Content Monetization Modeler Report

1. Executive Summary

This project focuses on predicting **advertisement revenue (USD)** for online video content using historical performance and engagement data. By applying regression techniques and proper data preprocessing, the model achieves strong predictive accuracy ($R^2 \approx 0.95$). The solution is further extended into an **interactive Streamlit application** that allows users to explore data insights and generate real-time predictions.

---

2. Problem Statement

Digital platforms rely heavily on advertisement revenue. Accurately predicting ad revenue helps:

- Content creators optimize videos
- Businesses forecast revenue
- Platforms understand key revenue drivers

The objective is to build a robust machine learning model that predicts `ad_revenue_usd` based on video engagement metrics and categorical attributes.

---

3. Dataset Overview

- **Records:** ~120,000 videos
- **Target Variable:** `ad_revenue_usd`

Features

Numerical

- views
- likes
- comments
- watch_time_minutes
- video_length_minutes
- subscribers
- engagement_rate

Categorical

- category (Entertainment, Gaming, Music, etc.)
- device (Mobile, TV, Tablet)
- country (IN, US, UK, etc.)

---

4. Workflow Architecture

1. Data Collection
2. Exploratory Data Analysis (EDA)
3. Data Cleaning & Preprocessing
4. Feature Engineering

5. Model Training
6. Model Evaluation
7. Streamlit App Development
8. Final Insights & Conclusion

---

5. Exploratory Data Analysis (EDA)

Key Analysis Performed

- Distribution plots for numerical features
- Correlation heatmap
- Outlier detection using boxplots
- Revenue comparison across categories and countries

Key Insights

- Watch time shows the highest correlation with ad revenue
- Views alone are not sufficient without engagement
- Category and country significantly influence revenue

---

6. Data Preprocessing

- Removed non-predictive columns (IDs)
- Handled missing values (~5%) using **median imputation**
- One-Hot Encoding applied to categorical features
- Checked multicollinearity using correlation matrix

- Used **regularization techniques** instead of feature removal

---

7. Model Development

Models Implemented

1. Linear Regression
2. Ridge Regression
3. Lasso Regression
4. ElasticNet Regression
5. Huber Regressor
6. RANSAC Regressor

Regularized models helped mitigate multicollinearity among engagement metrics.

---

8. Evaluation Metrics

The following metrics were used:

- **R^2 Score** – variance explained by the model
- **MAE (Mean Absolute Error)** – average absolute prediction error
- **RMSE (Root Mean Squared Error)** – penalizes larger errors

Final Model Performance

- **R^2 Score:** 0.95
- **MAE:** 3.11
- **RMSE:** 13.48

This indicates high accuracy with low prediction error.

Model	R^2 Score	MAE	RMSE
Linear Regression	**0.95257**	3.11191	13.48063
Lasso Regression	0.95259	3.07527	13.47847
Ridge Regression	0.95258	3.10714	13.47976
Huber Regressor	0.95259	**2.97133**	**13.47706**
RANSAC Regressor	0.95258	3.14676	13.47855
ElasticNet Regression	0.95258	3.07692	13.47852

---

9. Streamlit Application

App Features

- Dataset overview and statistics
- Interactive visualizations (distributions, correlations)

- User input panel for prediction
- Real-time ad revenue prediction

Streamlit Workflow

1. Load trained model and preprocessing pipeline
2. Accept user inputs (numerical + categorical)
3. Apply transformations
4. Display predicted ad revenue

Benefits

- Business-friendly interface
- No coding required for end users
- Easily deployable

---

10. Results & Business Impact

- Identifies key drivers of ad revenue
- Enables revenue forecasting before publishing content
- Helps creators focus on engagement over raw views

---

11. Conclusion

The project successfully delivers a high-performing regression-based ad revenue prediction system. Through structured data preprocessing, model evaluation, and

deployment via Streamlit, the solution is both **technically robust and business-ready**.
