Web Analytics
Online New Popularity

Advances in Data Science and Architecture

Prof. Srikanth Krishnamurthy

Team 11

Krutika Dedhia
Kinjal Gada
Ankur Vora

Northeastern University

INTRODUCTION

In this internet era, reading and sharing information have become the center of people's entertainment lives. Web Analytics is integral part of any online marketing plan. Analyzing your traffic and then finding ways to improve on it is the name of the game. These analytics that are tracked allow you to measure important information like sales and conversions, clicks, and page views. One can use web analytics applications to tailor website's content in order to make it more appealing to visitors or the type of people you want to visit your site! We have narrowed our Web Analytics domain to News Popularity. The concept of online news has been around much longer than the 90's Just because something is technologically feasible doesn't mean it will accepted/demanded. The demand stems from the quality of content whereas popularity of the news depends on various other factors like way of demonstrating, positivity, negativity, catchy title, no of shares, author, channel, topic etc. The need of web analytics arises here. It would allow us to accurately predict the popularity of news prior to its publication, for social media workers (authors, advertisers, etc). For the purpose of this paper, we intend to make use of a largely and recently collected dataset of news popularity with over 39000 articles from Mashable website, to first select informative features and then analyze and compare the performance of several machine learning algorithms

DATASOURCE LINK

https://archive.ics.uci.edu/ml/datasets/Online+News+Popularity

FEEL OF THE DATASET

Dataset gives details of each post which consists of 61 features. Details pertaining to posts on Mashable Website includes date, href details, positive/negative polarity of its over all post, sentimental polarity, title polarity, number of tokens in title, number of keywords, and so on. On analyzing the dataset, data cleaning was done wherever required, unwanted columns were deleted, new features were scraped and various machine learning algorithms were implemented with expanding to visualization in tableau.

Original Dataset:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	url	timedelta	n_tokens	n_tokens	n_unique	n_non_st	n_non_st	num_hre	num_self	num_img	num_vide	average_	num_key	data_char	data_char	data_char	data_char	data_char	data_char	kw_min_	kw_m
2	http://ma	731	12	219	0.663594	1	0.815385	4	2	1	0	4.680365	5	0	1	0	0	0	0	0	
3	http://ma	731	9	255	0.604743	1	0.791946	3	1	1	0	4.913725	4	0	0	1	0	0	0	0	
4	http://ma	731	9	211	0.57513	1	0.663866	3	1	1	0	4.393365	6	0	0	1	0	0	0	0	
5	http://ma	731	9	531	0.503788	1	0.665635	9	0	1	0	4.404896	7	0	1	0	0	0	0	0	
6	http://ma	731	13	1072	0.415646	1	0.54089	19	19	20	0	4.682836	7	0	0	0	0	1	0	0	
7	http://ma	731	10	370	0.559889	1	0.698198	2	2	0	0	4.359459	9	0	0	0	0	1	0	0	
8	http://ma	731	8	960	0.418163	1	0.549834	21	20	20	0	4.654167	10	1	0	0	0	0	0	0	
9	http://ma	731	12	989	0.433574	1	0.572108	20	20	20	0	4.617796	9	0	0	0	0	1	0	0	
10	http://ma	731	11	97	0.670103	1	0.836735	2	0	0	0	4.85567	7	0	0	0	0	1	0	0	
11	http://ma	731	10	231	0.636364	1	0.797101	4	1	1	1	5.090909	5	0	0	0	0	0	1	0	
12	http://ma	731	9	1248	0.49005	1	0.731638	11	0	1	0	4.617788	8	0	0	0	0	0	1	0	
13	http://ma	731	10	187	0.666667	1	0.8	7	0	1	0	4.657754	7	1	0	0	0	0	0	0	
14	http://ma	731	9	274	0.609195	1	0.707602	18	2	11	0	4.233577	8	0	0	0	0	0	0	0	
15	http://ma	731	9	285	0.744186	1	0.84153	4	2	0	21	4.34386	6	0	0	0	0	0	0	0	
16	http://ma	731	8	259	0.562753	1	0.644444	19	3	9	0	5.023166	7	0	0	0	0	0	0	0	
17	http://ma	731	12	682	0.459542	1	0.634961	10	0	1	0	4.620235	6	0	0	0	0	0	1	0	
18	http://ma	731	8	1118	0.512397	1	0.70977	26	18	12	1	4.703936	5	0	0	0	0	0	0	0	
19	http://ma	731	8	397	0.624679	1	0.805668	11	0	1	0	5.445844	6	0	0	1	0	0	0	0	
20	http://ma	731	11	103	0.68932	1	0.806452	3	1	1	0	4.84466	6	1	0	0	0	0	0	0	
21	http://ma	731	8	1207	0.410579	1	0.548969	24	24	42	0	4.716653	8	0	0	0	0	1	0	0	
22	http://ma	731	13	1248	0.390638	1	0.523388	21	19	20	0	4.686699	10	0	0	0	0	1	0	0	
23	http://ma	731	9	391	0.510256	1	0.65	9	2	1	1	5.296675	7	0	0	0	0	0	1	0	

OnlineNewsPopularity

Ready

OnlineNews.R

Source on Save

Run

Source

```r
library(RCurl)
library(XML)
library(stringr)

# read csv file
raw_data <- read.table("C:\\Users\\user\\Downloads\\ADS\\Final Project\\Dataset\\OnlineNewsPopularity\\
                       OnlineNewsPopularity\\OnlineNewsPopularity.csv", sep = ",",header=T, check.names = FALSE, stringsAsFactors = FALSE)

# Data Cleaning by removing unwanted columns
raw_data <- subset(raw_data, select = c(-n_unique_tokens, - n_non_stop_words,
                                        - n_non_stop_unique_tokens, - kw_min_min,- kw_max_min,
                                        - kw_avg_min, - kw_min_max,- kw_max_max,- kw_avg_max,
                                        - kw_min_avg,- kw_max_avg, - kw_avg_avg, - weekday_is_monday,
                                        - weekday_is_tuesday,- weekday_is_wednesday, - weekday_is_thursday,
                                        - weekday_is_friday, - weekday_is_saturday, - weekday_is_sunday,
                                        - is_weekend, - LDA_00, - LDA_01, - LDA_02, -  LDA_03,
                                        -  LDA_04))

# Extracting Date from URL field and make it in standard format
raw_data$Date <- strsplit(raw_data$url,"/")
raw_data$Year <- sapply(raw_data$Date, "[[", 4)
raw_data$Month <- sapply(raw_data$Date, "[[", 5)
raw_data$Day <- sapply(raw_data$Date, "[[", 6)
raw_data$Date <- paste(raw_data$Year,raw_data$Month,raw_data$Day,sep="")
raw_data$weekday <- weekdays(as.Date(raw_data$Date,'%Y%m%d'))

# Assign values to each day of week Sunday->0...Saturday->6)
raw_data$weekday <- as.POSIXlt(as.Date(raw_data$Date,'%Y%m%d'))$wday

# Find if day falls in weekend or no
raw_data$isWeekend <- ifelse(raw_data$weekday %in% c(0,6),1,0)

# Handling missing values for Type field to 'Others'
raw_data$Type <- ifelse(raw_data$data_channel_is_bus==1, 1,
                        ifelse(raw_data$data_channel_is_lifestyle==1, 2,
```

7:24 (Top Level) R Script

```r
                                ifelse(raw_data$data_channel_is_entertainment==1, 3,
                                       ifelse(raw_data$data_channel_is_socmed==1, 4,
                                              ifelse(raw_data$data_channel_is_tech==1, 5,
                                                     ifelse(raw_data$data_channel_is_world==1, 6, 7))))))

# Removing unwanted columns
raw_data <- subset(raw_data, select = c(- data_channel_is_lifestyle, - data_channel_is_entertainment, -  data_channel_is_bus, - data_channe

raw_data$shares1 <- log(raw_data$shares)
hist(raw_data$shares1)
# Write dataset to csv file
write.csv (raw_data, "C:\\Users\\user\\Downloads\\ADS\\Final Project\\Dataset\\NewsDataset_final.csv")

ds <- read.csv("C:\\Users\\user\\Downloads\\ADS\\Final Project\\Dataset\\NewsDataset_final.csv", stringsAsFactors = FALSE)

siteData <- function(N) {
  html<-getURL(N)
  doc = htmlParse(html, asText = TRUE)
  scraped<-c(url=N, type=xpathSApply(doc, "//*[@id='main']/div[1]/div/hgroup/h2", xmlValue))
  scraped<-c(scraped, topic=xpathSApply(doc, "//*[@id='main']/div[1]/div/div/div[2]/div/article/footer", xmlValue))
  scraped<-c(scraped, author=xpathSApply(doc, "//*[@id='main']/div[1]/div/div/div[2]/div/article/header/div[2]/span/span/a", xmlValue))
}
scrapedDf0<-lapply(ds[1:4000,2], siteData)
workWithScraped0<-scrapedDf0
res <- as.data.frame(t(stri_list2matrix(workWithScraped0)))
res<-res[,-5]
colnames(res) <- c("url", "tpye", "topics", "author")
res[,"topics"]<-gsub("\n","",res$topics)
res[,"topics"]<-gsub("Topics:","",res$topics)
res[,"topics"]<-trimws(res$topics)

scrapedDataSet<-merge(ds, res, by = "url")
scrapedDataSet<-scrapedDataSet[,-2]
scrapedDataSet1<-na.omit(scrapedDataSet)
write.csv (scrapedDataSet, "C:\\Users\\user\\Downloads\\ADS\\Final Project\\Dataset\\NewsDataset_final2.csv")
```

42:37 (Top Level) ÷ R Script ÷

After scraping, and converting in machine readable format:

B1 url

	B	C	L	M	N	R	S	T	Y	Z	AA	AB	AC	AD	AE	AF	AM	AN	AO
1	url	timedelta	self_refer	self_refer	self_refer	global_rat	rate_posit	rate_nega	min_nega	max_nega	title_subj	title_sent	abs_title_	abs_title_	shares	Date	tpye	topics	author
2	http://ma	731	496	496	496	0.013699	0.769231	0.230769	-0.6	-0.2	0.5	-0.1875	0	0.1875	593	20130107	Media	amazon, Entertain	Lauren Indvik
3	http://ma	731	0	0	0	0.015686	0.733333	0.266667	-0.125	-0.1	0	0	0.5	0	711	20130107	Media	Business, Media, T	Seth Fiegerman
4	http://ma	731	918	918	918	0.009479	0.857143	0.142857	-0.8	-0.13333	0	0	0.5	0	1500	20130107	Business	Apple, apps, Apps	Seth Fiegerman
5	http://ma	731	545	16000	3151.158	0.012127	0.860215	0.139785	-0.5	-0.05	0.454545	0.136364	0.045455	0.136364	505	20130107	Tech	apps, Apps and So	Emily Price
6	http://ma	731	8500	8500	8500	0.027027	0.52381	0.47619	-0.4	-0.1	0.642857	0.214286	0.142857	0.214286	855	20130107	Tech	CES, Gadgets, Mob	Lance Ulanoff
7	http://ma	731	545	16000	3151.158	0.016667	0.827957	0.172043	-0.5	-0.05	0	0	0.5	0	556	20130107	Gadgets	bodymedia, CES, H	Andrea Smith
8	http://ma	731	545	16000	3151.158	0.015167	0.846939	0.153061	-0.5	-0.05	1	0.5	0.5	0.5	891	20130107	Tech	Canon, CES, Gadge	Pete Pachal
9	http://ma	731	0	0	0	0.020619	0.6	0.4	-0.125	-0.125	0.125	0	0.375	0	3600	20130107	Tech	Cars, car of the fut	Stan Schroeder
10	http://ma	731	0	0	0	0.030303	0.5625	0.4375	-0.5	-0.1	0	0	0.5	0	710	20130107	U.S.	Politics, U.S., Worl	Alex Fitzpatrick
11	http://ma	731	10700	16200	13450	0.029197	0.636364	0.363636	-0.6	-0.05	1	-1	0.5	1	823	20130107	Watercoo	Alt Image Lead, Cc	Eric Larson
12	http://ma	731	770	22800	11785	0.052632	0.347826	0.652174	-0.7	-0.1	1	-1	0.5	1	10000	20130107	Watercoo	jokes, Video, Vide	Neha Prakash
13	http://ma	731	4800	4800	4800	0.011583	0.571429	0.428571	-0.1875	-0.05	0.75	0.55	0.25	0.55	761	20130107	Watercoo	Downton Abbey, g	Annie Colbert
14	http://ma	731	555	14000	3904.625	0.025939	0.613333	0.386667	-1	-0.05	0.566667	-0.1	0.066667	0.1	13600	20130107	Watercoo	Video, Videos, vir	Neha Prakash
15	http://ma	731	5000	5000	5000	0.009709	0.75	0.25	-0.125	-0.125	0.857143	-0.71429	0.357143	0.714286	5700	20130107	Marketing	Facebook, Health	Lauren Indvik
16	http://ma	731	545	16000	2829.542	0.011599	0.857143	0.142857	-0.5	-0.05	0.5	0	0	0	17100	20130107	Tech	Bluetooth speake	Samantha Murphy
17	http://ma	731	545	16000	3151.158	0.011218	0.858586	0.141414	-0.5	-0.05	0	0	0.5	0	2800	20130107	Tech	apps, Apps and So	Emily Price
18	http://ma	731	704	704	704	0.005115	0.777778	0.222222	-0.16667	-0.05	0	0	0.5	0	598	20130107	U.S.	Politics, federal tr	Alex Fitzpatrick
19	http://ma	731	545	16000	3151.158	0.016464	0.819048	0.180952	-0.5	-0.05	0	0	0.5	0	445	20130107	Tech	CES, Fujifilm, supe	Pete Pachal
20	http://ma	731	16100	16100	16100	0	1	0	0	0	0.45	0.4	0.05	0.4	1500	20130107	U.S.	Hillary Clinton, Ph	Alex Fitzpatrick
21	http://ma	731	2800	2800	2800	0.02459	0.4	0.6	-1	-0.125	0.7	-0.4	0.2	0.4	852	20130107	Business	Business, HTC, Mo	Todd Wasserman
22	http://ma	731	924	924	924	0.011278	0.5	0.5	-0.2	-0.05	0.9	0.4	0.4	0.4	783	20130107	Mobile	ascend d2, ascend	Samantha Murphy
23	http://ma	731	2500	2500	2500	0.009063	0.869565	0.130435	-0.5	-0.125	0	0	0.5	0	1500	20130107	Music	apps, Apps and So	Samantha Murphy

NewsDataset_final2

PROCEDURE

1. Data collection - Identify issues and/or opportunities for collecting data.
2. Data cleansing.
3. Data scraping - Scrape the data which was not available with original dataset which will help in useful analysis.
4. Data organization - Make or convert data into machine readable format.
5. Feature Selection - Select most affected features for prediction, classification and clustering in Azure
6. Data prediction - Perform prediction of number of shares for a given post and analyze results with using algorithms such as Two Class Decision Tree, Random Forest , Neural Network, and Poisson Regression.
7. Data classification - Classify all the news into “High Popular” and “Less Popular” class based on inputs using classification algorithms such as Random Forest, Two Class Decision Tree and Neural Network classification.
8. Data Clustering - Cluster the dataset into different clusters using K-means and Hierarchical Clustering algorithms.
9. Data Analysis
10. Visualization

Our dataset is provided by UCI machine learning repository, originally acquired and preprocessed by K.Fernandes et al. It extracts total of 39645 articles published in the years of 2013 and 2014 from Mashable website.

REGRESSION

Using Azure, implemented and analyzed regression algorithms to predict number of shares after feature selection on developed dataset. Algorithms which are implemented are Decision Forest (Random Forest), Two-Class Decision Tree Regression, Neural Network Regression, Poisson Regression . Based on least RMSE value.

Due to wide range of shares value, we also analyzed results after normalizing shares by taking the natural logarithm. However, results of RMSE were better predicted without normalization.

Feature Selected for regression

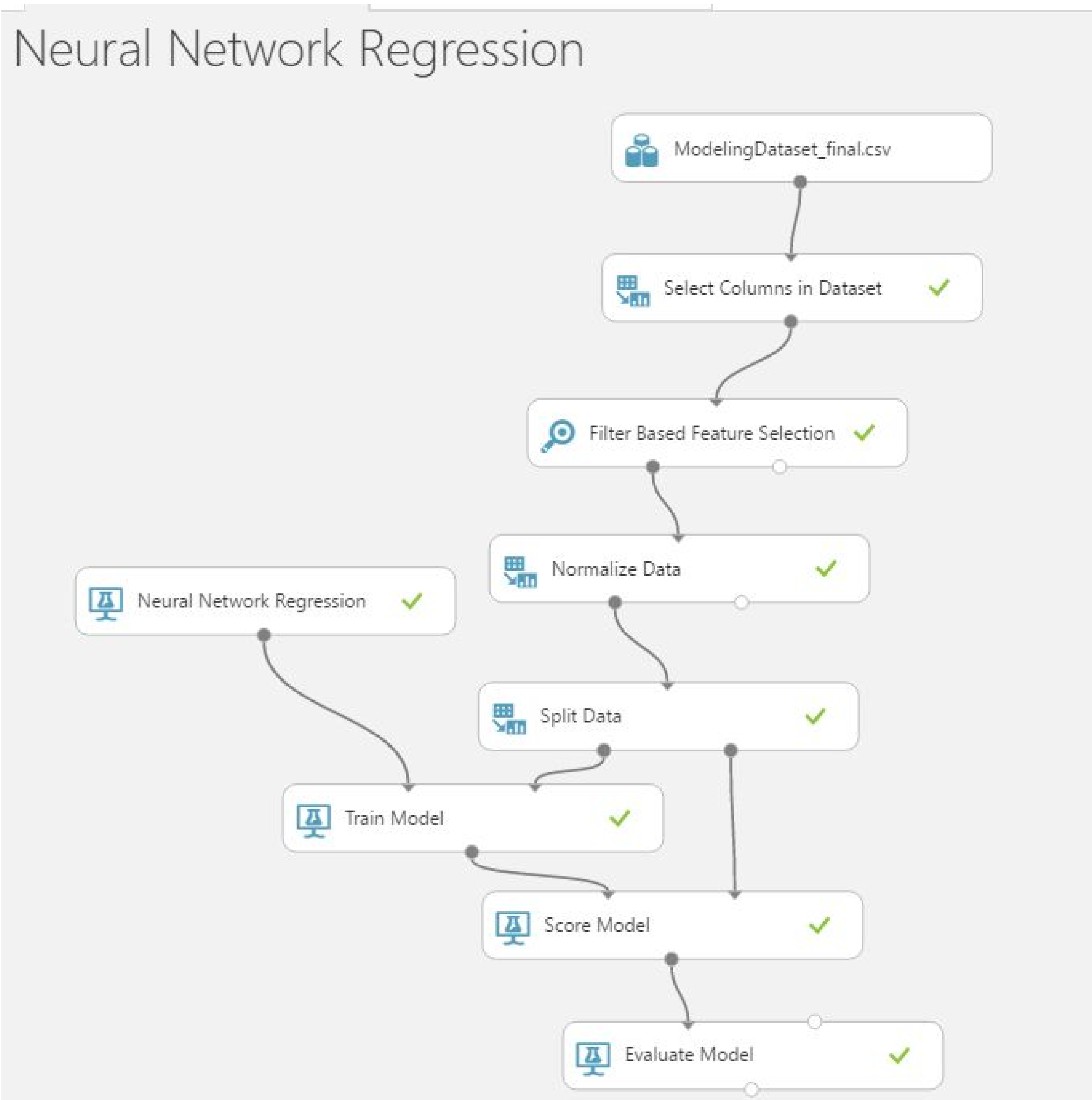
Boosted Decision Tree Regression › Filter Based Feature Selection › Filtered dataset

rows 39644 columns 10

shares	num_hrefs	num_imgs	Type	num_videos	num_keywords	isWeekend	Date	n_tokens_title	n_tokens_content
593	4	1	3	0	5	0	20130107	12	219
711	3	1	1	0	4	0	20130107	9	255
1500	3	1	1	0	6	0	20130107	9	211
1200	9	1	3	0	7	0	20130107	9	531
505	19	20	5	0	7	0	20130107	13	1072
855	2	0	5	0	9	0	20130107	10	370
556	21	20	2	0	10	0	20130107	8	960
891	20	20	5	0	9	0	20130107	12	989
3600	2	0	5	0	7	0	20130107	11	97
710	4	1	6	1	5	0	20130107	10	231
2200	11	1	6	0	8	0	20130107	9	1248
1900	7	1	2	0	7	0	20130107	10	187
823	18	11	7	0	8	0	20130107	9	274
10000	4	0	7	21	6	0	20130107	9	285

On analyzing the results for each algorithm Random Forest gives the lowest RMSE value. Below is the big picture of each algorithm implemented.

Neural Network Regression



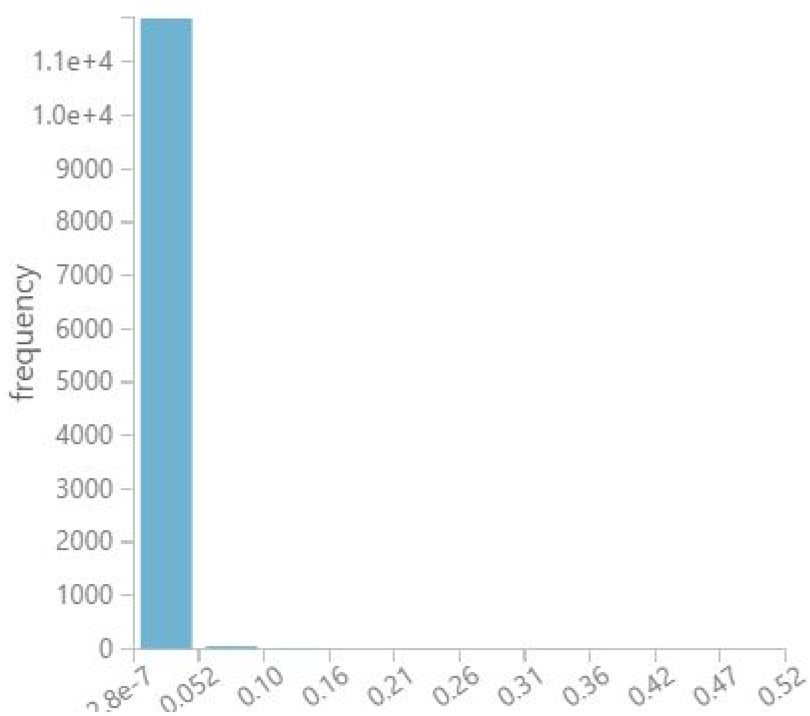
Result

Neural Network Regression › Evaluate Model › Evaluation results

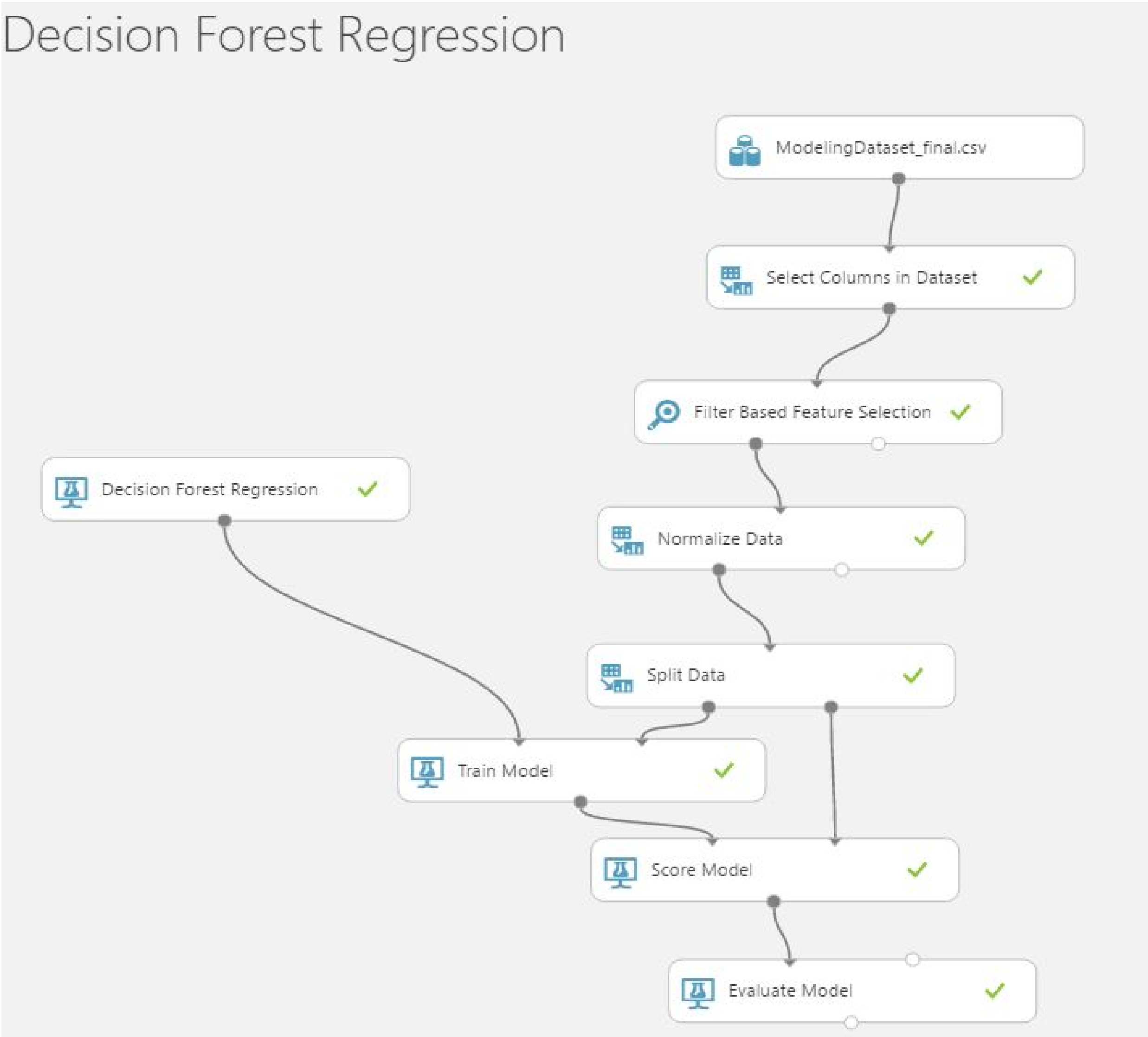
Metrics

Mean Absolute Error	0.002958
Root Mean Squared Error	0.011238
Relative Absolute Error	0.810867
Relative Squared Error	1.046367
Coefficient of Determination	-0.046367

Error Histogram



Decision Forest Regression (Random Forest)



Result

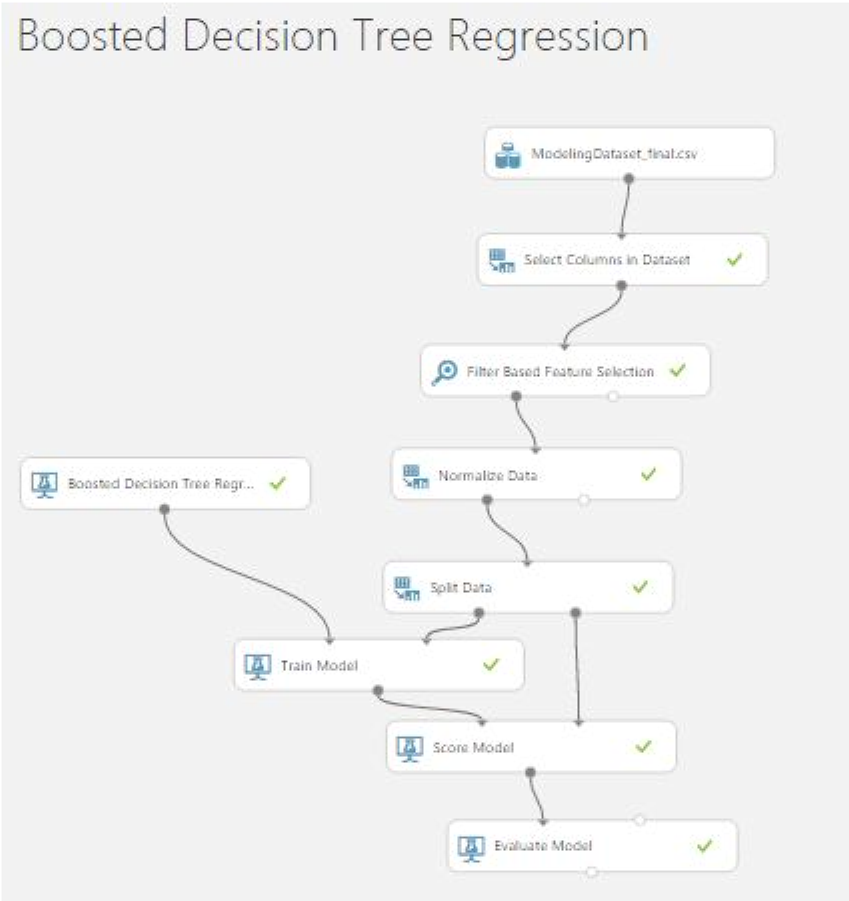
Decision Forest Regression › Evaluate Model › Evaluation results

rows 1

columns 6

Negative Log Likelihood	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Relative Squared Error	Coefficient of Determination
-37880.645535	0.00365	0.010906	1.000631	0.985514	0.014486

Two Class Boosted Decision Tree Regression

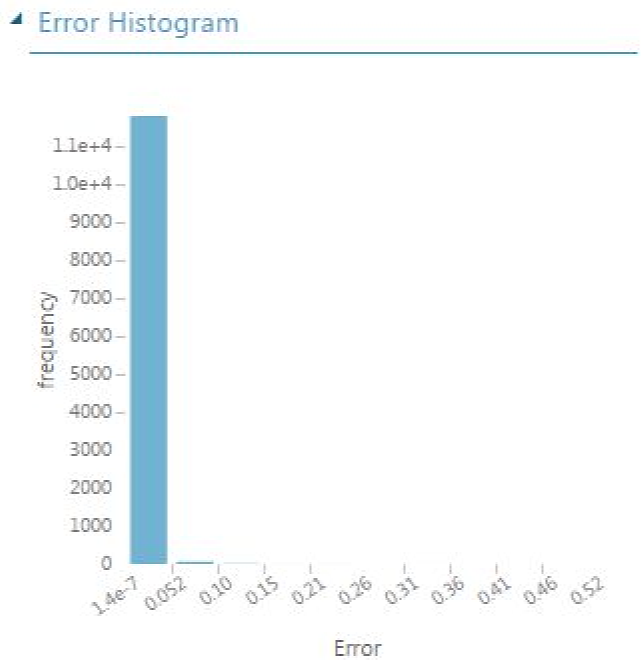


Result

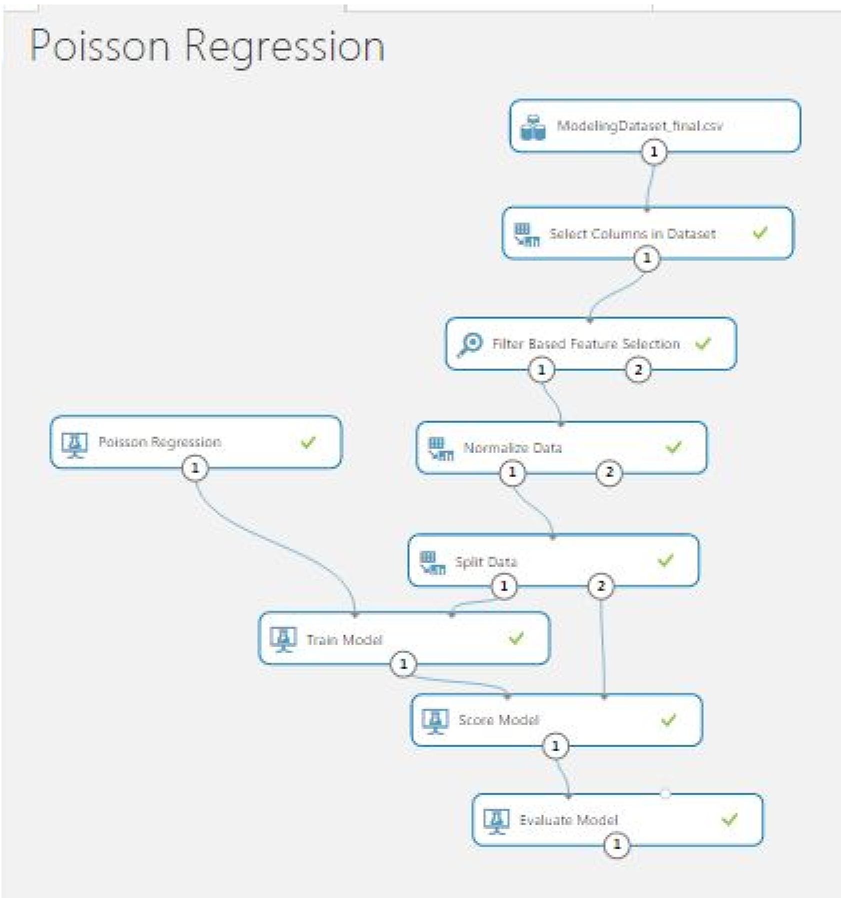
Boosted Decision Tree Regression › Evaluate Model › Evaluation results

Metrics

Mean Absolute Error	0.003956
Root Mean Squared Error	0.011911
Relative Absolute Error	1.084523
Relative Squared Error	1.175499
Coefficient of Determination	-0.175499


Error Histogram
frequency
1.1e+4
1.0e+4
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
1.4e-7
0.052
0.10
0.15
0.21
0.26
0.31
0.36
0.41
0.46
0.52
Error

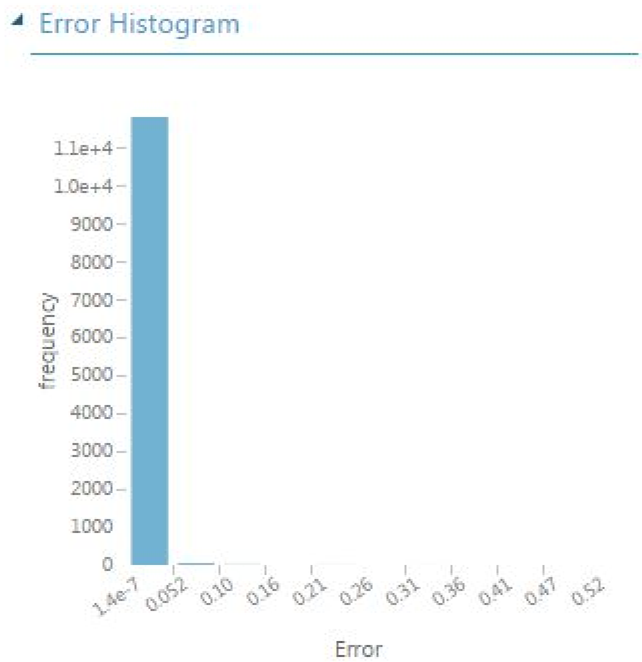
Poisson Regression



Result

Metrics

Mean Absolute Error	0.003661
Root Mean Squared Error	0.010973
Relative Absolute Error	1.003674
Relative Squared Error	0.997698
Coefficient of Determination	0.002302


Error Histogram
frequency
1.1e+4
1.0e+4
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
1.4e-7
0.052
0.10
0.16
0.21
0.26
0.31
0.36
0.41
0.47
0.52
Error

CLASSIFICATION

Classified the post is popular or not based on highest accuracy. Implemented classification using various algorithms Two Class Boosted Decision Tree, Random Forest and Neural Network. Highest accuracy was achieved with Two Class Boosted Decision Tree.

Features selected for classification were as below

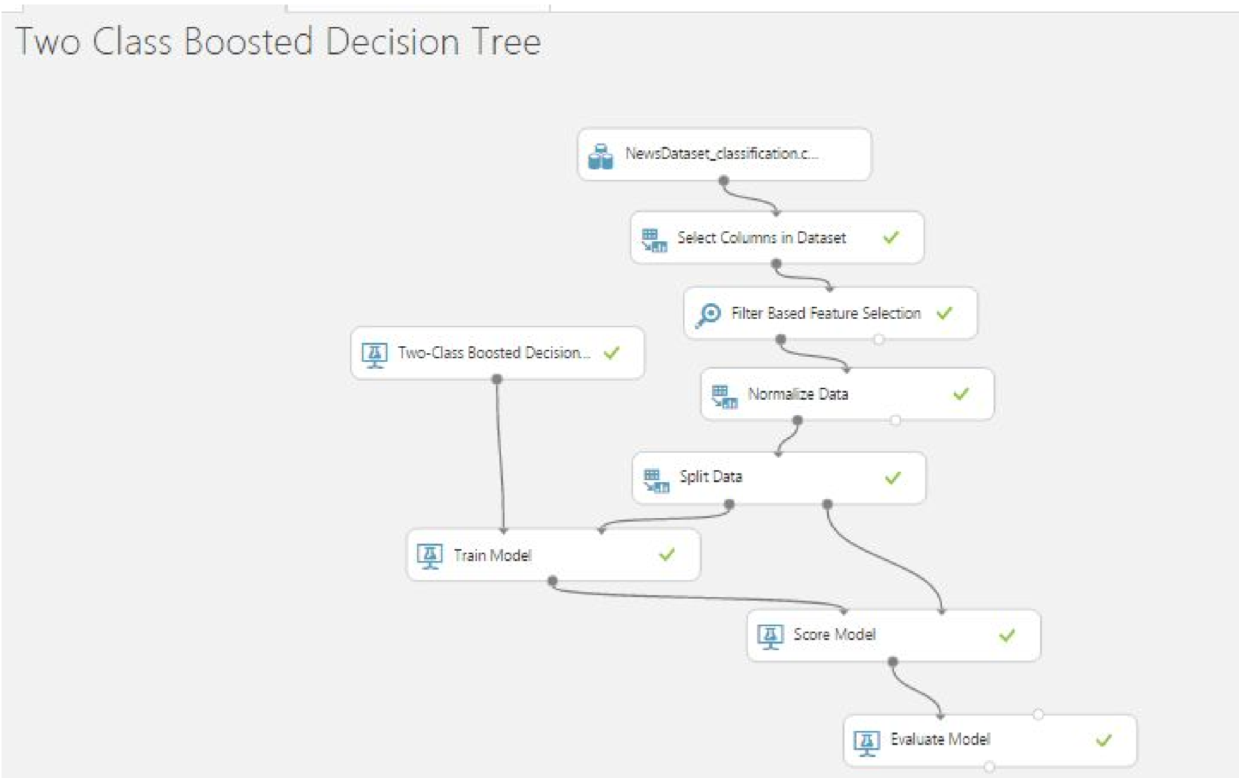
Neural Network Classification › Filter Based Feature Selection › Filtered dataset

rows 39643 columns 11

view as

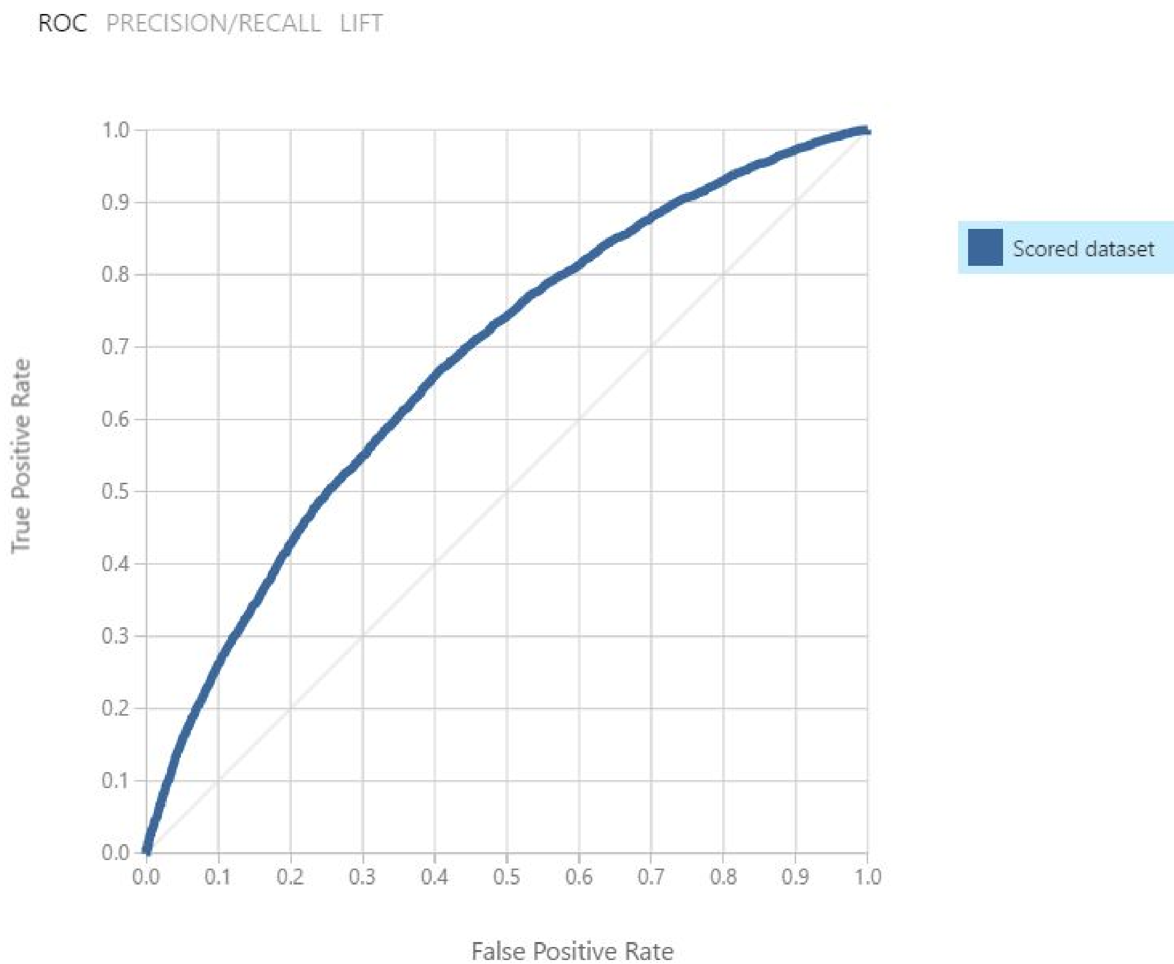
isPopular	isWeekend	num_hrefs	num_keywords	num_imgs	n_tokens_title	weekday	n_tokens_content	num_self_hrefs	Type	n_non_stop_words
Less Popular	0	4	5	1	12	1	219	2	3	1
Less Popular	0	3	4	1	9	1	255	1	1	1
High Popular	0	3	6	1	9	1	211	1	1	1
Less Popular	0	9	7	1	9	1	531	0	3	1
Less Popular	0	19	7	20	13	1	1072	19	5	1
Less Popular	0	2	9	0	10	1	370	2	5	1
Less Popular	0	21	10	20	8	1	960	20	2	1
Less Popular	0	20	9	20	12	1	989	20	5	1
High Popular	0	2	7	0	11	1	97	0	5	1
Less Popular	0	4	5	1	10	1	231	1	6	1
High Popular	0	11	8	1	9	1	1248	0	6	1
High Popular	0	7	7	1	10	1	187	0	2	1
Less Popular	0	18	8	11	9	1	274	2	7	1
High Popular	0	4	6	0	9	1	285	2	7	1

Two Class Boosted Decision Tree

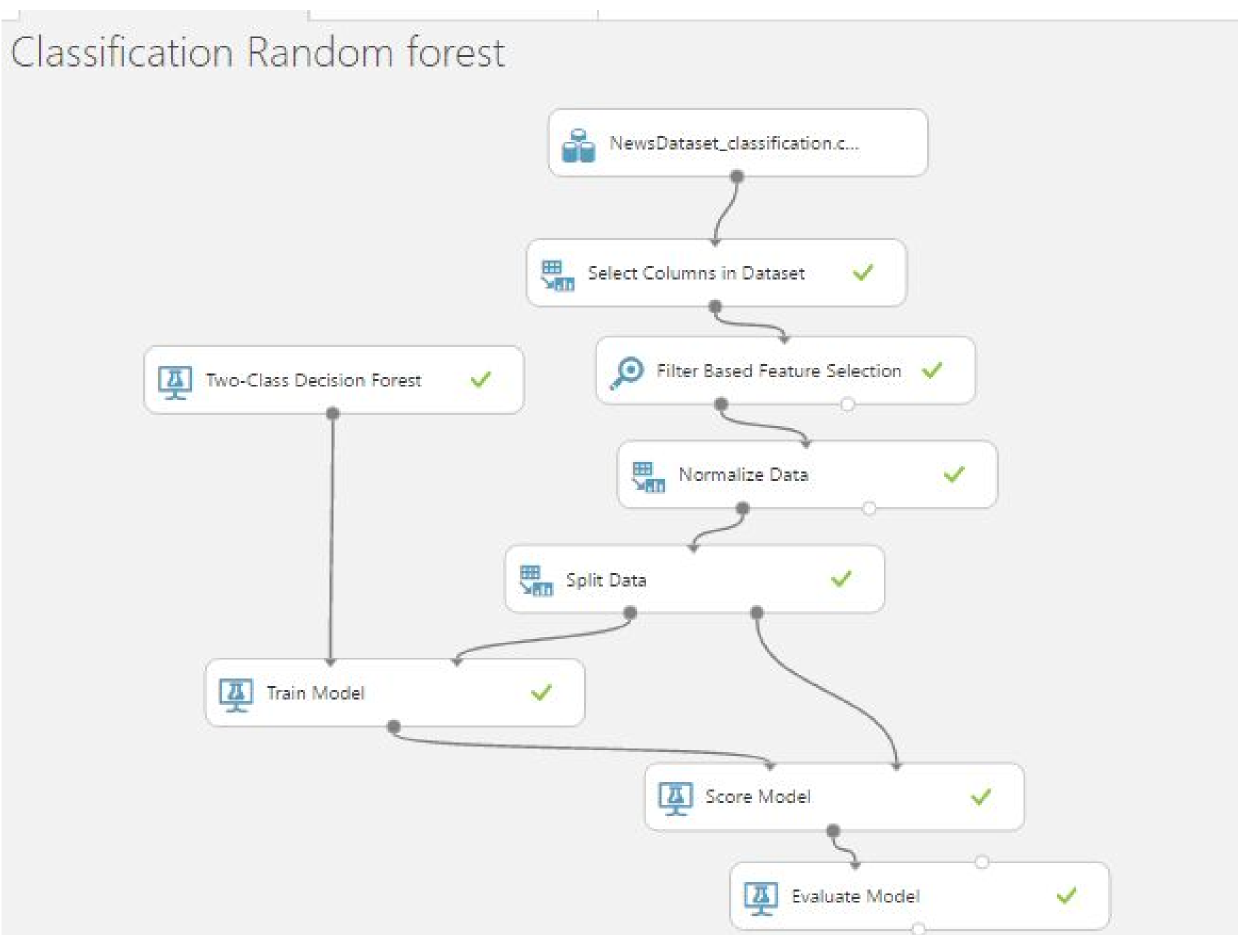


Result

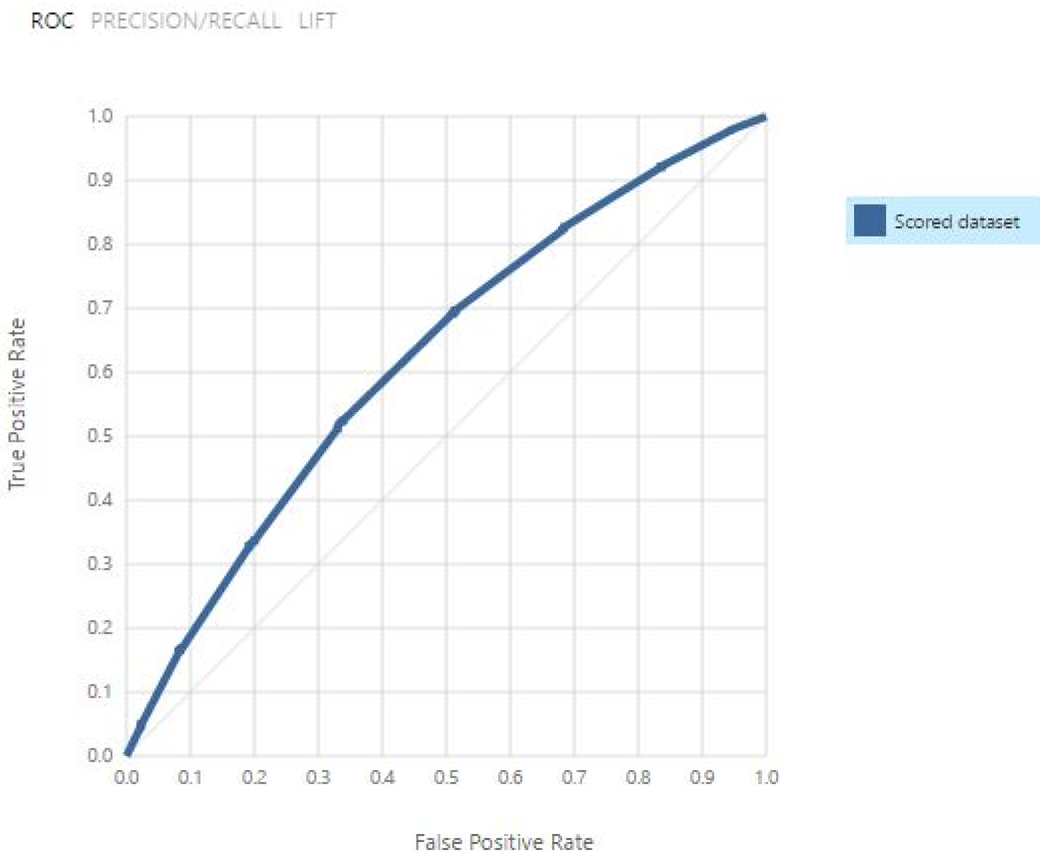
Two Class Boosted Decision Tree › Evaluate Model › Evaluation results



Random Forest Classification:

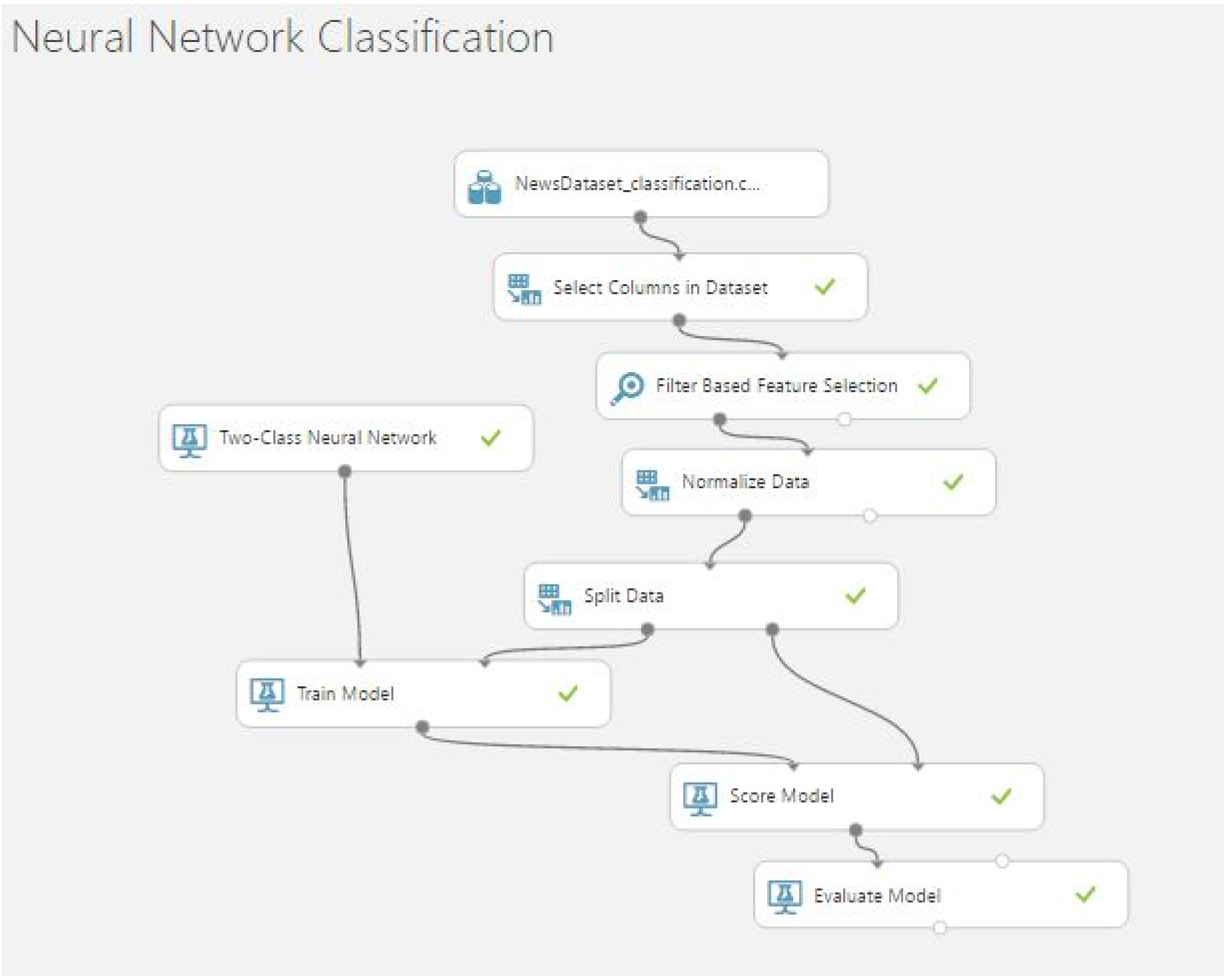


Result

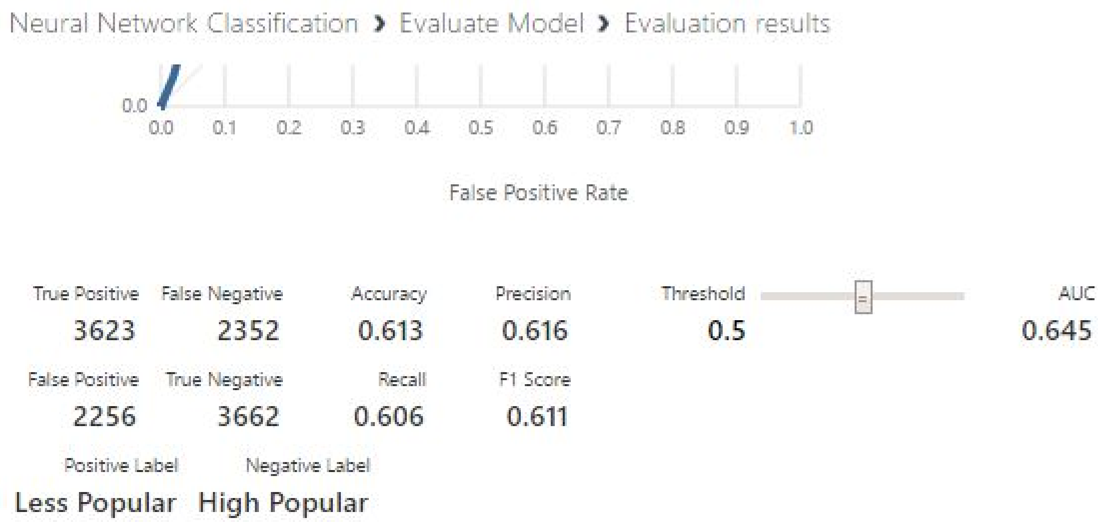


True Positive	False Negative	Accuracy	Precision	Threshold	AUC
3139	2836	0.593	0.610	0.5	0.622
False Positive	**True Negative**	**Recall**	**F1 Score**		
2009	3909	0.525	0.564		
Positive Label	**Negative Label**				
Less Popular	High Popular				

Neural Network



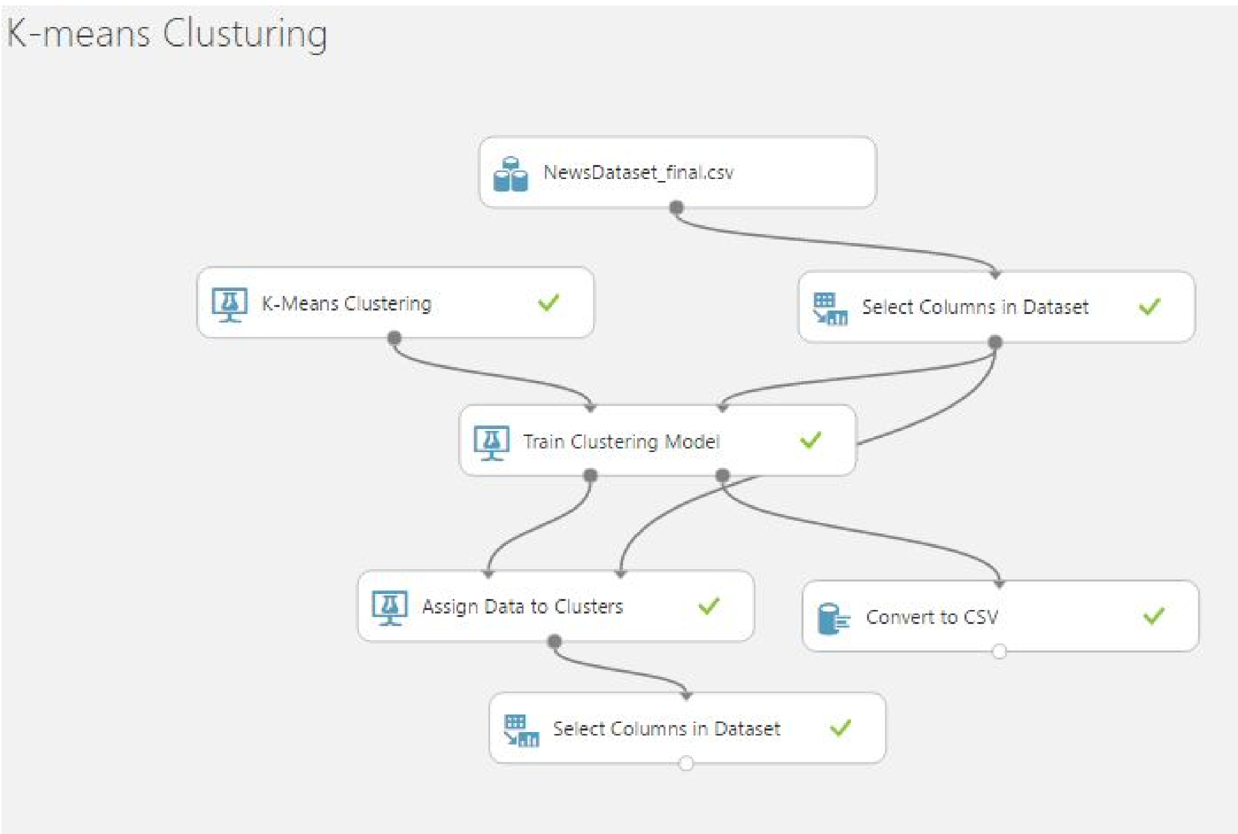
Result


Neural Network Classification › Evaluate Model › Evaluation results
0.0
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
False Positive Rate
True Positive 3623
False Negative 2352
Accuracy 0.613
Precision 0.616
Threshold 0.5
AUC 0.645
False Positive 2256
True Negative 3662
Recall 0.606
F1 Score 0.611
Positive Label Less Popular
Negative Label High Popular

CLUSTERING

Used K-means Clustering, defined clusters on Type of the post where number of clusters used are 3 (k = 3).
Determines the distance of articles based on a few parameters from the centroid of clusters.

K-Means



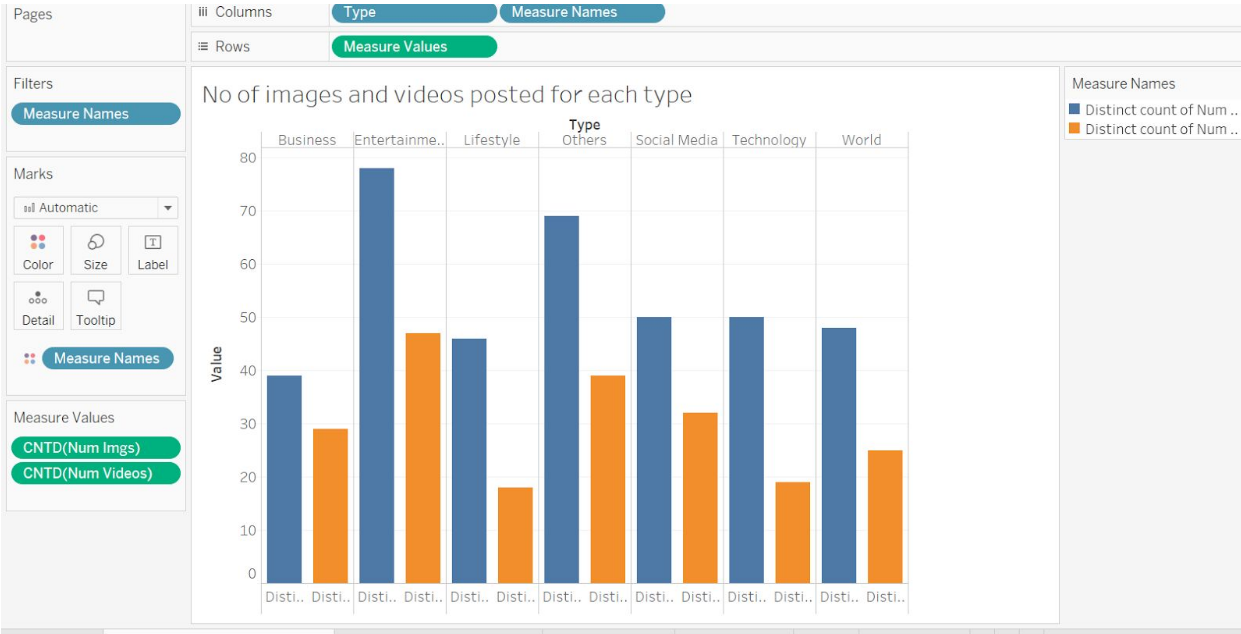
K-means Clusturing › Select Columns in Dataset › Results dataset

rows 39644 columns 2

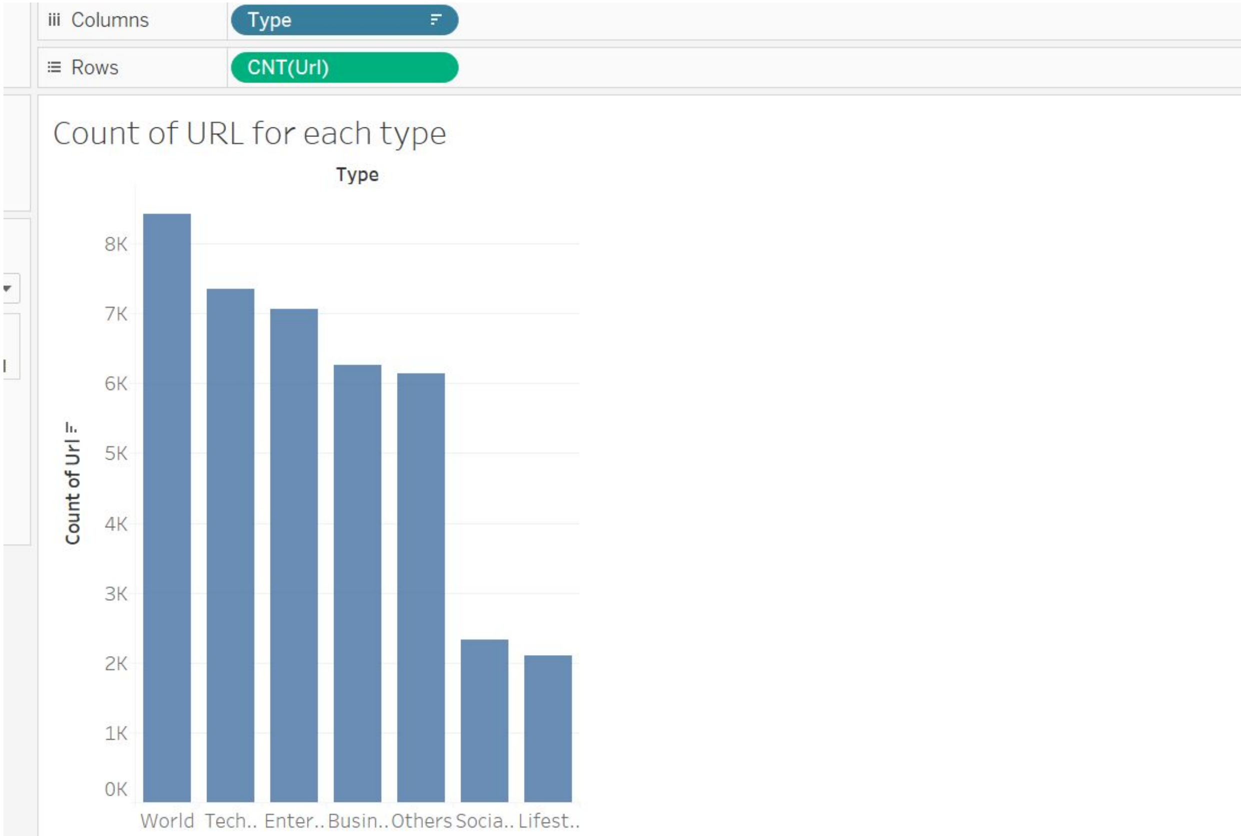
view as

Type	Assignments
3	2
1	1
1	1
3	2
5	0
5	0
2	1
5	0
5	0
6	0
6	0
2	1
7	0
7	0
7	0

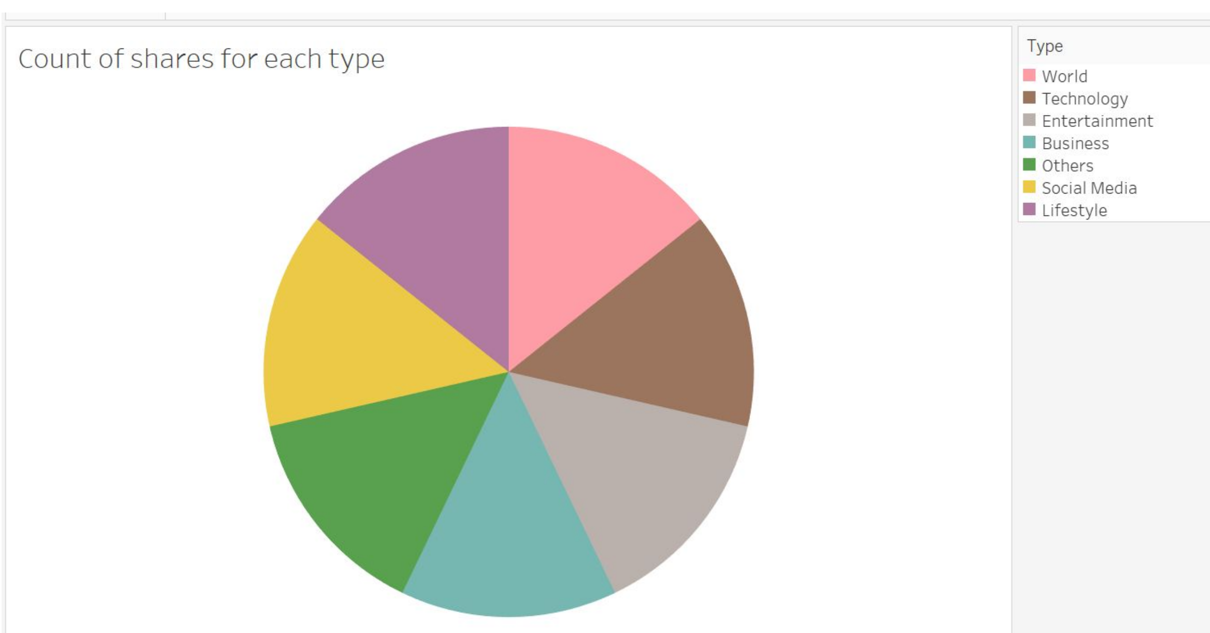
TABLEAU ANALYSIS



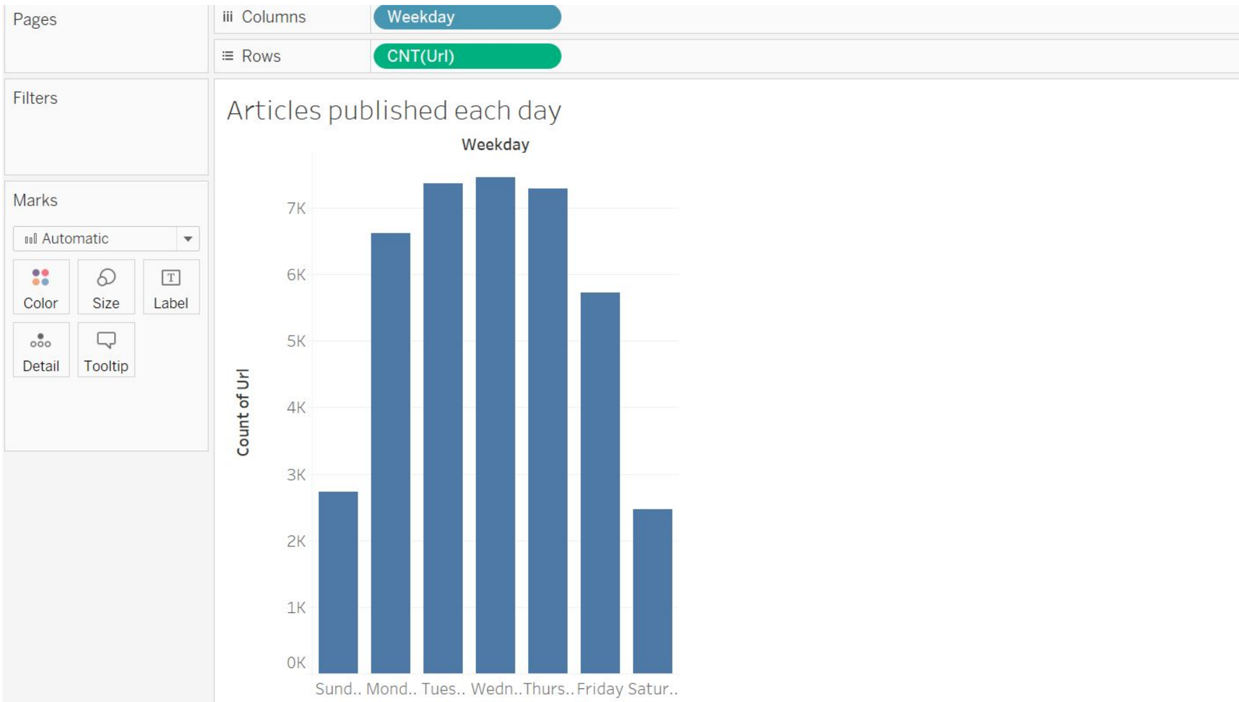
For each type, gives the statistics of images and videos posted.



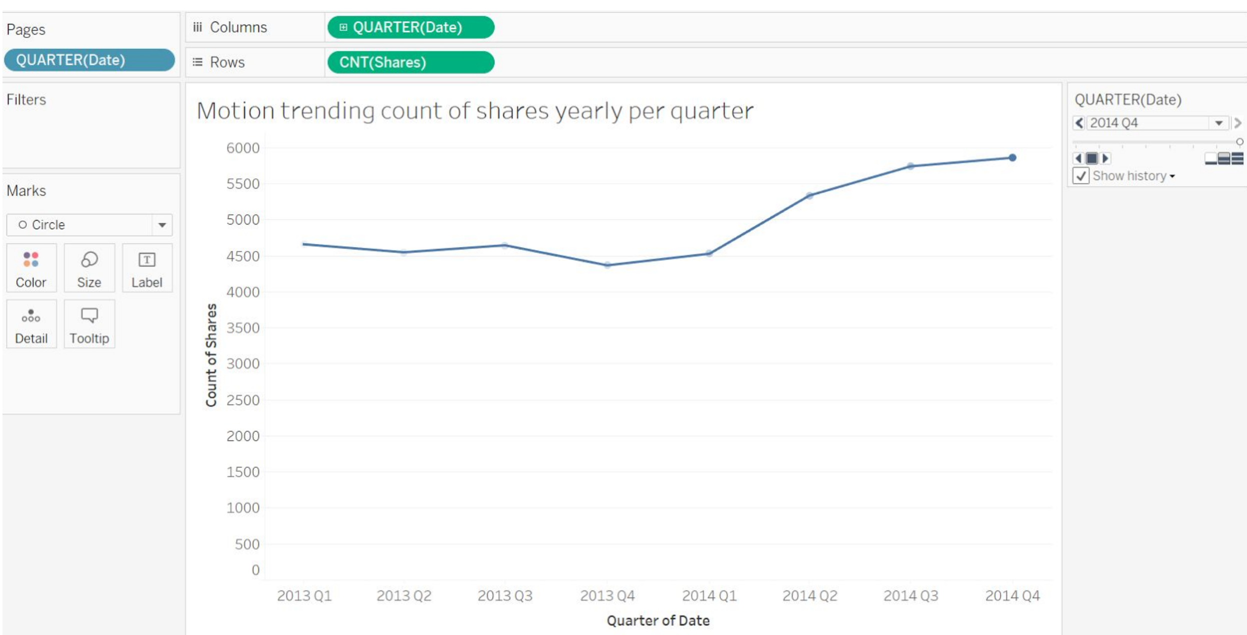
Gives the maximum number of URL's for type - world



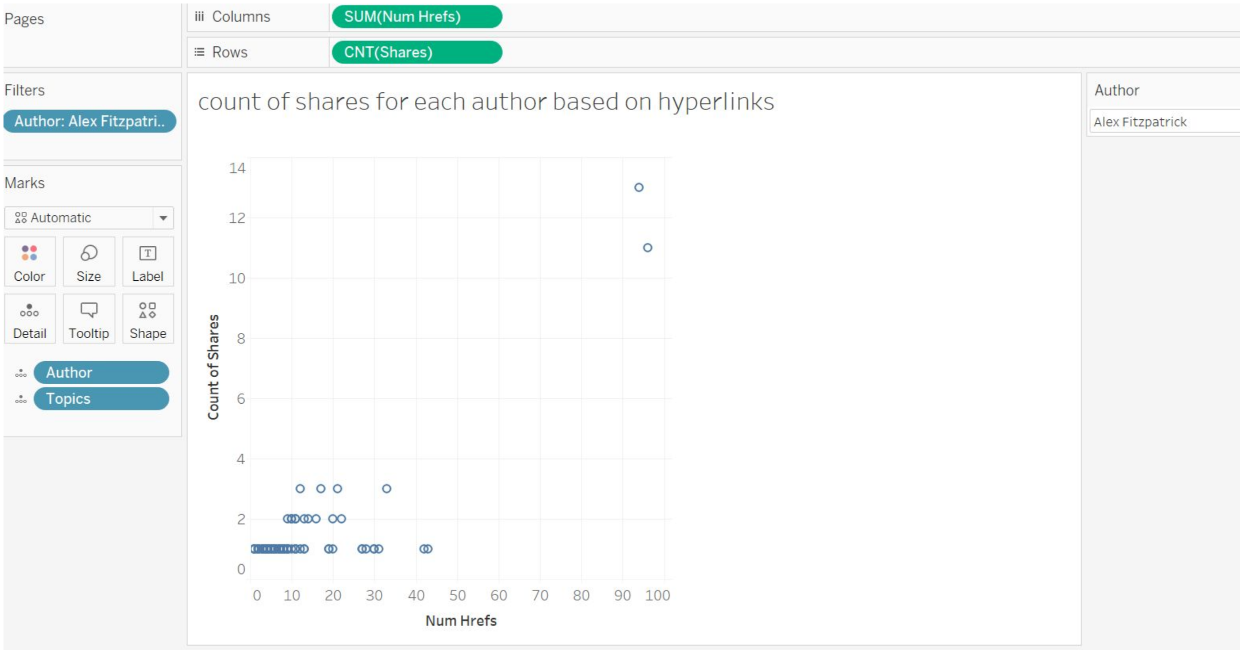
The above pie chart represents the count of shares for each type



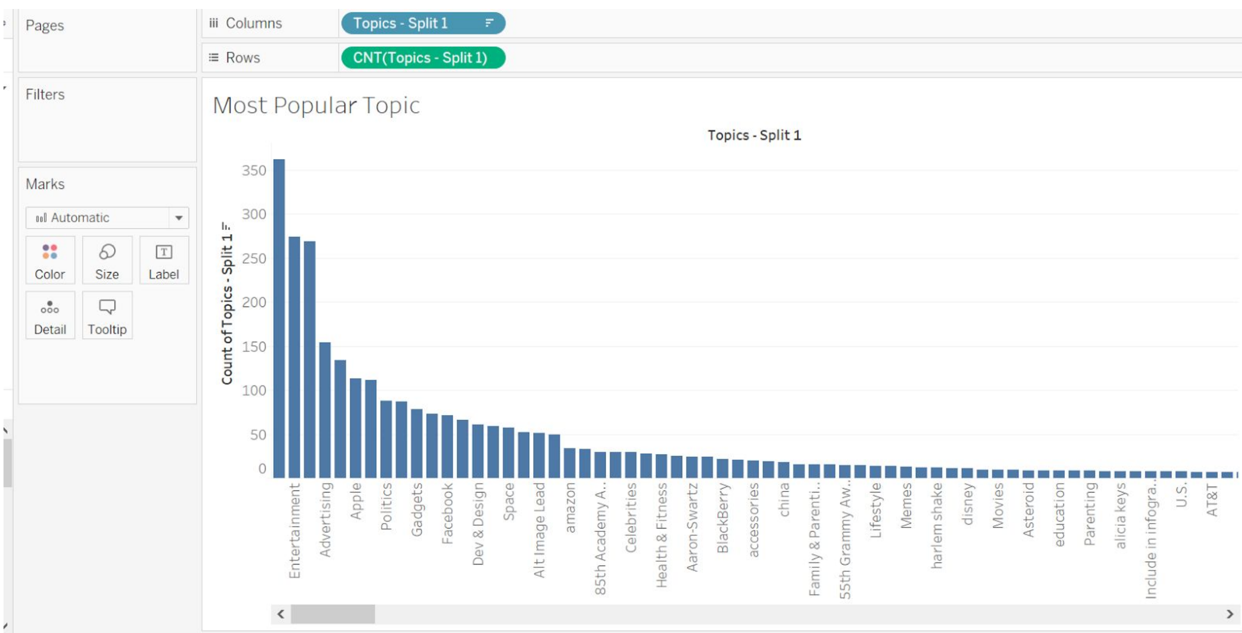
Observed that the maximum number of articles are published on Wednesday. So it is recommended to Mashable to publish articles on weekdays to become popular.



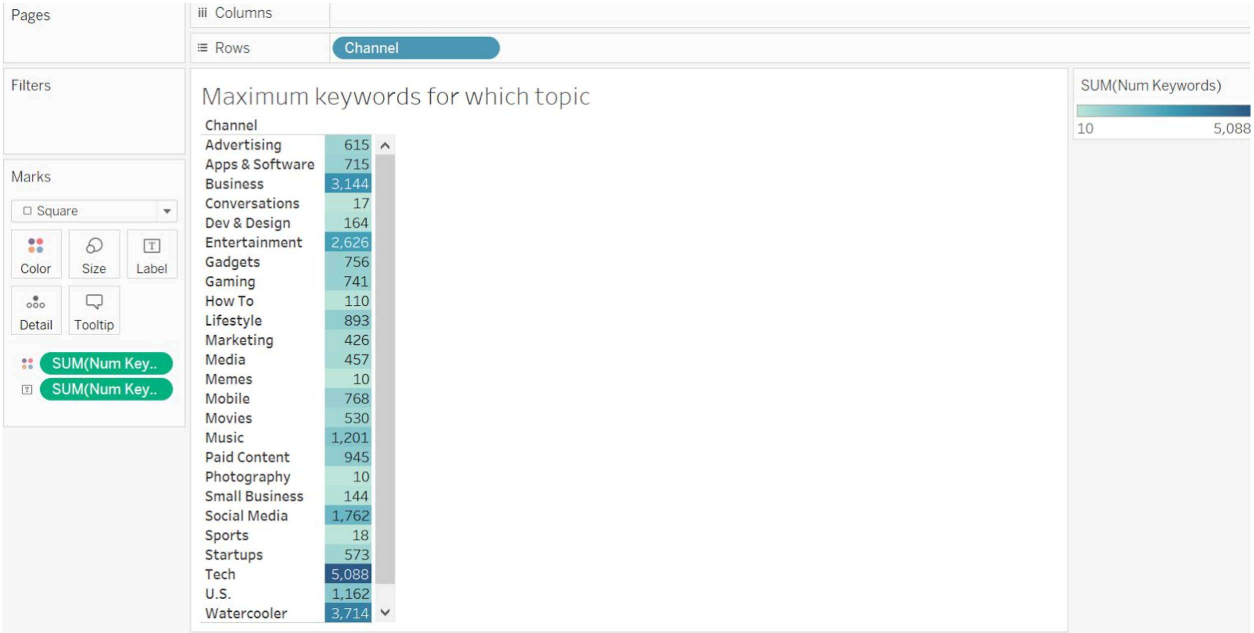
Trending **live** motion of count of shares yearly for each quarter. It can be run to view the live analysis.



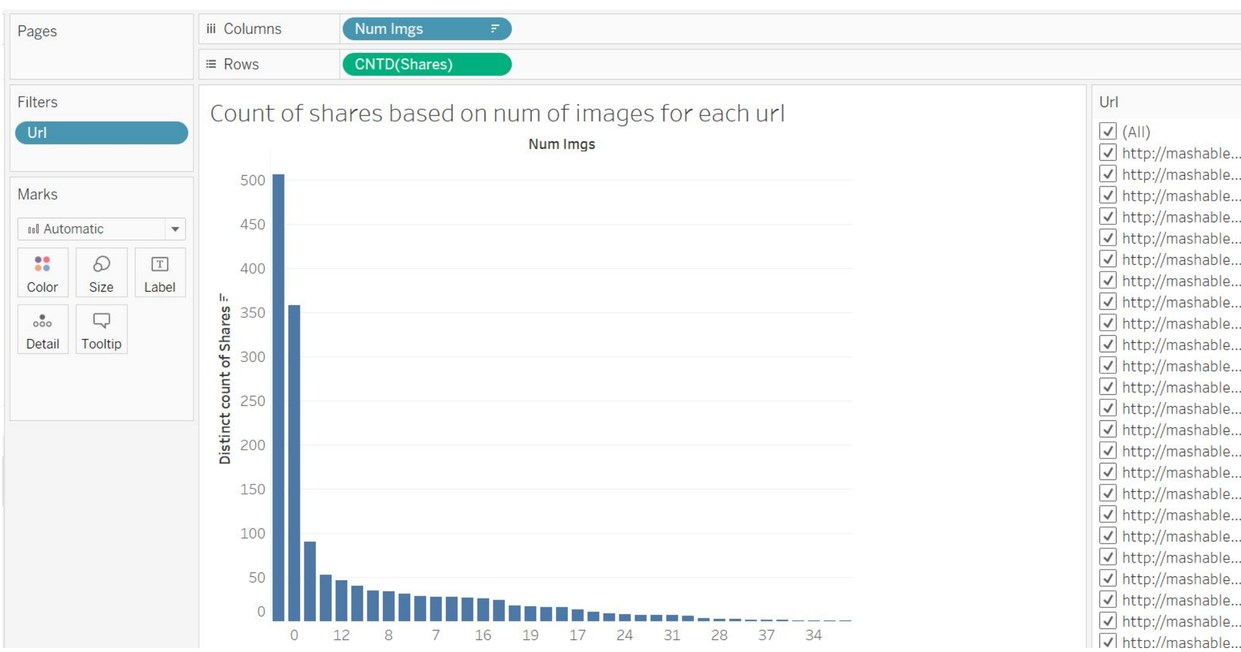
It gives the count of shares for each author based on the number of links on each article.



Determined that the most popular topic is entertainment.



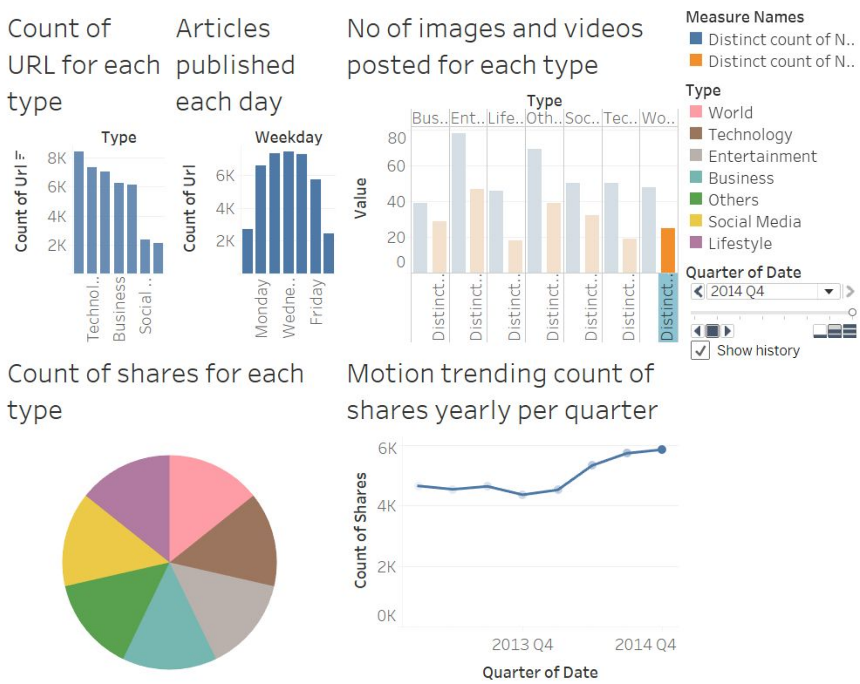
Determines that the maximum words are used for the topic Technology.

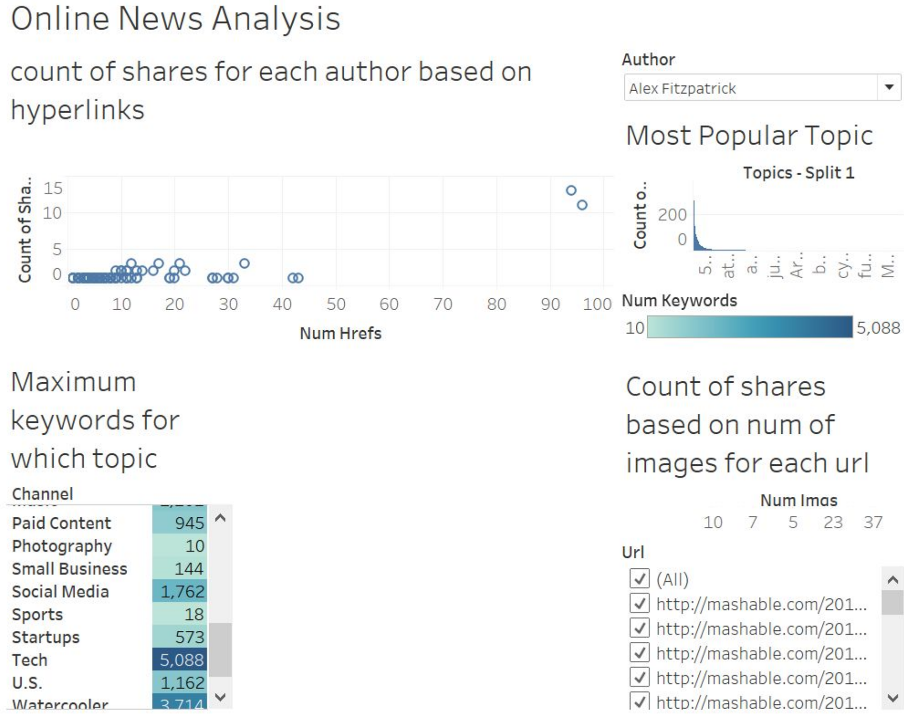


Gives the count of shares based on num of images for each URL.

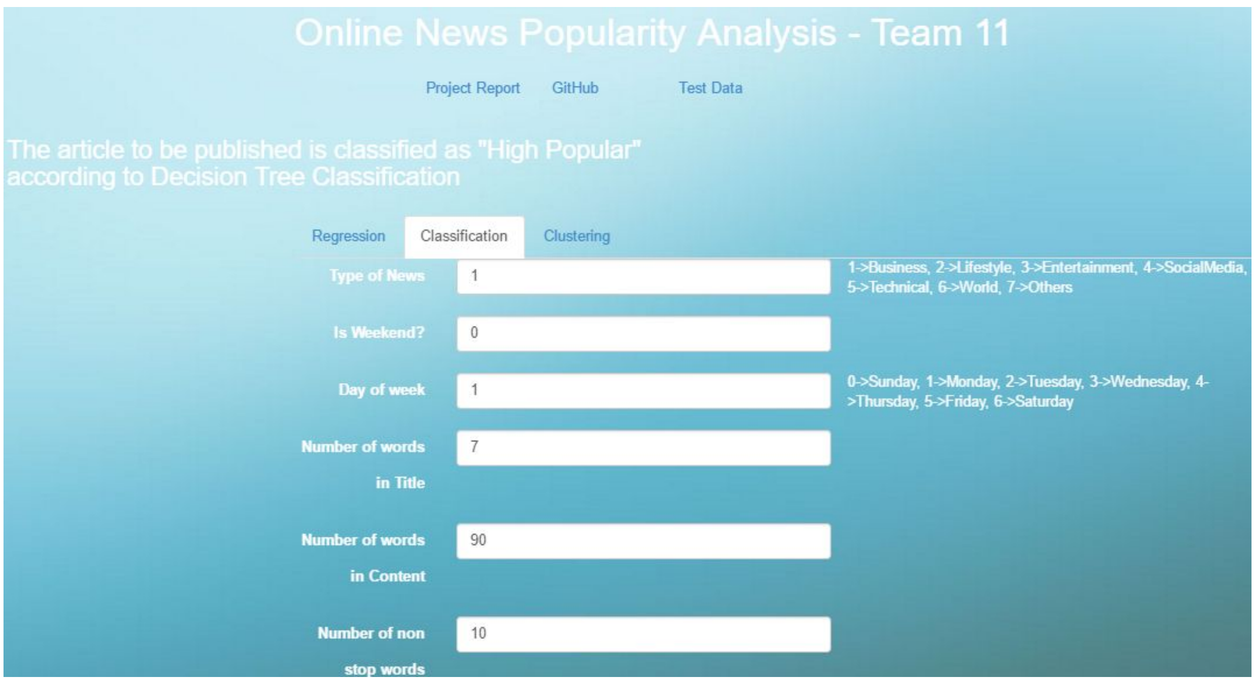
DASHBOARDS

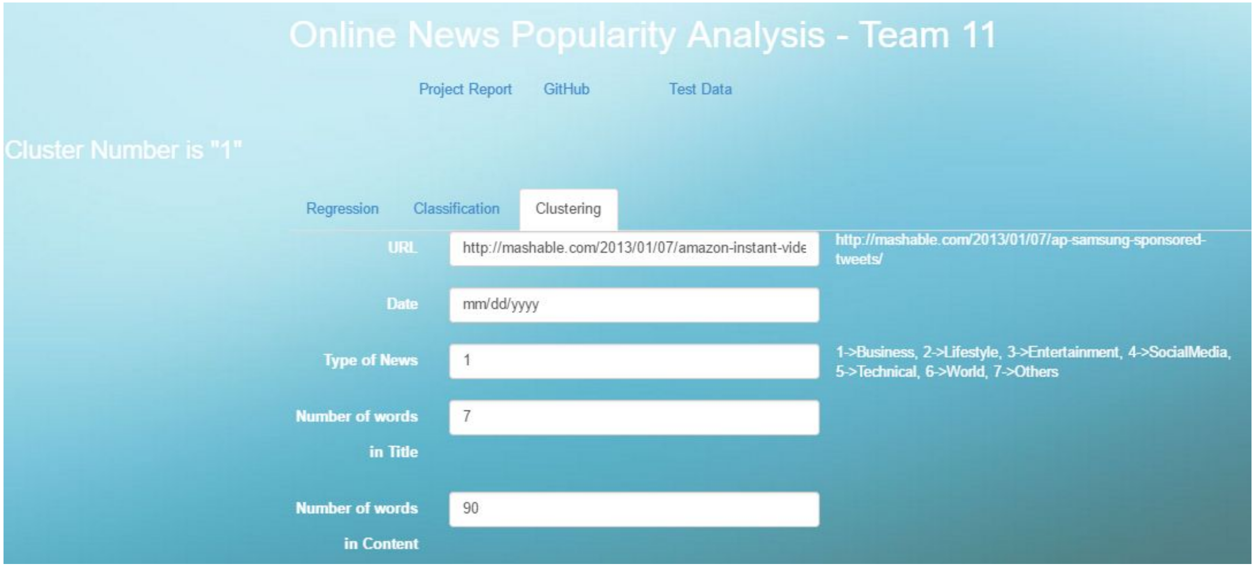
Dashboard shows complete analysis on how number of shares vary depending on some major factors like author, number of published posts each day, number of keywords for each topic.




Online News Analysis
count of shares for each author based on hyperlinks
Author
Alex Fitzpatrick
Most Popular Topic
Topics - Split 1
Count o..
Count of Sha..
Num Hrefs
Num Keywords
10
5,088
Maximum keywords for which topic
Channel
Paid Content 945
Photography 10
Small Business 144
Social Media 1,762
Sports 18
Startups 573
Tech 5,088
U.S. 1,162
Count of shares based on num of images for each url
Num Imgs
10 7 5 23 37
Url
(All)
http://mashable.com/201...
http://mashable.com/201...
http://mashable.com/201...
http://mashable.com/201...
http://mashable.com/201...

WEB INTERFACE




Online News Popularity Analysis - Team 11
Project Report
GitHub
Test Data
Cluster Number is "1"
Regression
Classification
Clustering
URL
http://mashable.com/2013/01/07/amazon-instant-vide
http://mashable.com/2013/01/07/ap-samsung-sponsored-tweets/
Date
mm/dd/yyyy
Type of News
1
1->Business, 2->Lifestyle, 3->Entertainment, 4->SocialMedia, 5->Technical, 6->World, 7->Others
Number of words in Title
7
Number of words in Content
90

CONCLUSION

Insights and Recommendations

To make the news more popular,

Increase:

- Amount of keywords.
- Number of linked embedded.
- Number of images.
- Reference articles with high popularity.
- A more subjective and positive title.

Time of publication:

- Postpone non-time sensitive articles (features etc.) to the weekend. Weekend receive more shares than weekdays.
- Focus more on social media articles during the weekdays.

Monday	Social > Lifestyle/ Tech > Business > World/ Entertainment
Tuesday	Social > Lifestyle/ Tech > Business > Entertainment > World
Wednesday	Social > Lifestyle/ Tech > Business > Entertainment > World
Thursday	Social > Lifestyle/ Tech > Business > Entertainment > World
Friday	Social > Lifestyle/ Tech > Business > World/ Entertainment
Saturday	Lifestyle/ Business/ Social/ Tech > Entertainment > World
Sunday	Lifestyle/ Business/ Social/ Tech > Entertainment > World

Channel:

- Editors may want to put more emphasis on articles of a specific channel.
- Social media> technology > lifestyle > business > entertainment > world

Next Step

- Continue to refine the model by including more independent variables.
- Extend the time interval, currently we only collected data for 2 years.
- Further subdivide news according to their topics and find what factors influence news with a particularly topic.

LINKS

Web UI Url – Web UI

Git Hub – GitHub

REFERENCES

https://archive.ics.uci.edu/ml/datasets/Online+News+Popularity

https://repositorium.sdum.uminho.pt/bitstream/1822/39169/1/main.pdf

http://mashable.com/