1. `XY_train.csv` in folder "Task 1" contains a dataset of **`Views`** against 6 input variables.

    a. Select and train a model to predict the views for a given video, given the features provided. - Done

    b. Explain your choice of model and why it is suitable for this problem.

- I chose the **XGBoost (Extreme Gradient Boosting)** model for predicting video views. This choice is driven by the following considerations:

- **Handling Non-Linear Relationships**: XGBoost is capable of capturing complex, non-linear relationships between input features and the target variable (views).
- **Feature Importance**: It provides insights into feature importance, allowing us to understand the significance of different predictors.
- **Scalability**: XGBoost is computationally efficient, even for large datasets, making it suitable for this task.
- P**erformance** : achieving high accuracy and robust predictions in regression problems.

    c. Comment on your model's performance as well as the significance of each feature.

- Performance Metrics:

- **Best RMSE**: 840,947.07 (after hyperparameter tuning at round 36).
- **Final Model RMSE**: 736,253.96.
- **Mean Absolute Error (MAE)**: 495,841.20.

These metrics indicate a good fit for the data, with the RMSE and MAE suggesting reasonably accurate predictions of views.

Feature Significance:

The feature importance chart from the XGBoost model highlights the following significant predictors:

1. **Subscribers**: The most influential feature, as subscribers represent a committed audience that directly impacts video views.
2. **Comments Added**: Indicates engagement levels, correlating with views.
3. **Shares**: Social sharing amplifies reach, boosting views.
4. **Likes (vs. Dislikes) Percentage**: A measure of audience approval and video quality.
5. **Impressions Click-Through Rate (CTR)**: Reflects how effectively the video attracts viewers from impressions.
6. **Shares per Subscriber**: A derived metric capturing the amplification effect of each subscriber.

The analysis shows that audience engagement metrics (subscribers, comments, and shares) play a pivotal role in predicting video views.

d. Provide predictions for views of videos within `X_test.csv,` as well as an estimation of your out-of-sample performance.\

- Using the XGBoost model, I generated predictions for the test dataset (X_test.csv). The model relies on features like subscribers, comments, shares, and click-through rates to estimate video views accurately. These predictions are based on patterns identified in the training data and reflect the relationships between these variables and views.

Out-of-Sample Performance

The model performed well on validation data, and its expected out-of-sample performance is as follows:

- **RMSE** (Root Mean Squared Error): ~736,000. This shows the average difference between predicted and actual views.
- **MAE** (Mean Absolute Error): ~495,000. This is the average size of the prediction errors, indicating the typical deviation from actual views.
- **Residual Analysis**:
  - The **mean residual** is ~65,760, suggesting a slight overestimation on average.
  - The **median residual** is ~-53,154, meaning most errors are close to zero.
  - The **standard deviation of residuals** (~733,311) aligns well with the RMSE, showing consistent variance in predictions.

Summary

The residuals are well-distributed around zero, with no noticeable patterns of bias, meaning the model is making consistent predictions across different ranges of views. While predictions for most videos are accurate, there is some variability for videos with extreme or unique features (e.g., viral content). Overall, the model demonstrates strong generalization and reliability for unseen data in the test set.

2. Using the audience retention graphs of the videos within the folder ‘’Task 2 & 3” predict the audience retention % of Video A at positions 0, 1, 2, 3, 4, 5 by inputting your answers in the yellow fields. Explain your reasoning and how you got to this number.

- The predicted audience retention percentages for Video A at positions 0, 1, 2, 3, 4, and 5 were determined using data analysis and regression modeling techniques. The data from Videos A-F, which includes audience retention percentages at various video positions, viewer engagement metrics (such as subscription status and audience type), and video position percentages, were used to understand the relationship between video position and audience retention.

Process:

1. **Data Insights**: The dataset revealed a common trend where audience retention decreases as the video progresses, a typical pattern in most video content.

2. **Model Selection**: I initially explored models like Decision Tree Regression and Random Forest Regression. While these models performed well, they introduced complexity and sometimes overfitting. Linear Regression, however, provided consistent and interpretable results, making it the best choice.

3. **Why Linear Regression**:

    - **Simplicity**: It offered a straightforward approach to understanding the relationship between video position and retention.
    - **Accuracy**: Linear regression captured the general decline in retention over time with a reasonable level of accuracy, as measured by performance metrics like R-squared and Mean Squared Error.

Final Predictions for Video A:

- **Position 0**: 92.94%
- **Position 1**: 91.80%
- **Position 2**: 90.67%
- **Position 3**: 89.54%
- **Position 4**: 88.41%
- **Position 5**: 87.27%

These predictions reflect a gradual decline in retention, consistent with common viewer behavior, where engagement tends to drop as the video progresses. The use of **video position** as a feature in the **linear regression model** allowed us to effectively predict how retention would change at specific points in Video A.

3. Using the scripts provided, come up with more effective wording for the intro of Video A to increase audience retention. Predict the audience retention with this new intro and explain your reasoning.

- New Introduction : **We all want to leave a memorable mark on the world, but not everyone gets it right. Imagine moments so absurd, they're almost unbelievable — the kind of tales that make you laugh, gasp, or shake your head in disbelief. Today, we're diving into some of the most hilariously bizarre and jaw-droppingly unexpected ways people have met their end. Curious? Let's explore these unforgettable and wildly entertaining stories together!.**

Reasoning : **Key Insight:** Retention bands indicate according to my Analysis:

- Low retention is often associated with **Fear**, **Anger**, and **Sadness**.
- Medium retention is distributed across all emotions but leans towards **Neutral** and **Fear**.
- High retention is sparse and includes positive or engaging emotions like **Joy**, **Surprise**, and **Disgust** (which can be captivating in specific contexts).

So I have analysed and conclude that :

Emotion Focus:

- Included **joy** and **curiosity-inducing** phrases like "hilariously bizarre" and "jaw-droppingly unexpected."
- Avoided words with negative connotations like "dumb" and "dim-witted," which may evoke **anger** or **disgust**.

Engagement Hook:

- Used open-ended questions ("Curious?") to engage viewers immediately.
- Crafted the tone to be lighter and more inclusive, appealing to a broader audience.
-

Retention-Enhancing Language:

- Leveraged words and phrases found in high-retention sections of other videos.

Result : Predicted Absolute Audience Retention: **89% , current intro audience retention is 77%**

Submission Instructions

- Provide a repository containing your code for each point above
- Send us your predictions for Task 1
- Send us your filled in `"X_test"` and `"Video A"` files for Tasks 2 and 3