YouTube Performance Analytics Report

Introduction

This project analyzes YouTube video performance metrics to identify factors that drive views and revenue. By examining a dataset containing 364 videos with 70 features, such as video duration, views, likes, shares, new subscribers, and estimated revenue, we aim to uncover relationships between these metrics and build a predictive model for revenue. The analysis involves **Exploratory Data Analysis** (EDA) to understand patterns and correlations, followed by a **Machine Learning (ML) model** to predict revenue based on key features. The insights gained can help content creators optimize their videos for higher engagement and earnings.

EDA Findings

The EDA was conducted using Pandas, Seaborn, and Matplotlib to explore the dataset and visualize relationships. Below are the key observations from the analysis, focusing on the correlation heatmap and other visualizations:

1. **Strong Correlation Between Views and Engagement Metrics**:
   - The correlation heatmap revealed that 'Views' has a strong positive correlation with 'Likes' (0.90), 'Dislikes' (0.78), 'Shares' (0.77), and 'New Subscribers' (0.80). This indicates that videos with higher view counts tend to receive more likes, dislikes, shares, and new subscribers, suggesting that viewership drives overall engagement.
   - For example, a video with 23,531 views (as seen in the dataset) had 924 likes and 54 new subscribers, supporting this trend.
2. **Moderate Correlation Between Views and Revenue**:
   - 'Views' and 'Estimated Revenue (USD)' have a moderate positive correlation of 0.36. This suggests that while higher views contribute to increased revenue, other factors (e.g., monetization rates or ad impressions) also play a role.
   - For instance, a video with 11,478 views earned $0.648, while another with 6,153 views earned $0.089, showing variability in revenue per view.
3. **Likes as a Strong Indicator of Revenue**:
   - 'Likes' and 'Estimated Revenue (USD)' have a relatively strong correlation of 0.43, higher than that of views. This implies that audience engagement through likes is a significant driver of revenue, possibly due to increased visibility or algorithmic promotion.
   - Videos with high likes (e.g., 924 likes for a video with $0.561 revenue) tend to earn more than those with fewer likes.
4. **Video Duration and View Percentage**:
   - 'Video Duration' has a strong negative correlation with 'Average View Percentage (%)' (-0.48). This suggests that longer videos are watched for a smaller percentage

    of their total length, indicating that shorter videos may retain viewer attention better.
    - For example, a 14-second video had an average view percentage of 103.05% (likely due to multiple views), while a 391-second video had 39.85%.
5. **Thumbnail Click-Through Rate (CTR) and Engagement**:
    - 'Video Thumbnail CTR (%)' has a moderate positive correlation with 'Views' (0.38) and 'Likes' (0.42). This indicates that compelling thumbnails drive more clicks and subsequent engagement.
    - A video with a high CTR of 27.66% achieved 23,531 views, highlighting the importance of thumbnail design.
6. **Day of Week Analysis**:
    - A boxplot of 'Estimated Revenue (USD)' by 'Day of Week' showed variability in revenue across days, with no clear day standing out as consistently high-earning. However, this suggests that the day of upload may have less impact on revenue compared to other factors like views or likes.
    - For example, videos uploaded on Thursday and Friday had median revenues around $4–$5, but outliers existed across all days.
7. **Scatter Plot Insights**:
    - A scatter plot of 'Views' vs. 'Video Duration' showed no strong linear relationship (correlation of -0.052), indicating that video length alone does not significantly affect view counts. Videos of varying durations (e.g., 14 seconds to 391 seconds) achieved high views, depending on other factors like content quality or promotion.

These findings highlight that engagement metrics (likes, shares, subscribers) and thumbnail effectiveness are critical drivers of views and revenue, while video duration impacts viewer retention.

ML Results

A Linear Regression model was built to predict 'Estimated Revenue (USD)' using five features: 'Video Duration', 'Views', 'Likes', 'Shares', and 'New Subscribers'. The dataset was split into 80% training and 20% testing sets, and the model was evaluated using Mean Squared Error (MSE) and R^2 Score.

- **Model Performance**:
    - **Mean Squared Error (MSE)**: 83.74. This indicates the average squared difference between predicted and actual revenue values. While the MSE is relatively high, it reflects the variability in revenue, which ranges from $0 to $103.12 in the dataset.
    - **R^2 Score**: 0.038. This low R^2 score suggests that the model explains only 3.8% of the variance in revenue, indicating poor predictive power. This could be due to the limited features used, non-linear relationships, or other unaccounted factors (e.g., ad types or viewer demographics).
- **Sample Prediction**:
    - For a hypothetical video with 500 seconds duration, 10,000 views, 500 likes, 50 shares, and 10 new subscribers, the model predicted a revenue of **$2.02**. This is

    plausible given the dataset's revenue range (mean of $8.85, median of $4.29), but the low R² score suggests caution in relying on this prediction.
    - A warning was noted during prediction (X does not have valid feature names), indicating that the input format for the sample video should match the training data's structure (e.g., using a DataFrame with column names).
- **Visualization**:
    - A scatter plot of actual vs. predicted revenue showed significant deviation from the ideal line (y=x), confirming the model's limited accuracy. Many predicted values were clustered around lower revenue figures, failing to capture higher revenue outliers.

The ML model's performance suggests that while it can provide rough estimates, additional features (e.g., 'Video Thumbnail CTR (%)' or ad-related metrics) or a more complex model (e.g., Random Forest) may improve accuracy.

Conclusion

The analysis reveals that views and likes strongly influence YouTube video revenue, with correlations of 0.36 and 0.43, respectively, with 'Estimated Revenue (USD)'. Engagement metrics like shares (0.36) and new subscribers (0.80 with views) also play a significant role, while thumbnail CTR (0.38 with views) highlights the importance of visual appeal. Video duration negatively affects view percentage (-0.48), suggesting shorter videos may retain viewers better. The Linear Regression model, with an R² score of 0.038 and MSE of 83.74, provides limited predictive power, indicating that revenue is influenced by factors beyond the selected features. Despite this, the model predicted $2.02 for a sample video, offering a starting point for revenue estimation. Future work could incorporate additional features (e.g., ad impressions), explore non-linear models, or analyze temporal trends to enhance predictions. These insights can guide content creators to focus on engagement and thumbnail optimization to maximize views and revenue.

Dashboard Link:

https://youtube-revenue-predictor-g56fvqedypcb8fclnu64pw.streamlit.app/