ISM6136 - Data Mining

Online News Popularity Analysis – FINAL PROJECT

4/23/2016

USF – Muma College of Business

SUBMITTED BY -

Sachin Kant Misra
Prashant Bhowmik
Jagpreet Singh Sethi
Renee Champagne

SUBMITTED TO –

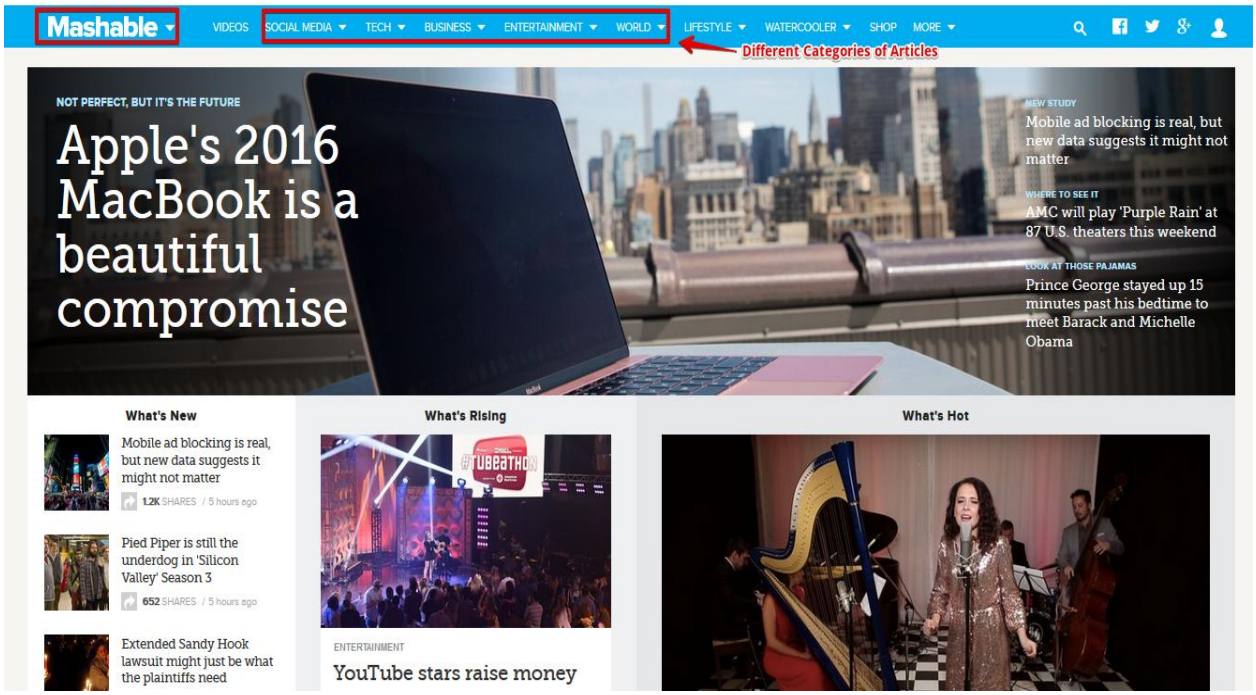
Prof. Athienitis
ISDS Department

TABLE OF CONTENTS

INTRODUCTION

The project “Online News Popularity” is based out of data collected from a website (www.mashable.com). The dataset summarizes a heterogeneous set of features about articles published by Mashable over a period of two years. The goal is to predict the number of shares in social networks (popularity).

The articles were published by Mashable (www.mashable.com) and their content as the rights to reproduce it belongs to them. Hence, this dataset does not share the original content but some statistics associated with it. The original content is publicly accessed and retrieved using the provided urls.



PROBLEM STATEMENT

1) To predict the number of shares in social networks (popularity).
2) To predict the popularity status of the article using threshold for number of shares i.e. popularity = "1" for shares >1400 and popularity = "0" for shares <1400.

Popular (Yes)	Popular (No)
Share > 1400	Shares <1400

3) To predict an ordinal outcome for popularity levels defined by "1" (High) for shares >2100, "2" (Medium) for shares <2100 and >1100, and "3" (Low) for shares <1100.

BQ	BR	BS
shares	popularity	popularity_level
823	0	3
10000	1	1
761	0	3
1600	1	2
13600	1	1

PopularLevel (Low)	PopularLevel (Medium)	PopularLevel (High)
Share <1100	Shares between 1100 and 2100	Shares > 2100

4) Visualize the dataset for various kinds of trend/insights found among the attributes of the Mashable article using tableau.

DATASET OVERVIEW

- The data set was acquired on 8th January' 2015.
- Total 39644 instances and 71 attributes.
  - 64 are independent predictors
  - 4 are non-predictive variables
  - 3 are target variables (Shares, Popularity, Popularity_level)

Below is the explanation for each of the attribute provided within the dataset:

1. *url*: URL of the article (**non-predictive**)
2. *year*: year of the article published
3. *month*: month of the article published
4. *timedelta*: Days between the article publication and the dataset acquisition (**non-predictive**)
5. *n_tokens_title*: Number of words in the title
6. *n_tokens_content*: Number of words in the content
7. *n_unique_tokens*: Rate of unique words in the content
8. *n_non_stop_words*: Rate of non-stop words in the content
9. *n_non_stop_unique_tokens*: Rate of unique non-stop words in the content
10. *num_hrefs*: Number of links
11. *num_self_hrefs*: Number of links to other articles published by Mashable
12. *num_imgs*: Number of images
13. *num_videos*: Number of videos
14. *average_token_length*: Average length of the words in the content
15. *num_keywords*: Number of keywords in the metadata
16. *Lifestyle*: Is data channel 'Lifestyle'?
17. *Entertainment*: Is data channel 'Entertainment'?
18. *Business*: Is data channel 'Business'?
19. *Social_Media*: Is data channel 'Social Media'?
20. *Technology*: Is data channel 'Tech'?
21. *World*: Is data channel 'World'?
22. *article_type*: categorical variable for the type of article published (**non-predictive**)
23. *kw_min_min*: Worst keyword (min shares)
24. *kw_max_min*: Worst keyword (max shares)
25. *kw_avg_min*: Worst keyword (avg shares)
26. *kw_min_max*: Best keyword (min shares)
27. *kw_max_max*: Best keyword (max shares)
28. *kw_avg_max*: Best keyword (avg shares)
29. *kw_min_avg*: Avg keyword (min shares)
30. *kw_max_avg*: Avg keyword (max shares)
31. *kw_avg_avg*: Avg keyword (avg shares)
32. *self_reference_min_shares*: Min shares of referenced articles in Mashable
33. *self_reference_max_shares*: Max shares of referenced articles in Mashable
34. *self_reference_avg_sharess*: Avg shares of referenced articles in Mashable
35. *Monday*: Was the article published on a Monday?
36. *Tuesday*: Was the article published on a Tuesday?
37. *Wednesday*: Was the article published on a Wednesday?
38. *Thursday*: Was the article published on a Thursday?
39. *Friday*: Was the article published on a Friday?
40. *Saturday*: Was the article published on a Saturday?

41. *Sunday*: Was the article published on a Sunday?
42. *is_weekend*: Was the article published on the weekend?
43. *Weekday*: Categorical variable for day of the week when the article was published (**non-predictive**)
44. *LDA_00*: Closeness to LDA topic 0
45. *LDA_01*: Closeness to LDA topic 1
46. *LDA_02*: Closeness to LDA topic 2
47. *LDA_03*: Closeness to LDA topic 3
48. *LDA_04*: Closeness to LDA topic 4
49. *global_subjectivity*: Text subjectivity
50. *global_sentiment_polarity*: Text sentiment polarity
51. *global_rate_positive_words*: Rate of positive words in the content
52. *global_rate_negative_words*: Rate of negative words in the content
53. *rate_positive_words*: Rate of positive words among non-neutral tokens
54. *rate_negative_words*: Rate of negative words among non-neutral tokens
55. *avg_positive_polarity*: Avg polarity of positive words
56. *min_positive_polarity*: Min polarity of positive words
57. *max_positive_polarity*: Max polarity of positive words
58. *avg_negative_polarity*: Avg polarity of negative words
59. *min_negative_polarity*: Min polarity of negative words
60. *max_negative_polarity*: Max polarity of negative words
61. *title_subjectivity*: Title subjectivity
62. *title_sentiment_polarity*: Title polarity
63. *abs_title_subjectivity*: Absolute subjectivity level
64. *abs_title_sentiment_polarity*: Absolute polarity level
65. *state*: State name from where the article was published tracked using IP address
66. *day_phase*: time of the day when the article was published
67. *avg_time_spent*: average time spent by users to read the article using log information from web
68. *device_type*: type of device used to read/share the article
69. *shares*: Number of shares (**target**)
70. *popularity*: activity measure of article over web using threshold value of number of shares (**target**); If shares > 1400 then popularity = "1" else popularity = "0"
71. *popularity_level*: categorical ordinal variable for popularity (**target**); for share <1100, value = "3" (LOW); for share >1100 and <2100, value = "2" (MEDIUM); for share >2100, value = "1" (HIGH)

Screenshot of the dataset:

1)

	A	B	C	D	E	F	G	H	I	J
1	url	year	month	timedelta	n_tokens_title	n_tokens_content	n_unique_tokens	n_non_stop_words	n_non_stop_unique_tokens	num_hrefs
2	http://mashable.com/2013/01/07/amazon-instant-video-browser/	2013	1	731	12	219	0.663594467	0.999999992	0.815384609	4
3	http://mashable.com/2013/01/07/ap-samsung-sponsored-tweets/	2013	1	731	9	255	0.604743081	0.999999993	0.791946303	3
4	http://mashable.com/2013/01/07/apple-40-billion-app-downloads/	2013	1	731	9	211	0.575129531	0.999999992	0.663865541	3
5	http://mashable.com/2013/01/07/astronaut-notre-dame-bcs/	2013	1	731	9	531	0.503787878	0.999999997	0.665634673	9
6	http://mashable.com/2013/01/07/att-u-verse-apps/	2013	1	731	13	1072	0.415645617	0.999999999	0.540889526	19
7	http://mashable.com/2013/01/07/beewi-smart-toys/	2013	1	731	10	370	0.559888578	0.999999995	0.698198195	2
8	http://mashable.com/2013/01/07/bodymedia-armbandgets-update/	2013	1	731	8	960	0.418162618	0.999999998	0.549833886	21
9	http://mashable.com/2013/01/07/canon-poweshot-n/	2013	1	731	12	989	0.433573635	0.999999998	0.572107765	20
10	http://mashable.com/2013/01/07/car-of-the-future-infographic/	2013	1	731	11	97	0.670103086	0.99999998	0.836734677	2
11	http://mashable.com/2013/01/07/chuck-hagel-website/	2013	1	731	10	231	0.636363634	0.999999993	0.797101443	4
12	http://mashable.com/2013/01/07/cosmic-events-doomsday/	2013	1	731	9	1248	0.490049751	0.999999999	0.731638417	11
13	http://mashable.com/2013/01/07/crayon-creatures/	2013	1	731	10	187	0.666666663	0.999999991	0.799999993	7
14	http://mashable.com/2013/01/07/creature-cups/	2013	1	731	9	274	0.6091954	0.999999994	0.707602335	18
15	http://mashable.com/2013/01/07/dad-jokes/	2013	1	731	9	285	0.744186044	0.999999995	0.84153005	4
16	http://mashable.com/2013/01/07/downton-abbey-tumblrs/	2013	1	731	8	259	0.562753034	0.999999994	0.644444441	19
17	http://mashable.com/2013/01/07/earth-size-planets-milky-way/	2013	1	731	12	682	0.459541984	0.999999997	0.634961438	10
18	http://mashable.com/2013/01/07/echo-game/	2013	1	731	8	1118	0.512396694	0.999999999	0.709770114	26
19	http://mashable.com/2013/01/07/entrepreneur-trends-2013/	2013	1	731	8	397	0.624678662	0.999999996	0.805668013	11
20	http://mashable.com/2013/01/07/facebook-sick-app/	2013	1	731	11	103	0.689320382	0.999999984	0.8064516	3
21	http://mashable.com/2013/01/07/felt-audio-pulse-speaker/	2013	1	731	8	1207	0.410579345	0.999999999	0.548969071	24
22	http://mashable.com/2013/01/07/ford-glympse/	2013	1	731	13	1248	0.390637611	0.999999999	0.523388116	21
23	http://mashable.com/2013/01/07/ftc-google-leaks/	2013	1	731	9	391	0.510256409	0.999999996	0.649999997	9
24	http://mashable.com/2013/01/07/fujifilm-50x-superzoom/	2013	1	731	11	1154	0.427304964	0.999999999	0.572815533	20
25	http://mashable.com/2013/01/07/hillary-clinton-helmet/	2013	1	731	11	125	0.674796742	0.999999987	0.797468344	1

OnlineNewsPopularity

2).

	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
1	num_self_hrefs	num_imgs	num_videos	average_token_length	num_keywords	Lifestyle	Entertainment	Business	Social Media	Technology	World	article_type	kw_min_min	kw_max_min	kw_avg_m
2	2	1	0	4.680365297	5	0	1	0	0	0	0	Entertainment	0	0	
3	1	1	0	4.91372549	4	0	0	1	0	0	0	Business	0	0	
4	1	1	0	4.393364929	6	0	0	1	0	0	0	Business	0	0	
5	0	1	0	4.404896422	7	0	1	0	0	0	0	Entertainment	0	0	
6	19	20	0	4.682835821	7	0	0	0	0	1	0	Technology	0	0	
7	2	0	0	4.359459459	9	0	0	0	0	1	0	Technology	0	0	
8	20	20	0	4.654166667	10	1	0	0	0	0	0	Lifestyle	0	0	
9	20	20	0	4.617795753	9	0	0	0	0	1	0	Technology	0	0	
10	0	0	0	4.855670103	7	0	0	0	0	1	0	Technology	0	0	
11	1	1	1	5.090909091	5	0	0	0	0	0	1	World	0	0	
12	0	1	0	4.617788462	8	0	0	0	0	0	1	World	0	0	
13	0	1	0	4.657754011	7	1	0	0	0	0	0	Lifestyle	0	0	
14	2	11	0	4.233576642	8	0	0	0	0	0	0	Lifestyle	0	0	
15	2	0	21	4.343859649	6	0	0	0	0	0	0	Lifestyle	0	0	
16	3	9	0	5.023166023	7	0	0	0	0	0	0	Lifestyle	0	0	
17	0	1	0	4.620234604	6	0	0	0	0	0	1	World	0	0	
18	18	12	1	4.703935599	5	0	0	0	0	0	0	Lifestyle	0	0	
19	0	1	0	5.445843829	6	0	0	1	0	0	0	Business	0	0	
20	1	1	0	4.844660194	6	1	0	0	0	0	0	Lifestyle	0	0	
21	24	42	0	4.716652858	8	0	0	0	0	1	0	Technology	0	0	
22	19	20	0	4.686698718	10	0	0	0	0	1	0	Technology	0	0	
23	2	1	1	5.296675192	7	0	0	0	0	0	1	World	0	0	
24	20	20	0	4.629982669	7	0	0	0	0	1	0	Technology	0	0	
25	1	1	0	4.824	6	0	0	0	0	0	1	World	0	0	

OnlineNewsPopularity

3).

	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	
1	kw_avg_min	kw_min_max	kw_max_max	kw_avg_max	kw_min_avg	kw_max_avg	kw_avg_avg	self_reference_min_shares	self_reference_max_shares	self_reference_avg_sharess	Monday	Tuesday	Wed
2	0	0	0	0	0	0	0	496	496	496	1	0	
3	0	0	0	0	0	0	0	0	0	0	1	0	
4	0	0	0	0	0	0	0	918	918	918	1	0	
5	0	0	0	0	0	0	0	0	0	0	1	0	
6	0	0	0	0	0	0	0	545	16000	3151.157895	1	0	
7	0	0	0	0	0	0	0	8500	8500	8500	1	0	
8	0	0	0	0	0	0	0	545	16000	3151.157895	1	0	
9	0	0	0	0	0	0	0	545	16000	3151.157895	1	0	
10	0	0	0	0	0	0	0	0	0	0	1	0	
11	0	0	0	0	0	0	0	0	0	0	1	0	
12	0	0	0	0	0	0	0	0	0	0	1	0	
13	0	0	0	0	0	0	0	0	0	0	1	0	
14	0	0	0	0	0	0	0	10700	16200	13450	1	0	
15	0	0	0	0	0	0	0	770	22800	11785	1	0	
16	0	0	0	0	0	0	0	4800	4800	4800	1	0	
17	0	0	0	0	0	0	0	0	0	0	1	0	
18	0	0	0	0	0	0	0	555	14000	3904.625	1	0	
19	0	0	0	0	0	0	0	0	0	0	1	0	
20	0	0	0	0	0	0	0	5000	5000	5000	1	0	
21	0	0	0	0	0	0	0	545	16000	2829.541667	1	0	
22	0	0	0	0	0	0	0	545	16000	3151.157895	1	0	
23	0	0	0	0	0	0	0	704	704	704	1	0	
24	0	0	0	0	0	0	0	545	16000	3151.157895	1	0	
25	0	0	0	0	0	0	0	16100	16100	16100	1	0	

OnlineNewsPopularity

4).

	AK	AL	AM	AN	AO	AP	AQ	AR	AS	AT	AU	AV	AW	AX	AY
1	Wednesday	Thursday	Friday	Saturday	Sunday	is_weekend	Weekday	LDA_00	LDA_01	LDA_02	LDA_03	LDA_04	global_subjectivity	global_sentiment_polarity	global_rate_positive
2	0	0	0	0	0	0	Monday	0.500331204	0.37827893	0.040004675	0.041262648	0.040122544	0.521617145	0.092561983	0.
3	0	0	0	0	0	0	Monday	0.799755687	0.050046675	0.050096252	0.050100673	0.050000712	0.341245791	0.148947811	0.04
4	0	0	0	0	0	0	Monday	0.217792289	0.033334457	0.033351425	0.033333536	0.682188294	0.702222222	0.323333333	0.05
5	0	0	0	0	0	0	Monday	0.028573216	0.419299642	0.494650826	0.028904718	0.028571598	0.429849687	0.100704666	0.04
6	0	0	0	0	0	0	Monday	0.02863281	0.028793552	0.028575185	0.028571675	0.885426778	0.513502123	0.281003476	0.07
7	0	0	0	0	0	0	Monday	0.022245276	0.306717576	0.022231278	0.02222429	0.626581581	0.437408649	0.071184192	0.0
8	0	0	0	0	0	0	Monday	0.020081666	0.114705387	0.020024369	0.020015328	0.82517325	0.514480301	0.268302724	0.08
9	0	0	0	0	0	0	Monday	0.022224357	0.150732973	0.243435476	0.022223603	0.561383591	0.543474234	0.29861347	0.08
10	0	0	0	0	0	0	Monday	0.458250415	0.028979431	0.028661883	0.029695859	0.454412412	0.538888889	0.161111111	0.03
11	0	0	0	0	0	0	Monday	0.040000096	0.040000026	0.839997207	0.040000633	0.040002038	0.313888889	0.051851852	0.03
12	0	0	0	0	0	0	Monday	0.025003564	0.28730114	0.40082932	0.261863749	0.025002227	0.482059813	0.102350146	0.03
13	0	0	0	0	0	0	Monday	0.02862788	0.028572877	0.028595689	0.028715367	0.885488188	0.477164502	0.15	0.02
14	0	0	0	0	0	0	Monday	0.150492535	0.025933599	0.025187709	0.304298398	0.494087759	0.534950029	0.100727513	0.05
15	0	0	0	0	0	0	Monday	0.03338623	0.033426586	0.033351578	0.866498538	0.033337068	0.50974359	-0.053084936	0.02
16	0	0	0	0	0	0	Monday	0.028780067	0.028814085	0.028574334	0.885144108	0.028687406	0.295175128	0.057299292	0.01
17	0	0	0	0	0	0	Monday	0.033333876	0.033333675	0.866662607	0.033335017	0.033334825	0.473285048	0.062226662	0.04
18	0	0	0	0	0	0	Monday	0.040080506	0.04014438	0.040027647	0.839741884	0.040005583	0.57962963	0.056378601	0.04
19	0	0	0	0	0	0	Monday	0.866666195	0.033333489	0.03333353	0.033333369	0.033333417	0.374313754	0.212487702	0.06
20	0	0	0	0	0	0	Monday	0.437373579	0.200363493	0.033456789	0.033403472	0.295402666	0.423611111	0.118055556	0.02
21	0	0	0	0	0	0	Monday	0.025000582	0.025163367	0.025000551	0.025000235	0.899835265	0.539251643	0.288259179	0.06
22	0	0	0	0	0	0	Monday	0.020068747	0.020004525	0.020018508	0.020008425	0.919899795	0.506535132	0.290440393	0.06
23	0	0	0	0	0	0	Monday	0.028774389	0.028577206	0.680662392	0.028573875	0.233412138	0.284210526	0.033333333	0.01
24	0	0	0	0	0	0	Monday	0.311931164	0.232677833	0.02857502	0.028574292	0.398241691	0.533657622	0.268667865	0.07
25	0	0	0	0	0	0	Monday	0.033333533	0.033335451	0.69998514	0.200010579	0.033335296	0.396401515	0.210795455	

OnlineNewsPopularity

5).

	AZ	BA	BB	BC	BD	BE	BF	BG	BH
1	global_rate_negative_words	rate_positive_words	rate_negative_words	avg_positive_polarity	min_positive_polarity	max_positive_polarity	avg_negative_polarity	min_negative_polarity	max_negative_p
2	0.01369863	0.769230769	0.230769231	0.378636364	0.1	0.7	-0.35	-0.6	
3	0.015686275	0.733333333	0.266666667	0.286914601	0.033333333	0.7	-0.11875	-0.125	
4	0.009478673	0.857142857	0.142857143	0.495833333	0.1	1	-0.466666667	-0.8	-0.13
5	0.020715631	0.666666667	0.333333333	0.385965171	0.136363636	0.8	-0.36969697	-0.6	-0.16
6	0.012126866	0.860215054	0.139784946	0.411127435	0.033333333	1	-0.220192308	-0.5	
7	0.027027027	0.523809524	0.476190476	0.350609996	0.136363636	0.6	-0.195	-0.4	
8	0.016666667	0.827956989	0.172043011	0.402038567	0.1	1	-0.224479167	-0.5	
9	0.015166835	0.846938776	0.153061224	0.427720492	0.1	1	-0.242777778	-0.5	
10	0.020618557	0.6	0.4	0.566666667	0.4	0.8	-0.125	-0.125	
11	0.03030303	0.5625	0.4375	0.298412698	0.1	0.5	-0.238095238	-0.5	
12	0.020833333	0.648648649	0.351351351	0.404480069	0.1	1	-0.415064103	-1	
13	0.010695187	0.714285714	0.285714286	0.435	0.2	0.7	-0.2625	-0.4	
14	0.02919708	0.636363636	0.363636364	0.375510204	0.2	0.7	-0.310416667	-0.6	
15	0.052631579	0.347826087	0.652173913	0.4575	0.16	1	-0.337888889	-0.7	
16	0.011583012	0.571428571	0.428571429	0.249090909	0.136363636	0.5	-0.138690476	-0.1875	
17	0.039589443	0.557377049	0.442622951	0.34307041	0.05	0.6	-0.220149912	-0.6	
18	0.025939177	0.613333333	0.386666667	0.504528986	0.1	1	-0.401436782	-1	
19	0.010075567	0.866666667	0.133333333	0.381847319	0.033333333	1	-0.144642857	-0.2	
20	0.009708738	0.75	0.25	0.277777778	0.033333333	0.5	-0.125	-0.125	
21	0.011599006	0.857142857	0.142857143	0.426568491	0.1	1	-0.226785714	-0.5	
22	0.011217949	0.858585859	0.141414141	0.408244716	0.1	1	-0.206547619	-0.5	
23	0.00511509	0.777777778	0.222222222	0.15	0.05	0.35	-0.108333333	-0.166666667	
24	0.016464471	0.819047619	0.180952381	0.416342998	0.1	1	-0.230263158	-0.5	
25	0	1	0	0.281060606	0.1	0.6	0	0	

OnlineNewsPopularity

6).

	BH	BI	BJ	BK	BL	BM	BN	BO	BP	BQ	BR	BS
1	max_negative_polarity	title_subjectivity	title_sentiment_polarity	abs_title_subjectivity	abs_title_sentiment_polarity	state	day_phase	avg_time_spent	device_type	shares	popularity	popularity_level
2	-0.2	0.5	-0.1875	0	0.1875	Virginia	Evening	15	Mobile	593	0	3
3	-0.1	0	0	0.5	0	New York	Evening	20	Mobile	711	0	3
4	-0.133333333	0	0	0.5	0	Vermont	Night	7	Tablet	1500	1	2
5	-0.166666667	0	0	0.5	0	Washington	Night	4	Mobile	1200	0	2
6	-0.05	0.454545455	0.136363636	0.045454545	0.136363636	Michigan	Night	1	PC	505	0	3
7	-0.1	0.642857143	0.214285714	0.142857143	0.214285714	California	Night	1	PC	855	0	3
8	-0.05	0	0	0.5	0	Florida	Night	15	PC	556	0	3
9	-0.05	1	0.5	0.5	0.5	California	Night	7	PC	891	0	3
10	-0.125	0.125	0	0.375	0	Missouri	Evening	5	Mobile	3600	1	1
11	-0.1	0	0	0.5	0	New York	Morning	1	Mobile	710	0	3
12	-0.1	0	0	0.5	0	Massachusetts	Afternoon	1	Mobile	2200	1	1
13	-0.125	0	0	0.5	0	Delaware	Afternoon	1	Mobile	1900	1	2
14	-0.05	1	-1	0.5	1	California	Afternoon	5	Mobile	823	0	3
15	-0.1	1	-1	0.5	1	New York	Morning	8	Mobile	10000	1	1
16	-0.05	0.75	0.55	0.25	0.55	Florida	Night	12	Mobile	761	0	3
17	-0.05	0.75	-0.25	0.25	0.25	Illinois	Morning	5	Mobile	1600	1	2
18	-0.05	0.566666667	-0.1	0.066666667	0.1	Illinois	Afternoon	8	Tablet	13600	1	1
19	-0.1	0	0	0.5	0	Massachusetts	Night	1	Mobile	3100	1	1
20	-0.125	0.857142857	-0.714285714	0.357142857	0.714285714	Washington	Night	1	Tablet	5700	1	1
21	-0.05	0.5	0	0	0	California	Night	11	Mobile	17100	1	1
22	-0.05	0	0	0.5	0	Missouri	Night	5	Tablet	2800	1	1
23	-0.05	0	0	0.5	0	Maryland	Night	4	Mobile	598	0	3
24	-0.05	0	0	0.5	0	New York	Evening	5	Mobile	445	0	3
25	0	0.45	0.4	0.05	0.4	California	Afternoon	1	Mobile	1500	1	2

OnlineNewsPopularity

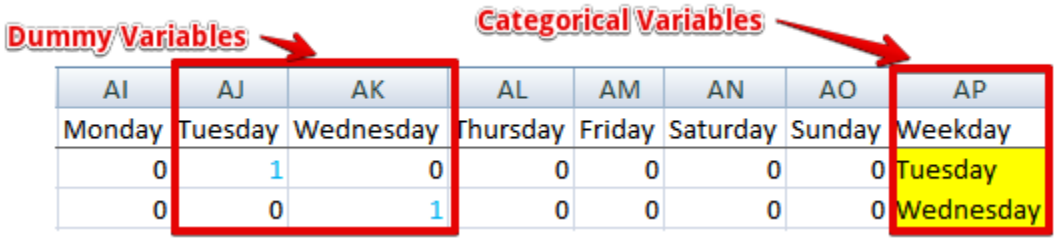
DATA CLEANING AND PRE-PROCESSING

The dataset didn't contain any missing values. However, certain categorical columns had to be sparsed so as to run linear regression model as our first given target variable – "shares" is a continuous numerical variable.

1. **Extracting Year and Month from "url" attribute:**
   We have used excel formula e.g. "=MID(A2,21,4)" to extract year and "=MID(A2,26,2)" to extract month from the "url" attribute.

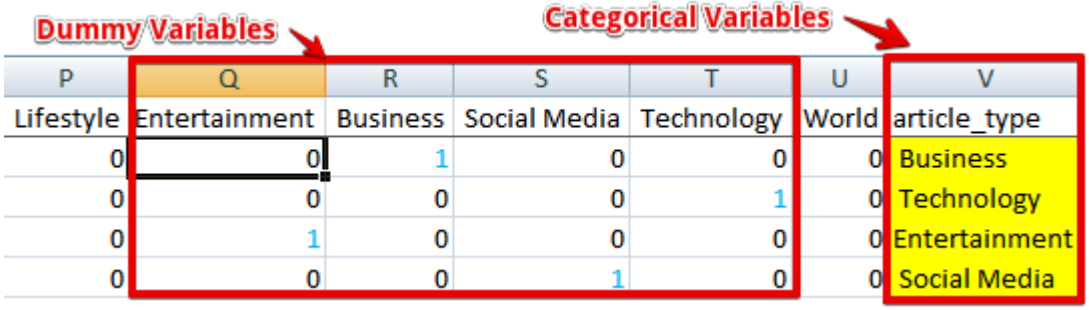
	A	B	C
1	url	year	month
2	http://mashable.com/2013/01/07/amazon-instant-video-browser/	2013	1

2. **Sparsing 'Weekdays' attribute into Dummy variables:**
   We have used excel formula e.g. "=INDEX(AJ$1:AO$1,MATCH(MAX(AJ2:AM2),AJ2:AM2,0))" to sparse the categorical variable "Weekday" into dummy variable sets.



AI	AJ	AK	AL	AM	AN	AO	AP
Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday	Weekday
0	1	0	0	0	0	0	Tuesday
0	0	1	0	0	0	0	Wednesday

3. **Sparsing'article_type' attribute into Dummy variables:**
   We have used excel formula e.g. "=INDEX(P$1:U$1,MATCH(MAX(P2:U2),P2:U2,0))" to sparse the categorical variable "Weekday" into dummy variable sets.



P	Q	R	S	T	U	V
Lifestyle	Entertainment	Business	Social Media	Technology	World	article_type
0	0	1	0	0	0	Business
0	0	0	0	1	0	Technology
0	1	0	0	0	0	Entertainment
0	0	0	1	0	0	Social Media

4. **Target variable 'popularity' and 'popularity_level' based on attribute 'shares':**

We have used below excel formulae to create two more categorical target variable e.g.

(I) "=IF(BQ>1400,1,0)" – for 'popularity' where value = '1' for shares>1400 and value = '0' otherwise.

(II) "=IF(BQ>2100,1,(IF(BQ=>1100 and <=2100),2,3))" – for 'popularity' where value = '1' for shares>2100 and value = '2' for shares between 1100 and 2100 and value = '3' for shares < 1100.

BQ	BR	BS
shares	popularity	popularity_level
459	0	3
1400	0	2
6400	1	1

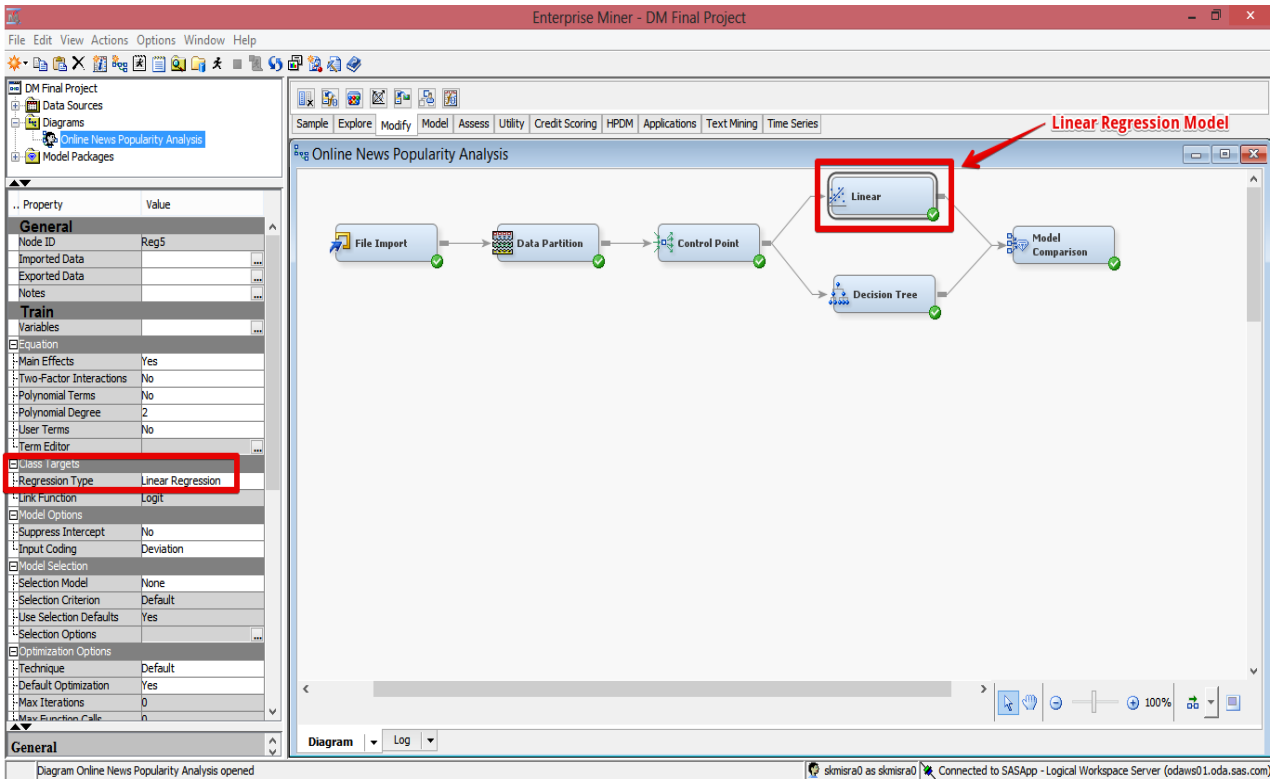
DATA MODELING AND CONCLUSIONS

In order to execute models for prediction, we have used SAS Enterprise Miner tool.

Problem 1: To predict the number of Mashable article shares.

Approach 1: Use Kitchen sink model on Linear Regression Algorithm

Screenshot of the SAS Dashboard:



Result:

Buttons Used	Target Variable	Model	Result
File Import, Data Partition, Control Point, Linear Regression	Shares	Linear	Model Fit Statistics R-Square 0.0242 Adj R-Sq 0.0199 AIC 519812.0351 BIC 519815.1482 SBC 520832.6825 C(p) 124.0000

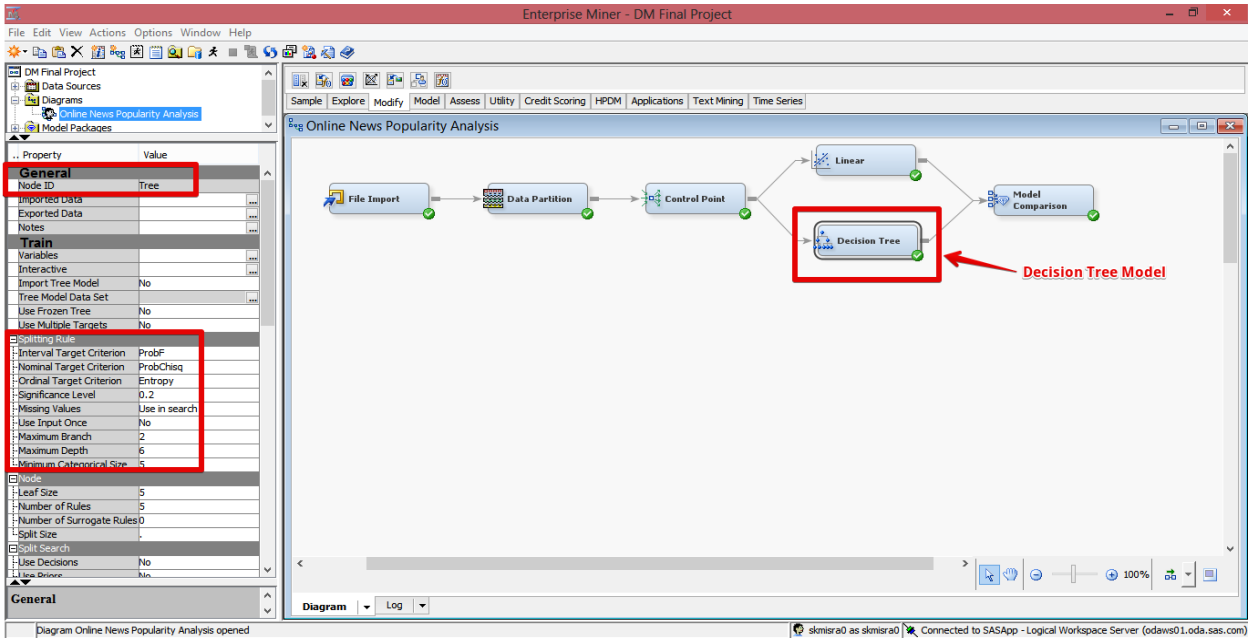
Conclusion: Since our first goal is to predict 'shares' that is a continuous variable, for input we removed certain categorical variables that were added for extended prediction as explained above in data pre-processing section – 'Weekday', 'article_type', 'popularity', 'popularity_level'.

Note: We also have removed attributes ‘url’ and ‘timedelta’ that were assumed non-predictive in the beginning.

After Execution we found that the R-Square value is only 0.0242 i.e. only 2% of the variance in share attribute is explained by given independent predictor variables. This is too low for a model to be considered for prediction.

Approach 2: Use Kitchen sink model on Decision Tree Algorithm

Screenshot of the SAS Dashboard:



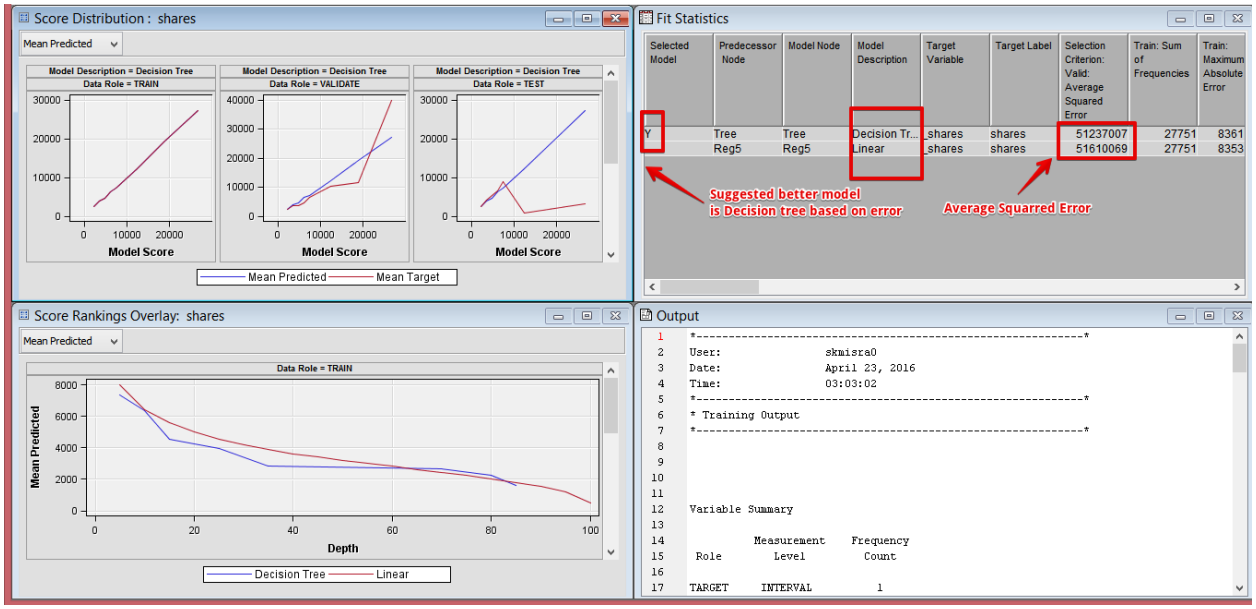
Result:

Buttons Used	Target Variable	Model	Result
File Import, Data Partition, Control Point, Decision Tree	Shares	Decision Tree	

Fit Statistics	Statistics Label	Train	Validation	Test
NOBS	Sum of Frequencies	27751.00	7929.00	3964.00
MAX	Maximum Absolute Error	836112.18	190412.18	645712.18
SSE	Sum of Squared Errors	3.7619029E12	406258225661.72	1.078877E12
ASE	Average Squared Error	135559183.47	51237006.64	272168774.54
RASE	Root Average Squared Error	11642.99	7158.00	16497.54
DIV	Divisor for ASE	27751.00	7929.00	3964.00
DFT	Total Degrees of Freedom	27751.00	.	.

Conclusion: we also ran Decision Tree for probabilistic estimation of the share values, for which again, we received very high Average Squared error = 51237007.

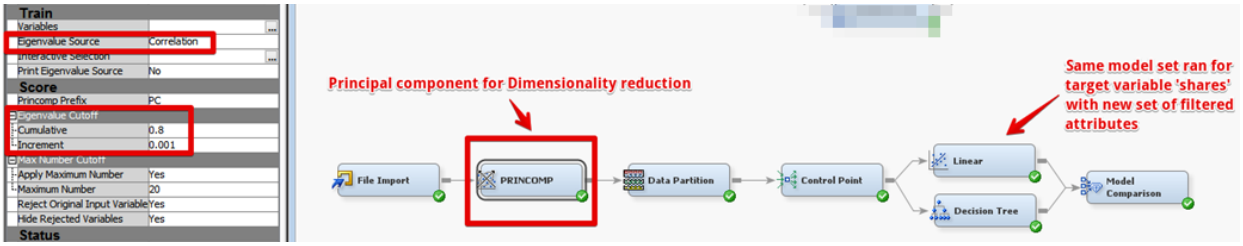
Model Comparison Result:



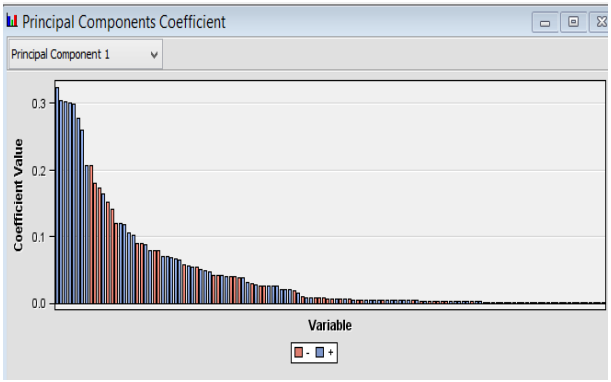
FINAL RESULT:*Based on these statistics, we can clearly see that predicting a continuous target variable is not fruitful using these set of attributes and thus we decided to go for refining our attribute set using 'Principal Component Analysis'.*

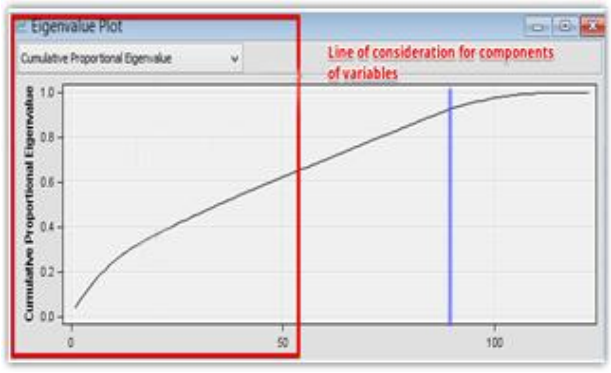
Approach 3: Use Principal Component Analysis and Linear Regression Algorithm

Screenshot of the SAS Dashboard:



Following graphs shows set of variables selected using Principal Component Analysis technique based on maximum variance achieved in components. We have selected 80% as the cumulative cut-off score and accordingly selected minimum PC components.





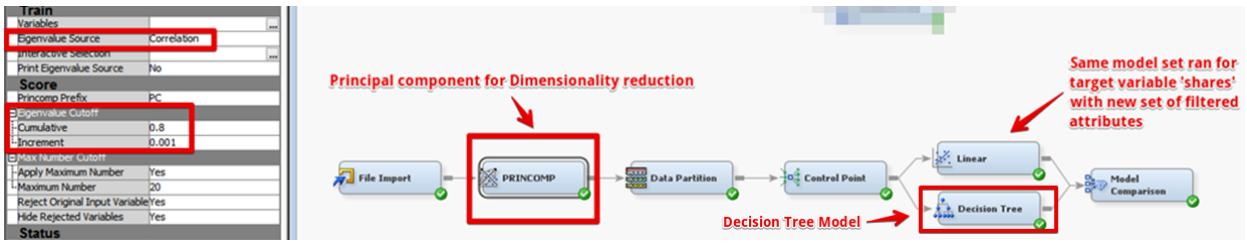
Result:

Buttons Used	Target Variable	Model	Result
File Import, Principal Component, Data Partition, Control Point, Linear Regression	Shares	Linear	Model Fit Statistics R-Square 0.0143 Adj R-Sq 0.0136 AIC 519887.1583 BIC 519889.1901 SBC 520060.0098 C(p) 21.0000

Conclusion: Even after adding Dimensionality reduction algorithm like 'Principal Component Analysis' when we ran with similar set of attributes as in execution set 1, we found that our models went worse with the R-Square value as low as 0.0143 i.e. only 1.4% of the variance in share attribute is explained by given independent predictor variables. This is again too low for a model to be considered for prediction.

Approach 4: Use Principal Component Analysis and Decision Tree Algorithm

Screenshot of the SAS Dashboard:

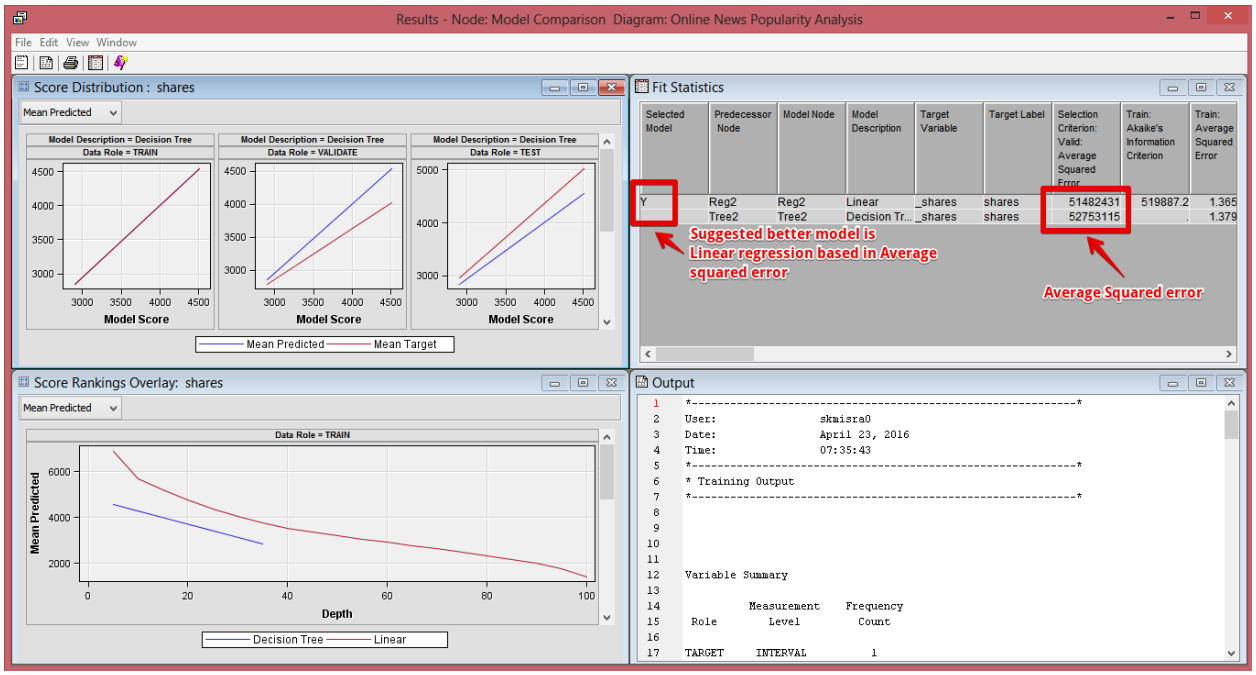


Result:

Buttons Used	Target Variable	Model	Result
File Import, Principal Component, Data Partition, Control Point, Decision Tree	Shares	Decision Tree	Fit Statistics Target=_shares Target Label=' shares' Fit Statistics / Statistics Label / Train / Validation / Test _NOBS_ Sum of Frequencies 27751.00 7929.00 3964.00 _MAX_ Maximum Absolute Error 838751.77 193051.77 648351.77 _SSE_ Sum of Squared Errors 3.8277178E12 418279446733.66 1.0897152E12 _ASE_ Average Squared Error 137930805.99 52753114.73 274902920.60 _RASE_ Root Average Squared Error 11744.39 7263.13 16580.20 _DIV_ Divisor for ASE 27751.00 7929.00 3964.00 _DFT_ Total Degrees of Freedom 27751.00 . .

Conclusion: we also ran Decision Tree for probabilistic estimation of the share values, for which again, we received very high Average Squared error = 52753114.

Model Comparison Result:



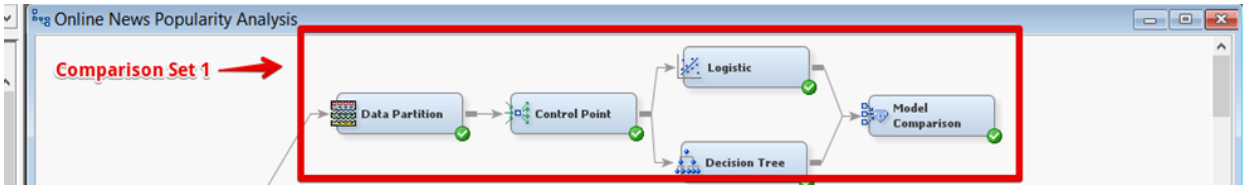
FINAL RESULT: ***Based on these statistics, we can finally conclude that predicting a continuous target variable is not fruitful using this dataset and thus we decided to create categorical (binary) variable 'popularity' as target variable and try executing logistic regression and other classification models.***

REASONING: ***Similar is the case with stock price prediction example from the book 'Data Science for Business', where exact stock price value prediction cannot be made. In such situations, we opt for binary categorical target variable keeping a threshold value from continuous target variable as base value for prediction of 'SURGE' or 'PLUNGE' in the target.***

Problem2:To predict binary target variable 'Popularity'

Approach 1: Use Kitchen sink model on Linear Regression Algorithm and Decision Tree Algorithm

Screenshot of the SAS Dashboard:



Result:

Buttons Used	Target Variable	Model	Result
File Import, Data Partition, Control Point, Logistic Regression **(SET 1)**	Popularity (binary = '0','1')	Logistic	Fit Statistics / Statistics Label / Train / Validation / Test _AIC_ Akaike's Information Criterion 34819.38 . . _ASE_ Average Squared Error 0.22 0.22 0.22 _AVERR_ Average Error Function 0.62 0.63 0.62 _DFE_ Degrees of Freedom for Error 27625.00 . . _DFM_ Model Degrees of Freedom 123.00 . . _DFT_ Total Degrees of Freedom 27748.00 . . _DIV_ Divisor for ASE 55496.00 15856.00 7936.00 _ERR_ Error Function 34573.38 9997.71 4947.38 _FPE_ Final Prediction Error 0.22 . . _MAX_ Maximum Absolute Error 1.00 1.00 0.94 _MSE_ Mean Square Error 0.22 0.22 0.22 _NOBS_ Sum of Frequencies 27748.00 7928.00 3968.00 _NW_ Number of Estimate Weights 123.00 . . _RASE_ Root Average Sum of Squares 0.47 0.47 0.47 _RFPE_ Root Final Prediction Error 0.47 . . _RMSE_ Root Mean Squared Error 0.47 0.47 0.47 _SBC_ Schwarz's Bayesian Criterion 35831.78 . . _SSE_ Sum of Squared Errors 12004.77 3475.71 1722.20 _SUMW_ Sum of Case Weights Times Freq 55496.00 15856.00 7936.00 _MISC_ Misclassification Rate 0.34 0.35 0.34
File Import, Data Partition, Control Point, Decision Tree**(SET 1)**	Popularity (binary = '0','1')	Decision Tree	Fit Statistics / Statistics Label / Train / Validation / Test _NOBS_ Sum of Frequencies 27748.00 7928.00 3968.00 _MISC_ Misclassification Rate 0.35 0.36 0.37 _MAX_ Maximum Absolute Error 0.83 0.83 0.83 _SSE_ Sum of Squared Errors 12367.35 3593.78 1813.95 _ASE_ Average Squared Error 0.22 0.23 0.23 _RASE_ Root Average Squared Error 0.47 0.48 0.48 _DIV_ Divisor for ASE 55496.00 15856.00 7936.00 _DFT_ Total Degrees of Freedom 27748.00 . .

Confusion Matrices:

Logistic Regression:

Event Classification Table

Data Role=TRAIN Target=popularity Target Label=' '

False Negative	True Negative	False Positive	True Positive
5041	9591	4465	8651

Data Role=VALIDATE Target=popularity Target Label=' '

False Negative	True Negative	False Positive	True Positive
1492	2745	1271	2420

Decision Tree:

Event Classification Table

Data Role=TRAIN Target=popularity Target Label=' '

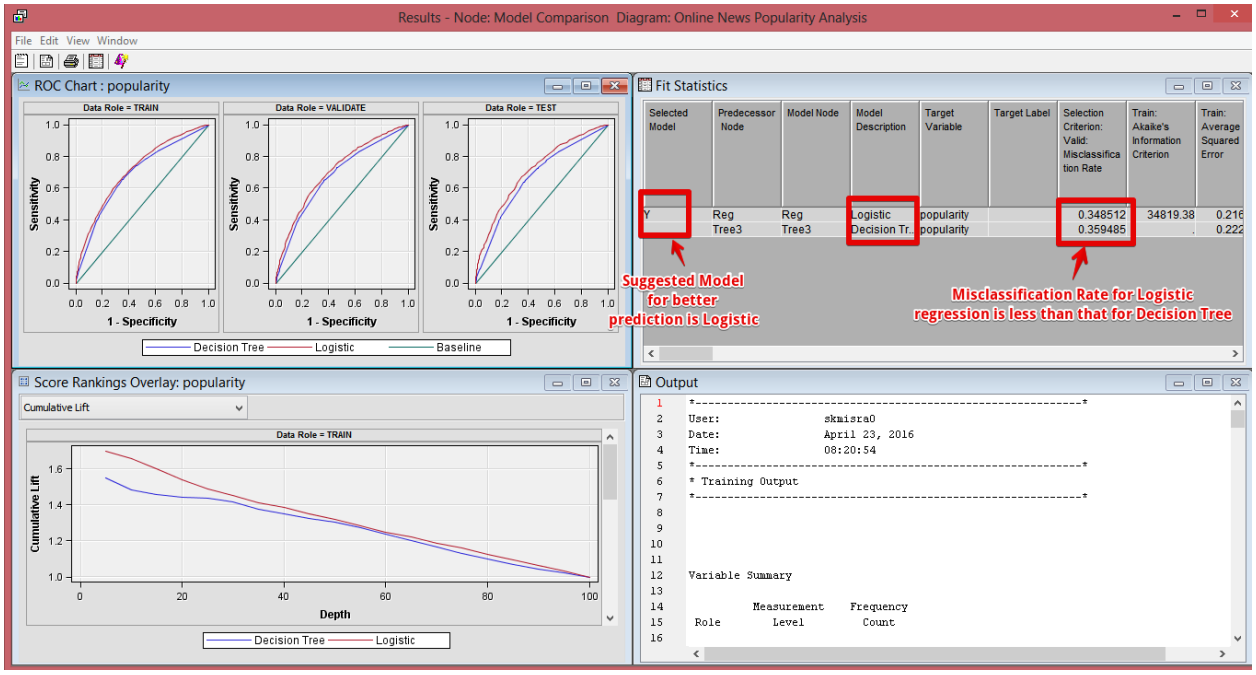
False Negative	True Negative	False Positive	True Positive
4700	9066	4990	8992

Data Role=VALIDATE Target=popularity Target Label=' '

False Negative	True Negative	False Positive	True Positive
1379	2545	1471	2533

Model Comparison Results:

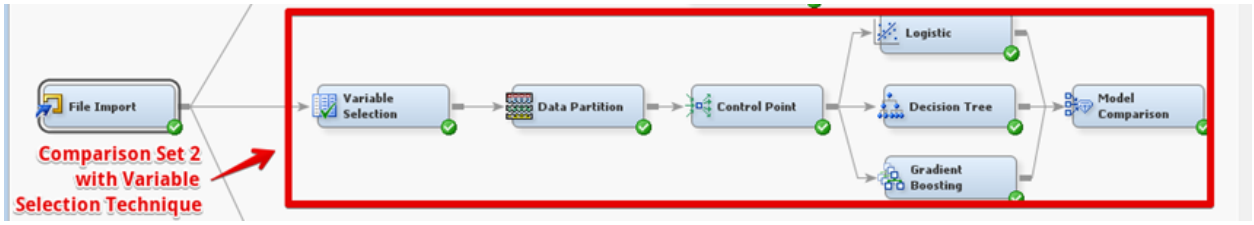
FOR SET 1:



Conclusion: Without implementing any dimensionality reduction technique, we have found better prediction results for binary target variable 'popularity' from Logistic Regression model. This is because the misclassification rate from logistic regression =0.348512 while that from Decision Tree =0.359485. This can alternatively be stated as **the prediction accuracy of logistic regression model is ~65%** which is considerably good for forecasting.

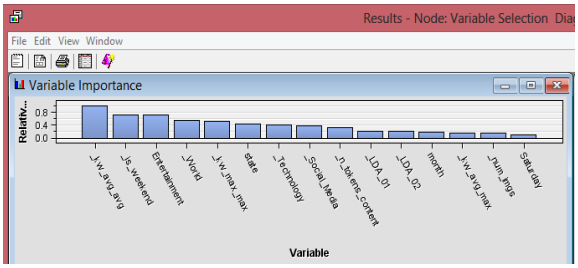
Approach 2: Use Variable Selection on Logistic Regression, Decision Tree and Gradient Boosting Algorithm

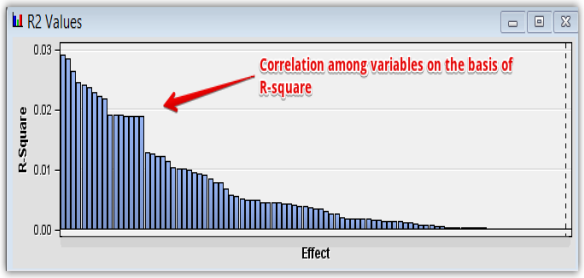
Screenshot of the SAS Dashboard:



Following graphs shows set of variables selected using R-Square and Chi-Square criteria from Variable Selection technique block.

We can see that number of effective variables have drastically reduced from 71 variables to 15 variables only.





Result:

Buttons Used	Target Variable	Model	Result
File Import, **Variable Selection**, Data Partition, Control Point, Logistic Regression **(SET 2)**	Popularity (binary = '0','1')	Logistic	Fit Statistics / Statistics Label / Train / Validation / Test _AIC_ Akaike's Information Criterion 35515.58 . . _ASE_ Average Squared Error 0.22 0.22 0.22 _AVERR_ Average Error Function 0.64 0.64 0.64 _DFE_ Degrees of Freedom for Error 27736.00 . . _DFM_ Model Degrees of Freedom 12.00 . . _DFT_ Total Degrees of Freedom 27748.00 . . _DIV_ Divisor for ASE 55496.00 15856.00 7936.00 _ERR_ Error Function 35491.58 10166.84 5073.81 _FPE_ Final Prediction Error 0.22 . . _MAX_ Maximum Absolute Error 1.00 1.00 0.98 _MSE_ Mean Square Error 0.22 0.22 0.22 _NOBS_ Sum of Frequencies 27748.00 7928.00 3968.00 _NW_ Number of Estimate Weights 12.00 . . _RASE_ Root Average Sum of Squares 0.47 0.47 0.47 _RFPE_ Root Final Prediction Error 0.47 . . _RMSE_ Root Mean Squared Error 0.47 0.47 0.47 _SBC_ Schwarz's Bayesian Criterion 35614.35 . . _SSE_ Sum of Squared Errors 12401.94 3560.32 1776.83 _SUMW_ Sum of Case Weights Times Freq 55496.00 15856.00 7936.00 _MISC_ Misclassification Rate 0.36 0.37 0.36
File Import, **Variable Selection**, Data Partition, Control Point, Decision Tree**(SET 2)**	Popularity (binary = '0','1')	Decision Tree	Fit Statistics / Statistics Label / Train / Validation / Test _NOBS_ Sum of Frequencies 27748.00 7928.00 3968.00 _MISC_ Misclassification Rate 0.35 0.36 0.37 _MAX_ Maximum Absolute Error 0.83 0.83 0.76 _SSE_ Sum of Squared Errors 12524.76 3623.41 1837.75 _ASE_ Average Squared Error 0.23 0.23 0.23 _RASE_ Root Average Squared Error 0.48 0.48 0.48 _DIV_ Divisor for ASE 55496.00 15856.00 7936.00 _DFT_ Total Degrees of Freedom 27748.00 . .

File Import, **Variable Selection**, Data Partition, Control Point, Gradient Boosting**(SET 2)**	Popularity (binary = '0','1')	Gradient Boosting	Fit Statistics

Fit Statistics	Statistics Label	Train	Validation	Test
NOBS	Sum of Frequencies	27748.00	7928.00	3968.00
SUMW	Sum of Case Weights Times Freq	55496.00	15856.00	7936.00
MISC	Misclassification Rate	0.37	0.38	0.38
MAX	Maximum Absolute Error	0.71	0.71	0.71
SSE	Sum of Squared Errors	12723.41	3666.54	1824.79
ASE	Average Squared Error	0.23	0.23	0.23
RASE	Root Average Squared Error	0.48	0.48	0.48
DIV	Divisor for ASE	55496.00	15856.00	7936.00
DFT	Total Degrees of Freedom	27748.00	.	.

Confusion Matrices:

Logistic:

Event Classification Table

Data Role=TRAIN Target=popularity Target Label=' '

False Negative	True Negative	False Positive	True Positive
5152	9128	4928	8540

Data Role=VALIDATE Target=popularity Target Label=' '

False Negative	True Negative	False Positive	True Positive
1488	2575	1441	2424

Decision Tree:

Event Classification Table

Data Role=TRAIN Target=popularity Target Label=' '

False Negative	True Negative	False Positive	True Positive
4467	8756	5300	9225

Data Role=VALIDATE Target=popularity Target Label=' '

False Negative	True Negative	False Positive	True Positive
1325	2486	1530	2587

Gradient Boosting:

Event Classification Table

Data Role=TRAIN Target=popularity Target Label=' '

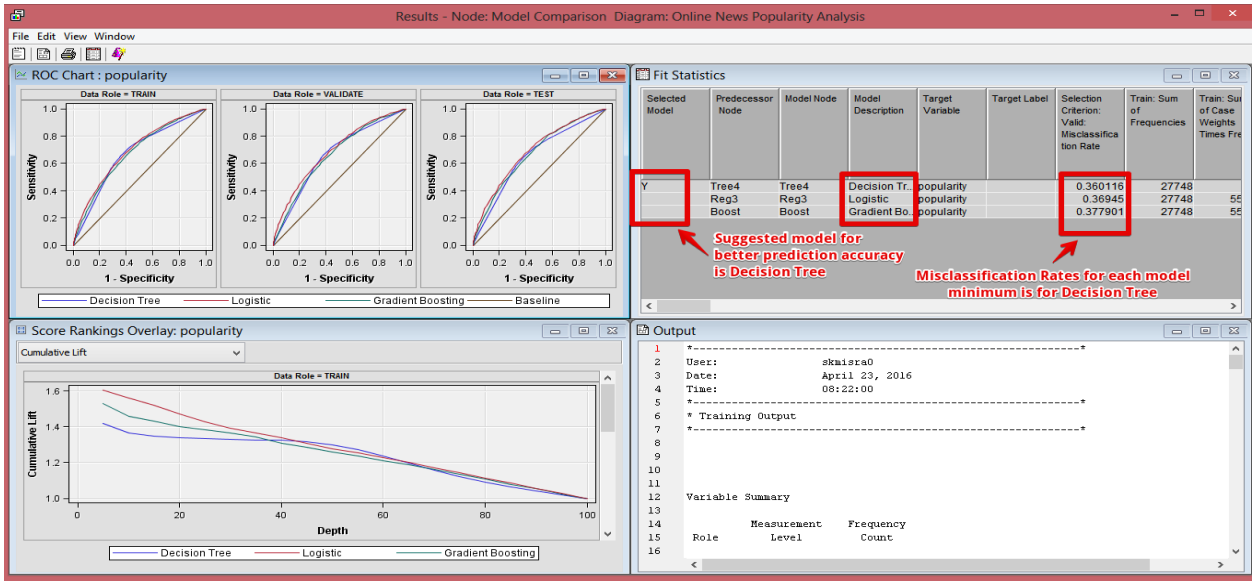
False Negative	True Negative	False Positive	True Positive
5339	9051	5005	8353

Data Role=VALIDATE Target=popularity Target Label=' '

False Negative	True Negative	False Positive	True Positive
1545	2565	1451	2367

Model Comparison Results:

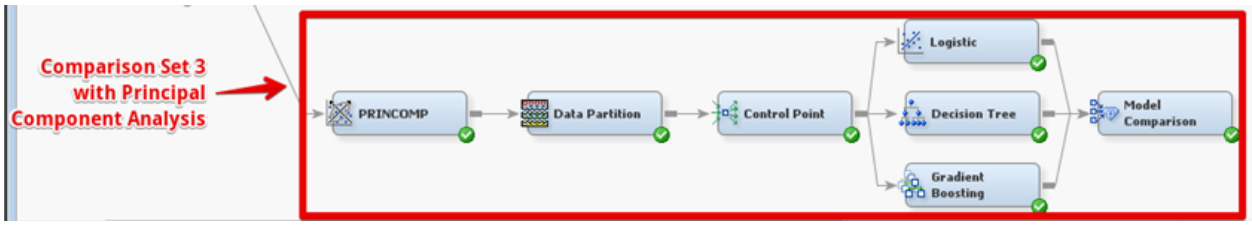
FOR SET 2:



Conclusion: When implementing dimensionality reduction technique "Variable Selection" using R-Square and Chi-Square values both, we have found better prediction results for binary target variable 'popularity' from Decision Tree model. This is because the misclassification rate from Decision Tree =0.360116 while that from Logistic Regression =0.36945 and that from Gradient Boosting = 0.377901. This can alternatively be stated as **the prediction accuracy of decision tree model is ~63%** which is considerably good for forecasting.

Approach 3: Use Principal Component Analysis on Logistic Regression, Decision Tree and Gradient Boosting Algorithm

Screenshot of the SAS Dashboard:



Result:

Buttons Used	Target Variable	Model	Result
File Import, **Principal Component**, Data Partition, Control Point, Logistic Regression **(SET 3)**	Popularity (binary = '0','1')	Logistic	
File Import, **Principal Component**, Data Partition, Control Point, Decision Tree **(SET 3)**	Popularity (binary = '0','1')	Decision Tree	
File Import, **Principal Component**, Data Partition, Control Point, Gradient Boosting **(SET 3)**	Popularity (binary = '0','1')	Gradient Boosting	

Fit Statistics

Statistics	Statistics Label	Train	Validation	Test
AIC	Akaike's Information Criterion	35978.58	.	.
ASE	Average Squared Error	0.23	0.23	0.23
AVERR	Average Error Function	0.65	0.65	0.65
DFE	Degrees of Freedom for Error	27727.00	.	.
DFM	Model Degrees of Freedom	21.00	.	.
DFT	Total Degrees of Freedom	27748.00	.	.
DIV	Divisor for ASE	55496.00	15856.00	7936.00
ERR	Error Function	35936.58	10346.61	5123.80
FPE	Final Prediction Error	0.23	.	.
MAX	Maximum Absolute Error	1.00	1.00	0.96
MSE	Mean Square Error	0.23	0.23	0.23
NOBS	Sum of Frequencies	27748.00	7928.00	3968.00
_NW_	Number of Estimate Weights	21.00	.	.
RASE	Root Average Sum of Squares	0.48	0.48	0.48
RFPE	Root Final Prediction Error	0.48	.	.
RMSE	Root Mean Squared Error	0.48	0.48	0.48
SBC	Schwarz's Bayesian Criterion	36151.43	.	.
SSE	Sum of Squared Errors	12601.75	3635.82	1801.88
_SUMW	Sum of Case Weights Times Freq	55496.00	15856.00	7936.00
MISC	Misclassification Rate	0.37	0.38	0.38

Fit Statistics

Statistics	Statistics Label	Train	Validation	Test
NOBS	Sum of Frequencies	27748.00	7928.00	3968.00
MISC	Misclassification Rate	0.37	0.38	0.41
MAX	Maximum Absolute Error	0.87	0.87	0.87
SSE	Sum of Squared Errors	12704.96	3652.83	1860.22
ASE	Average Squared Error	0.23	0.23	0.23
RASE	Root Average Squared Error	0.48	0.48	0.48
DIV	Divisor for ASE	55496.00	15856.00	7936.00
DFT	Total Degrees of Freedom	27748.00	.	.

Fit Statistics

Statistics	Statistics Label	Train	Validation	Test
NOBS	Sum of Frequencies	27748.00	7928.00	3968.00
_SUMW	Sum of Case Weights Times Freq	55496.00	15856.00	7936.00
MISC	Misclassification Rate	0.37	0.37	0.38
MAX	Maximum Absolute Error	0.77	0.76	0.76
SSE	Sum of Squared Errors	12680.50	3639.70	1828.09
ASE	Average Squared Error	0.23	0.23	0.23
RASE	Root Average Squared Error	0.48	0.48	0.48
DIV	Divisor for ASE	55496.00	15856.00	7936.00
DFT	Total Degrees of Freedom	27748.00	.	.

Confusion Matrices:

Logistic:

Event Classification Table

Data Role=TRAIN Target=popularity Target Label=' '

False Negative	True Negative	False Positive	True Positive
5132	9026	5030	8560

Data Role=VALIDATE Target=popularity Target Label=' '

False Negative	True Negative	False Positive	True Positive
1533	2566	1450	2379

Decision Tree:

Event Classification Table

Data Role=TRAIN Target=popularity Target Label=' '

False Negative	True Negative	False Positive	True Positive
4258	7938	6118	9434

Data Role=VALIDATE Target=popularity Target Label=' '

False Negative	True Negative	False Positive	True Positive
1235	2259	1757	2677

Gradient Boosting:

Event Classification Table

Data Role=TRAIN Target=popularity Target Label=' '

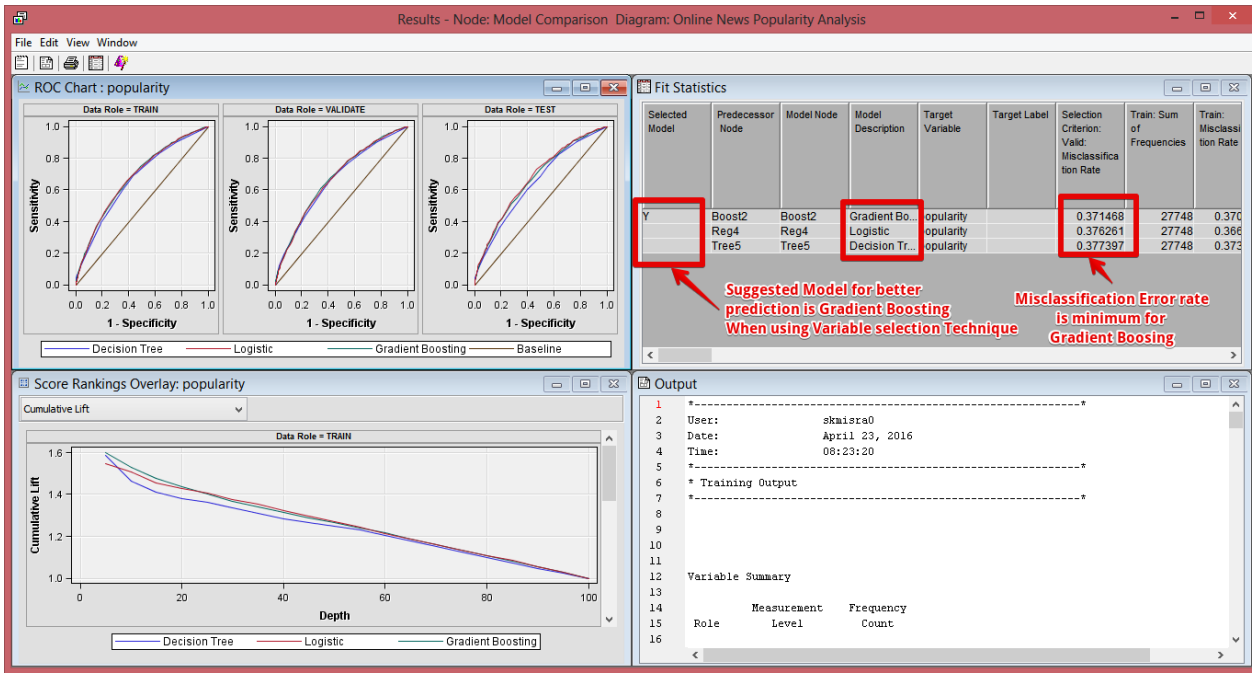
False Negative	True Negative	False Positive	True Positive
4838	8623	5433	8854

Data Role=VALIDATE Target=popularity Target Label=' '

False Negative	True Negative	False Positive	True Positive
1400	2471	1545	2512

Model Comparison Results:

FOR SET 3:



Conclusion: When implementing dimensionality reduction technique "Principal Component", we have found better prediction results for binary target variable 'popularity' from Gradient Boosting model. This is because the misclassification rate from Gradient boosting = 0.371468 while that from Logistic Regression = 0.376261 and that from Decision Tree = 0.377397. This can alternatively be stated as **the prediction accuracy of gradient boosting model is ~63%** which is considerably good for forecasting.

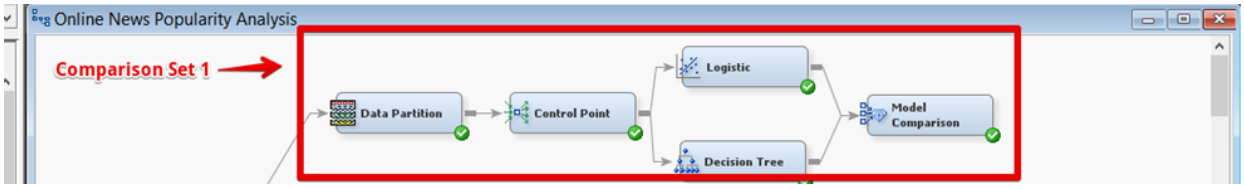
FINAL RESULT:***Based on these statistics, we can finally conclude that predicting a binary target variable to check whether the article published is 'popular' or 'not popular' we must go for Decision Tree with Variable Selection dimensionality reduction technique.***

REASONING:***Although with a kitchen sink model, we achieved ~65% accuracy using logistic regression, the model seems too complex. However, on compromising only ~1% accuracy, we built a model with less input variables keeping our model simple. Thus, we prefer Decision Tree with Variable Selection dimensionality reduction technique over any other model.***

Problem3:To predict ordinal outcome for 'Popularity_level'

Approach 1: Use Kitchen sink model on Logistic Regression Algorithm and Decision Tree Algorithm

Screenshot of the SAS Dashboard:



Result:

Buttons Used	Target Variable	Model	Result
File Import, Data Partition, Control Point, Logistic Regression **(SET 1)**	Popularity_level (ordinal = '1','2','3')	Logistic	
File Import, Data Partition, Control Point, Decision Tree **(SET 1)**	Popularity_level (ordinal = '1','2','3')	Decision Tree	

Fit Statistics

Fit Statistics	Statistics Label	Train	Validation	Test
AIC	Akaike's Information Criterion	56478.17	.	.
ASE	Average Squared Error	0.20	0.20	0.20
AVERR	Average Error Function	0.68	0.68	0.68
DFE	Degrees of Freedom for Error	55374.00	.	.
DFM	Model Degrees of Freedom	124.00	.	.
DFT	Total Degrees of Freedom	55498.00	.	.
DIV	Divisor for ASE	83247.00	23781.00	11904.00
ERR	Error Function	56230.17	16166.10	8052.59
FPE	Final Prediction Error	0.20	.	.
MAX	Maximum Absolute Error	1.00	1.00	0.97
MSE	Mean Square Error	0.20	0.20	0.20
NOBS	Sum of Frequencies	27749.00	7927.00	3968.00
_NW_	Number of Estimate Weights	124.00	.	.
RASE	Root Average Sum of Squares	0.45	0.45	0.45
RFPE	Root Final Prediction Error	0.45	.	.
RMSE	Root Mean Squared Error	0.45	0.45	0.45
SBC	Schwarz's Bayesian Criterion	57584.76	.	.
SSE	Sum of Squared Errors	16871.99	4845.52	2417.53
SUMW	Sum of Case Weights Times Freq	83247.00	23781.00	11904.00
MISC	Misclassification Rate	0.51	0.52	0.52

Fit Statistics

Fit Statistics	Statistics Label	Train	Validation	Test
NOBS	Sum of Frequencies	27749.00	7927.00	3968.00
MISC	Misclassification Rate	0.52	0.53	0.54
MAX	Maximum Absolute Error	0.94	1.00	1.00
SSE	Sum of Squared Errors	17436.27	5017.38	2532.61
ASE	Average Squared Error	0.21	0.21	0.21
RASE	Root Average Squared Error	0.46	0.46	0.46
DIV	Divisor for ASE	83247.00	23781.00	11904.00
DFT	Total Degrees of Freedom	55498.00	.	.

Confusion Matrices:

Logistic Regression:

Event Classification Table

Data Role=TRAIN Target=popularity_level Target Label=' '

False Negative	True Negative	False Positive	True Positive
2549	10022	7415	7763

Data Role=VALIDATE Target=popularity_level Target Label=' '

False Negative	True Negative	False Positive	True Positive
757	2775	2207	2188

Decision Tree:

Event Classification Table

Data Role=TRAIN Target=popularity_level Target Label=' '

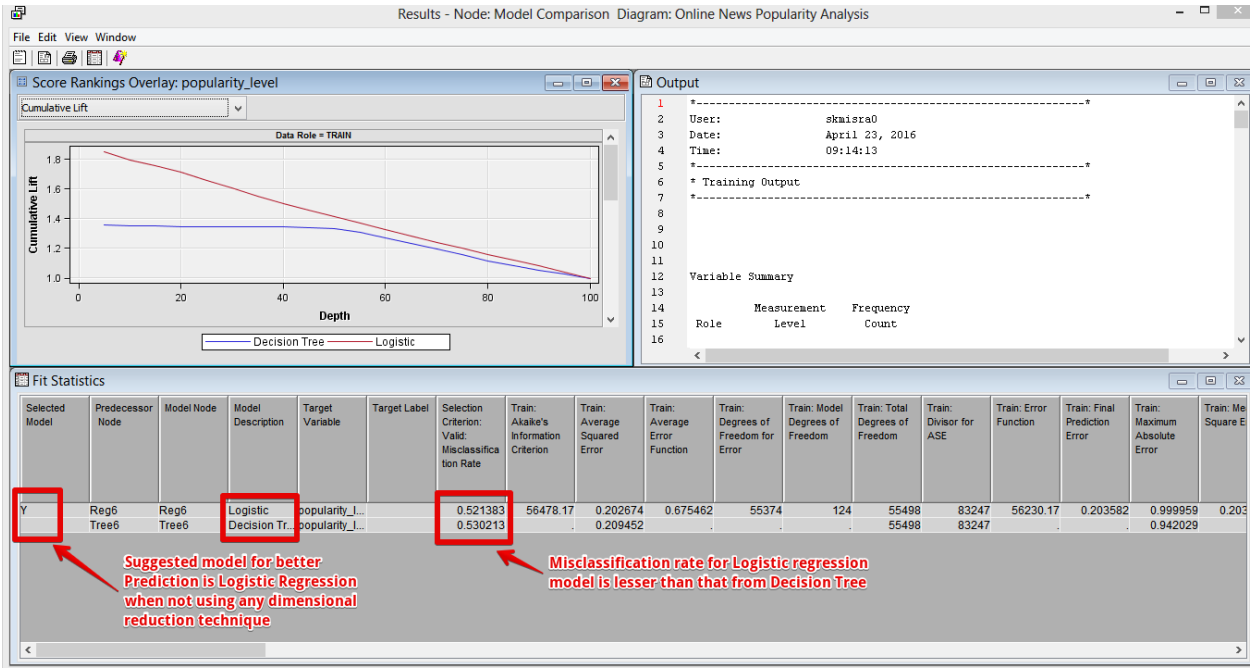
False Negative	True Negative	False Positive	True Positive
3090	9951	7486	7222

Data Role=VALIDATE Target=popularity_level Target Label=' '

False Negative	True Negative	False Positive	True Positive
869	2818	2164	2076

Model Comparison Results:

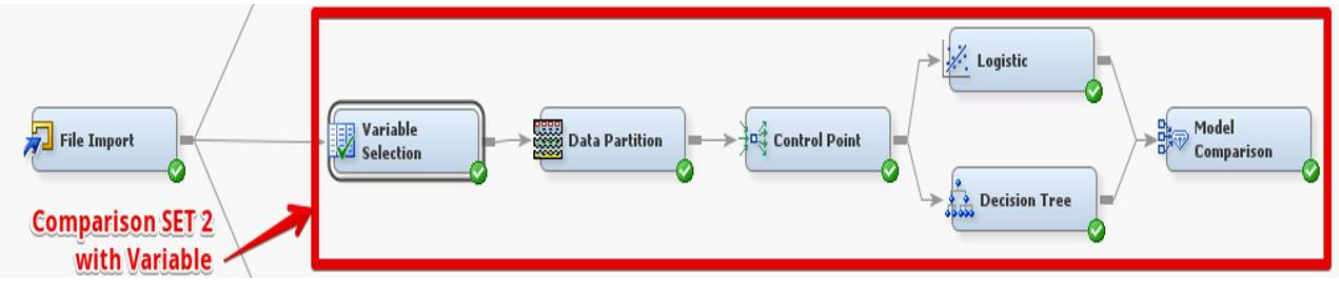
FOR SET 1:



Conclusion: Without implementing any dimensionality reduction technique, we have found better prediction results for ordinal categorical target variable 'popularity_level' from Logistic Regression model. This is because the misclassification rate from logistic regression =0.521383 while that from Decision Tree =0.530213. This can alternatively be stated as **the prediction accuracy of logistic regression model is ~48%** which is considerably good for forecasting.

Approach 2: Use Variable Selection on Logistic Regression Algorithm and Decision Tree Algorithm

Screenshot of the SAS Dashboard:



Result:

Buttons Used	Target Variable	Model	Result
File Import, **Variable Selection**, Data Partition, Control Point, Logistic Regression **(SET 2)**	Popularity_level (ordinal = '1','2','3')	Logistic	
File Import, **Variable Selection**, Data Partition, Control Point, Decision Tree **(SET 2)**	Popularity_level (ordinal = '1','2','3')	Decision Tree	

Fit Statistics	Statistics Label	Train	Validation	Test
AIC	Akaike's Information Criterion	57368.23	.	.
ASE	Average Squared Error	0.21	0.21	0.21
AVERR	Average Error Function	0.69	0.69	0.69
DFE	Degrees of Freedom for Error	55486.00	.	.
DFM	Model Degrees of Freedom	12.00	.	.
DFT	Total Degrees of Freedom	55498.00	.	.
DIV	Divisor for ASE	83247.00	23781.00	11904.00
ERR	Error Function	57344.23	16374.19	8201.50
FPE	Final Prediction Error	0.21	.	.
MAX	Maximum Absolute Error	1.00	1.00	0.99
MSE	Mean Square Error	0.21	0.21	0.21
NOBS	Sum of Frequencies	27749.00	7927.00	3968.00
_NW_	Number of Estimate Weights	12.00	.	.
RASE	Root Average Sum of Squares	0.45	0.45	0.46
RFPE	Root Final Prediction Error	0.46	.	.
RMSE	Root Mean Squared Error	0.46	0.45	0.46
SBC	Schwarz's Bayesian Criterion	57475.32	.	.
SSE	Sum of Squared Errors	17232.82	4921.14	2467.49
_SUMW	Sum of Case Weights Times Freq	83247.00	23781.00	11904.00
MISC	Misclassification Rate	0.52	0.53	0.53

Fit Statistics	Statistics Label	Train	Validation	Test
NOBS	Sum of Frequencies	27749.00	7927.00	3968.00
MISC	Misclassification Rate	0.53	0.53	0.54
MAX	Maximum Absolute Error	0.93	1.00	1.00
SSE	Sum of Squared Errors	17562.02	5039.45	2529.25
ASE	Average Squared Error	0.21	0.21	0.21
RASE	Root Average Squared Error	0.46	0.46	0.46
DIV	Divisor for ASE	83247.00	23781.00	11904.00
DFT	Total Degrees of Freedom	55498.00	.	.

Confusion Matrices:

Logistic:

Event Classification Table

Data Role=TRAIN Target=popularity_level Target Label=' '

False Negative	True Negative	False Positive	True Positive
2562	9363	8074	7750

Data Role=VALIDATE Target=popularity_level Target Label=' '

False Negative	True Negative	False Positive	True Positive
748	2678	2304	2197

Decision Tree:

Event Classification Table

Data Role=TRAIN Target=popularity_level Target Label=' '

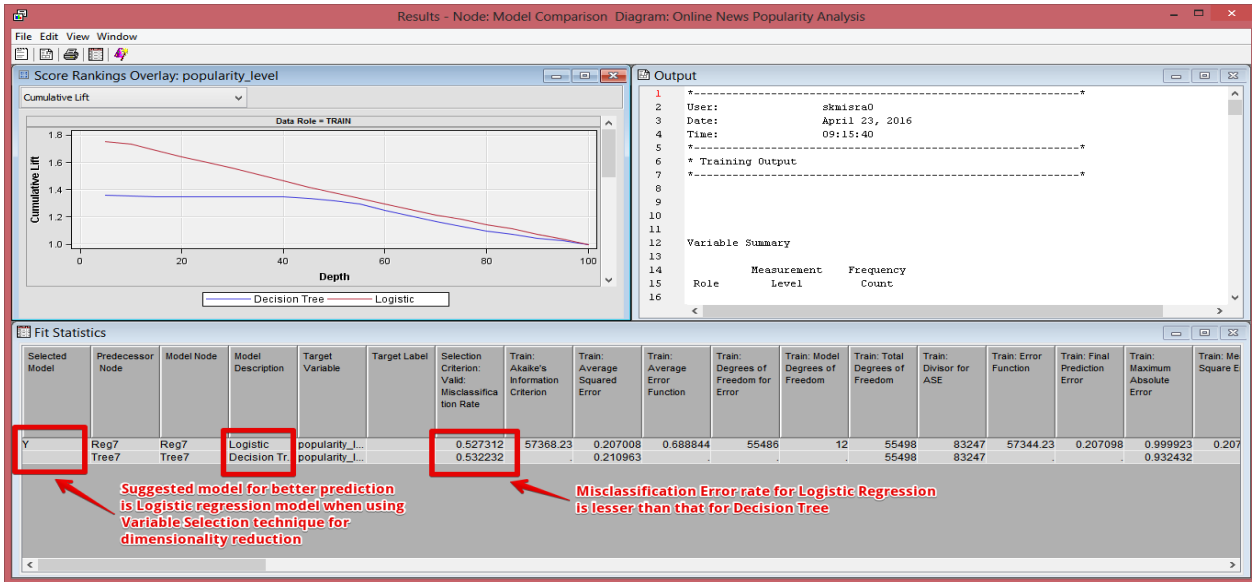
False Negative	True Negative	False Positive	True Positive
3034	9629	7808	7278

Data Role=VALIDATE Target=popularity_level Target Label=' '

False Negative	True Negative	False Positive	True Positive
850	2748	2234	2095

Model Comparison Results:

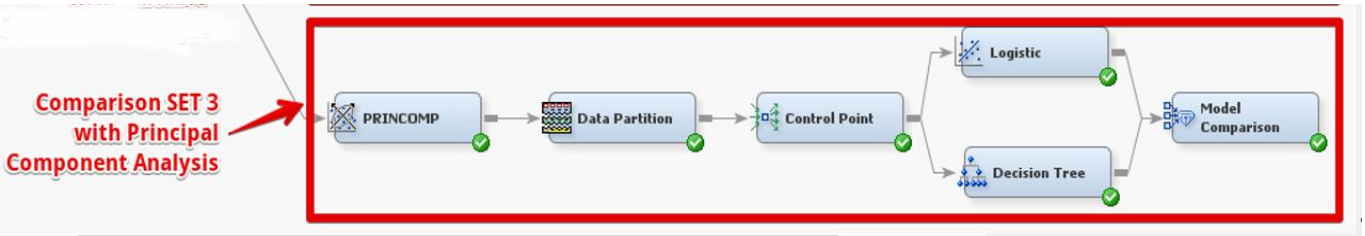
FOR SET 2:



Conclusion: When implementing dimensionality reduction technique "Variable Selection" using R-Square and Chi-Square values both, we have found better prediction results for ordinal categorical target variable 'popularity_level' from Logistic Regression model. This is because the misclassification rate from Logistic regression = 0.527312 while that from Decision Tree = 0.532232. This can alternatively be stated as **the prediction accuracy of logistic regression model is ~47%** which is considerably good for forecasting.

Approach 3: Use Principal Component Analysis on Logistic Regression and Decision Tree Algorithm

Screenshot of the SAS Dashboard:



Result:

Buttons Used	Target Variable	Model	Result
File Import, **Principal Component**, Data Partition, Control Point, Logistic Regression **(SET 3)**	Popularity_level (ordinal = '1','2','3')	Logistic	
File Import, **Principal Component**, Data Partition, Control Point, Decision Tree **(SET 3)**	Popularity_level (ordinal = '1','2','3')	Decision Tree	

Fit Statistics	Statistics Label	Train	Validation	Test
AIC	Akaike's Information Criterion	57914.16	.	.
ASE	Average Squared Error	0.21	0.21	0.21
AVERR	Average Error Function	0.70	0.70	0.69
DFE	Degrees of Freedom for Error	55476.00	.	.
DFM	Model Degrees of Freedom	22.00	.	.
DFT	Total Degrees of Freedom	55498.00	.	.
DIV	Divisor for ASE	83247.00	23781.00	11904.00
ERR	Error Function	57870.16	16598.88	8245.43
FPE	Final Prediction Error	0.21	.	.
MAX	Maximum Absolute Error	1.00	1.00	0.97
MSE	Mean Square Error	0.21	0.21	0.21
NOBS	Sum of Frequencies	27749.00	7927.00	3968.00
_NW_	Number of Estimate Weights	22.00	.	.
RASE	Root Average Sum of Squares	0.46	0.46	0.46
RFPE	Root Final Prediction Error	0.46	.	.
RMSE	Root Mean Squared Error	0.46	0.46	0.46
SBC	Schwarz's Bayesian Criterion	58110.49	.	.
SSE	Sum of Squared Errors	17413.78	4994.03	2481.83
SUMW	Sum of Case Weights Times Freq	83247.00	23781.00	11904.00
MISC	Misclassification Rate	0.53	0.54	0.54

Fit Statistics	Statistics Label	Train	Validation	Test
NOBS	Sum of Frequencies	27749.00	7927.00	3968.00
MISC	Misclassification Rate	0.54	0.54	0.56
MAX	Maximum Absolute Error	0.87	0.87	0.87
SSE	Sum of Squared Errors	17621.73	5061.77	2546.43
ASE	Average Squared Error	0.21	0.21	0.21
RASE	Root Average Squared Error	0.46	0.46	0.46
DIV	Divisor for ASE	83247.00	23781.00	11904.00
DFT	Total Degrees of Freedom	55498.00	.	.

Confusion Matrices:

Logistic Regression:

Event Classification Table

Data Role=TRAIN Target=popularity_level Target Label=' '

False Negative	True Negative	False Positive	True Positive
2607	9016	8421	7705

Data Role=VALIDATE Target=popularity_level Target Label=' '

False Negative	True Negative	False Positive	True Positive
748	2549	2433	2197

Decision Tree:

Event Classification Table

Data Role=TRAIN Target=popularity_level Target Label=' '

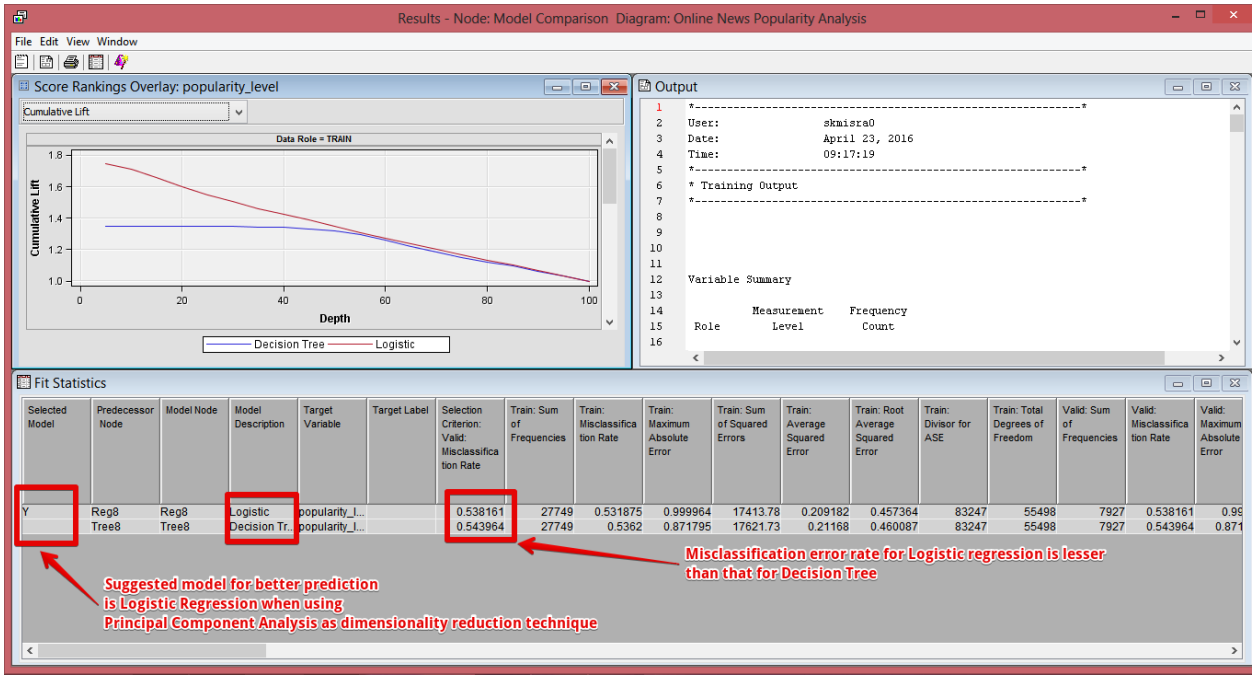
False Negative	True Negative	False Positive	True Positive
2769	9210	8227	7543

Data Role=VALIDATE Target=popularity_level Target Label=' '

False Negative	True Negative	False Positive	True Positive
811	2595	2387	2134

Model Comparison Results:

FOR SET 3:



Conclusion: When implementing dimensionality reduction technique "Principal Component Analysis", we have found better prediction results for ordinal categorical target variable 'popularity_level' from Logistic Regression model. This is because the misclassification rate from Logistic Regression = 0.538161 while that from Decision Tree = 0.543964. This can alternatively be stated as **the prediction accuracy of Logistic Regression model is ~46%** which is considerably good for forecasting.

FINAL RESULT:*Based on these statistics, we can finally conclude that predicting an ordinal categorical target variable to check whether the article published is of 'high popularity' or 'medium popularity' or 'low popularity' we must go for Logistic regression modeling withVariable Selection dimensionality reduction technique with an accuracy of ~47%.*

REASONING: ***Although with a kitchen sink model, we achieved ~48% accuracy using logistic regression, the model seems too complex. However, on compromising only ~1% accuracy, we built a model with less input variables keeping our model simple. Thus, we prefer Logistic Regression with Variable Selection dimensionality reduction technique over any other model.***

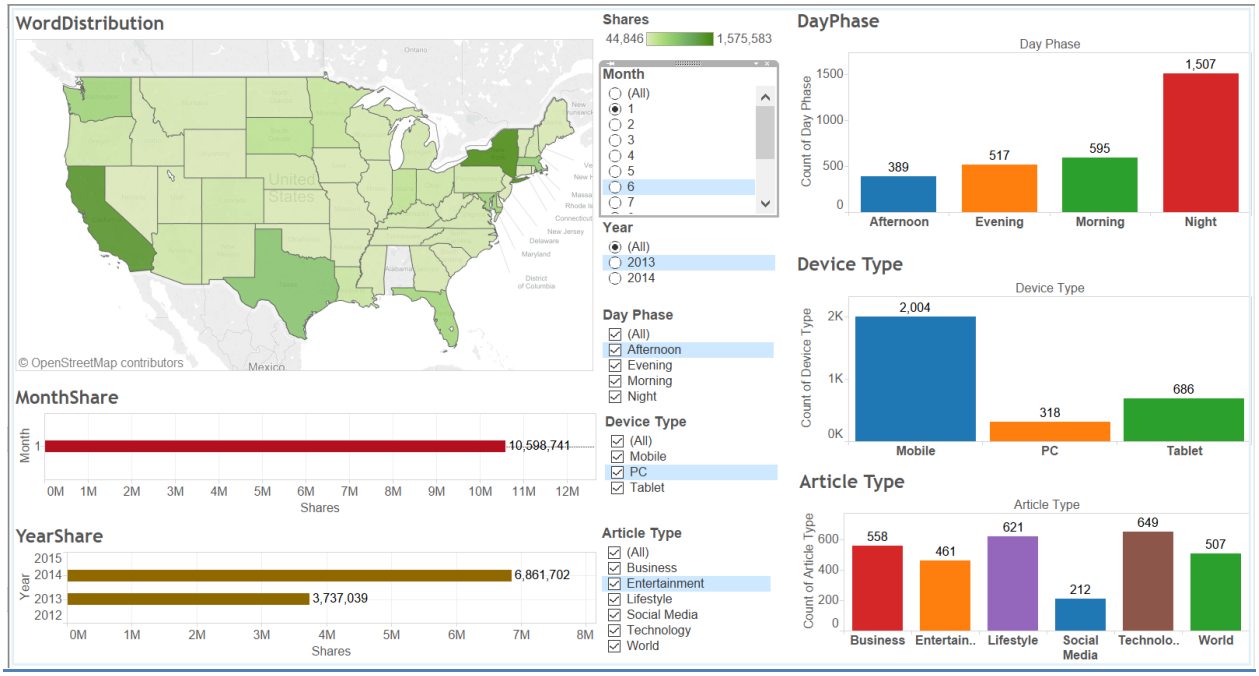
VISUALIZATION USING TABLEAU

Our Online News Popularity – UCI dataset has couple of columns over which insightful information could be extracted. To do so we created a Tableau Dashboard.

Link to the Public Tableau Dashboard:

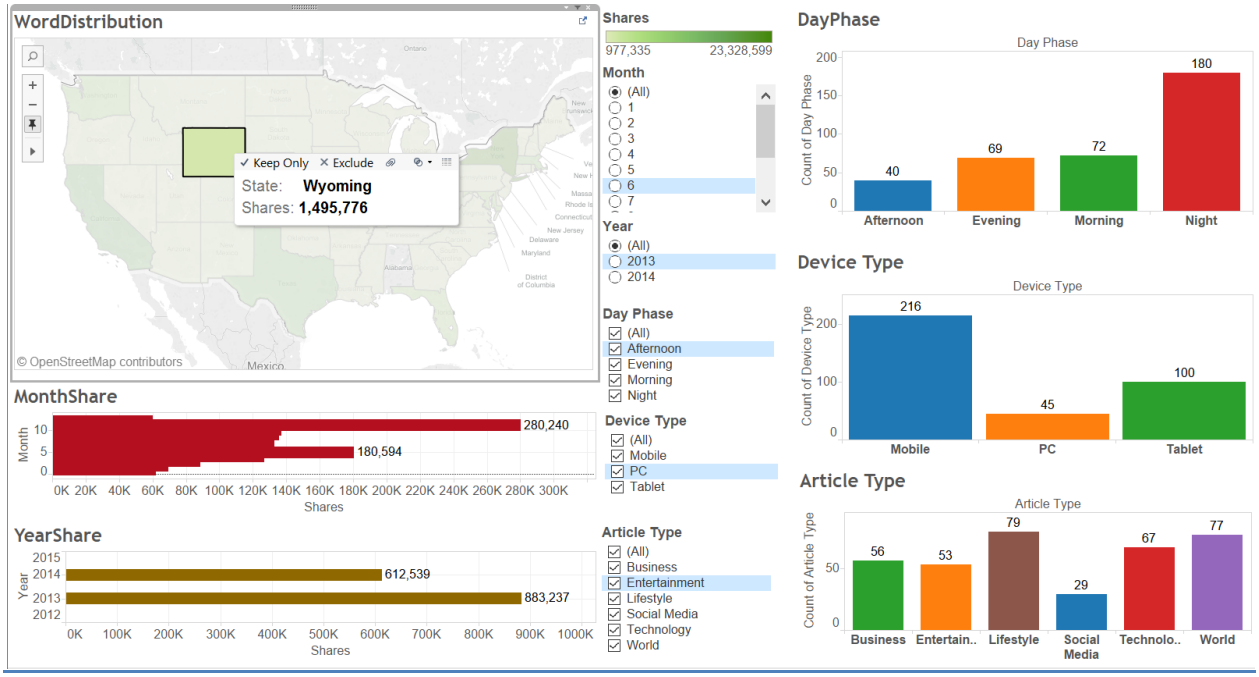
https://public.tableau.com/profile/jagpreet#!/vizhome/Book1_10486/Dashboard1

Insight1: CES Conference by CNET in Jan makes people share more tech articles.



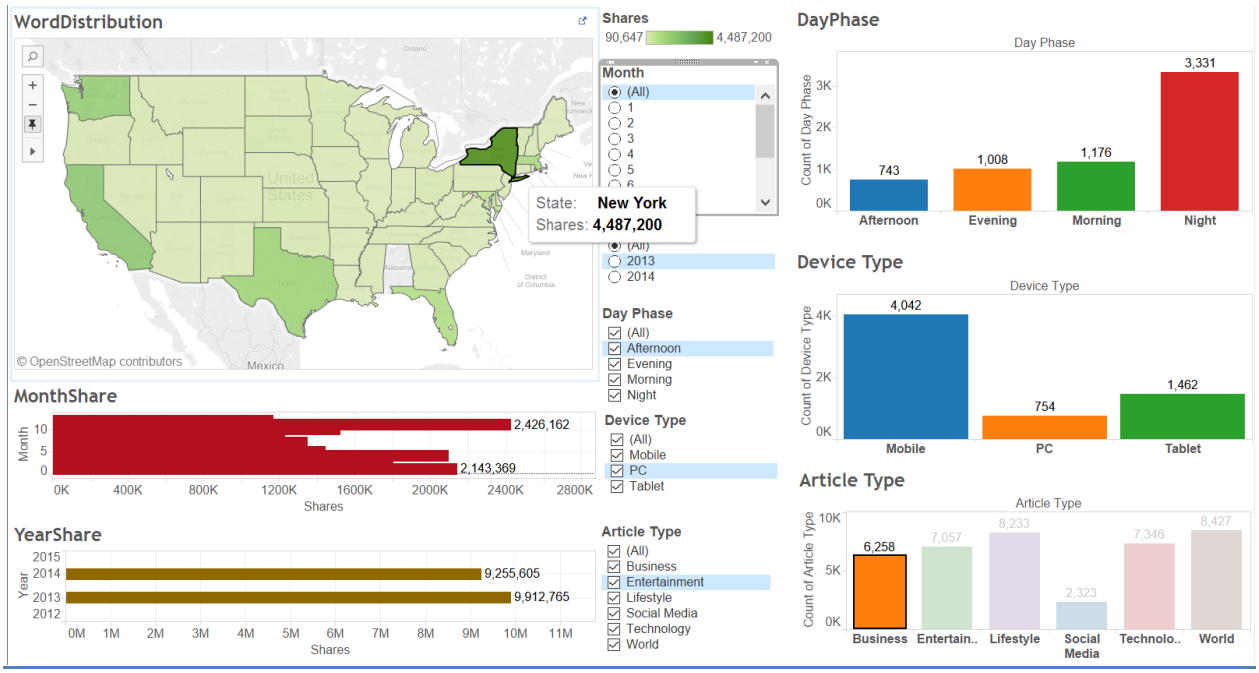
- January 2014 has almost twice the number of article shares as it was in January 2013.
- People prefer to read Mashable articles on Mobile device as compared to Tablet or PC.
- Most of the authors of an article are based in cities like California, Texas, New York and Massachusetts. It is also observed that they prefer to post their articles during Night hours.
- In January, people share maximum Technology related articles as compared to any other category. Probably, this can be justified with the fact that CES technology product launch conference by CNET happens annually in the month of January, so people like to stay active and share more articles on technology.

Insight2: Christmas holiday and Black Friday week, make people visit Lifestyle related Mashable article even more



- Authors in Wyoming publish more of Lifestyle related articles than any other category. This clearly signifies they have a keen interest in this domain.
- Also, it is observed that such articles are shared more during last few months of a year and this is exactly when heavy discount week like Black Friday and Labor Day falls.
- People prefer to read and share more lifestyle related articles before making an actual purchase of latest trendy clothes by Levis, Louis Vuitton etc brands.

Insight 3: New York is the city of Business-minded crowd



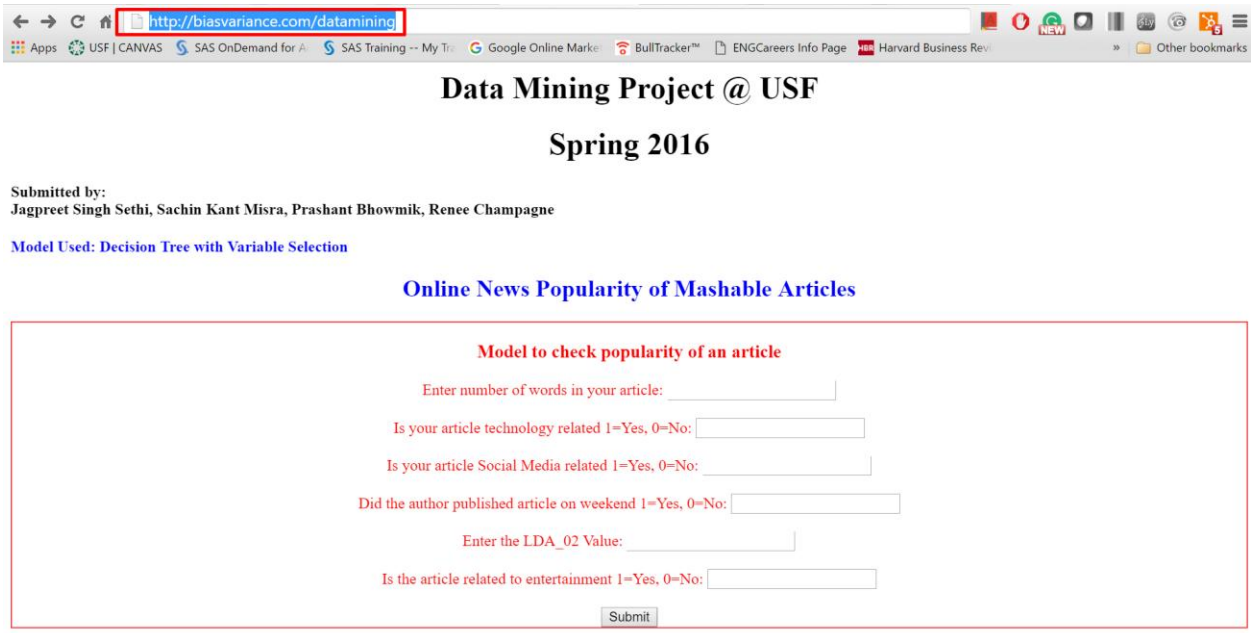
- Most of the authors in New York work and publish articles on Mashable.com late night.
- Since, New York is the hub of financial companies with largest NYSE stock market, people working in top MNCs prefer to read and share “Business” centric Mashable articles with friends.

Thus, we can conclude that “New York is the City of Business-minded crowd”.

MODEL IMPLEMENTATION

- **Model Implemented:** Decision with Variable Selection for predicting popularity of the Mashable articles using IF ELSE logic in PHP.
- **Technology used:** Amazon Web Server
- **Website Link:**http://biasvariance.com/datamining

Following is the relevant Input Screen:



Following is the relevant Output Screen: