Unlocking YouTube Channel Performance Secrets

Domain: Data Analytics

Tools & Technologies Used:

- Languages: Python
- Tools/IDEs: Jupyter Notebook, Visual Studio Code
- Libraries: pandas, numpy, matplotlib, seaborn, scikit-learn, isodate, joblib

Project Overview:

This project analyzes the performance of a YouTube channel by exploring extensive video analytics data including engagement, monetization, and audience behavior. It includes exploratory data analysis (EDA), visualization, and machine learning modeling to uncover trends and predict revenue. The dataset includes metrics such as video duration, views, likes, revenue, subscriber behavior, and more.

Objectives:

- Explore and clean YouTube video performance data.
- Analyze correlations between video attributes and revenue.
- Create visualizations to interpret data insights.
- Engineer new features like engagement rate and revenue per view.
- Build a predictive model for estimating revenue.

Data Preparation & Feature Engineering:

- Converted video duration to seconds.
- Checked and removed missing values.
- Extracted new features such as:
  - Revenue per View = Estimated Revenue / Views
  - Engagement Rate = (Likes + Shares + Comments) / Views * 100
- Selected relevant numerical and categorical features for modeling.

Exploratory Data Analysis:

- Analyzed distribution of video durations and revenues.
- **Top Performing Videos:** Identified videos with highest estimated revenue.
- **Correlation Analysis:**
  - Strong positive correlation between **Views** and **Revenue**.
  - Moderate correlation between **Engagement Rate** and **Revenue**.

- **Visualization Tools Used:** Matplotlib & Seaborn for histograms, scatter plots, and heatmaps.

Machine Learning Model:

- Used Random Forest Regressor to predict Estimated Revenue (USD).
- Input features:
  - Views
  - Subscribers
  - Likes
  - Shares

- Comments
- Engagement Rate

Model Training

A Random Forest Regressor was selected for its ability to handle non-linear relationships and provide feature importance insights.

- **Data Split:** 80% training, 20% testing.
- **Hyperparameter Tuning:** Performed using GridSearchCV to optimize parameters such as n_estimators, max_depth, and min_samples_split.
- **Cross-Validation:** 5-fold cross-validation was applied to ensure model generalization.
- **Training Objective:** Predict Estimated Revenue (USD) based on selected performance metrics including Views, Engagement Rate, Likes, Shares, Comments, and Subscribers.

Model Evaluation

- **Mean Squared Error (MSE):** 0.0411
- **Root Mean Squared Error (RMSE):** 0.4828
- **R^2 Score (Test Set):** 0.9700
- **Cross-Validation R^2 Scores:** [0.6945, 0.9144, 0.9624, 0.9395, 0.9129]

- **Average CV R^2:** 0.8847

Insights

- Revenue is positively correlated with views and engagement.
- Feature importance plot highlights Views and Engagement Rate as top contributors.
- Identified content strategies that drive monetization through high ad impressions and average view duration.

Conclusion

This analysis identified views as the strongest driver of YouTube revenue, with engagement rate also contributing significantly. The Random Forest model achieved a high R^2 of 0.97 and strong cross-validation results, showing reliable predictive power. These insights highlight that focusing on increasing views and audience engagement can greatly enhance monetization potential.