Youtube Data Analysis & views prediction

Group #6
Appidi Moni
Mohammad Feroz Ahmad Qureshi
Ravina Ingole Ringo
Sai Gowtham Reddy Kallu
Sudip Adhikari

Business Understanding

- Helps ad campaigns.
- Aids in answering business queries such as which titles are most loved and disliked on YouTube based on region, category.
- Helps content creators to pick a user-based relevant topic .

Highlights

- Data Visualization using Power BI
- Implementation of ETL process
- YouTube View Predictions using Machine Learning

Data Source and Distributions

- Dataset is collected from kaggle. Below is the link for the reference.

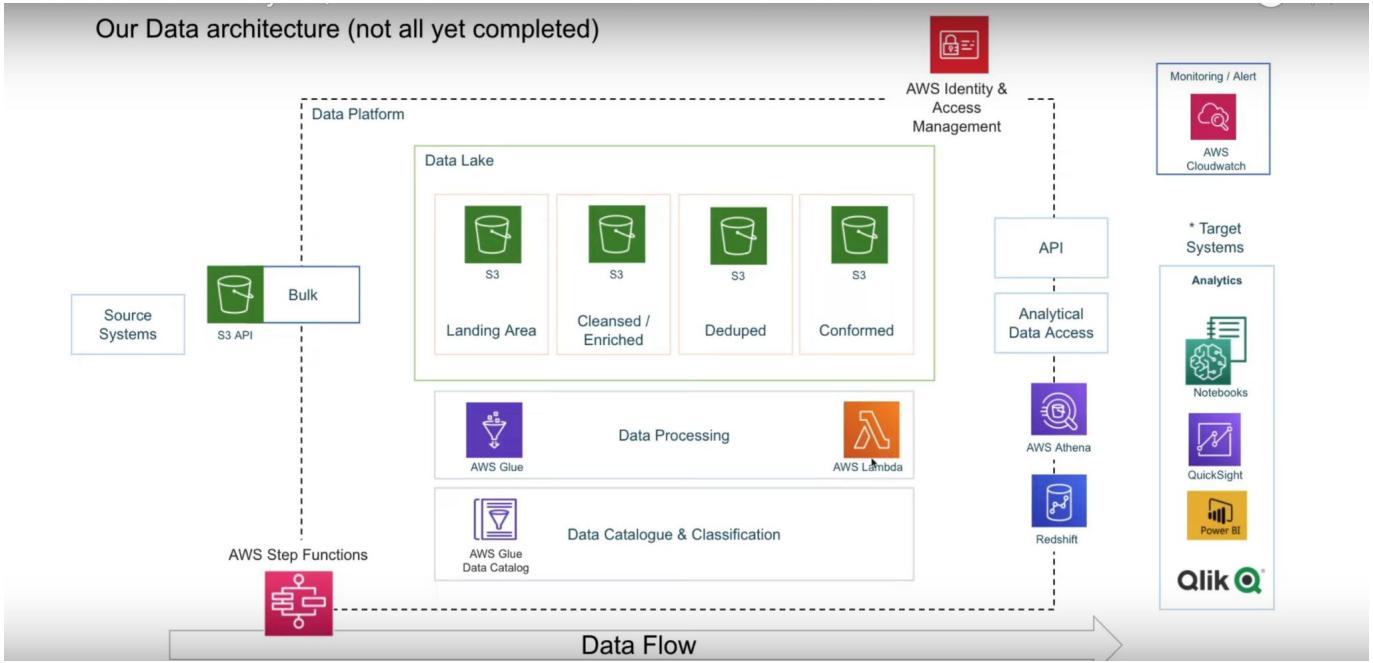
  https://www.kaggle.com/datasets/datasnaek/youtube-new

- It consist of below features:

video_id	category_id
trending_date	publish_time
title	tags
channel_title	views
likes	dislikes
comment_count	thumbnail_link
comments_disabled	ratings_disabled
video_error_or_removed	description

Our Process

- This is the chart of workflow.



S3 Bucket for data storage:

Buckets (15) Info

Buckets are containers for data stored in S3. Learn more

Find buckets by name

Name
aws-glue-assets-434479421845-us-east-1
aws-logs-434479421845-us-east-1
awssparkify3.1
experimental1.0
feroz.qureshi.cs79a.bucket
final-project-analytics-bucket
final-project-youtube-cleansed-useast1-dev
final-project-youtube-raw-useast1-athena-job
final-project-youtube-raw-useast1-dev
jsonobjects
logs.feroz.qureshi.cs79a.bucket
mllib1.0.1
myawssparkify
sparkify5.0.1
www.feroz.qureshi.cs79a.bucket

Databases and running crawlers

Databases (8)
View and manage all available databases

Filter databases

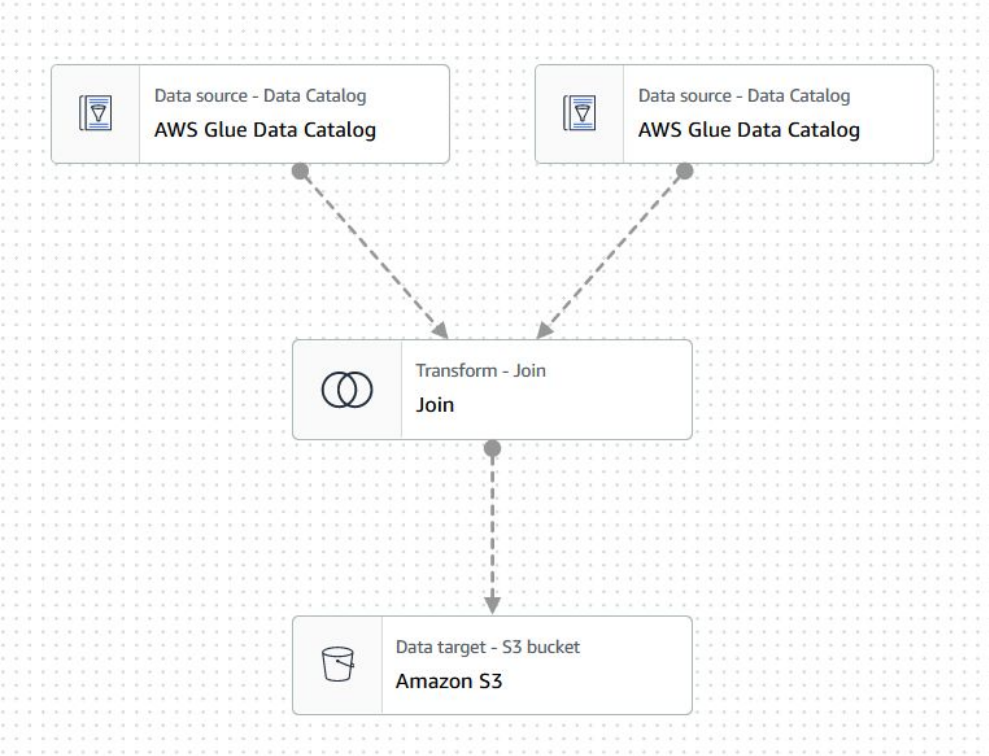
Name	Desc
analytics_data	-
csv_to_parquet	-
default	-
final-project-csv-data	-
final-project-parquet-data	-
final-project-youtube-raw	-
json_to_parquet	-
testing-parquet	-

Crawlers (3) Info
View and manage all available crawlers.

Filter crawlers

Name	State
cleaned_p...	Ready
csv_to_par...	Ready
final-proje...	Ready

ETL process in Glue studio



Data Understanding

- DataFrame has 15 columns and data of 3 regions (US, canada, UK)
- View column is our target feature.
- There’re 48 duplicates and 169 null values in the dataset.

video_id	category_id
trending_date	publish_time
title	tags
channel_title	views
likes	dislikes
comment_count	thumbnail_link
comments_disabled	ratings_disabled
video_error_or_removed	description

Data Preparation

- The `publish_time`, `publish_date`, `description`, `tags`, `title`, `channel_title` features are removed.
- Features with numerical data are filled in with the median value of each feature.
- The transformation log is carried out on features with numerical data values to convert them to normal / almost normal distributions
- Then we fill in the features that have missing values by using the mode of the feature (the `comments_disabled`, `video_error_or_removed`, and `ratings_disabled` features.

Modeling

Performed below steps for machine Learning Youtube views Prediction

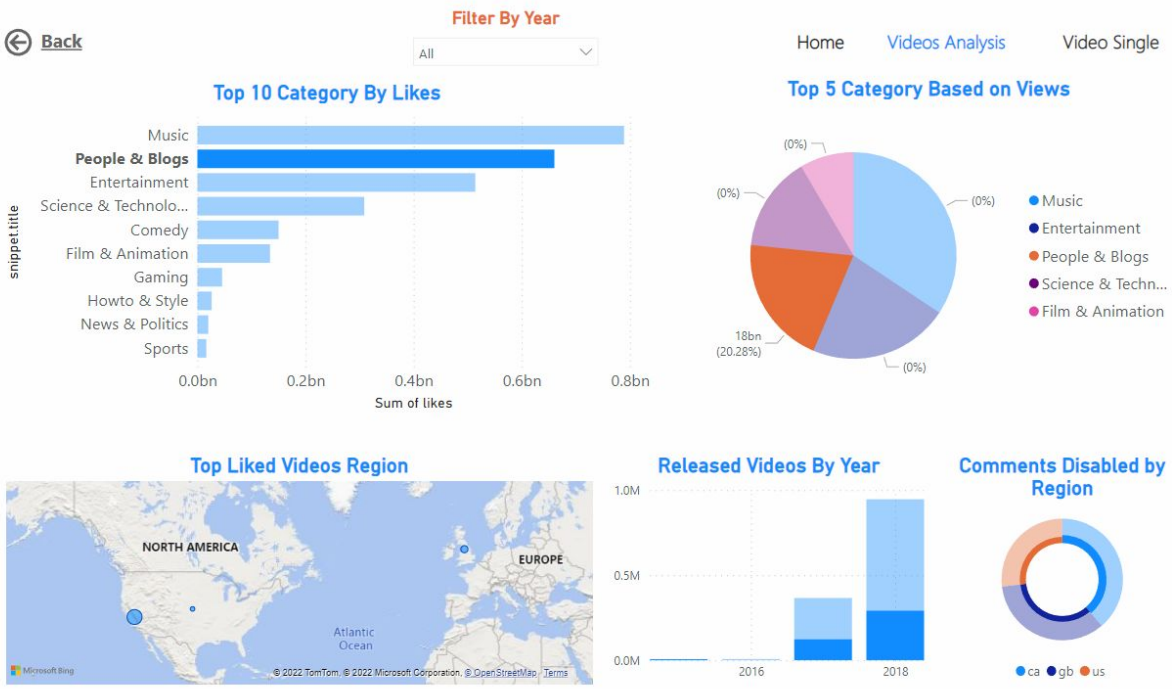
- Exploratory Data Analysis
- Data Pre-processing
- Model Building with hyperparameter tuning
- Validations on test datasets

Evaluation

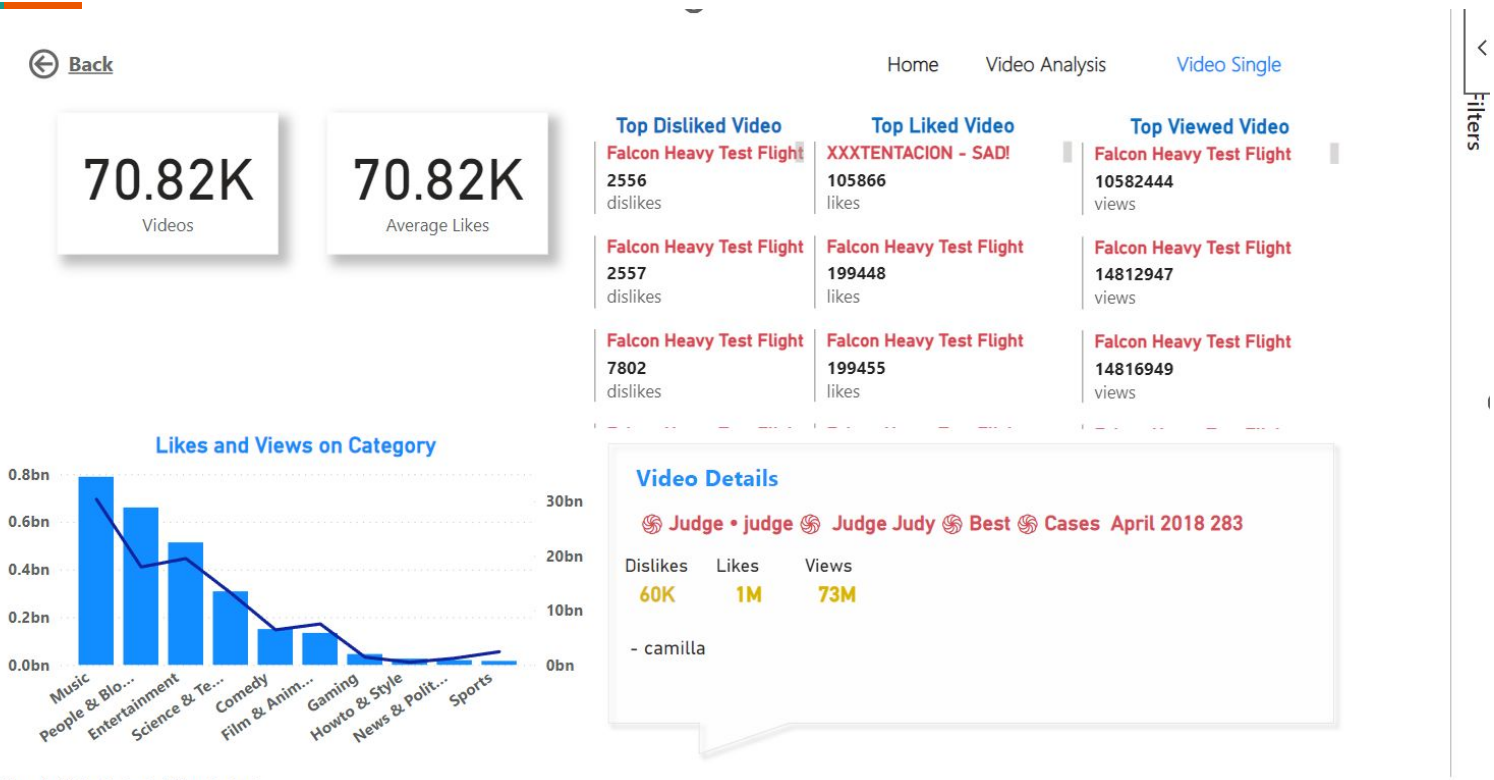
- Random Forest with MAE 0.00, RMSE 0.01, and R2 of 0.96 makes Random Forest the best model so far
- Decision Tree is the second best sequence model after Random Forest with a slightly smaller R2 value of 0.92
- Ridge Regularization is the next best model with MAE values of 0.01, RMSE 0.01, and R2 of 0.77
- A very influential feature is the number of likes and dislikes of a video

Model	MAE	RMSE	R2 Score
Regressor	0.01	0.01	0.77
ridge_model	0.01	0.01	0.77
Fit Lasso Regularization Model	0.01	0.03	0.77
Fit Elastic Net Regularization Model	0.01	0.03	0.77
Fit Decision Tree Model	0	0.01	0.93
Fit Random Forest Model	0	0.01	0.96
Fit Support Vector Regressor Model	0.09	0.09	0.07

Visualizations



Visualizations



Any Queries?

Thank You!!!