Introduction
YouTube is a world-renowned platform used to share videos on a wide range of topics, including entertainment, music, gaming, lifestyle, and many more. In this digital age, many people have shifted away from traditional 9-5 jobs and become full-time content creators, posting videos for consumption by audiences around the world. With the potential to make substantial earnings, YouTube has become a lucrative platform for creators.

For my final project, I aim to build a predictive model that forecasts earnings by category. This model will be useful for both current and aspiring content creators who want to maximize their earnings and engagement. It will help them make informed decisions about the type of content to create, thus increasing their chances of gaining popularity and sustaining their brand. Additionally, this model can assist brands in identifying the best creators for partnerships based on predicted earnings. I will use a regression task to predict the highest earnings by category.

Data Description
The data used comes from the **Global YouTube Statistics** dataset on Kaggle. It contains 995 rows and 28 columns, which provide a comprehensive overview of YouTube channels. The columns include:

- **rank**: Position of the YouTube channel based on the number of subscribers
- **Youtuber**: Name of the YouTube channel
- **subscribers**: Number of subscribers to the channel
- **video views**: Total views across all videos on the channel
- **category**: Category or niche of the channel
- **Title**: Title of the YouTube channel
- **uploads**: Total number of videos uploaded on the channel
- **Country**: Country where the YouTube channel originates
- **Abbreviation**: Abbreviation of the country
- **channel_type**: Type of the YouTube channel (e.g., individual, brand)
- **video_views_rank**: Ranking of the channel based on total video views
- **country_rank**: Ranking of the channel based on the number of subscribers within its country
- **channel_type_rank**: Ranking of the channel based on its type (individual or brand)
- **video_views_for_the_last_30_days**: Total video views in the last 30 days
- **lowest_monthly_earnings**: Lowest estimated monthly earnings from the channel
- **highest_monthly_earnings**: Highest estimated monthly earnings from the channel
- **lowest_yearly_earnings**: Lowest estimated yearly earnings from the channel
- **highest_yearly_earnings**: Highest estimated yearly earnings from the channel
- **subscribers_for_last_30_days**: Number of new subscribers gained in the last 30 days
- **created_year**: Year when the YouTube channel was created

- **created_month**: Month when the YouTube channel was created
- **created_date**: Exact date of the YouTube channel's creation
- **Gross tertiary education enrollment (%)**: Percentage of the population enrolled in tertiary education in the country
- **Population**: Total population of the country
- **Unemployment rate**: Unemployment rate in the country
- **Urban_population**: Percentage of the population living in urban areas
- **Latitude**: Latitude coordinate of the country's location
- **Longitude**: Longitude coordinate of the country's location

A preliminary glance at the dataset shows that it contains some null values, so it will require some cleaning. Furthermore, not every column is necessary for predicting the target variable (highest yearly earnings), so some will be removed. For my model, I decided to use the following features:

- **'subscribers'**, **'video views'**, **'uploads'**, **'video_views_for_the_last_30_days'**, **'lowest_monthly_earnings'**, **'highest_monthly_earnings'**, **'lowest_yearly_earnings'**, **'subscribers_for_last_30_days'**, **'log_earnings'**, **'Country_Abr'**, and **'channel_type'**.

This selection was based on the fact that audience engagement metrics like subscribers and video views are key factors in YouTube's algorithm for determining how much a creator is paid. Additionally, channels with more subscribers and views often indicate that their category appeals to a larger audience. Features like monthly and yearly earnings help to highlight trends in the maximum earnings potential within each category. The country feature was included because content from certain countries may earn more due to regional ad rates.

Once a new dataframe was created with the relevant columns, I explored the dataset. Upon inspecting the target variable (highest yearly earnings), I found that the data was highly skewed to the left, with a maximum of $110,600,000.00 and a minimum of $2.00. This could cause problems since linear regression models are sensitive to the scale of the target variable and input features. Outliers can skew the mean squared error significantly. To correct this and help the model run more efficiently, I took the logarithm of the highest yearly earnings and created a new column, 'log_earnings'. This resulted in a slightly right-skewed but more normally distributed target variable.

In evaluating the relationship between the features and the target, I observed that video views for the last 30 days, lowest monthly earnings, highest monthly earnings, and lowest yearly earnings showed a positive correlation with the target variable. The subscribers feature showed no direct correlation, although channels with the highest number of subscribers tend to earn more. However, these outliers were far from most of the data points. Video views are clustered around the middle, but channels with the most views still see the highest earnings. For uploads, most data is clustered to the right, indicating that channels with many uploads can see both high and

low earnings. As for the two categorical features—country and channel type—videos from India and the United States earn more than videos from other countries, and Entertainment and Music channels tend to earn more than other categories.

Models and Methods

To predict highest yearly earnings, I used 4 different regression models (Baseline Model, Linear Regression, K-Nearest Neighbors, Random Forest) and evaluated which performed the best. The data was spilt 80-20, with 80% of the data being used to train the model and 20% being used to test the model.

Baseline Model

The baseline model is a reference to help assess the performance of the more complex models. To create the baseline model, I created an array of predictions that are equal to the mean of the log highest yearly earnings and calculated the mean squared error between those and the true value of log highest earnings.

Linear Regression

I chose a Linear Regression because the coefficients will help explain how much each feature contributes to the predicted earnings. Moreover, when the relationship between features and targets are linear this model performs well. To create the Linear Regression model, the following steps were taken:

1. Create X (all categorical and numerical features) and y (log earnings).
2. Split the data 80-20
3. Encoded the categorical features
4. Made a pipeline for regression model
5. Fit the pipeline
6. Calculator MSE for training and testing data

K-Nearest Neighbors

I chose K-Nearest Neighbors because it makes predictions on the closest data points in the feature and does not make assumptions about the underlying distribution. To create the K-Nearest Neighbors model, the following steps were taken:

1. Create X (all categorical and numerical features) and y (log earnings).
2. Split the data 80-20
3. Encoded the categorical features
4. Make a pipeline for K-Nearest Neighbors model

5. Define grid of hyperparameters for number of neighbors
6. Perform grid search with cross validation
7. Fit grid search
8. Find the best parameter
9. Use the best parameter to estimate the model
10. Calculator MSE for training and testing data

Random Forest

Random Forest works really well at capturing non-linear relationships. This also helps to capture complex interaction between features and is less likely to overfit than a single decision tree.

1. Create X (all categorical and numerical features) and y (log earnings)
2. Split the data 80-20
3. Encoded the categorical features
4. Make a pipeline for random forest regressor model
5. Define grid of hyperparameters for number of estimators and max depth
6. Perform grid-search w/ cross validation
7. Fit grid search
8. Find the best parameter
9. Use the best parameter to estimate the model
10. Calculator MSE for training and testing data

Results

Baseline Model

The baseline model sets the starting point for performance comparison. With a mean squared error of 6.155, it performed poorly in comparison to all other models, as expected, since it's a naive prediction without learning from the data.

Linear Regression

The Linear Regression model performs reasonably well but not as well as KNN and Random Forest. Its performance is acceptable, but there's room for improvement with a mean squared error of 0.542 on the training data and 0.521 on the test data. The Linear Regression model shows reasonable performance with a slight difference between training and testing MSE, indicating a relatively low risk of overfitting. The MSE values are significantly lower than the baseline (6.155), suggesting that the model is capturing some patterns in the data.

K-Nearest Neighbors

The KNN model performs exceptionally well. It produced the lowest Mean Squared Error on both training (0.022) and test sets(0.004). The small difference between training and test MSE suggests that the model generalizes well to unseen data, avoiding overfitting. It's significantly better than both the baseline and the other models, indicating strong performance on the task.

Random Forest

The Random Forest model also performs very well, with identical Mean squared Error values for both training (0.005) and testing (0.005) data. This indicates that the model fits well to the training data and generalizes well to the test data, without overfitting. The performance is strong, though slightly higher than KNN on the test set.

Limitations

1. Logged Earnings
   The current results are based on the logged highest yearly earnings.

2. Data Availability
   The model doesn't include any data advertisements and payments for increased visibility. A large portion of earnings comes from Youtube's AdSense, in which creators get paid for how many people view and click on the ads in their videos. Moreover, payment from in video brand promotions are not included.

3. Dynamic Environment
   Youtube trends and revenue models can change, so the model needs to be regularly updated to remain effective.

Improvements

To increase the capabilities of my data, I'd like to include the following features:

1. Social Media & External Factors
   Channels with high social media activity have better engagement and higher potential earnings. In addition,Viral videos typically see huge spikes in views and ad revenue, which can significantly impact yearly earnings.

2. Country & Region-Specific Factors
   Some countries have higher advertising rates than others. For example, content targeted at U.S. or European audiences tends to generate higher ad revenue. Features such as GDP, unemployment rate, consumer spending, and average income in the country could provide insight into how much people in a particular country might be able to spend on goods and services, impacting ad revenue.

3. Time of Year (Seasonal Factors)
   Earnings can vary depending on the time of year. For instance, many YouTubers experience higher earnings during the holiday season, an election year or summertime when advertisers tend to spend more on ads and people may be consuming more content.