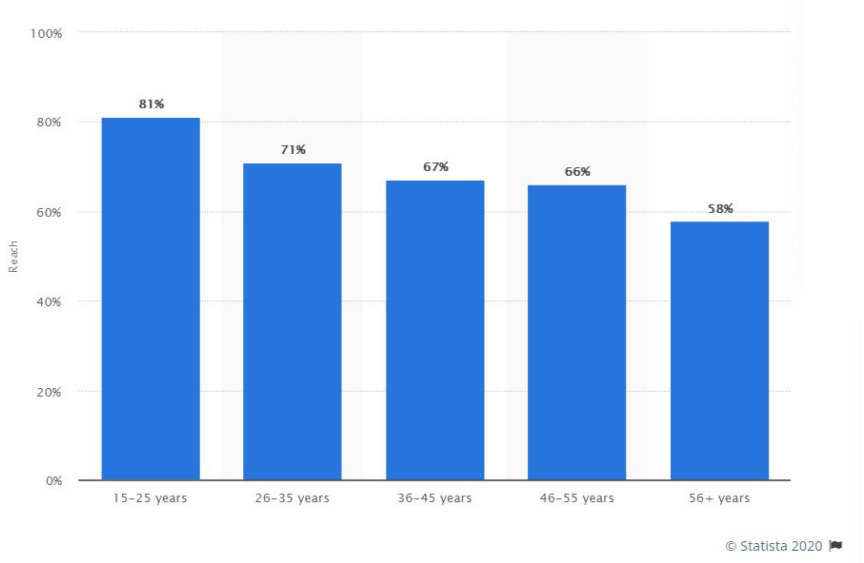
Prediction on Trending YouTube Videos

Kahang Ngau

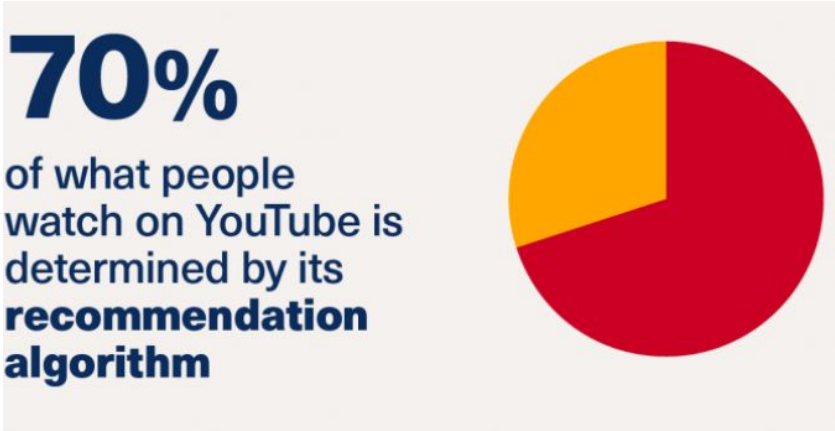
Introduction

- **1 billion** videos are watched on Youtube every single day in U.S.
- Female users are 38% and **male users are 62%.**
- The graph below shows the distribution of different age range of people spend time on YouTube every single day.



Introduction

- YouTube has a list of the top trending videos that measures user's' interaction, such as the number of views/likes/comments/shares.
- In this project, we want to explore among the trending videos, what factor(s) can predict the trending videos get the likes/dislikes/comments.



Methodology

Goals - 1. Use machine learning models to predict ‘likes’.

2. Keep track of models’ performance by conducting ‘RMSE’ and ‘R2’ evaluation.

3. Conduct feature engineering to find the most importance features.

Processes - Python / PySpark

Materials - CSV file downloaded from kaggle.com/YouTube

Technology - data preprocessing, NLP analysis, data visualization, train-test-split data, linear, decision tree, and random forest regression

Data Cleansing & Extraction

Variables to keep - 'publish_year', 'publish_month', 'publish_quarter', 'publish_dayofweek', 'publish_hour', 'category_id', 'views', 'likes', 'dislikes', 'comment_count', 'comments_disabled', 'ratings_disabled', 'video_error_or_removed', 'popular_word'

Variables to Drop - 'video_id', 'trending_date', 'publish_time', 'tag', 'channel_title', 'title', 'description', 'thumbnail_link'

- Extracted the value of year, quarter, month, dayofweek, hour from 'publish_time' column.
- Conducted NLP analysis on tokenizing 'tag', 'title' and 'channel_title' columns.
- Found the top 10 most frequent words.

text	tokens	most_common	popular_word
WE WANT TO TALK ABOUT OUR MARRIAGECaseyNeistat...	[want, talk, marriagecaseyneistatshantell, mar...	want	False
The Trump Presidency: Last Week Tonight with J...	[trump, presidency, last, week, tonight, john,...	trump	False
Racist Superman | Rudy Mancuso, King Bach & Le...	[racist, superman, rudy, mancuso, king, bach, ...	mancuso	False
Nickelback Lyrics: Real or Fake?Good Mythical ...	[nickelback, lyric, real, fake, good, mythical...	nickelback	False
I Dare You: GOING BALD!? nigahiga"ryan"|"higa"|...	[dare, going, bald, nigahiga, ryan, higa, higa...	dare	False

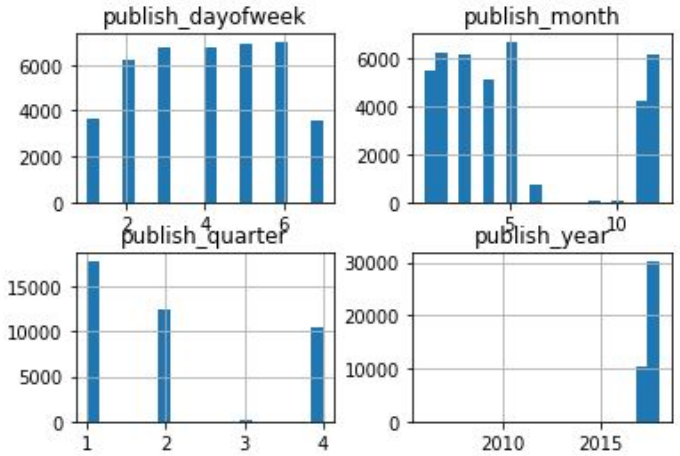
Word	Frequency
makeup	725
late	340
cat	316
trailer	285
news	234
show	221
star	219
movie	207
react	200
black	193

Analysis

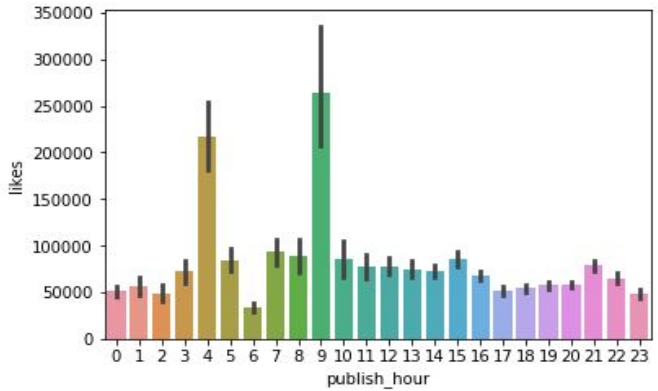
Correlation between all features to likes

```
Correlation to likes for  publish_year 0.06489273235963669
Correlation to likes for  publish_month -0.01689284679274735
Correlation to likes for  publish_quarter -0.014355534245320666
Correlation to likes for  publish_dayofweek 0.021693932429804004
Correlation to likes for  publish_hour -0.04529574054352491
Correlation to likes for  category_id -0.17392077195292174
Correlation to likes for  views 0.8491765212088963
Correlation to likes for  likes 1.0
Correlation to likes for  dislikes 0.4471864632166012
Correlation to likes for  comment_count 0.8030568578359273
Correlation to likes for  comments_disabled -0.028917523269866255
Correlation to likes for  ratings_disabled -0.020888209357161805
Correlation to likes for  video_error_or_removed -0.0026407555837714893
Correlation to likes for  popular_word -0.03281748682245744
```

Distribution of published dayofweek, month, quarter, and year

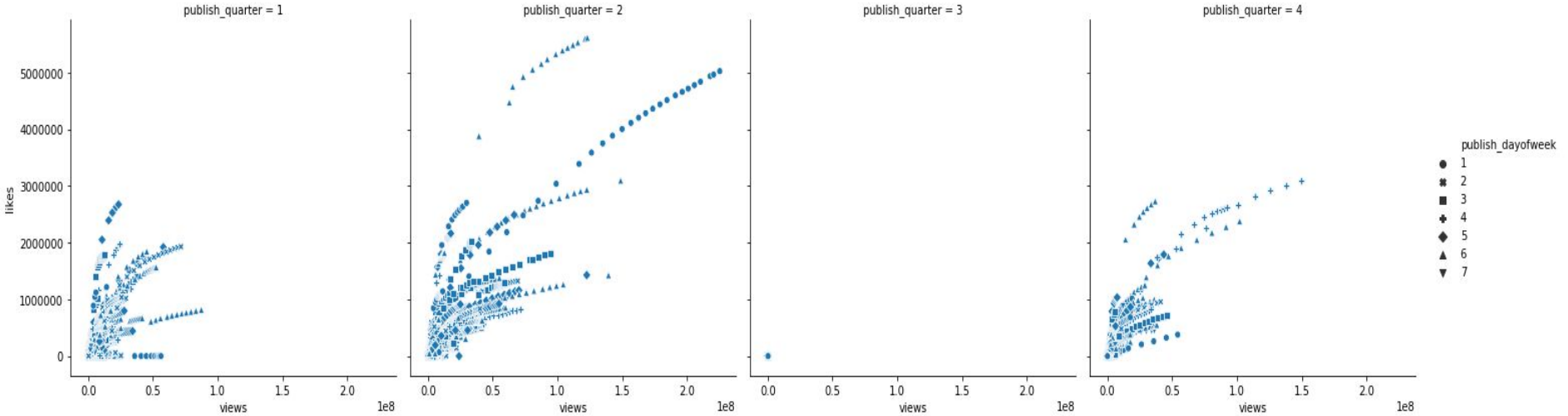


Distribution of publish_hour



More on Analysis

Views vs Likes among dayofweek distributed on quarters

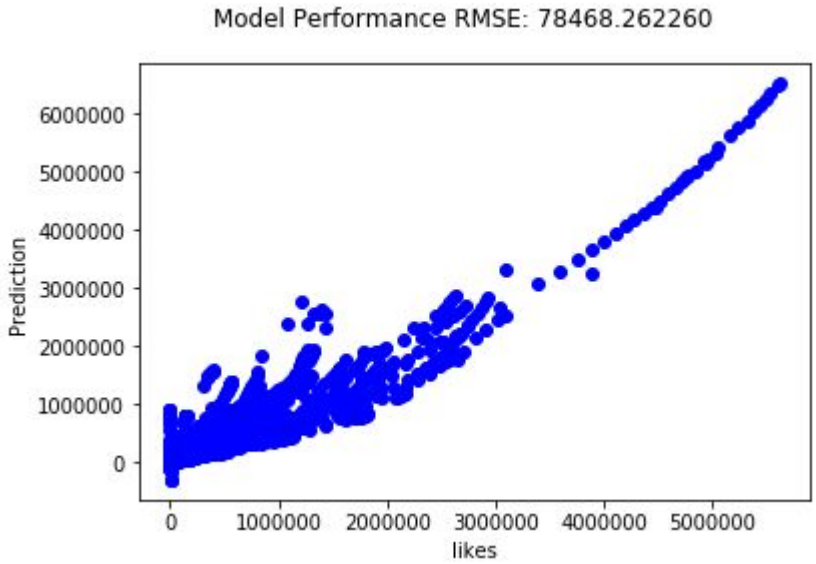


Results - Linear

```
RMSE is 78468.26225960495
R2 is 0.8758320590091677
```

- Converting all boolean type of data into integer type (0 and 1).
- First conducted machine learning model : **Linear Regression Model**

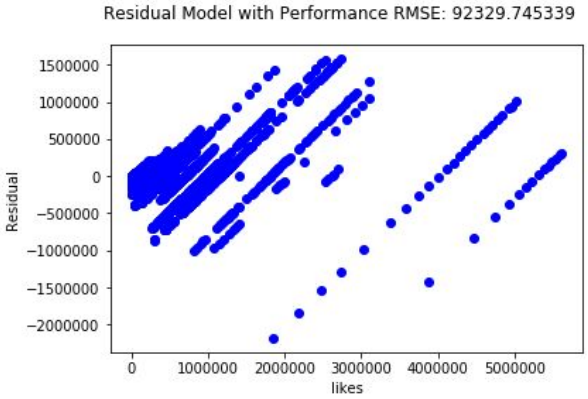
	feature	coefficients
0	publish_year	3364.640779
1	publish_month	2793.543244
3	publish_dayofweek	822.035956
4	publish_hour	191.208605
8	comment_count	3.853985
6	views	0.017658
7	dislikes	-1.967947
5	category_id	-1167.214029
11	video_error_or_removed	-6574.256070
2	publish_quarter	-7266.364985
9	comments_disabled	-8703.393732
12	popular_word	-15261.954468
10	ratings_disabled	-71219.648402

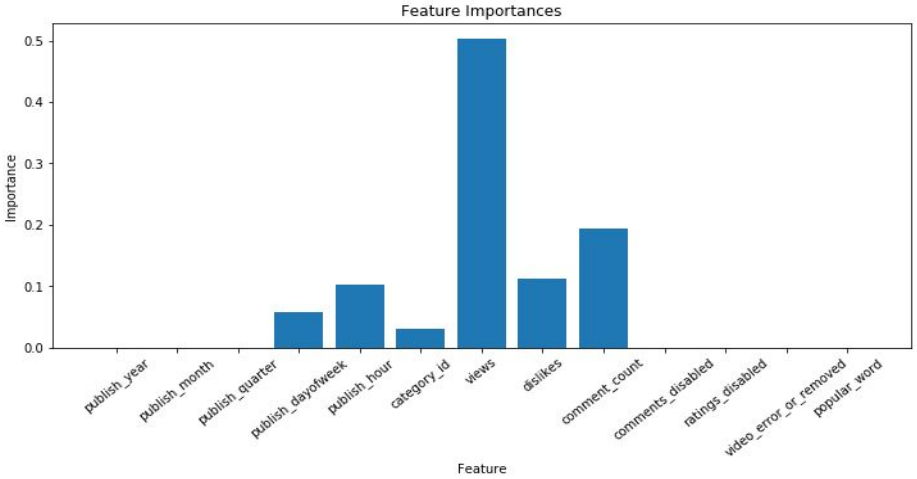


Decision Tree

- Conduct Decision Tree Regression model, with MaxBin 40.

```
RMSE is 92201.52095038704
R2 is 0.8285657546307985
```





	feature	importance
8	comment_count	0.632317
6	views	0.195621
3	publish_dayofweek	0.069893
4	publish_hour	0.043563
5	category_id	0.030750
7	dislikes	0.027855
0	publish_year	0.000000
1	publish_month	0.000000
2	publish_quarter	0.000000
9	comments_disabled	0.000000
10	ratings_disabled	0.000000
11	video_error_or_removed	0.000000
12	popular_word	0.000000

Decision Tree - Hyperparameter Tuning

- Conduct Hyperparameter Tuning on setting the ParamGrid and Cross Validation.

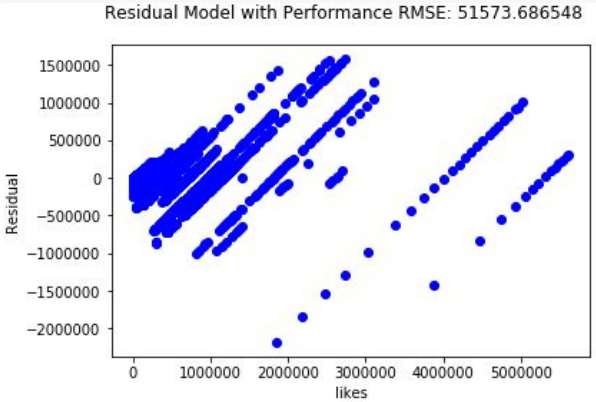
```
paramGrid = (ParamGridBuilder()
             .addGrid(dt.maxDepth, [2, 5, 10])
             .addGrid(dt.maxBins, [10, 20, 40, 80, 100])
             .build())

cv = CrossValidator(estimator=dt, evaluator=evaluator, estimatorParamMaps=paramGrid,
                    numFolds=3, parallelism=4, seed=345)
```

```
RMSE is 51573.68654842105
R2 is 0.9463613594300854
```

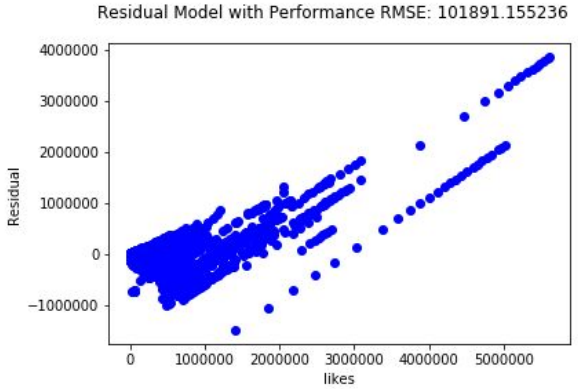
Best combination:

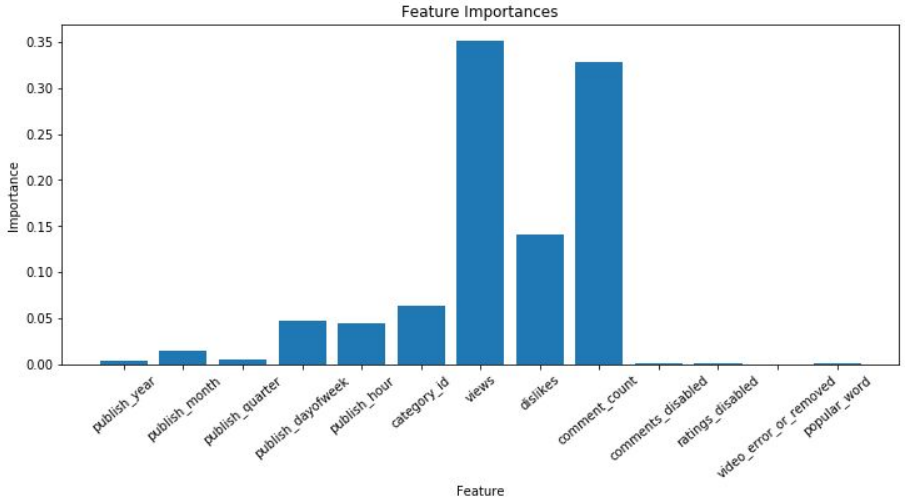
- ❏ MaxDepth: 10
- ❏ MaxBins: 80



Random Forest

RMSE is 101891.15523644455
R2 is 0.79063967577177





	feature	importance
6	views	0.351620
8	comment_count	0.327768
7	dislikes	0.140892
5	category_id	0.063606
3	publish_dayofweek	0.046731
4	publish_hour	0.044798
1	publish_month	0.014333
2	publish_quarter	0.005496
0	publish_year	0.003685
12	popular_word	0.000699
9	comments_disabled	0.000193
10	ratings_disabled	0.000179
11	video_error_or_removed	0.000000

Random Forest - Hyperparameter Tuning

- Conduct Hyperparameter Tuning on setting the ParamGrid and Cross Validation.

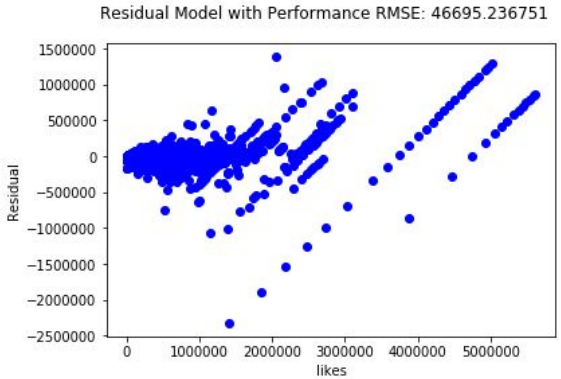
```
paramGrid_rf = (ParamGridBuilder()
            .addGrid(rf.numTrees, [10, 20, 50])
            .addGrid(rf.maxDepth, [5, 10, 15])
            .build())

# apply cross validation with numFolds=3, parallelism=4.
cv_rf = CrossValidator(estimator=rf, evaluator=evaluator, estimatorParamMaps=paramGrid_rf,
                   numFolds=2, parallelism=2, seed=345)
```

```
RMSE is 46695.23675105105
R2 is 0.9560289766693215
```

Best combination:

- ❏ numTrees: 50
- ❏ maxDepth: 15



Conclusion

- Random Forest Regression model after hyperparameter tuning does the best among other models.
- Two features appeared as important on all three models: views & comment_count.
- Based on the Linear regression model, we can see the strong positive correlation(cor>.6) between several variables: likes & views, comment_count & views, likes & comment_count, comment_count & dislikes.
- Friday as the day of the week when videos get the most views, so if you post a video on that day, the chance of the video being seen is relatively higher than you post on other days of the week.