

# PREDICTION AND SENTIMENT ANALYSIS ON YOUTUBE VIDEOS

**Rohan Nagalkar**  
**(10403259)**

**Module Guide**  
**Terri Hoare**

## **Content**

Introduction

Business Understanding

Data Understanding

Data preparation

Modelling

Evaluation and Deployment

Conclusion

## Introduction

**Abstract-** This project mainly based upon a prediction system that takes video likes, dislikes comments and category as input and predict the views of the video. We have also done sentiment analysis based on the comments available on every video. We have used several data classifying techniques to build a trained classifier and to classify our test data. Using this cla to Variety magazine, “To determine the year’s top-trending videos, YouTube uses a combination of factors including measuring users interactions (number of views, shares, comments and likes). Note that they’re not the most-viewed videos overall for the calendar year”. Top performers on the YouTube trending list are music videos (such as the famously virile “Gangnam Style”), celebrity and/or reality TV performances, and the random dude-with-a-camera viral videos that YouTube is well-known for.

Possible uses for this dataset could be:

- Sentiment analysis 9+ in a variety of forms
- Categorising YouTube videos based on their comments and statistics.
- Training ML algorithms like GLM to predict their views.
- Analysing what factors affect how popular a YouTube video will be.

The main goal of the project is to predict the views of the video posted recently so that it will be helpful for the video poster whether it is a company or a person just uploading a video.

We have done sentiment analysis using Python and R as our csv size was too much we were facing issues to load csv file using Rstudio so we had to use Python for reading csv and getting the unique videos from a csv.

There will be enormous benefits of this application like which video can be used to post advertisement.

Also if we take an example of Gangnam style video there is a history of a video that because of Nine quintillion of views the youtube application got crashed

On 1 December, Google posted a statement saying: "We never thought a video would be watched in numbers greater than a 32-bit integer... but that was before we met Psy." Google, which owns YouTube, later told website **The Verge** that engineers "saw this coming a couple months ago and updated our systems to prepare for it".

YouTube now uses a 64-bit integer for its video counter, which means videos have a maximum viewer count of 9.22 quintillion.

So this eventually will be used not only for generating revenue but also will be helpful in preventing application from generating errors.

We have used **CRISP-DM** methodology for our application

This methodology integrates the Security, Trust, Efficiency and Freedom infringement (S-T-E-Fi) dimensions in its evaluation stage. This is a highly innovative approach, as certification has, to date, primarily focused on the assessment of technical requirements for security systems.

## **CRISP-DM Business Understanding:**

### **Objectives:**

Following are the objectives from the datasets given

- Analyse the people's comments and get the sentiments analysis from the user's comments.
- Getting the statistical view of the youtube video to analyze the factors
- Using some machine learning algorithms like GLMs to predict the views of the video.
- Analysing what the other factors are that affect the youtube trending video

### **Evaluating the circumstance:**

Nowadays Youtube is a very trending platform, which has a very huge user and it increases day by days. Even now televisions and the other social media platform is replaced by the youtube hence for the industries and the new youtube creators need some statistical analysis of the trending videos to either promote the product or to generate the great content.

The big problem in front of them is how to get good content?

What is the people opinion about the different category?

Which video can be viral and why?

### **The aim of data mining:**

Base on the Trending video data, i.e. its views, likes, dislikes and comment it is beneficial to analyse some statistical parameter which can tell the secrets behind the achievements of the popularity of the videos. The analysis will help the commercial industries about their product reviews earlier, from their promo video also for youtube creators the trending topics and category who are seeking to get more subscribers. This analysis will help the commercial industries to get the weakness of their product even to improve some of the business strategies and the weakness of the product. For the youtube creators, what are parameters factors are affecting the parameter to trend the video

**Data :**

Data collection part is done through [Kaggle](#)

The dataset is about the trending video from the USA and Great Britain

This dataset has the 4 CSV file which is one the data of the two countries data which is US and Canada, and the remain 2 are comments of the two countries.

The data have the following parameters:

video\_id

title

channel\_title

category\_id

tags

views

likes

dislikes

comment\_total

thumbnail\_link

date

**Data Investigation:**

The videos have around Great Britain database have 30581 videos, and also there are 44 categories and 6680 channels In the United, State database have the 40567 videos and 44 categories and 7000 channels data

**Data Attribute:**

The data contain mixed data which include the unstructured and not English data; hence there are must need of the data cleaning process.

**Data Preparation**

As there were almost no missing values(can be counted on fingers) present in data we deleted the rows which were having no values or empty values.

Removing outlier is also a part of data preparation, there were few outliers in the dataset like a video has around lakhs of views but not a single like or dislike or a single comment for that video. So we removed such data from our data set.

**The aim of data mining:**

Base on the Trending video data, i.e. its views, likes, dislikes and comment it is beneficial to analyse some statistical parameter which can tell the secrets behind the achievements of the popularity of the videos. The analysis will help the commercial industries about their product reviews earlier, from their promo video also for youtube creators the trending topics and category who are seeking to get more subscribers. This analysis will help the commercial industries to get the weakness of their product even to improve some of the business strategies and the weakness of the product. For the youtube creators, what are parameters factors are affecting the parameter to trend the video

## Modeling

How to choose a model in rapidminer using Auto model.

With following two factors we can choose an algorithm:

1. Performance: closer the correlation to 1
2. Run time: Less the time required for processing better the model.

The actual representation of data is shown below:

The screenshot displays the RapidMiner Studio interface. The main window shows a data mining project titled "data mining project". The left sidebar lists the project details: "data mining project", "Local Repository/data mining project", "Rows: 7,931", "Columns: 21", and "Last Change: None". The main area displays a table of data with the following columns: video\_id, title, channel\_title, category\_id, tags, views, likes, dislikes, comment\_tot..., and thumbnail\_li... The table contains 7,931 rows of data. The bottom status bar indicates "7,931 rows - 21 columns (15 nominal, 6 numerical)".

video_id	title	channel_title	category_id	tags	views	likes	dislikes	comment_tot...	thumbnail_li...
Category	Category	Category	Number	Category	Number	Number	Number	Number	Category
j12OHQH0HoQ	Live Apple Ev...	Apple Event	28	apple events...	7426393	78240	13548	705	https://i.ytim...
AqokkXoa7uE	Holly and Phill...	This Morning	24	this morningli...	494203	2651	1309	0	https://i.ytim...
YPVcg4SW0z4	My DNA Test ...	emmablackery	24	emmablackery...	142819	13119	151	1141	https://i.ytim...
T_PuZBdT2IM	getting into a ...	ProZD	1	skit[korean]la...	1580028	65729	1529	3598	https://i.ytim...
NsjsmgmbCfc	Baby Name C...	Sprinkleofglitter	26	sprinkleofglitt...	40592	5019	57	490	https://i.ytim...
zZ2CLmvqIXg	REVEALED - F...	Rio Ferdinand...	24	Rio Ferdinand...	317696	9449	135	464	https://i.ytim...
w8fAellnPns	Juicy Chicken ...	You Suck At C...	26	how tocookin...	479291	23935	638	1941	https://i.ytim...
LcZ2AuvvXNA	Nerf Bow Tric...	Dude Perfect	17	dude perfect...	10532409	384841	7547	23849	https://i.ytim...
lamseWekh8Y	Weekend in B...	FleurDeVlog	22	fleurdevlogfle...	65453	2761	33	223	https://i.ytim...
QB8GaO89c8MI	Radiohead - Lift	Radiohead	10	radiohead[lift]...	751743	42272	358	3250	https://i.ytim...
r4RfsfD6ZHM	iPhone leaks (...)	jackfilms	23	jackfilms[lyia...	1121715	65988	591	51025	https://i.ytim...
gFIUXqcOula	Flinch w/ Fifth...	The Late Late ...	24	James Corden...	199102	14916	324	1098	https://i.ytim...



For Sentiment analysis we used python for getting unique video Ids and applied api from R to get the sentiments of each videos.

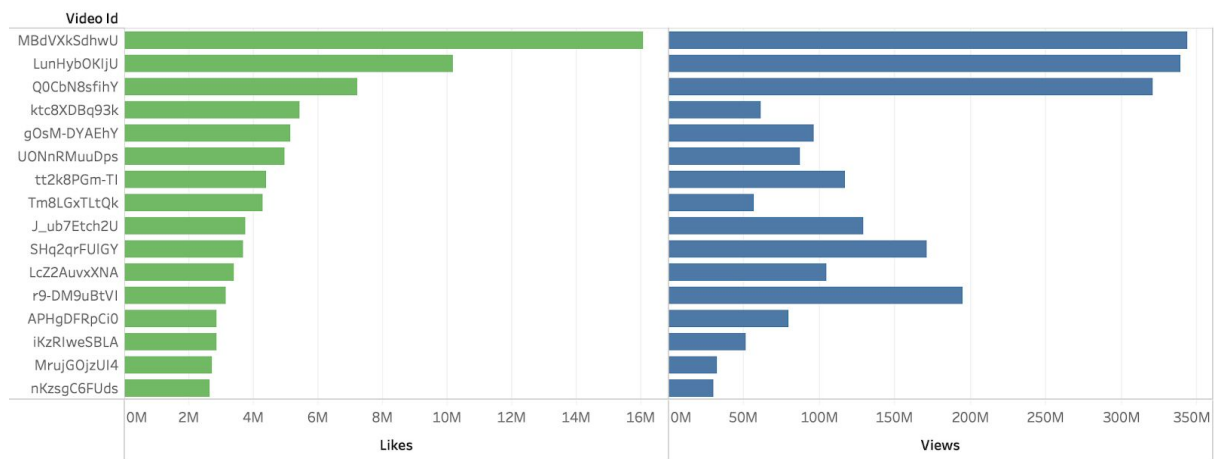
With these we got 7 types of emotions and two sentiments including

Angry, anticipation, disgust, joy, sadness, surprise, positive and negative

## Exploratory Data Analysis Visualizations :

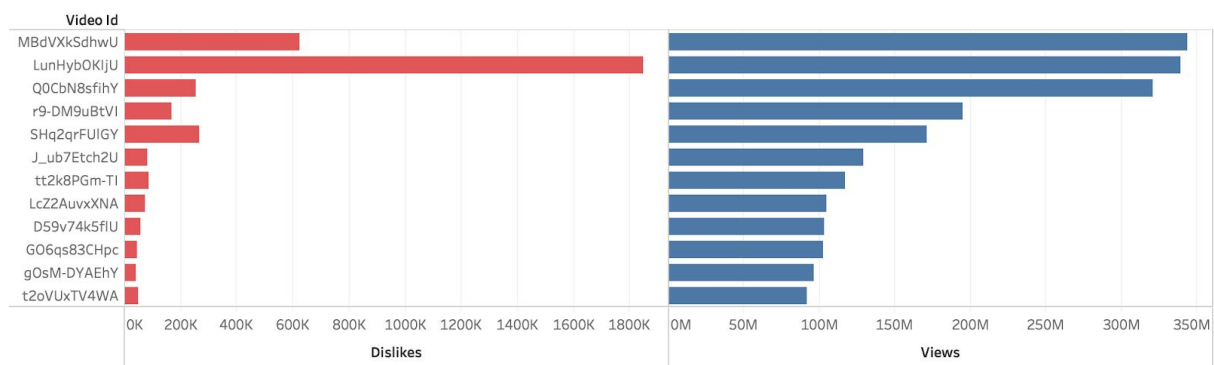
Bar chart of the Likes vs Views:

Total no of likes vs Views



Sum of Likes and sum of Views for each Video Id. The view is filtered on Video Id, which keeps 16 of 1,699 members.

Dislike videos vs Views



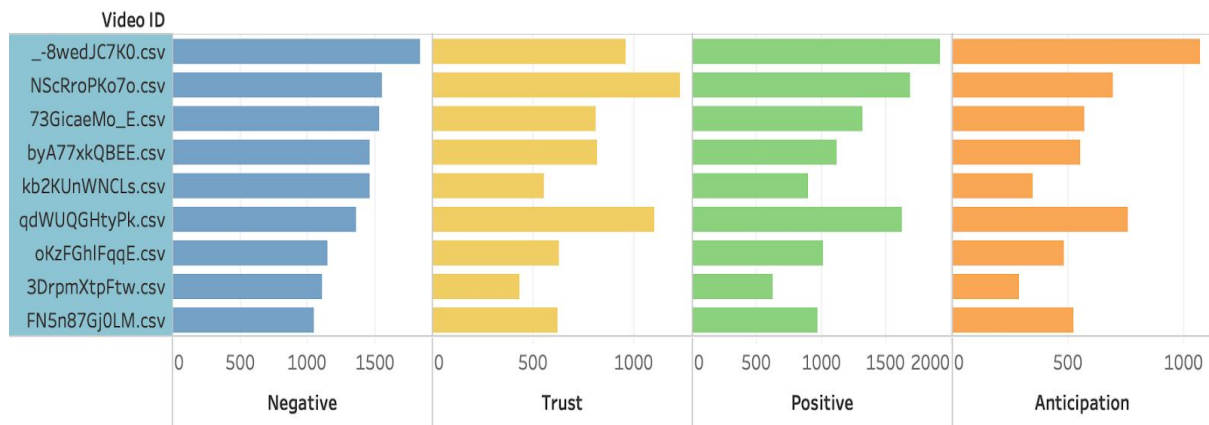
Sum of Dislikes and sum of Views for each Video Id. The view is filtered on Video Id, which keeps 12 of 1,699 members.

## Sentiment Analysis

For Sentiment analysis we used python for getting unique video Ids and applied api from R named `get_nrc_sentiment` from `syuzhet` package to get the sentiments of each videos. With this package we got 7 types of emotions and two sentiments including

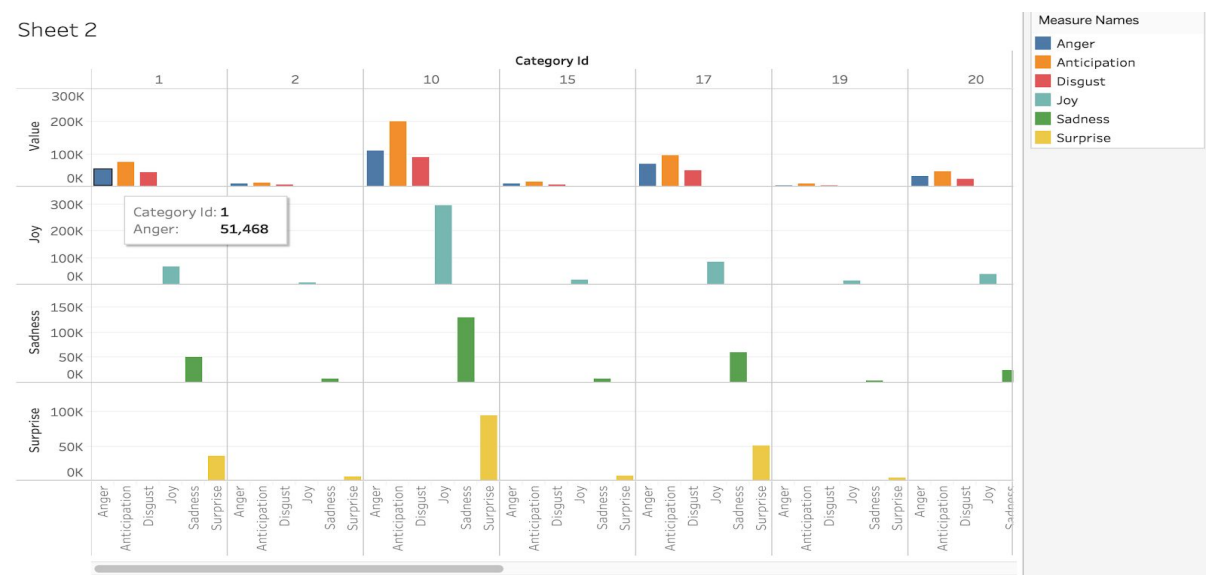
Angry, anticipation, disgust, joy, sadness, surprise, positive and negative

Sheet 1



Sum of Negative, sum of Trust, sum of Positive and sum of Anticipation for each Video ID. The view is filtered on Video ID, which keeps 9 of 1,654 members.

Sheet 2

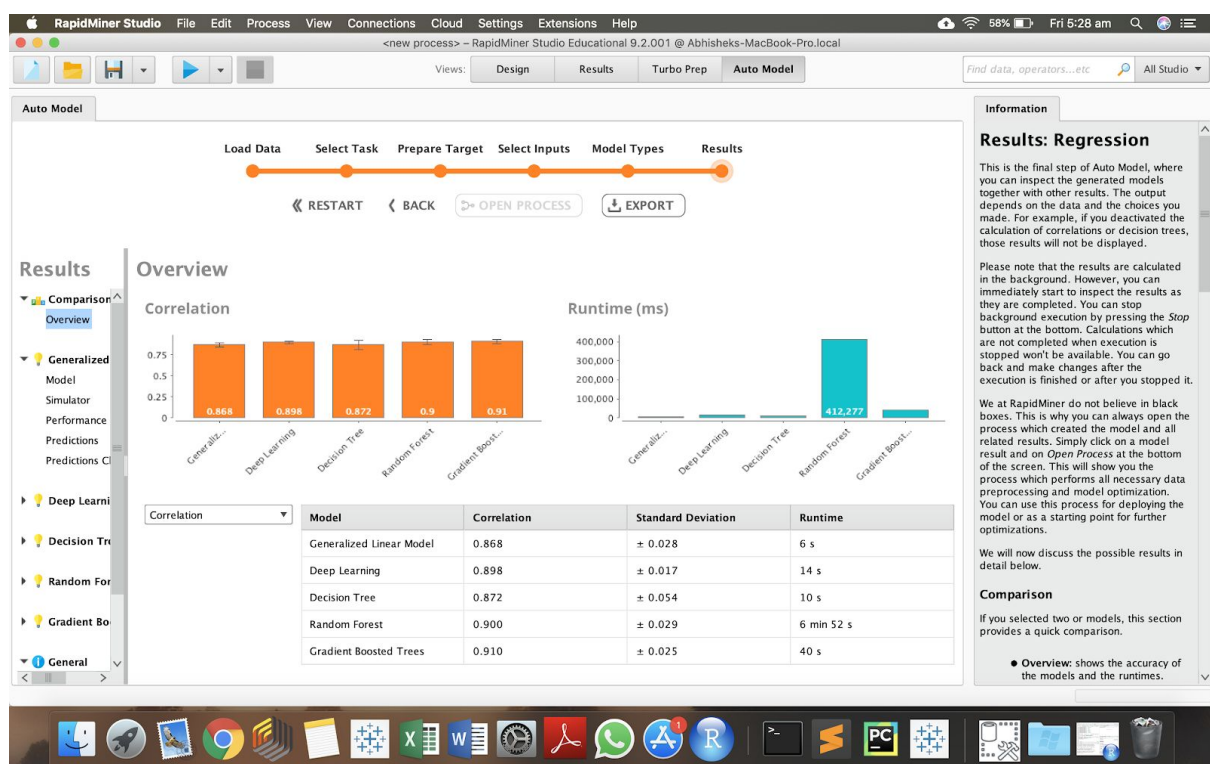


R code:

Auto Model:

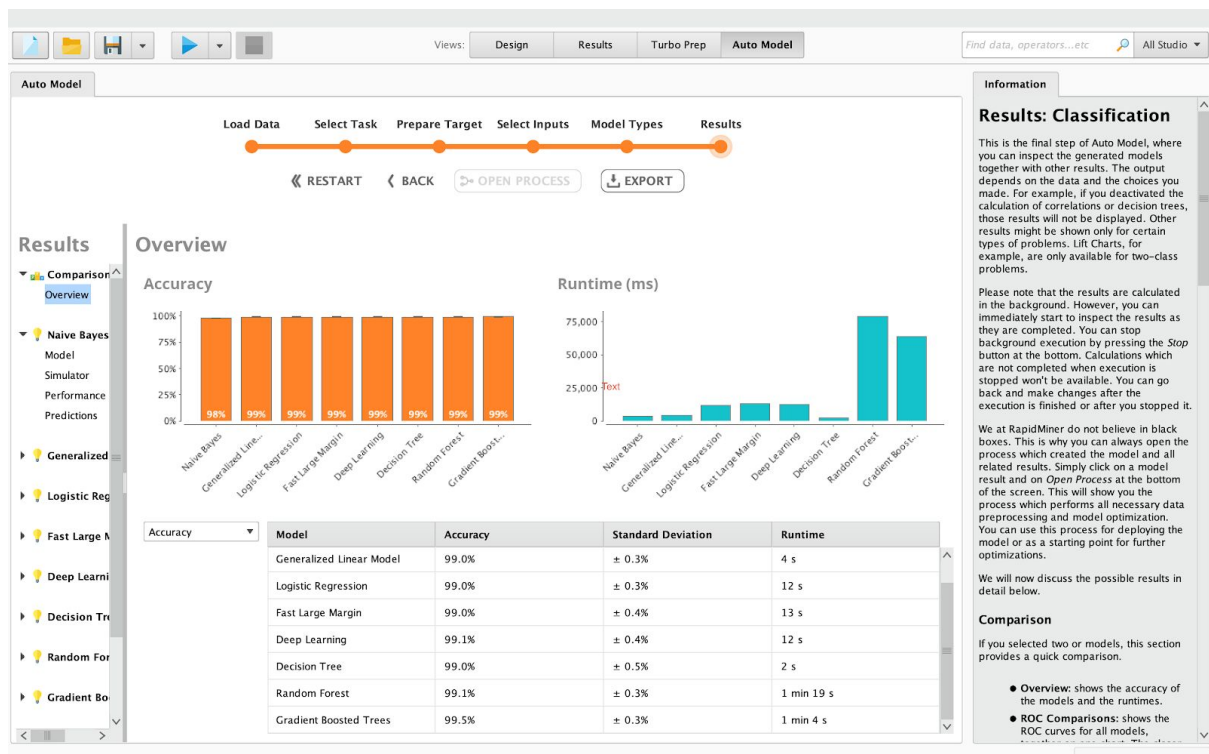
## Modelling and Evaluation

1) Regression:



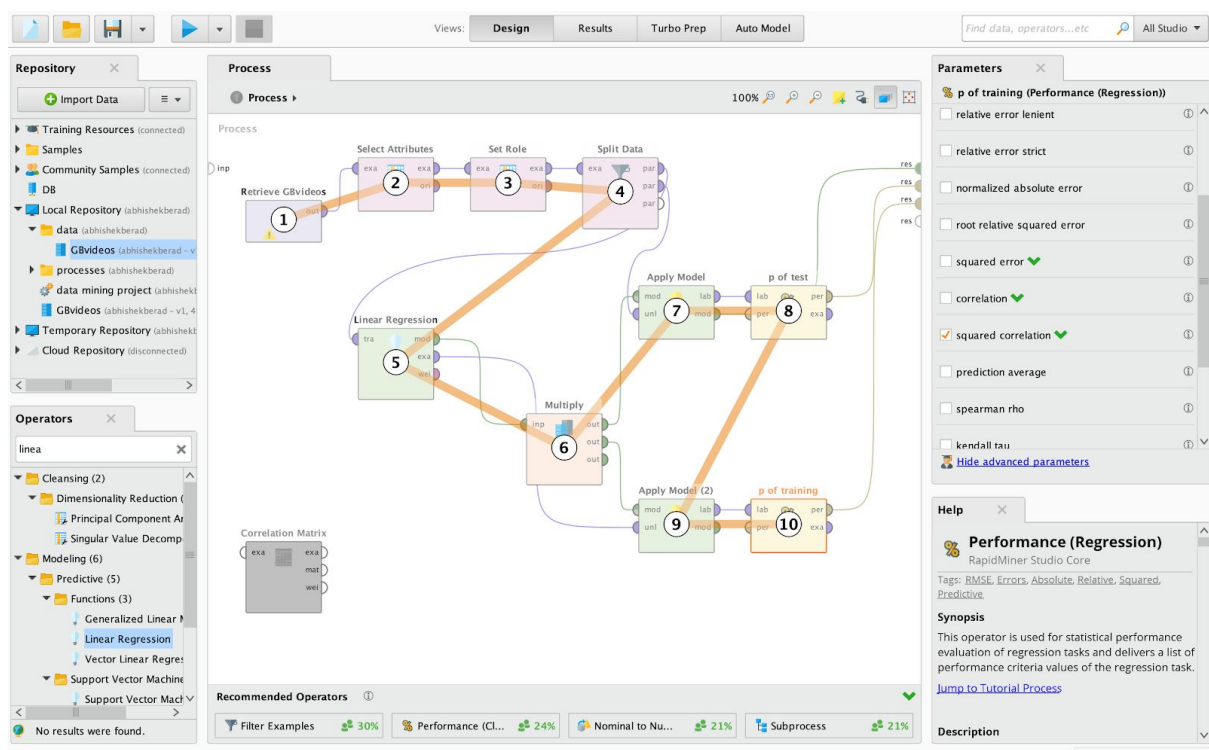
After applying auto model we got better correlation for Gradient boost but the runtime was very low for GLM, so we proceeded with GLM algorithm for our views prediction.

## 2) Classification over regression:

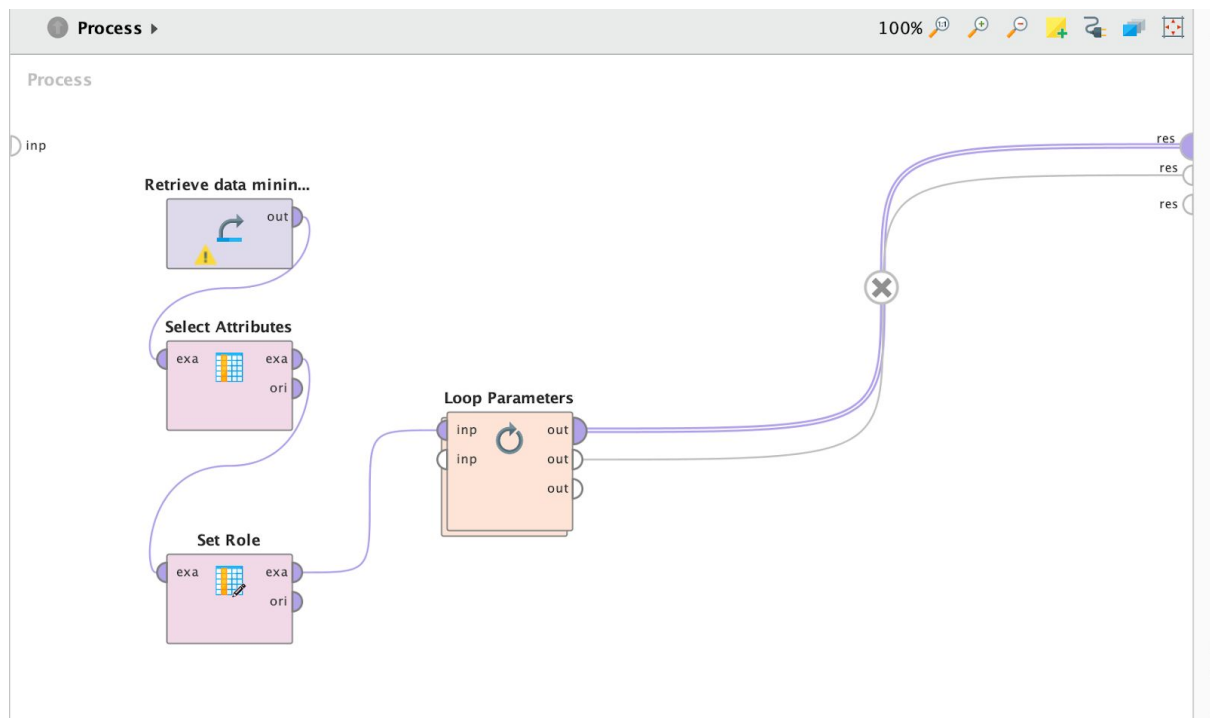


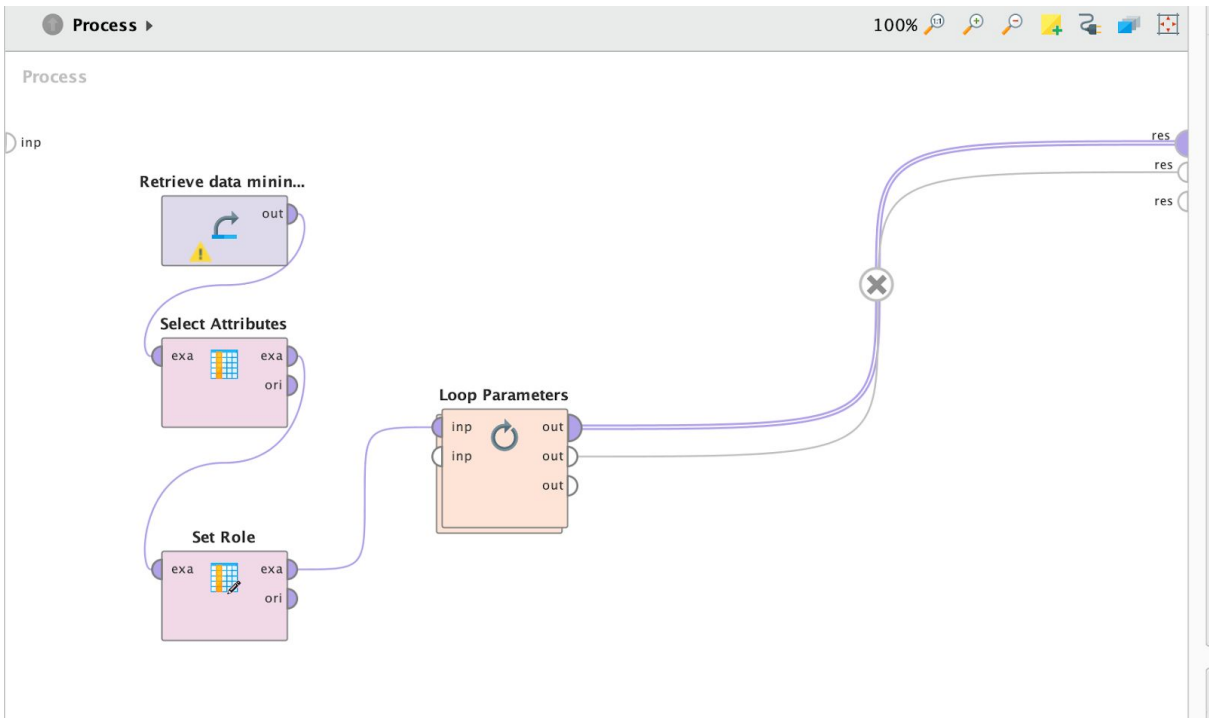
We were not getting accurate prediction when we applied GLM but when we applied classification on our data we got better prediction of the views.

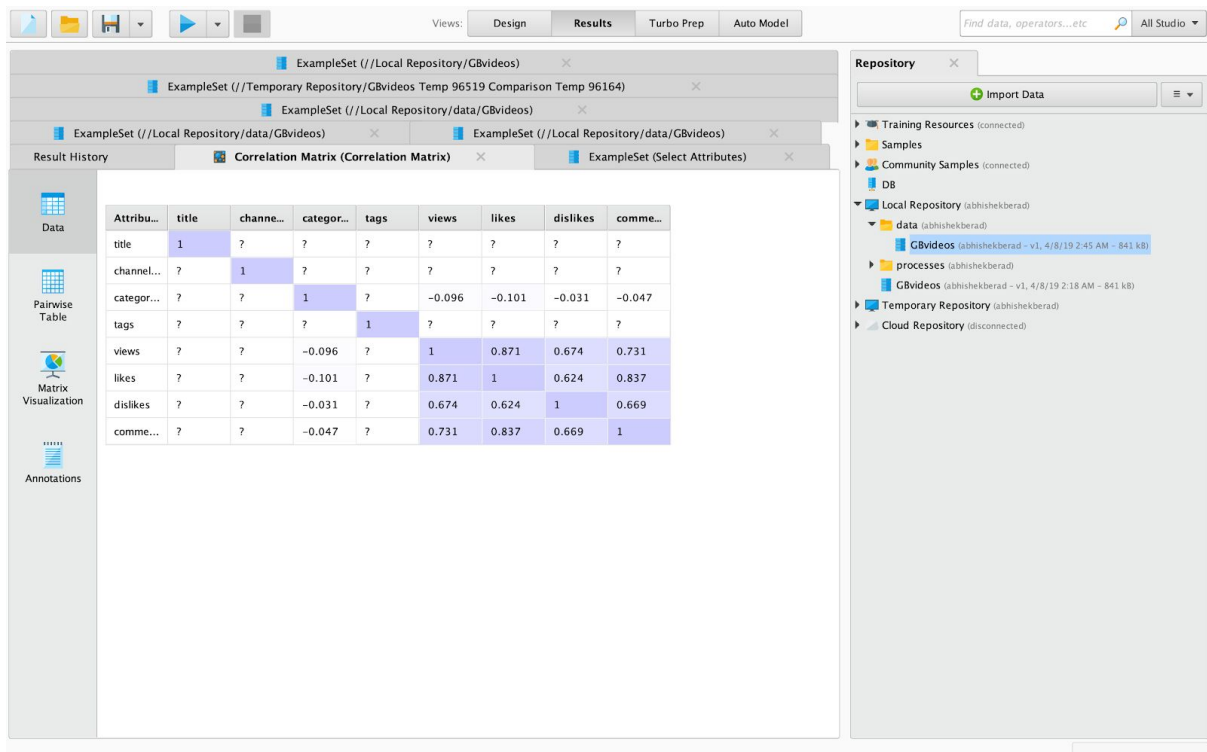
Following is the process we got when we applied auto model:



Following is the process we did in Rapidminer:







This is the final prediction we got after applying GLM without classification

Row No.	views	prediction(...	likes	dislikes	comment_t...
1	40592	268984.355	5019	57	490
2	479291	729845.984	23935	638	1941
3	483360	525432.937	16251	245	1588
4	1691734	1136793.851	39633	1775	5191
5	1936216	1853538.162	74528	1059	6552
6	1701667	1556973.140	53795	8517	36290
7	189389	488768.519	15787	104	2479
8	4659935	743169.527	17509	2542	2157
9	153427	351989.653	8915	91	1080
10	109101	255337.272	4209	109	462
11	4615562	2570896.500	104889	2180	11355
12	218536	415951.414	10659	432	1424
13	2524251	523954.453	17100	676	5311
14	136677	156969.528	490	206	1841
15	214126	375837.499	9278	139	349
16	807265	715826.885	26518	522	5996
17	58169	203580.567	1906	73	198

## Conclusion:

While working on this project we not only got to learn Rapidminer but also we got a chance to explore Databricks and azure blobstore.

We stored a csv file on the azure storage and we wrote a databricks notebook which will fetch the csv and apply GLM application and give the prediction based on the algorithm.