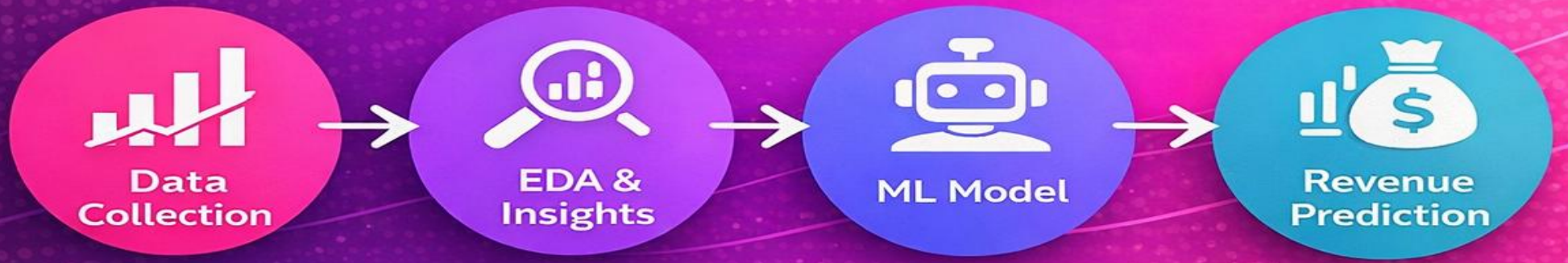
Content Monetization Modeler

Social Media Analytics | Machine Learning Project



Data Collection → EDA & Insights → ML Model → Revenue Prediction

Presented by **Abitha Jesuraj**

Domain: Data Analytics / Machine Learning

End-to-End ML-Driven Monetization Analytics Project

Project Overview

End-to-End ML Pipeline

End-to-end machine learning regression project

Revenue Prediction

Predicts YouTube ad revenue (ad_revenue_usd)

Feature Utilization

Uses video performance, engagement, and contextual features

Model Deployment

Final model deployed using **Streamlit** for real-time prediction

Regression-based ML model focused on monetization analytics

Project Objectives

Ad Revenue Modeling

Build a regression-based machine learning model to predict YouTube ad revenue (ad_revenue_usd)

EDA & Revenue Insights

Analyze video performance data using EDA to identify key revenue drivers

Feature Extraction

Apply data preprocessing and feature engineering to improve model accuracy

Model Evaluation & Deployment

Evaluate multiple regression models and deploy the final model using **Streamlit**

Clear objectives centered on driving **monetization** analytics through ML regression.

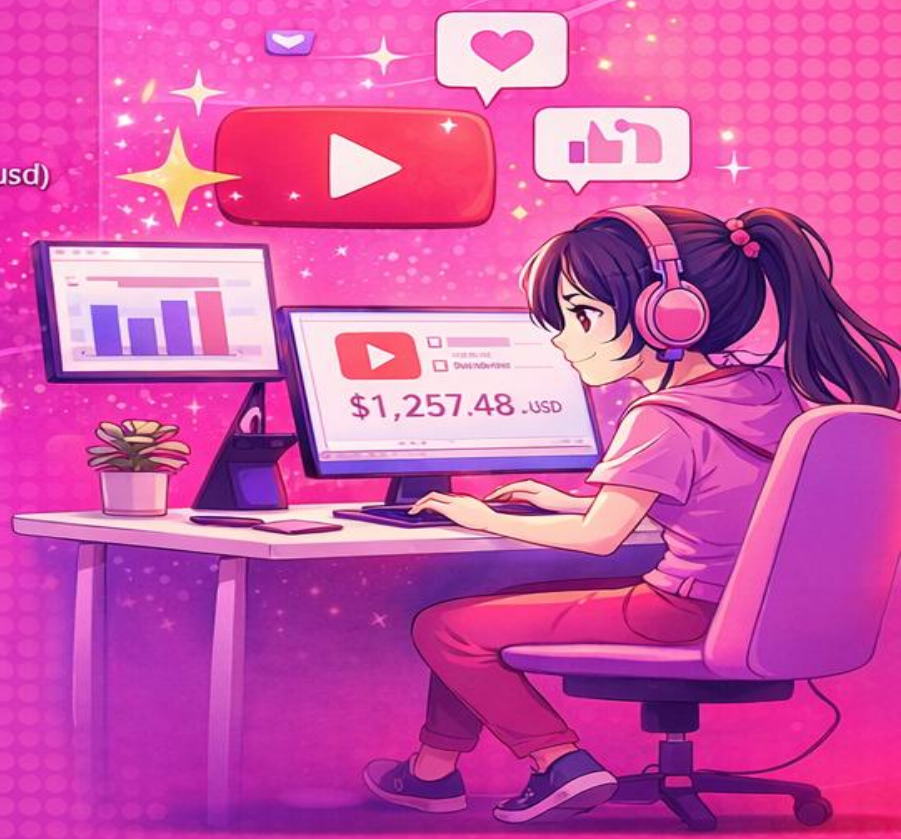
Problem Statement

YouTube creators and media companies **depend** heavily on **ad revenue** for income.

Ad revenue is influenced by multiple factors such as **views, engagement,** and watch **time**, making manual **prediction** difficult.

There is a need for a **data-driven machine learning** **solution** to accurately **predict YouTube** ad revenue (ad_revenue_usd)

This project addresses the problem by building and deploying a regression-based predictive model.



Dataset: YouTube Monetization Modeler (Synthetic)

- The model was trained on a comprehensive synthetic dataset designed to replicate real-world YouTube performance metrics and monetization outcomes.

Scale & Scope

Approx. 122,000 rows of video performance data, providing a robust foundation for modeling.

Target Variable

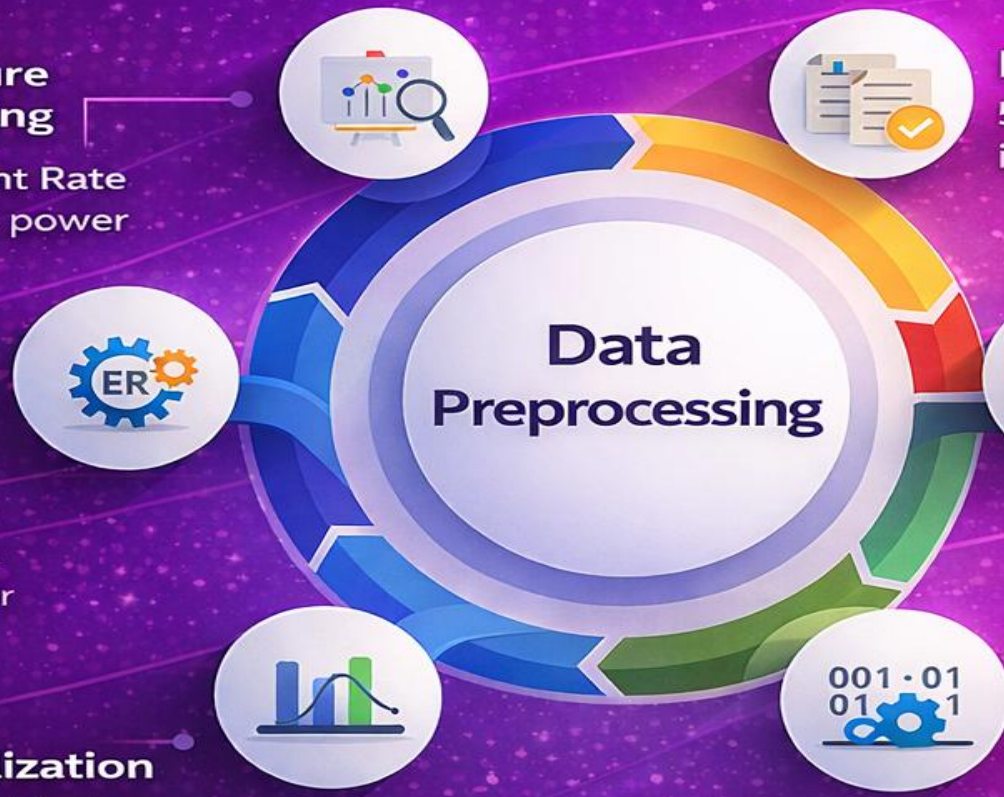
The predicted variable is ad_revenue_usd, serving as the core monetization metric.

Key Dataset Features

Feature Category	Columns
Engagement Metrics	Views, Likes, Comments, Watch Time (minutes)
Content Attributes	Video Length (minutes), Category
Channel Metrics	Subscriber Count
Contextual Features	Device Type, Country
Derived Feature	Engagement Rate = (Likes + Comments) / Views

Rigorous Data Preprocessing Pipeline

Ensuring high data quality and feature readiness to build accurate, reliable, and stable regression models for ad revenue prediction.



Feature Engineering

Added Engagement Rate feature for predictive power

Missing Values

5% of missing values imputed for data integrity

Feature Engineering

Added Engagement Rate feature for predictive power

Data Deduplication

2% of duplicate records removed to reduce bias

Normalization

Scaled numerical features for model stability

Categorical Encoding

Categorical features encoded using one-hot encoding

Exploratory Data Analysis (EDA) Insights

Initial analysis confirmed strong relationships between content consumption metrics and ad revenue, helping establish key hypotheses for model development.

Primary Drivers

A strong positive correlation was **observed between views, watch time, and ad revenue**, identifying them as **primary drivers** of **monetization**.

Categorical Variation

Revenue distribution exhibited **noticeable variation across different content categories** and geographical regions, indicating the impact of **contextual factors** on monetization.

Subscriber Impact

Subscriber count demonstrated **a moderate yet consistent influence** on **ad revenue**, suggesting that established audiences tend to generate higher monetization potential.

Engagement Factor

The engineered **Engagement Rate** emerged as a **highly significant predictor** of monetization performance, reinforcing highe morte genration.

Outlier Management

Outllers were **systematically analyzed** and **treated using appropriate techniques** (such as capping or transformation) to enhance model robus-tness and **stability**.

Comparative Model Building Strategy

A comparative modeling approach was adopted to identify the most suitable regression model while balancing interpretability and predictive accuracy.

Linear Regression

- Used as a baseline model
- Provides **high interpretability and quick** performance benchmarking

Ridge Regression

- Applied L2 regularization
- Mitigates multicollinearity and **stabilizes coefficient estimates**

Lasso Regression

- Applied L1 regularization
- Enables feature selection and **promotes model simplicity**

Random Forest Regressor

- Ensemble-based approach
- Effectively captures non-linear relationships and **improves**

Gradient Boosting Regressor

- High-performance boosting technique
- Maximizes predictive power through sequential error correction

Model Validation Strategy

- Dataset split into 80% training data and 20% validation/testing data
- Ensured objective and unbiased model evaluation

Model Performance Metrics

Evaluation focused on **R-squared (R^2)**, **Root Mean Squared Error (RMSE)**, and **Mean Absolute Error (MAE)** to assess model fit and prediction errors

Model	R^2 Score	RMSE	MAE
Linear Regression	**0.9526**	**13.48**	**3.12**
Ridge Regression	**0.9526**	13.48	3.12
Lasso Regression	0.9526	13.47	3.12
Random Forest Regressor	0.9521	13.55	3.70
Gradient Boosting Regressor	0.9518	13.58	4.07

Conclusion: Linear Regression as the Optimal Model

Given **comparable R^2 values** and **low error metrics** across models, **Linear Regression** was selected as the optimal model due to its strong predictive performance combined **maximum interpretability**, making it most suitable for business decision-making.

Streamlit Deployment: Revenue Predictor App

The trained regression model was deployed as an **interactive Streamlit web application**, enabling real-time **YouTube** ad revenue prediction through a simple and intuitive user interface.

Interactive Input

Users provide **key video performance metrics such as views**, Likes, Comments, Subscribers, and **related inputs** through a clean web interface.

Automatic Feature Calculation

The application **automatically** computes **derived features**, including **Engagement Rate**, within the backend to ensure consistency with the trained model.

Model Inference

A pre-trained **Linear Regression** model (saved using **Joblib**) is loaded within the app to generate instant predictions.

Real-Time Prediction

The application displays the estimated YouTube Ad Revenue (USD) in real time, providing actionable insights for content creators and media planners.

Key Insights

Views and watch time are the strongest drivers of **YouTube** ad **revenue**, showing a clear positive relationship with monetization.

Engagement Rate emerged as a highly influential feature, highlighting the importance of audience interaction beyond raw view counts.

Subscriber count has a moderate but consistent impact, indicating that established channels tend to achieve higher revenue stability.

Content category and geography significantly infuence **revenue** distribution, reflecting differences in audience behavior and CPM rates.

Simple, interpretable models such as **Linear Regression** can perform competitively with complex ensemble models for this use case.

Conclusion

- **Successfully** developed an **end-to-end** machine learning **solution** to predict **YouTube ad revenue** using video performance and engagement metrics.
- Identified **Views, Watch Time, Engagement Rate,** and **Subscriber Count** as the key drivers of **monetization**.
- **Compared multiple regression models** and selected **Linear Regression** due to its strong predictive performance and high interpretability.
- Deployed the final model using **Streamlit,** enabling **real-time revenue prediction** and **practical** business usage.
- **Demonstrated** a complete **analytics-to-deployment** workflow, applicable to real-world social media monetization scenarios.