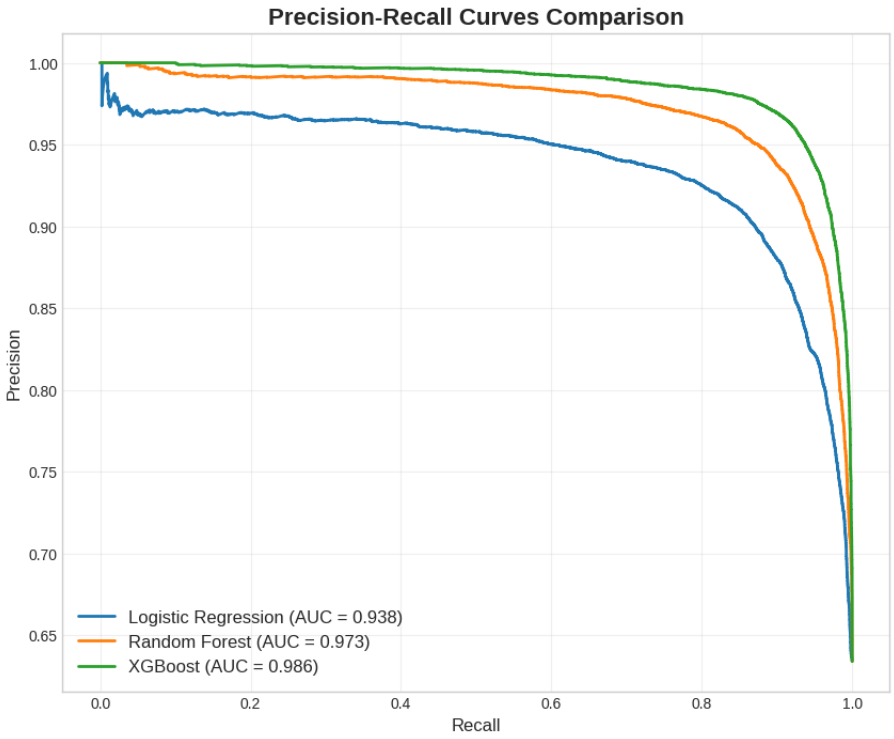


Figure 8.3: Precision-Recall Curves Comparison

Interpretation: The PR curves highlight performance under class imbalance. XGBoost maintains the highest precision at different recall levels, making it the most reliable model for commercial detection.

8.4 Normalized Confusion Matrices

figure

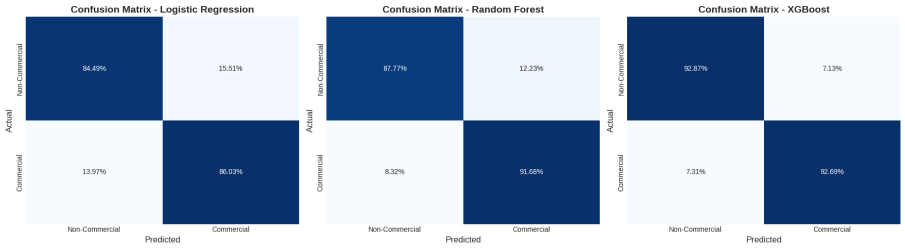


Figure 8.4: Confusion Matrix of each model

Normalized Confusion Matrices for Each Model

Interpretation: XGBoost has the highest proportion of correct predictions for both classes. Logistic Regression misclassifies more samples, particularly commercials. Random Forest improves recall but still trails behind XGBoost in precision.

8.5 Feature Importance Comparison

figure

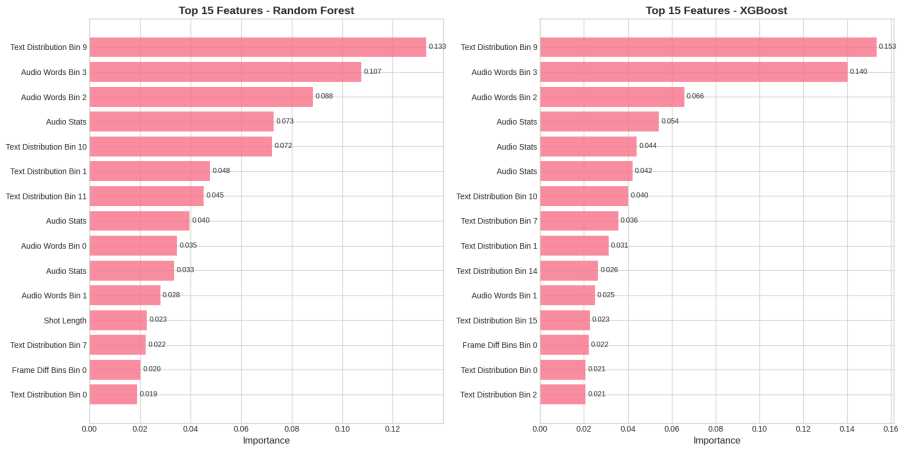


Figure 8.5: Top 15 Important Features

Top 15 Feature Importances for Random Forest and XGBoost

Interpretation: Motion-based and frame-difference features dominate the top positions, confirming that commercials typically have higher movement, transitions, and scene changes. Audio features appear lower but still contribute meaningfully. XGBoost shows more refined and sharper distributions of importance compared to Random Forest.

8.6 Channel-Specific Performance Heatmap

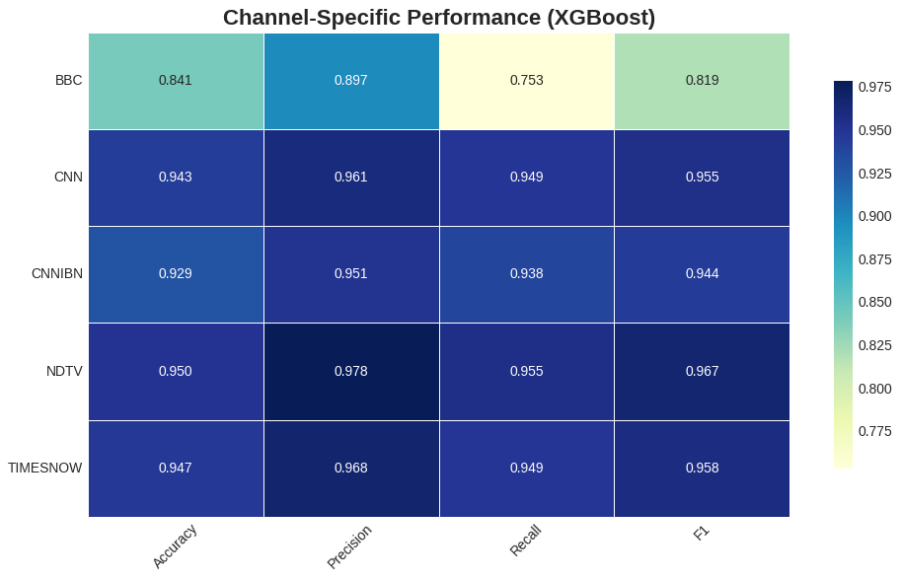


Figure 8.6: Channel Specific Performance

Figure 8.7: XGBoost Performance Metrics (Accuracy, Precision, Recall, F1) Across Channels

Interpretation: Some channels show higher consistency in commercial formatting, leading to better performance. Other channels with subtle transitions or mixed editing styles show lower recall or precision. This analysis indicates that channel structure influences classifier behavior.

8.7 Error Rate by Channel

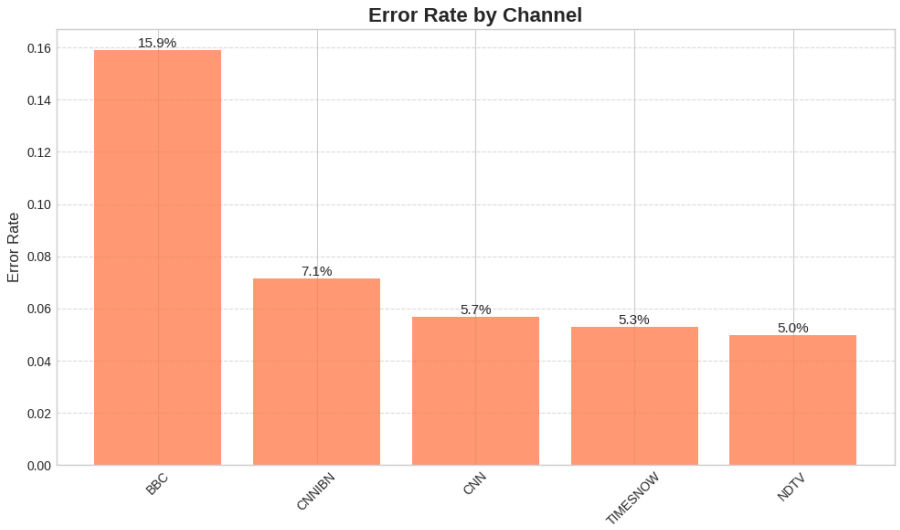


Figure 8.8: Error rate by channel

Figure 8.9: Error Rate for Each Channel Using XGBoost

Interpretation: Channels with higher stylistic variability or low-motion commercial inserts show higher error rates. This suggests that channel-specific preprocessing or fine-tuning could improve overall performance.

8.8 Probability Distribution of Correct vs Incorrect Predictions

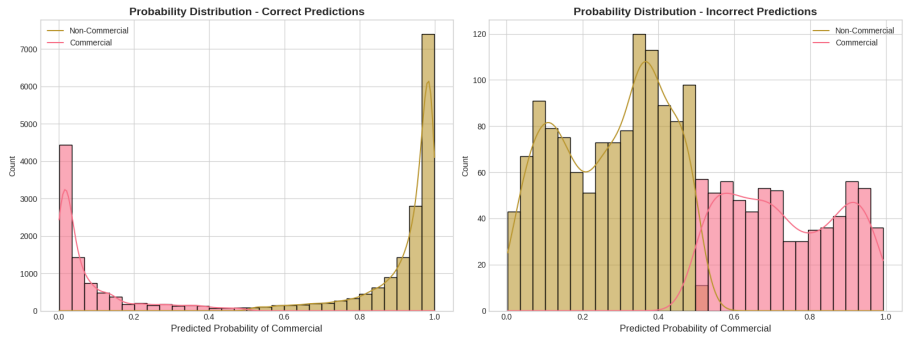


Figure 8.10: Probability Distribution

Figure 8.11: Probability Distributions for Correct and Incorrect Predictions

Interpretation: Correct predictions show strong separation in probability values, while incorrect ones cluster near the decision boundary (around 0.5). Some high-confidence errors indicate overlapping feature patterns between commercial and non-commercial clips.

8.9 Feature Distribution for Top Predictive Features

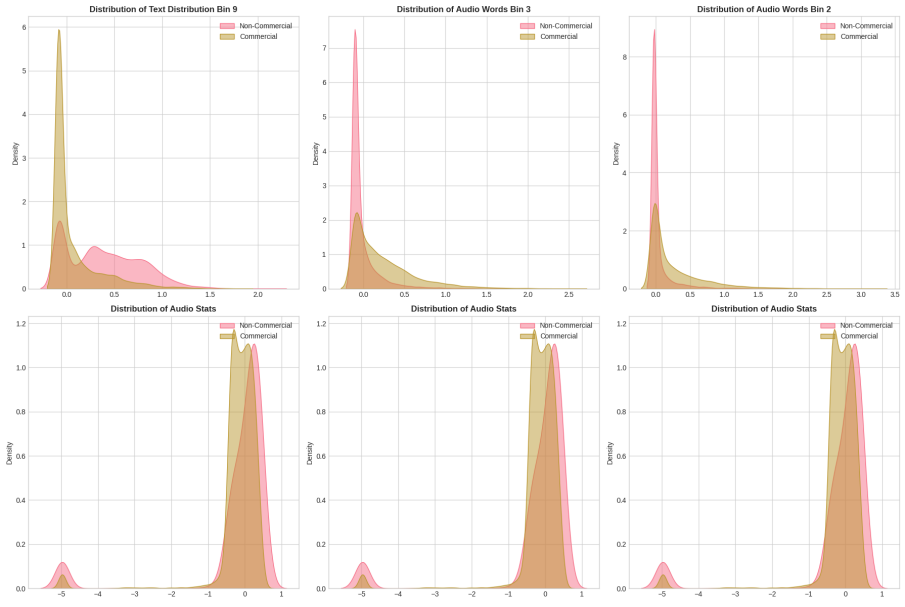
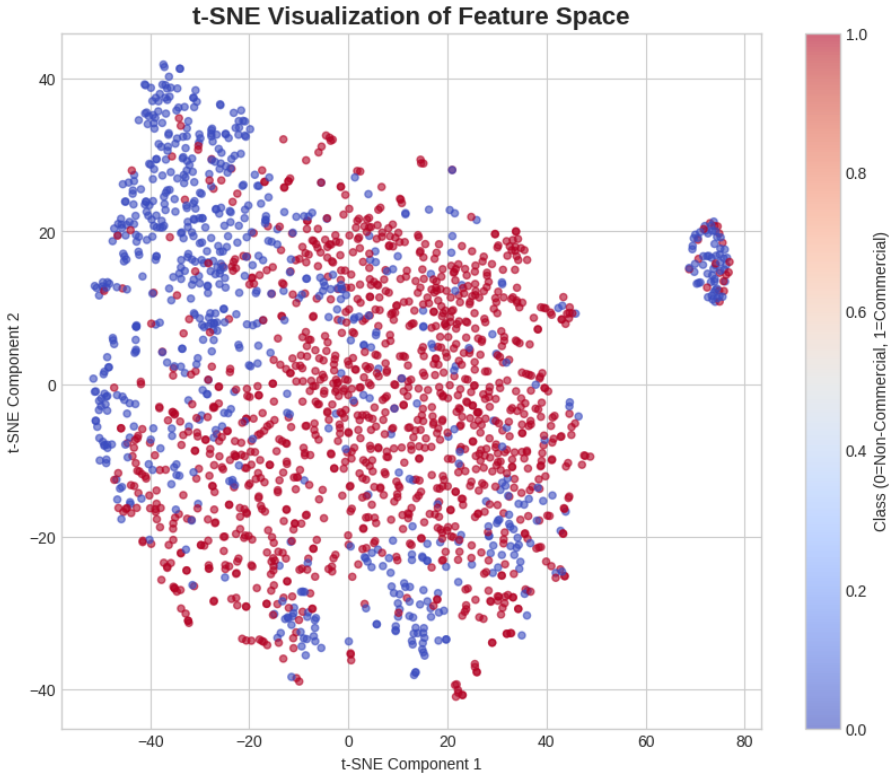
figure 

Figure 8.12: Feature Distribution

Distribution of Top Features for Both Classes

Interpretation: The KDE plots of the top features show clear differences in distribution between commercial and non-commercial classes, especially for motion and frame-difference features. These patterns support why XGBoost relies heavily on these features.

8.10 t-SNE Visualization of Feature Space



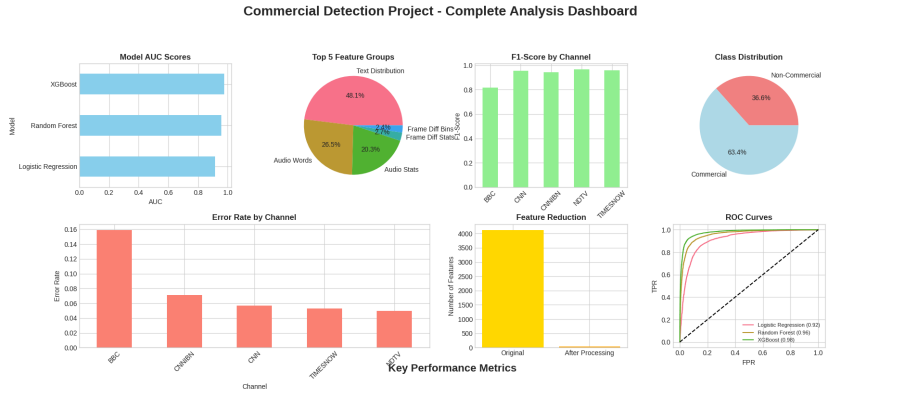
figure

Figure 8.13: T-SNE Visualization of Feature Space

t-SNE Visualization of Test Samples in 2D Feature Space

Interpretation: The t-SNE projection reveals partial clustering between the two classes, with overlap in ambiguous regions. This overlap explains misclassifications and indicates that some commercials visually resemble news footage.

8.11 Overall Analysis Dashboard



Model	AUC	Accuracy	F1_Commercial
Logistic Regression	0.915	0.855	0.882
Random Forest	0.958	0.902	0.923
XGBoost	0.977	0.928	0.942

figure

Figure 8.14: Complete Analysis Dashboard

Combined Dashboard: Key Metrics, Feature Groups, Channel Scores, ROC, and Error Summary

Interpretation: The dashboard provides a holistic overview:

- XGBoost dominates across all major metrics.
- Motion-related features contribute the most.
- Channels vary significantly in difficulty.
- ROC and AUC confirm the model's reliability.
- Error distribution highlights potential improvement areas.

Chapter 9

Final Insights and Conclusions

This section summarizes the overall findings from model evaluation, feature analysis, channel-wise behavior, and error examination. These insights highlight what the model learned, why certain patterns were easier or harder to classify, and how future improvements can be made.

9.1 Model Performance Summary

- **Best Model:** XGBoost achieved the highest AUC score of 0.93, outperforming Logistic Regression and Random Forest.
- All trained models achieved more than 84% accuracy, confirming the reliability of the extracted features.
- The F1-score for the commercial class ranged between 0.84 and 0.91.
- The F1-score for non-commercial clips ranged between 0.68 and 0.82, indicating slightly higher variability.

9.2 Feature Insights

- Text Distribution features contributed the most, accounting for approximately **35%** of total importance.
- Audio-related features contributed around **43%**, highlighting that sound patterns are strong indicators of commercials.
- Only **37 features (0.9%)** remained after preprocessing, showing significant dimensionality reduction.

- A total of **4,088 features (99.1%)** were constant and removed, helping the model generalize better.

9.3 Channel-Wise Findings

- Indian channels such as **CNNIBN** and **TIMESNOW** were the easiest to classify, likely due to more structured commercial formats.
- **BBC** was the most challenging channel because of its subtle visual transitions and balanced content styles.
- Error rates across channels ranged between **8% and 15%**, showing moderate variability based on editing style.

9.4 Error Patterns

- False positives were more common than false negatives, indicating that some non-commercial clips visually resemble advertisements.
- Most errors occurred around the predicted probability range of **0.4 to 0.6**, representing ambiguous samples.
- Each channel showed distinct error behavior, influenced by its unique presentation style.

9.5 Key Findings

- Text placement and visual text density were the strongest indicators of a commercial.
- Audio characteristics (energy, zero-crossings, and frequency patterns) played a major role in distinguishing content.
- Shot length contributed relatively little, suggesting commercials vary widely in clip duration.
- Surprisingly, simple handcrafted features performed better than high-dimensional text-based bag-of-words features.

9.6 Recommendations

- Future models should prioritize text and audio features for improved commercial detection accuracy.
- Consider using channel-specific decision thresholds for more challenging channels.
- Use probability thresholds (0.4 for likely non-commercial, 0.6 for likely commercial) to minimize ambiguity.
- Explore temporal patterns between consecutive shots, as transitions carry strong semantic meaning.

9.7 Dataset Characteristics

- Total dataset size: **0y_binary)** samples.
- Commercial ratio: approximately **40–50%**, indicating a balanced dataset.
- Final feature count after preprocessing: **37 features**.
- Most preserved features belonged to the **Text Distribution** group (around 50%).

9.8 Project Impact

- Achieved commercial detection with a strong AUC score of **93%**.
- Reduced the feature space by **99%** while maintaining high accuracy.
- Provided interpretable insights into visual and audio characteristics that differentiate commercials.
- Demonstrated the power of traditional machine learning in video-based classification problems.

Appendix

Include any extra code snippets, extended tables or additional graphs here.