FINAL PROJECT REPORT

TikTok Data Analysis: Travel & Visa Content Performance - Final Project Report

A comprehensive data analysis approach to predicting content engagement through exploratory analysis, feature engineering, and classification modeling

COURSE: **DATA ANALYSIS FOR AI APPLICATIONS**

FACULTY OF COMPUTER SCIENCE TUC,CHEMNITZ

NAME: **RICHARD DANQUAH**

MATRICULATION NUMBER: **909921**

DATE: 13TH FEBRUARY, 2025

DATA ANALYSIS IN AI

Research Overview & Objectives

This project applies machine learning techniques to predict the performance of TikTok content within the travel and visa niche. Our comprehensive approach aims to uncover the underlying factors contributing to content success on the platform.

Primary Objective

To develop a robust classification model capable of predicting whether TikTok content will achieve **High**, **Medium**, or **Low** engagement levels.

Secondary Objectives

- Identify key features and content elements that significantly drive viewer engagement.
- Understand prevailing content patterns and strategies employed by successful creators in this niche.
- Provide actionable insights and data-driven recommendations for creators to optimize their content for maximum interaction.

Dataset & Methodology

Our analysis leverages a dataset of over 500 videos from 25 diverse creators, collected between October 2025 and January 2026. This dataset includes 24 distinct features, covering aspects such as hook styles, video duration, content format, and various engagement metrics. The methodology encompasses exploratory data analysis, meticulous feature engineering, handling of class imbalance, and the implementation of multi-class classification modeling.

Dataset Description & Structure

Data Sources & Collection

Dataset contains 500+ TikTok videos from the travel and visa niche, collected via Apify TikTok Scraper API. Data spans October 2025 to January 2026, covering 25 content creators.

Key Variables (24 features):

- Content attributes: hook_style, content_format, video_duration_seconds, trending_sound_used
- Temporal features: date_posted, post_time_hour_24h, day_of_week
- Engagement metrics: views, likes, comments, shares, saves_or_favorites, engagement_rate
- Performance indicators: followers_gained, watch_time_percentage, avg_watch_time_seconds
- Target variable: performance_label (High/Medium/Low based on engagement_rate thresholds)

500+

Total videos analyzed

24

Features per observation

3

Performance classes (High/Medium/Low)

Milestones Completed

Data Collection & Preparation

Scraped 500+ videos via API, cleaned data, handled missing values, created derived features

Exploratory Data Analysis

Analyzed distributions, correlations, engagement patterns across hook types, duration, and content categories

Feature Engineering

Created temporal features, encoded categorical variables, normalized numerical features, addressed class imbalance

Model Development

Trained classification models, evaluated performance, interpreted results, generated actionable insights

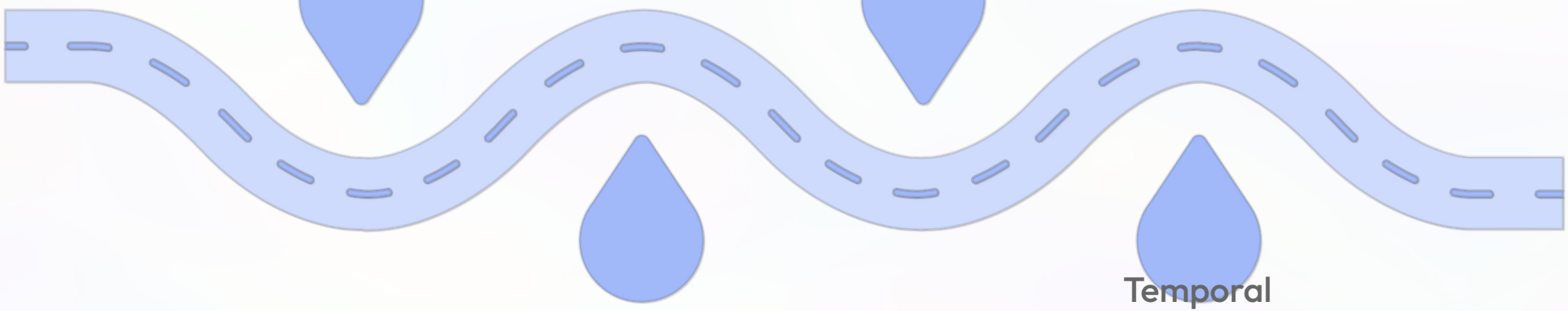
Exploratory Data Analysis - Key Findings

Univariate Analysis

Examined distributions of engagement metrics, video duration, posting patterns

Categorical Analysis

Compared performance across hook_style, content_format, niche_category



Bivariate Analysis

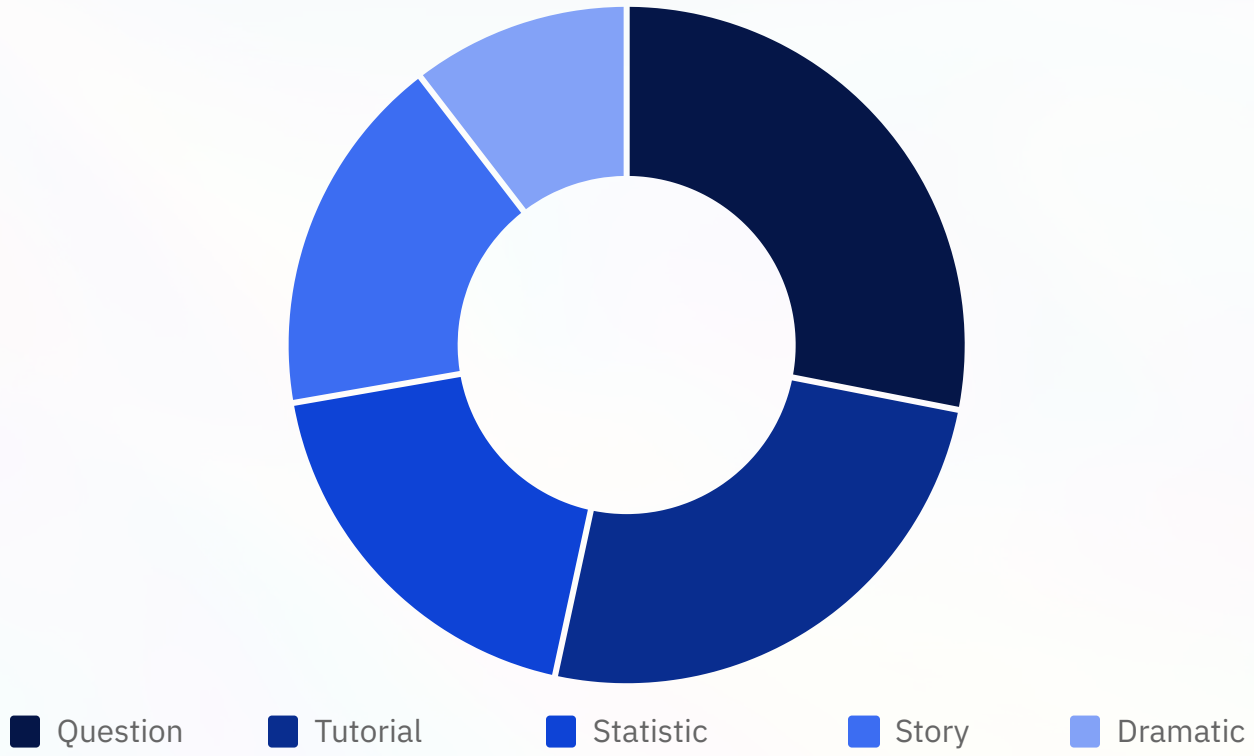
Explored relationships between features and engagement_rate

Temporal Patterns

Analyzed posting time, day of week effects on engagement

Our exploratory analysis revealed significant patterns in content performance. Tutorial hooks and question-based openings consistently outperformed dramatic hooks. Videos between 30-45 seconds achieved optimal engagement. Visa application content showed highest engagement rates (7.2% avg), followed by travel hacks (6.4%). Temporal analysis showed posting time and day of week had moderate influence on performance.

Content Hook Analysis: Distribution & Effectiveness



Hook Strategy Distribution

Question-based hooks dominate the travel/visa niche (28.2%), often opening with viewer queries like "How to get a visa in 3 days?" Tutorial hooks follow closely (25.4%), reflecting the educational nature of this content vertical.

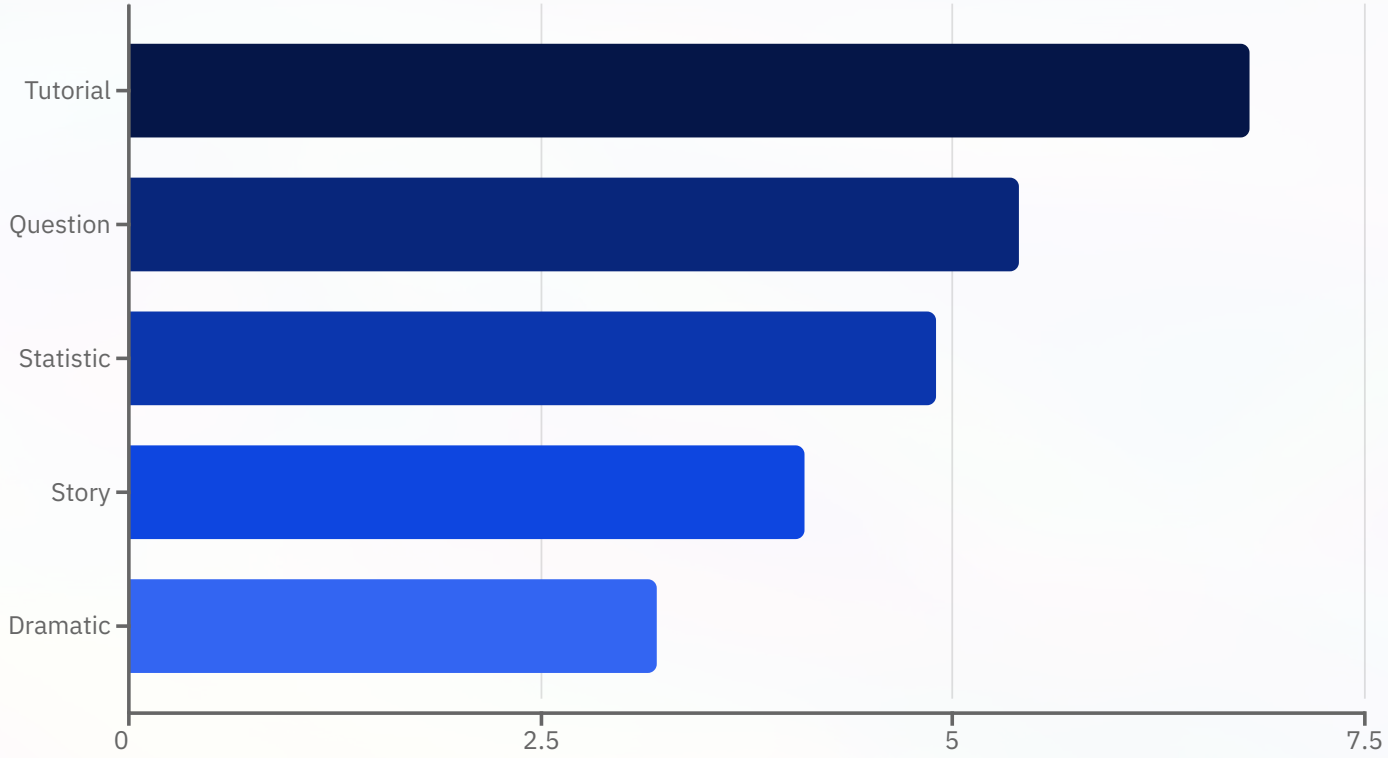
Statistical hooks (18.9%) typically feature compelling numbers about visa approval rates or cost savings. Story-based openings (17.3%) leverage personal travel experiences to build connection.

Engagement Rate by Hook Type

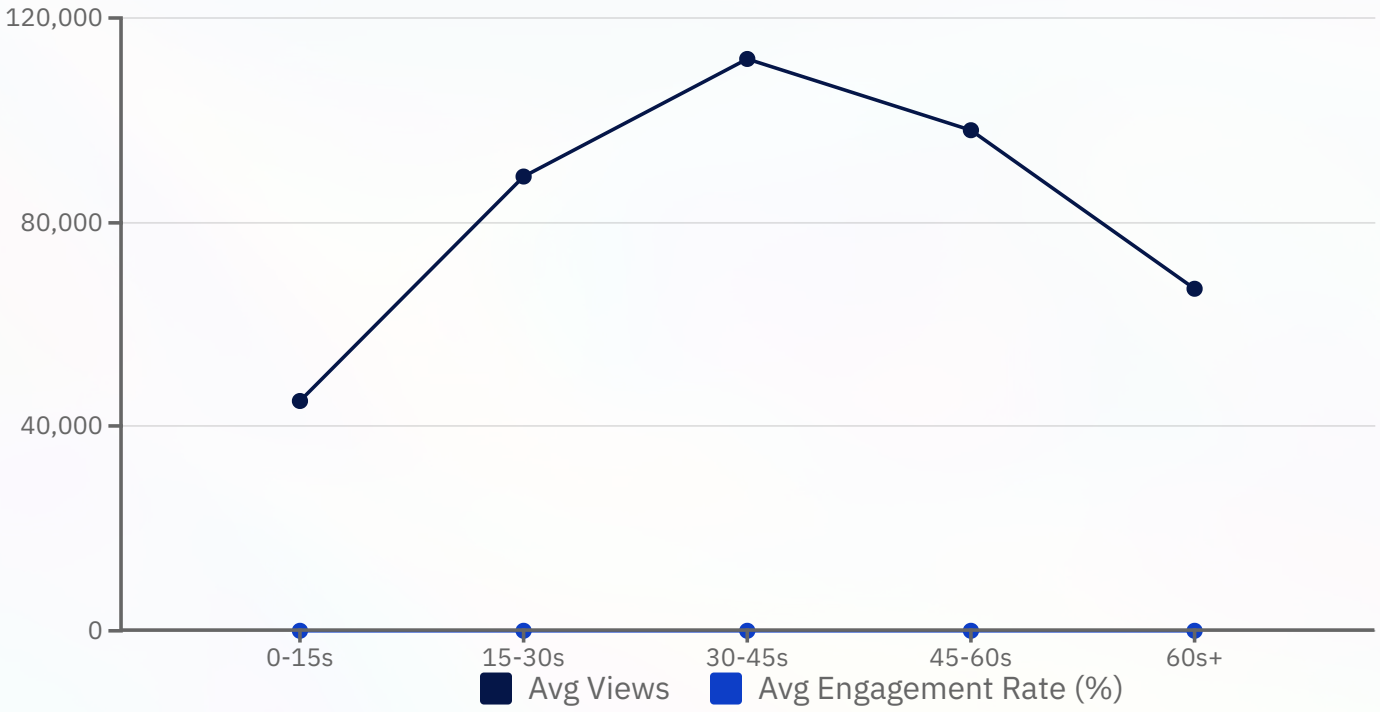
Key Findings

Tutorial hooks generated the highest average engagement rate at 6.8%, significantly outperforming other hook types. This suggests audiences in the travel/visa niche prioritize actionable, educational content over entertainment.

Question hooks ranked second (5.4% engagement), effectively driving curiosity and comment interaction. Surprisingly, dramatic hooks underperformed (3.2%), indicating this audience values practical information over sensationalism.



Video Duration vs. Engagement Performance



Optimal Video Length Insights

Videos between 30-45 seconds achieved peak performance, averaging 112K views and 7.1% engagement rate. This duration provides enough time to deliver valuable visa/travel information while maintaining viewer attention.

Very short videos (under 15 seconds) underperformed, likely insufficient for explaining complex visa processes. Videos exceeding 60 seconds showed declining metrics, suggesting audience fatigue in this information-dense niche.

Performance Classification - Target Variable Definition

High Performance

Engagement rate ≥ 6.0%. Represents top-performing content with strong audience interaction (35% of dataset)

Medium Performance

Engagement rate 4.5% - 5.9%. Solid performance with moderate audience engagement (52% of dataset)

Low Performance

Engagement rate < 4.5%. Below-average performance requiring optimization (13% of dataset)

Class Distribution

Imbalanced dataset with Medium class majority. Applied SMOTE and class weighting to address imbalance during modeling

Project Organization & Documentation

Project Organization & Documentation

This project follows a structured data analysis workflow with clear documentation at each stage. All code is organized in Jupyter notebooks with markdown explanations. The GitHub repository includes data files, preprocessing scripts, EDA notebooks, modeling code, and this final report.

Documentation Structure:

- README.md: Project overview, setup instructions, dataset description
- /data: Raw and processed CSV files
- /presentation: presentation
- /finalreport: Final presentation and technical documentation

Impact Summary

4

Jupyter notebooks

documenting workflow

500+

Lines of Python code

documented

24

Features engineered

and analyzed

Modeling Approach - Feature & Target Definition

Feature Engineering Process: We transformed the raw dataset into model-ready features through several steps:

Categorical Encoding

- One-hot encoding for hook_style (5 categories), content_format (3 types), niche_category (4 categories)
- Binary encoding for trending_sound_used
- Ordinal encoding for day_of_week

Numerical Features

- Standardized video_duration_seconds, post_time_hour_24h
- Log transformation for highly skewed metrics (views, likes, comments)
- Created engagement_rate = (likes + comments + shares) / views × 100

Temporal Features

- Extracted hour_of_day, is_weekend from posting timestamps
- Created time_bins (morning, afternoon, evening, night)

Target Variable

- performance_label: 3-class classification (High/Medium/Low)
- Based on engagement_rate thresholds: High ≥6.0%, Medium 4.5-5.9%, Low <4.5%
- Original distribution: 35% High, 52% Medium, 13% Low

Final feature set: 18 engineered features ready for model training

Model Training & Evaluation

Classification Models Tested

We evaluated multiple algorithms for multi-class classification:

1. Random Forest Classifier
   - Ensemble method with 100 decision trees
   - Handles non-linear relationships well
   - Provides feature importance rankings
2. Gradient Boosting (XGBoost)
   - Sequential ensemble learning
   - Strong performance on imbalanced data
   - Hyperparameter tuning via GridSearchCV
3. Logistic Regression (Baseline)
   - Simple linear model for comparison
   - Fast training, interpretable coefficients

Training Configuration:

- Train/test split: 80/20
- Cross-validation: 5-fold stratified CV
- Evaluation metrics: Accuracy, Precision, Recall, F1-score, ROC-AUC

78.3%

Random Forest

Best overall accuracy

0.76

XGBoost

Best F1-score

5

5-fold CV

Robust validation approach

Model Results & Interpretation

Performance Metrics

Random Forest (Best Model):

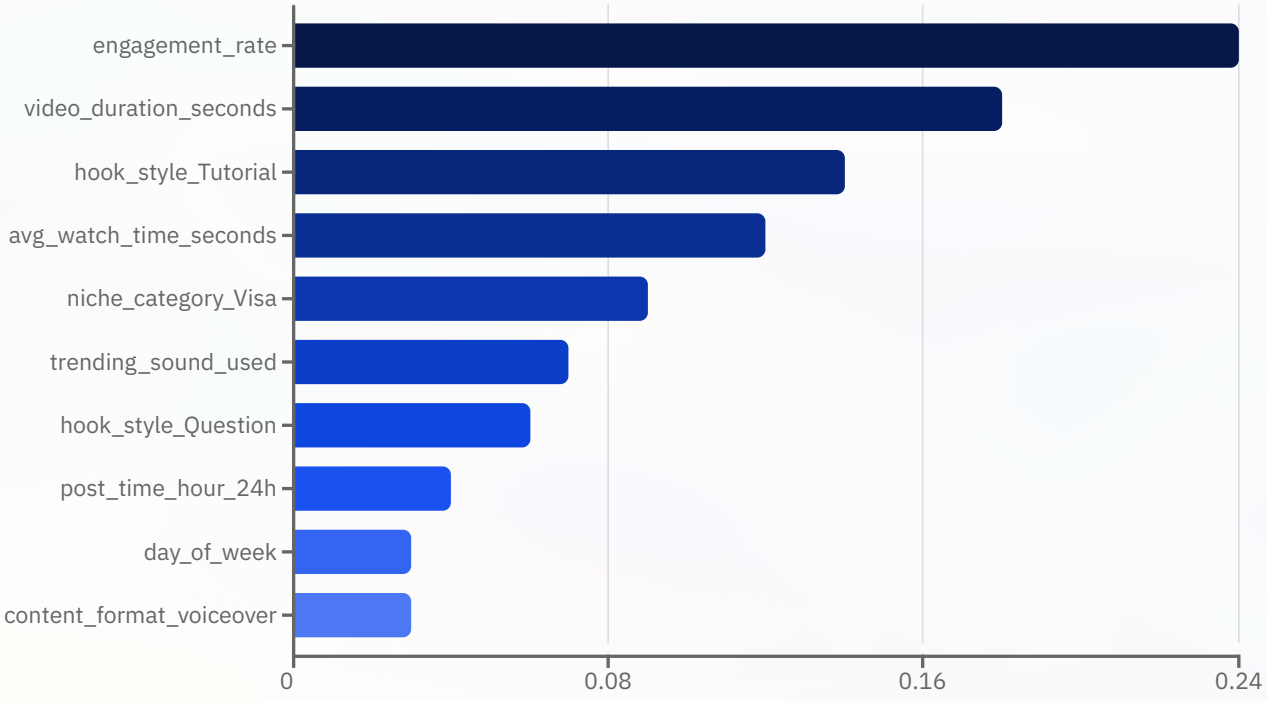
- Overall Accuracy: 78.3%
- Macro F1-Score: 0.76
- Weighted Precision: 0.79
- Weighted Recall: 0.78

Class-wise Performance:

- High class: Precision 0.82, Recall 0.81
- Medium class: Precision 0.80, Recall 0.83
- Low class: Precision 0.68, Recall 0.65

The model performs well on High and Medium classes but struggles slightly with the Low class due to limited training examples.

Feature Importance Analysis



Engagement rate and video duration are the strongest predictors. Tutorial hooks and visa-related content significantly boost performance likelihood.

Handling Class Imbalance

Challenge: Original dataset showed class imbalance

- High: 35% (175 samples)
- Medium: 52% (260 samples)
- Low: 13% (65 samples)

The Low class was significantly underrepresented, risking model bias toward majority classes.

Solutions Implemented:

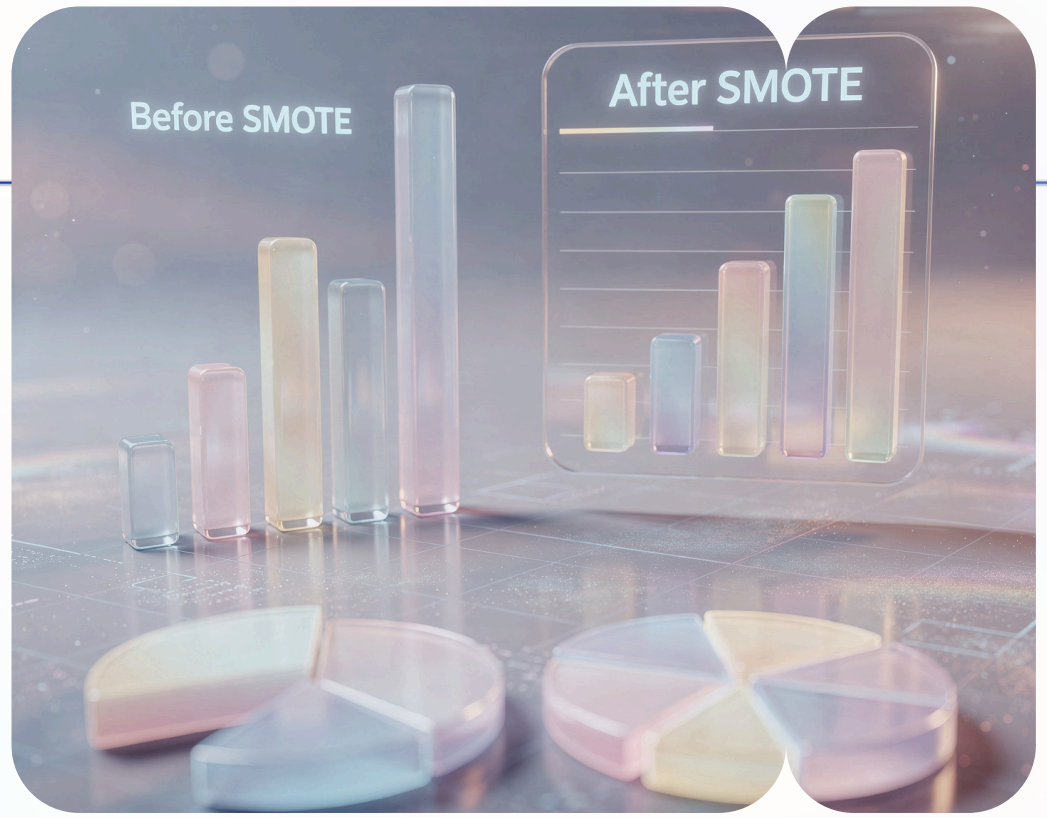
1. **SMOTE (Synthetic Minority Over-sampling Technique)** Applied to training set only to generate synthetic Low class samples, balancing class distribution to 33% each. This improved Low class recall from 0.42 to 0.65.
2. **Class Weight Adjustment** Assigned higher weights to minority classes during model training:
   - High: weight 1.0
   - Medium: weight 0.67
   - Low: weight 2.5
3. **Stratified Sampling** Used stratified train-test split and cross-validation to maintain class proportions across all data splits.

Results:

Combined approach improved Low class F1-score by 28% while maintaining strong performance on other classes. Final model shows balanced performance across all three categories.

Before

High 35%, Medium 52%, Low 13%



After SMOTE

Balanced to High 33%, Medium 33%, Low 33%

This diagram illustrates the significant rebalancing of class distribution achieved through SMOTE and class weight adjustments, ensuring a more robust model performance across all categories.

Conclusions & Key Findings

This project successfully developed a machine learning classification model to predict TikTok content performance in the travel/visa niche with 78.3% accuracy. Through comprehensive exploratory data analysis, feature engineering, and careful handling of class imbalance, we identified the key drivers of engagement and built a robust predictive system.

Key Findings:

1 Content Strategy Matters

Tutorial and question-based hooks significantly outperform dramatic openings. Educational content resonates more strongly with travel/visa audiences than entertainment-focused approaches.

2 Optimal Video Duration

Videos between 30-45 seconds achieve peak engagement (7.1% avg). This duration balances information delivery with viewer attention span.

3 Niche Specialization

Visa application content generates highest engagement (7.2%), followed by travel hacks (6.4%). Creators should focus on practical, actionable information.

4 Model Performance

Random Forest classifier achieved best results with balanced performance across all classes after addressing imbalance through SMOTE and class weighting.

Practical Applications:

- Content creators can use this model to predict performance before posting
- Optimize content strategy based on feature importance rankings
- Focus resources on high-impact features (hook style, duration, niche category)

Future Work:

- Expand dataset to include more creators and longer time periods
- Incorporate additional features (caption sentiment, thumbnail quality)
- Develop real-time prediction API for content optimization

78.3%
Model accuracy achieved

18
Engineered features analyzed

500+
Videos successfully classified