ONLINE ARTICLES POPULARITY

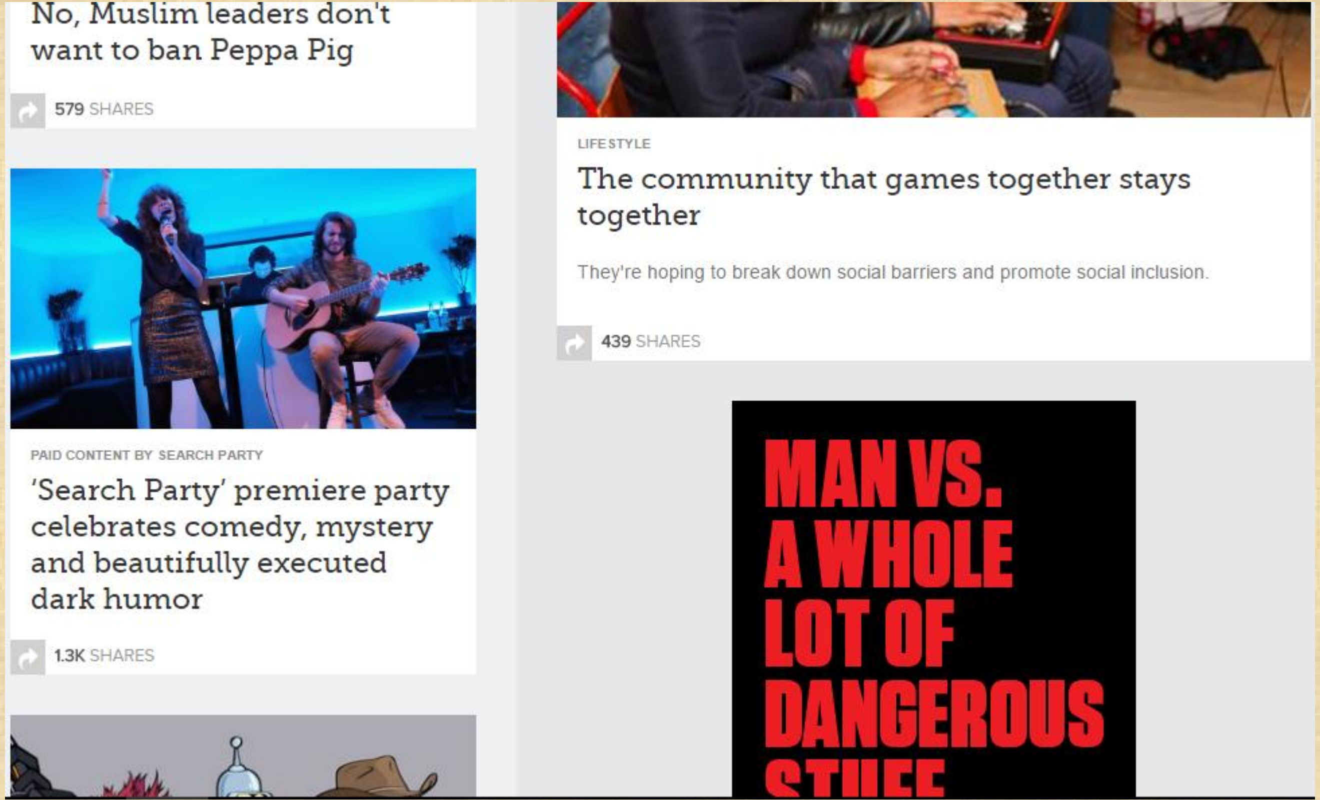
BADAR ALMARI

MAYANK MAKAN

MEGHANA KASULA

SHAKHLO KHODJAEVA

Why this Dataset?



No, Muslim leaders don't want to ban Peppa Pig

579 SHARES

PAID CONTENT BY SEARCH PARTY

'Search Party' premiere party celebrates comedy, mystery and beautifully executed dark humor

1.3K SHARES

LIFESTYLE

The community that games together stays together

They're hoping to break down social barriers and promote social inclusion.

439 SHARES

More on the Data set

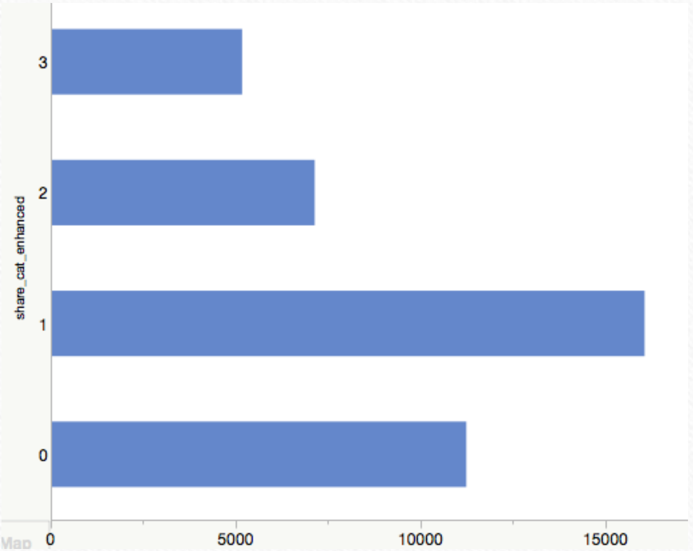
- Created to analyze the number of shares depending on the attributes and predict if an article will be popular on the internet or not.
- 39,644 observations
- 61 attributes
- Mashable website: collected over a 2 year period from Jan 2013 - Jan 2015
- No missing values, but some topics were unclassified
- Target: number of shares

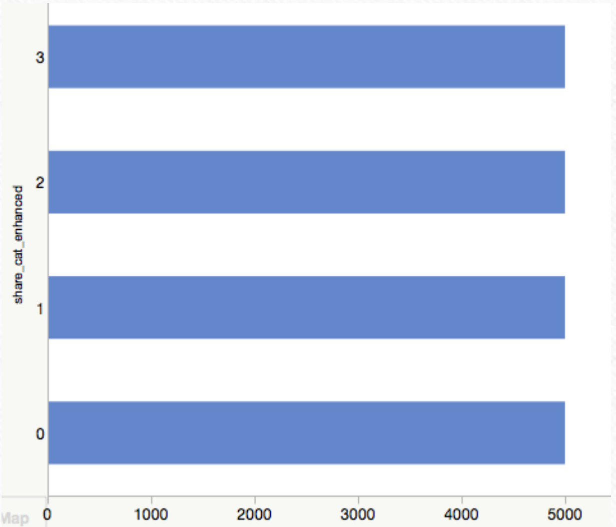
Approach we would follow

- Sample
- Explore
- Modify
- Modelling
- Assess

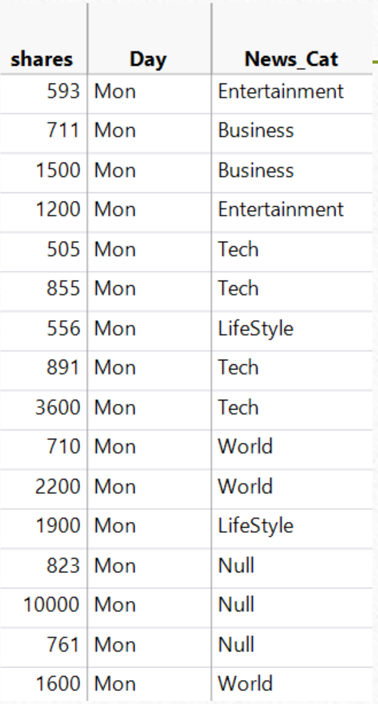
Sample

Sampling is done to make balance in the labeling as some models ignore

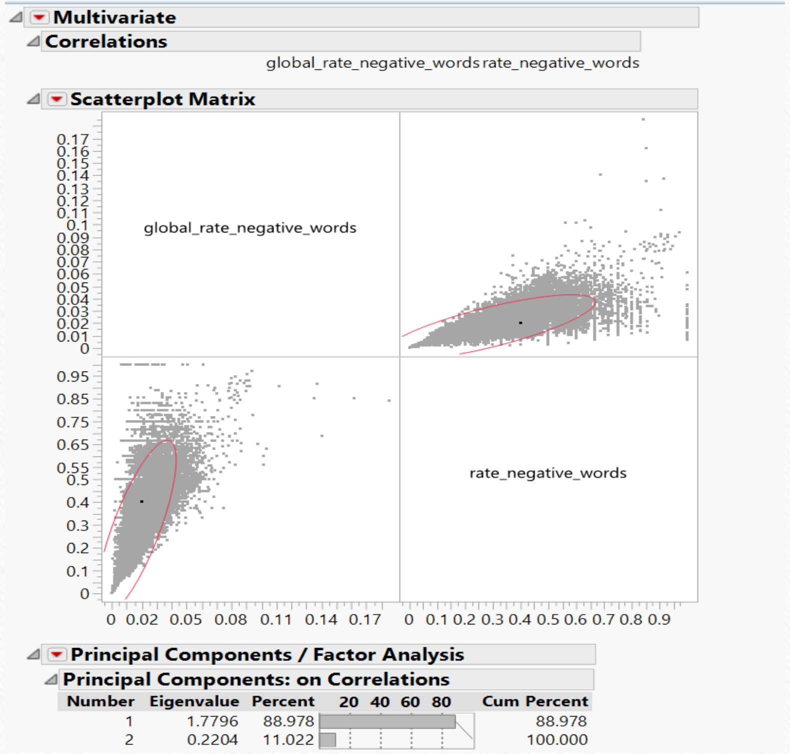




Modifications



shares	Day	News_Cat
593	Mon	Entertainment
711	Mon	Business
1500	Mon	Business
1200	Mon	Entertainment
505	Mon	Tech
855	Mon	Tech
556	Mon	LifeStyle
891	Mon	Tech
3600	Mon	Tech
710	Mon	World
2200	Mon	World
1900	Mon	LifeStyle
823	Mon	Null
10000	Mon	Null
761	Mon	Null
1600	Mon	World



Multivariate

Correlations

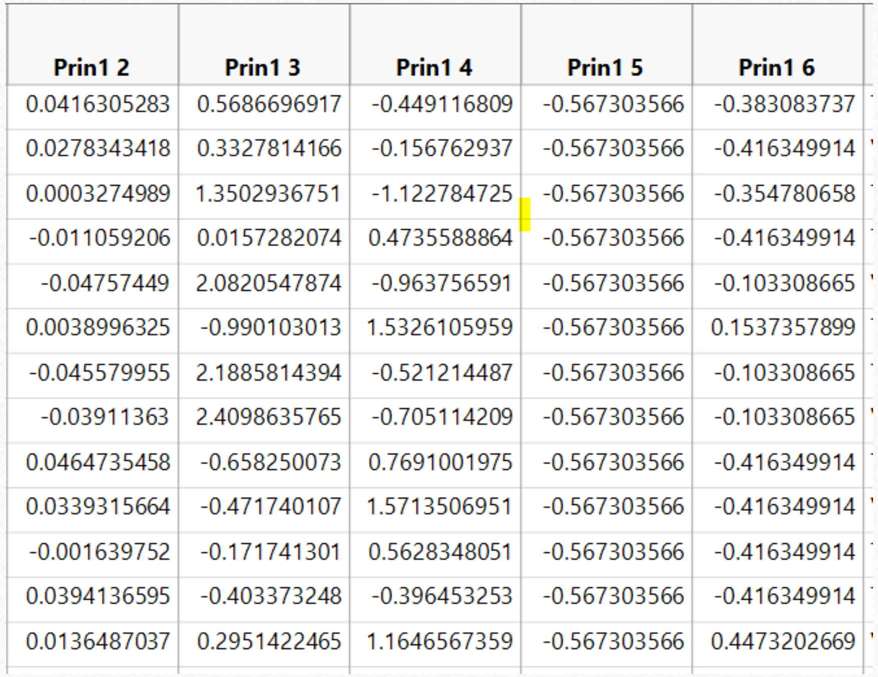
global_rate_negative_words rate_negative_words

Scatterplot Matrix

Principal Components / Factor Analysis

Principal Components: on Correlations

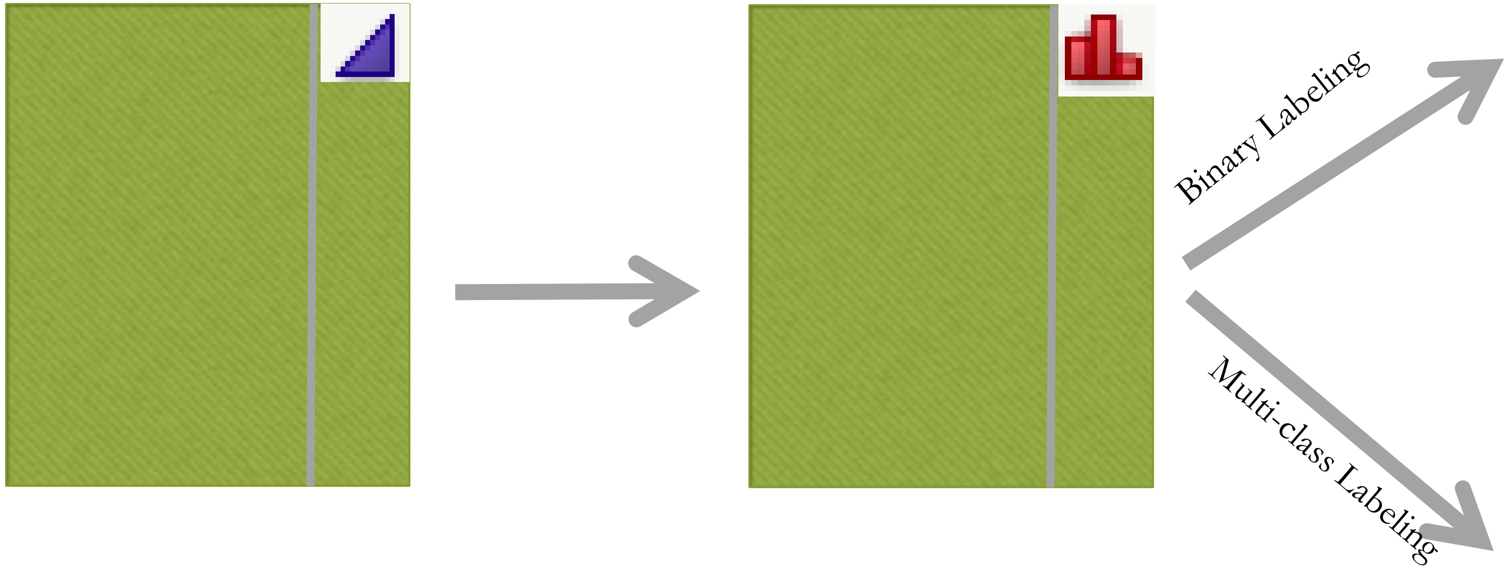
Number	Eigenvalue	Percent	20 40 60 80	Cum Percent
1	1.7796	88.978		88.978
2	0.2204	11.022		100.000



Prin1 2	Prin1 3	Prin1 4	Prin1 5	Prin1 6
0.0416305283	0.5686696917	-0.449116809	-0.567303566	-0.383083737
0.0278343418	0.3327814166	-0.156762937	-0.567303566	-0.416349914
0.0003274989	1.3502936751	-1.122784725	-0.567303566	-0.354780658
-0.011059206	0.0157282074	0.4735588864	-0.567303566	-0.416349914
-0.04757449	2.0820547874	-0.963756591	-0.567303566	-0.103308665
0.0038996325	-0.990103013	1.5326105959	-0.567303566	0.1537357899
-0.045579955	2.1885814394	-0.521214487	-0.567303566	-0.103308665
-0.03911363	2.4098635765	-0.705114209	-0.567303566	-0.103308665
0.0464735458	-0.658250073	0.7691001975	-0.567303566	-0.416349914
0.0339315664	-0.471740107	1.5713506951	-0.567303566	-0.416349914
-0.001639752	-0.171741301	0.5628348051	-0.567303566	-0.416349914
0.0394136595	-0.403373248	-0.396453253	-0.567303566	-0.416349914
0.0136487037	0.2951422465	1.1646567359	-0.567303566	0.4473202669

Target Manipulation

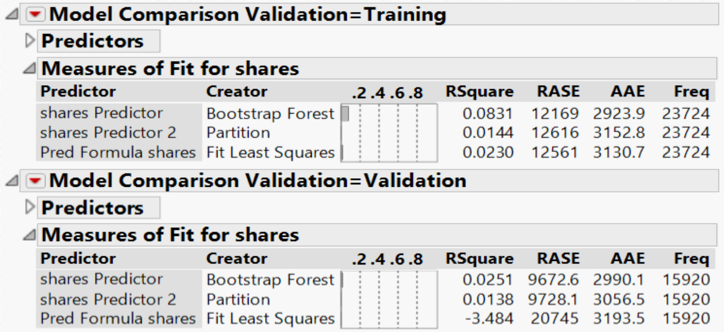
T = Threshold/Cutoff
L = Label 📊
n = # of categories



```
If ◢ < T
    L = 0
Else
    L = 1
```

```
If ◢ < T1
    L = 0
Else if ◢ < T2
    L = 1
…
Else
    L = n-1
```

Models (Continuous Target)



Model Comparison Validation=Training

Predictors

Measures of Fit for shares

Predictor	Creator	.2 .4 .6 .8	RSquare	RASE	AAE	Freq
shares Predictor	Bootstrap Forest		0.0831	12169	2923.9	23724
shares Predictor 2	Partition		0.0144	12616	3152.8	23724
Pred Formula shares	Fit Least Squares		0.0230	12561	3130.7	23724

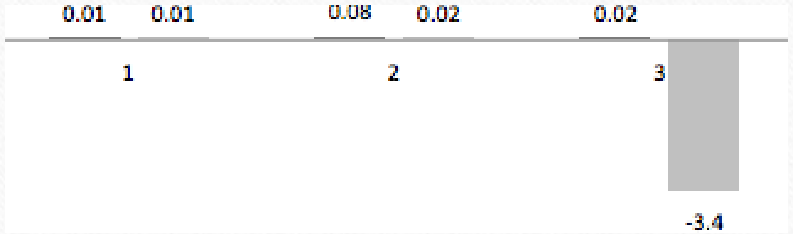
Model Comparison Validation=Validation

Predictors

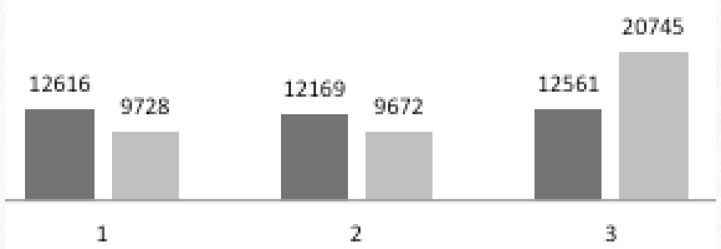
Measures of Fit for shares

Predictor	Creator	.2 .4 .6 .8	RSquare	RASE	AAE	Freq
shares Predictor	Bootstrap Forest		0.0251	9672.6	2990.1	15920
shares Predictor 2	Partition		0.0138	9728.1	3056.5	15920
Pred Formula shares	Fit Least Squares		-3.484	20745	3193.5	15920

Models: Decision Tree, Bootstrap Forest and Fit least Square

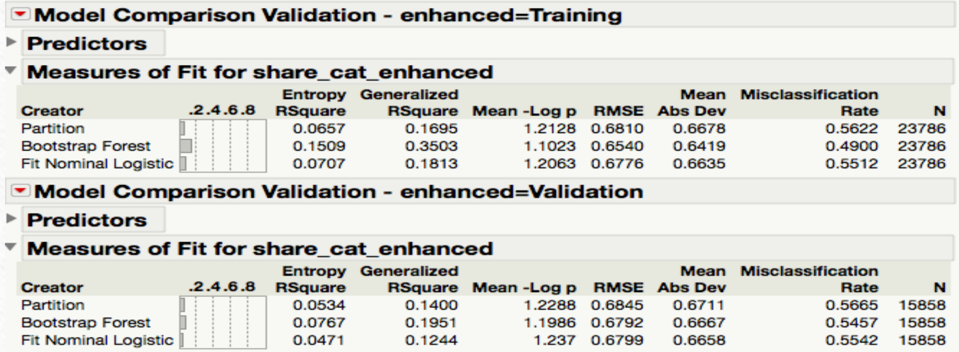


RSquare



RASE

Models (Categorical Target - Multi-class)



Model Comparison Validation - enhanced=Training

Predictors

Measures of Fit for share_cat_enhanced

Creator	.2.4.6.8	Entropy RSquare	Generalized RSquare	Mean -Log p	RMSE	Mean Abs Dev	Misclassification Rate	N
Partition		0.0657	0.1695	1.2128	0.6810	0.6678	0.5622	23786
Bootstrap Forest		0.1509	0.3503	1.1023	0.6540	0.6419	0.4900	23786
Fit Nominal Logistic		0.0707	0.1813	1.2063	0.6776	0.6635	0.5512	23786

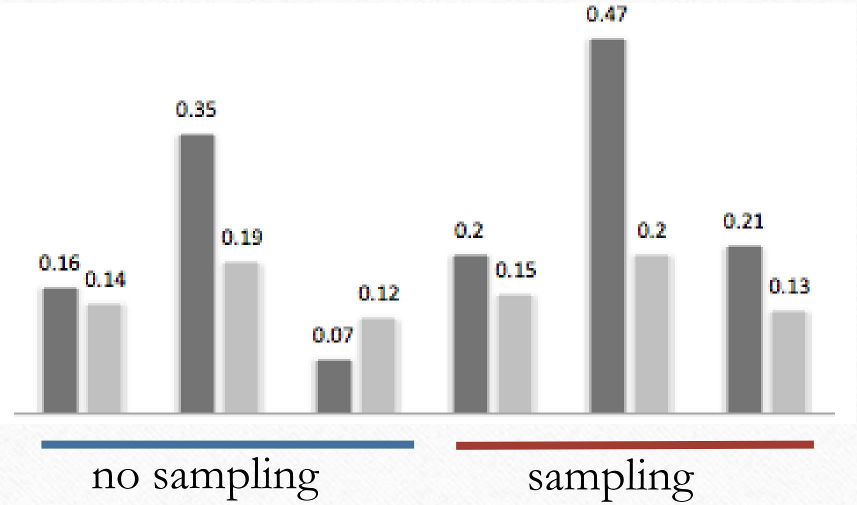
Model Comparison Validation - enhanced=Validation

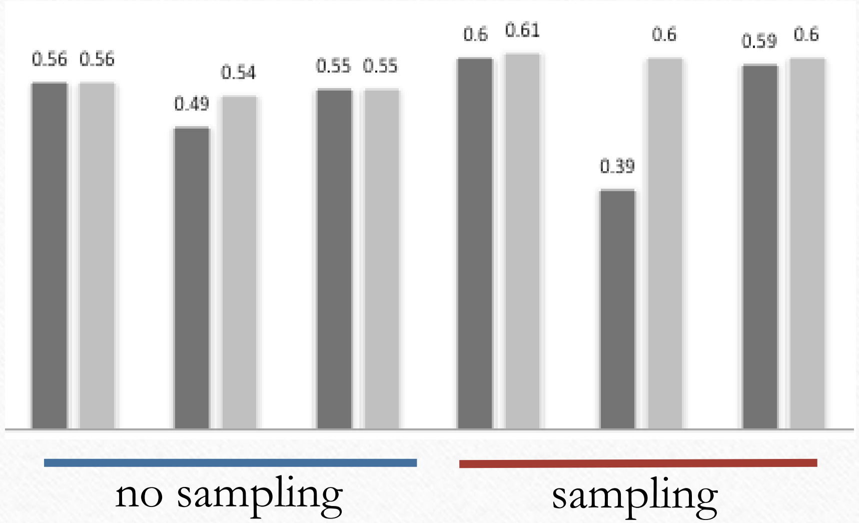
Predictors

Measures of Fit for share_cat_enhanced

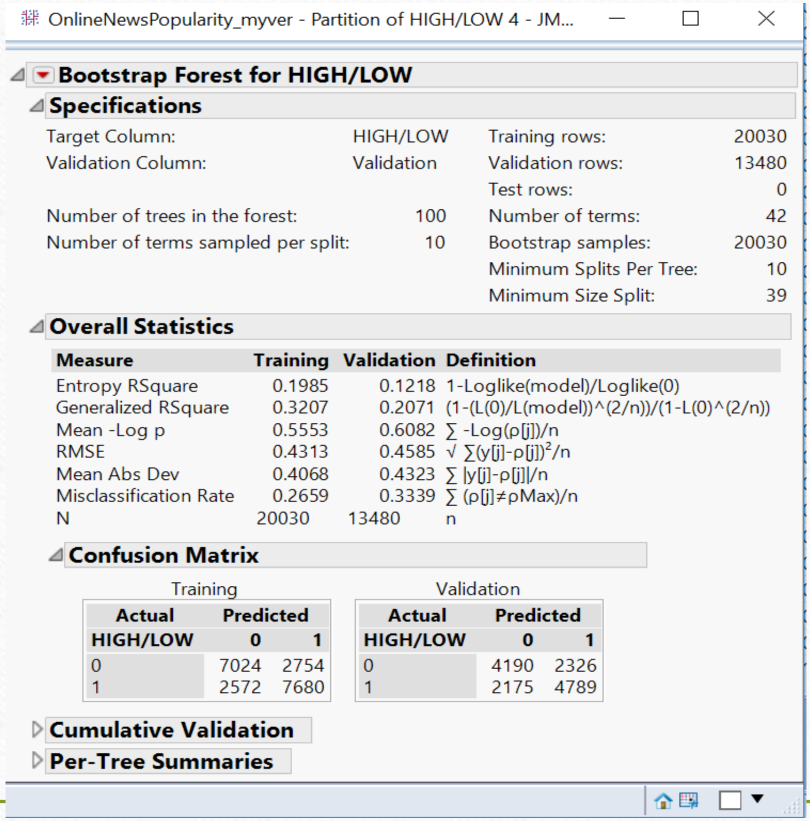
Creator	.2.4.6.8	Entropy RSquare	Generalized RSquare	Mean -Log p	RMSE	Mean Abs Dev	Misclassification Rate	N
Partition		0.0534	0.1400	1.2288	0.6845	0.6711	0.5665	15858
Bootstrap Forest		0.0767	0.1951	1.1986	0.6792	0.6667	0.5457	15858
Fit Nominal Logistic		0.0471	0.1244	1.237	0.6799	0.6658	0.5542	15858

Models: Decision Tree, Bootstrap Forest and Logistic Regression.





Models (Categorical Target – Binary Class)



OnlineNewsPopularity_myver - Partition of HIGH/LOW 4 - JM...

Bootstrap Forest for HIGH/LOW

Specifications

Target Column:	HIGH/LOW	Training rows:	20030
Validation Column:	Validation	Validation rows:	13480
		Test rows:	0
Number of trees in the forest:	100	Number of terms:	42
Number of terms sampled per split:	10	Bootstrap samples:	20030
		Minimum Splits Per Tree:	10
		Minimum Size Split:	39

Overall Statistics

Measure	Training	Validation	Definition
Entropy RSquare	0.1985	0.1218	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.3207	0.2071	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.5553	0.6082	∑ -Log(ρ[j])/n
RMSE	0.4313	0.4585	$\sqrt{\sum(y[j]-\rho[j])^2/n}$
Mean Abs Dev	0.4068	0.4323	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.2659	0.3339	∑ (ρ[j]≠ρMax)/n
N	20030	13480	n

Confusion Matrix

Training

Actual	Predicted	
HIGH/LOW	0	1
0	7024	2754
1	2572	7680

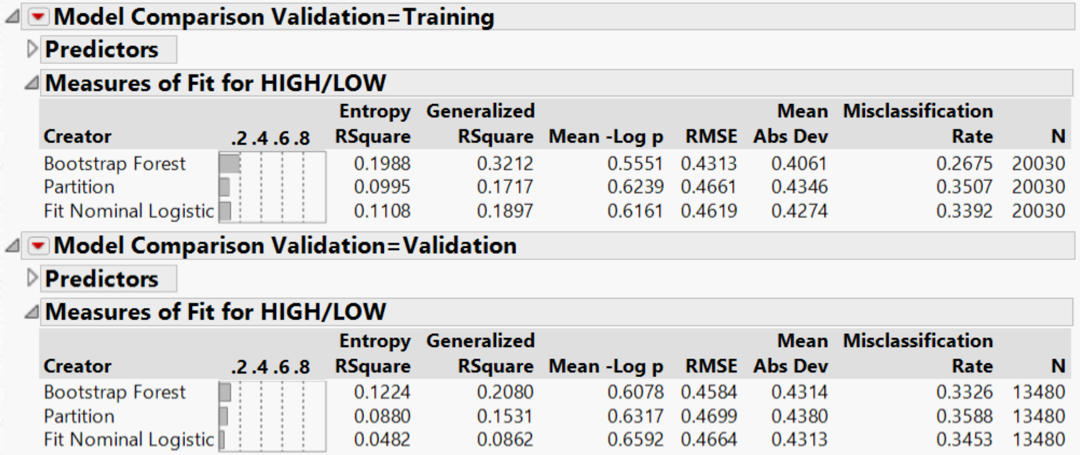
Validation

Actual	Predicted	
HIGH/LOW	0	1
0	4190	2326
1	2175	4789

Cumulative Validation

Per-Tree Summaries

Models: Decision Tree, Bootstrap Forest and Logistic Regression



Model Comparison Validation=Training

Predictors

Measures of Fit for HIGH/LOW

Creator	.2 .4 .6 .8	Entropy RSquare	Generalized RSquare	Mean -Log p	RMSE	Mean Abs Dev	Misclassification Rate	N
Bootstrap Forest		0.1988	0.3212	0.5551	0.4313	0.4061	0.2675	20030
Partition		0.0995	0.1717	0.6239	0.4661	0.4346	0.3507	20030
Fit Nominal Logistic		0.1108	0.1897	0.6161	0.4619	0.4274	0.3392	20030

Model Comparison Validation=Validation

Predictors

Measures of Fit for HIGH/LOW

Creator	.2 .4 .6 .8	Entropy RSquare	Generalized RSquare	Mean -Log p	RMSE	Mean Abs Dev	Misclassification Rate	N
Bootstrap Forest		0.1224	0.2080	0.6078	0.4584	0.4314	0.3326	13480
Partition		0.0880	0.1531	0.6317	0.4699	0.4380	0.3588	13480
Fit Nominal Logistic		0.0482	0.0862	0.6592	0.4664	0.4313	0.3453	13480

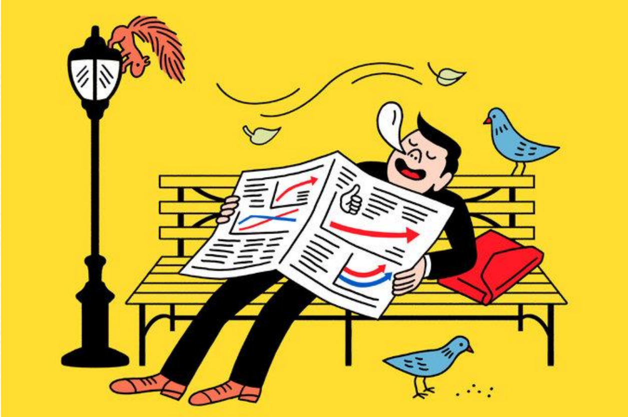
The Assessment

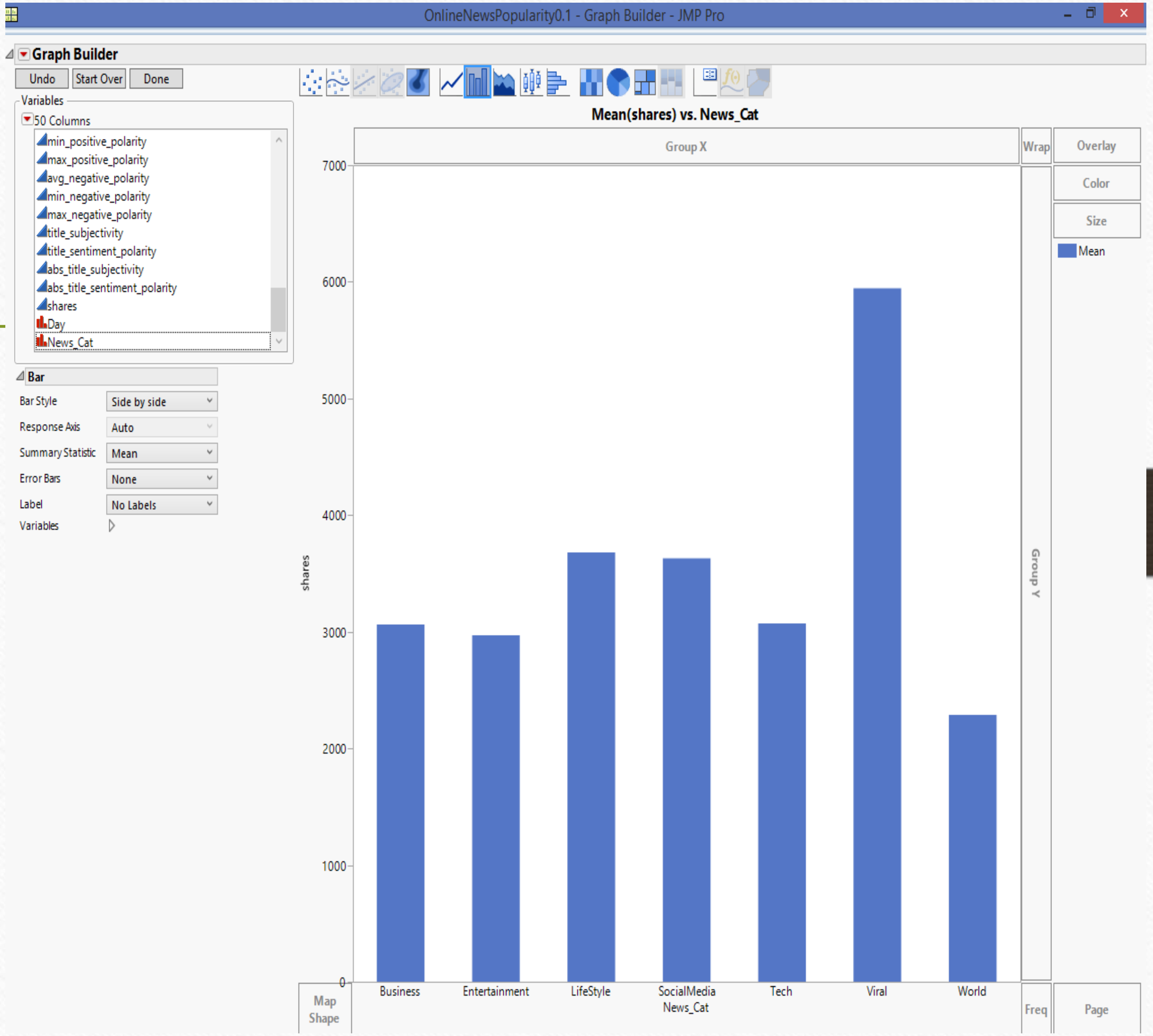
- Our best model is Bootstrap forest with misclassification 26.75.
- Based of given variables Mashable can predict whether article will get high share and low share.
- Mashable can do some changes in its article accordingly to get high shares and making website more profitable

Data Insights

Channel:

- Most popular topic is Viral,
- followed by lifestyle and social media
- Least popular topic is World News





OnlineNewsPopularity0.1 - Graph Builder - JMP Pro

Graph Builder

Bar Style: Side by side

Response Axis: Auto

Summary Statistic: Mean

Error Bars: None

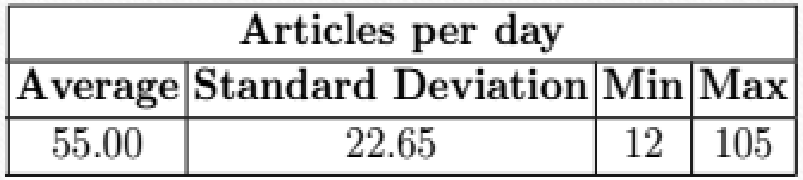
Label: No Labels

Variables

Data Insights

No. of keywords:

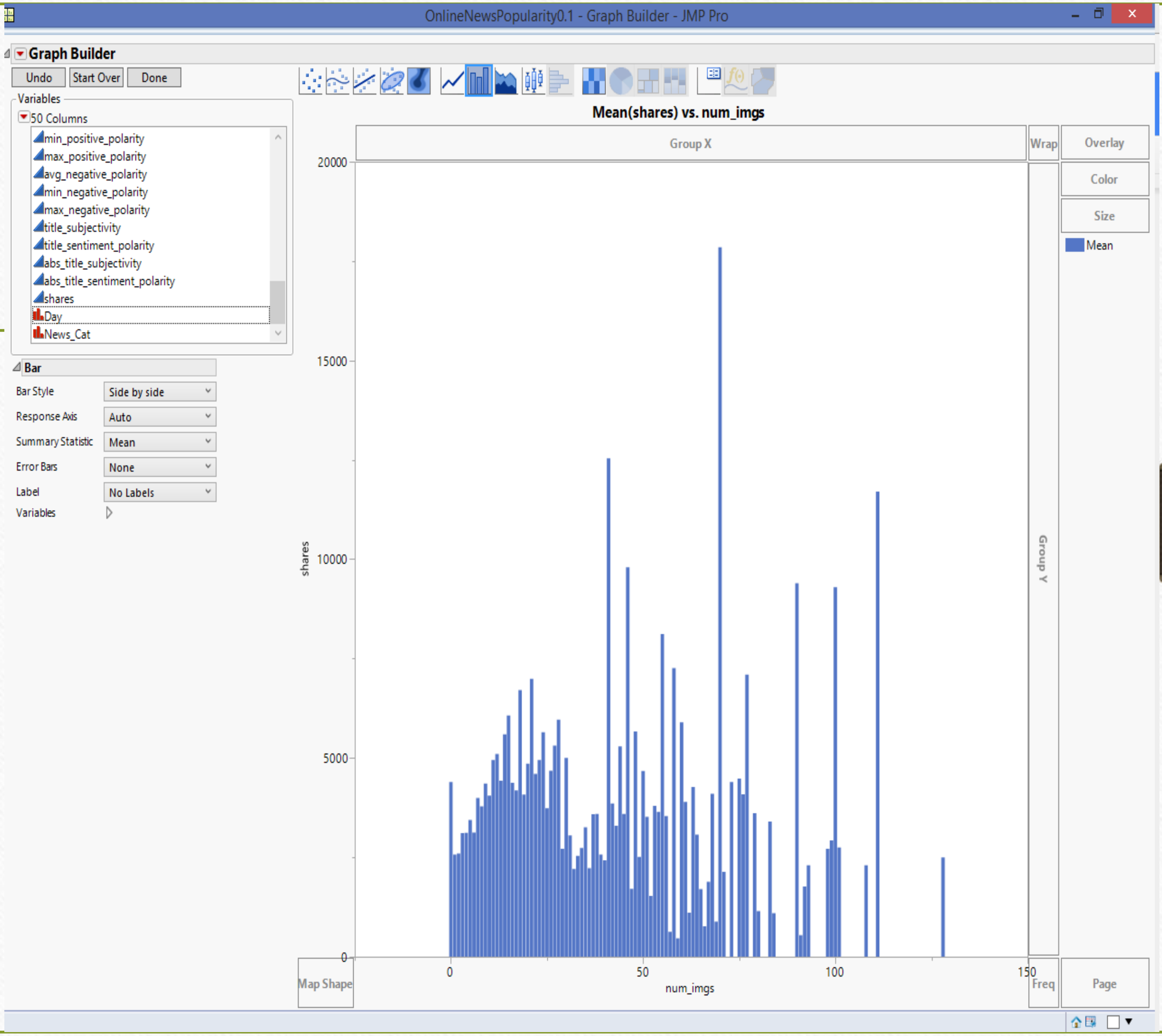
- Generally between 5 to 10.



Articles per day			
Average	Standard Deviation	Min	Max
55.00	22.65	12	105

No. of images

- The number of shares are dense when number of Images are between 0-50.



OnlineNewsPopularity0.1 - Graph Builder - JMP Pro

Graph Builder

Undo | Start Over | Done

Variables

50 Columns

- min_positive_polarity
- max_positive_polarity
- avg_negative_polarity
- min_negative_polarity
- max_negative_polarity
- title_subjectivity
- title_sentiment_polarity
- abs_title_subjectivity
- abs_title_sentiment_polarity
- shares
- Day
- News_Cat

Bar

Bar Style: Side by side

Response Axis: Auto

Summary Statistic: Mean

Error Bars: None

Label: No Labels

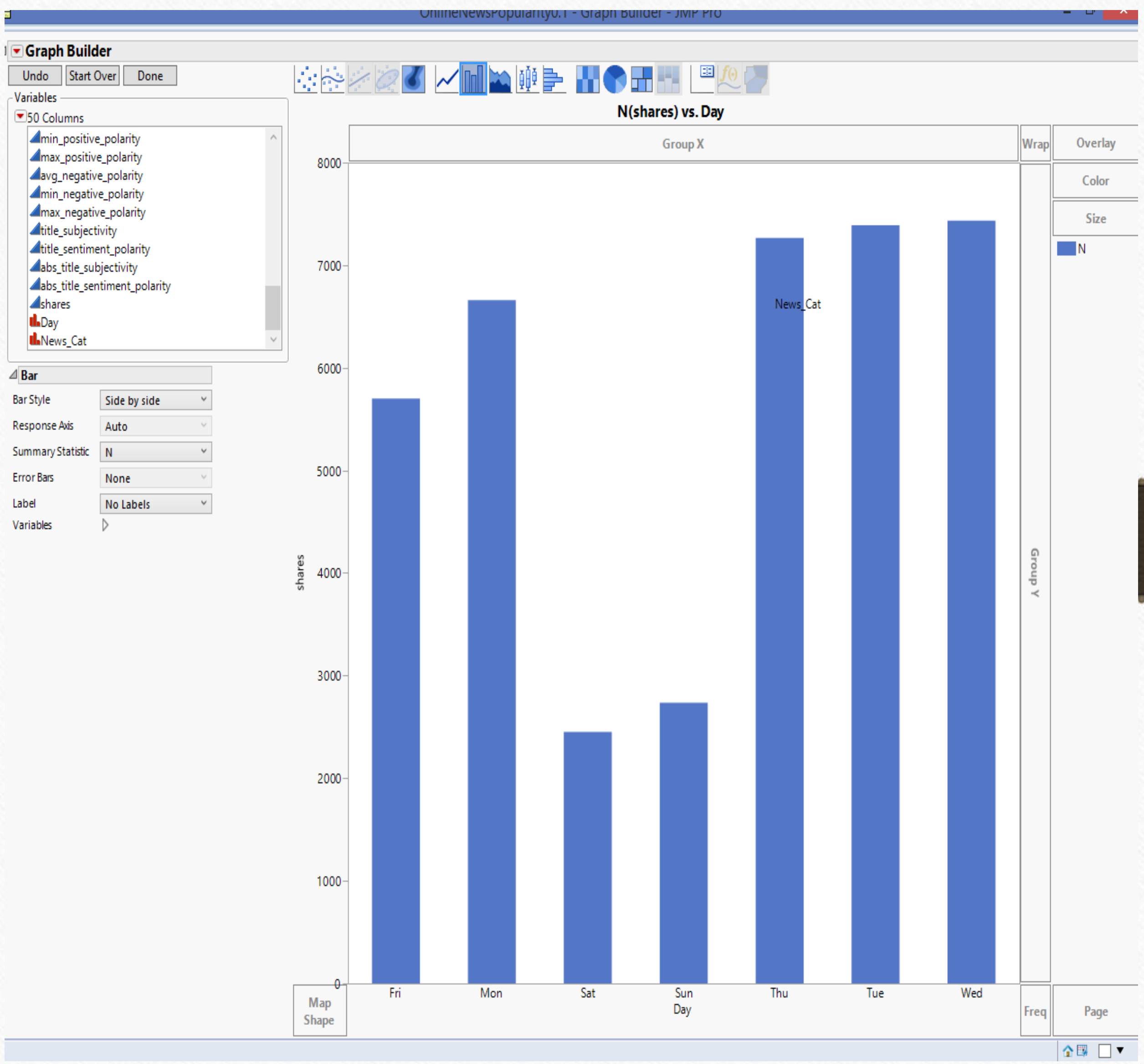
Variables

Data Insights

Publication Day:

- Most articles published - Tuesday, Wednesday, and Thursday.
- Least articles published - Weekends.

OFFICE!!



Bar Style: Side by side
Response Axis: Auto
Summary Statistic: N
Error Bars: None
Label: No Labels
Variables

The Business Value-How can we use this to our advantage

For Mashable

- Publish during the week rather than weekend
- Publish about viral topics, social media articles and avoid world news
- Publish articles closer to the topic (minimize impurity)
- Consider adding ads in the peak time, and related to the topic.

For Researchers

- Always identify your attributes
- To get more accurate results, get data about the number of likes and comments and the time.
- number of tweets or hashtags, number of URL mentions and to understand the

THANK YOU
MERCI
OBRIGADO
ASANTE
СПАСИБО
GRAZIE
WELALIN
CHOKRANE
RAIBH MAITH AGAT
MULTUMESC
NIRRINGRAZZJAK
MOCHCHAKKERAM
MATUR NUWUN
SPASIBO
KHUBLEI
МНОГО ВИ БЛАГОДАРЯ
VINAKA
ZIKOMO
QUYANAQ
SAAGHA XALDA HWA
THANK YOU VERY MUCH
どうも
MATONDO
SUWUN
EIO
MAAKE
БАЯРЛАААА
MOCHCHAKKERAM
ZIKOMO
SPASIBO
МНОГО ВИ БЛАГОДАРЯ
QUYANAQ
MATUR NUWUN