DATA MINING

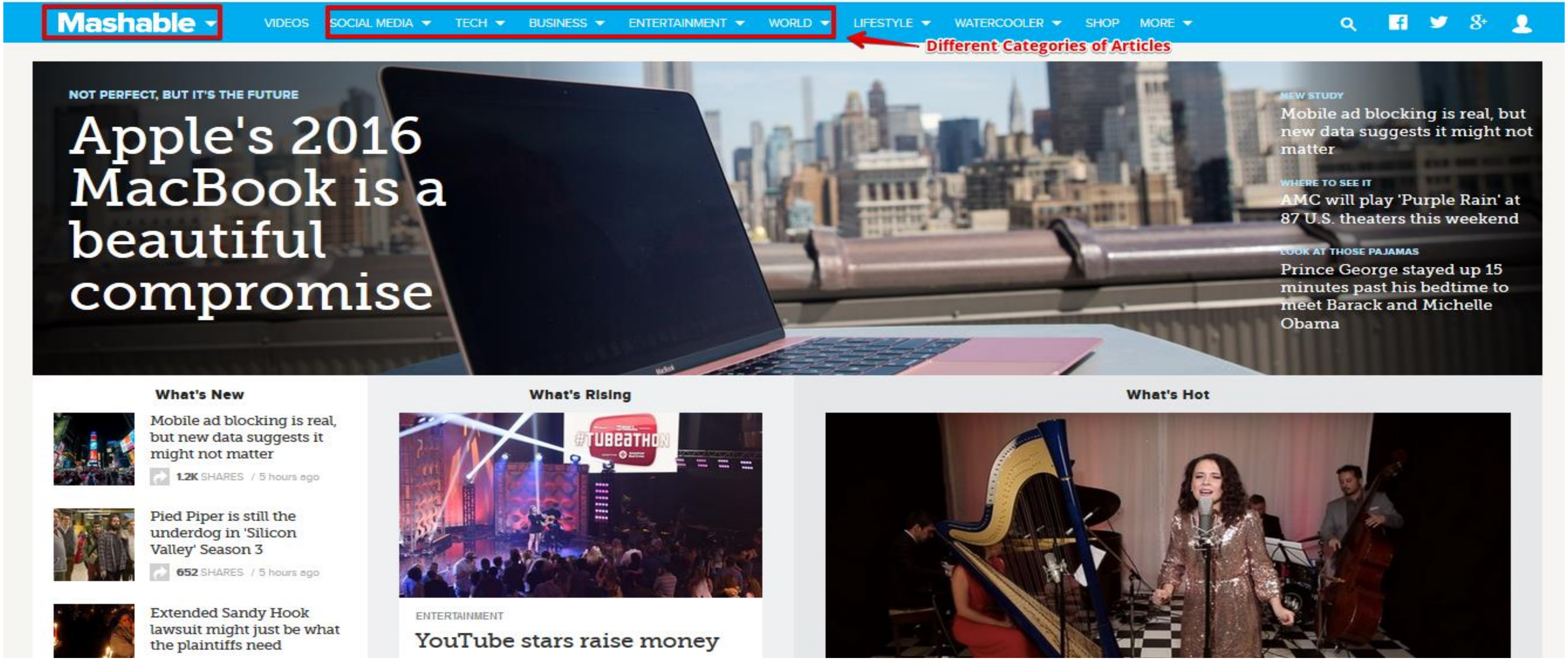
ONLINE NEWS POPULARITY

Sachin Kant Misra
Prashant Bhowmik
Jagpreet Singh Sethi
Renee Champagne

ONLINE NEWS POPULARITY

- ✓ Introduction
- ✓ Problem Statement
- ✓ Dataset Overview
- ✓ Data Cleaning and Pre-Processing
- ✓ Data Modeling and Conclusions
- ✓ Problem 1: To predict the number of Mashable article shares
- ✓ Problem 2: To predict binary target variable ‘Popularity’
- ✓ Problem 3: To predict ordinal outcome for ‘Popularity_level’
- ✓ Visualization using Tableau
  - ❖ Insight 1: How is the distribution of News articles in the month – January
  - ❖ Insight 2: Before shopping on Black Friday, people read and share lot of articles.
  - ❖ Insight 3: New York is the city of Business
- ✓ Model Implementation on Amazon Web Server

INTRODUCTION - MASHABLE WEBSITE



Different Categories of Articles

NOT PERFECT, BUT IT'S THE FUTURE

Apple's 2016 MacBook is a beautiful compromise

NEW STUDY

Mobile ad blocking is real, but new data suggests it might not matter

WHERE TO SEE IT

AMC will play 'Purple Rain' at 87 U.S. theaters this weekend

LOOK AT THOSE PAJAMAS

Prince George stayed up 15 minutes past his bedtime to meet Barack and Michelle Obama

What's New

Mobile ad blocking is real, but new data suggests it might not matter

1.2K SHARES / 5 hours ago

Pied Piper is still the underdog in 'Silicon Valley' Season 3

652 SHARES / 5 hours ago

Extended Sandy Hook lawsuit might just be what the plaintiffs need

What's Rising

ENTERTAINMENT

YouTube stars raise money

What's Hot

PROJECT AIM

- To predict the number of shares of Mashable article.
- To predict the popularity status of the article

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

- To predict an ordinal outcome for popularity levels

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

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

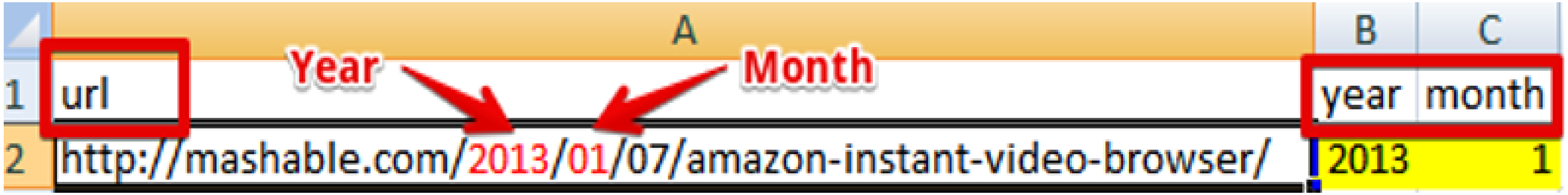
DATASET OVERVIEW

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

DATA CLEANING-EXTRACT YEAR & MONTH

1. **Extracting Year and Month from “url” attribute:**

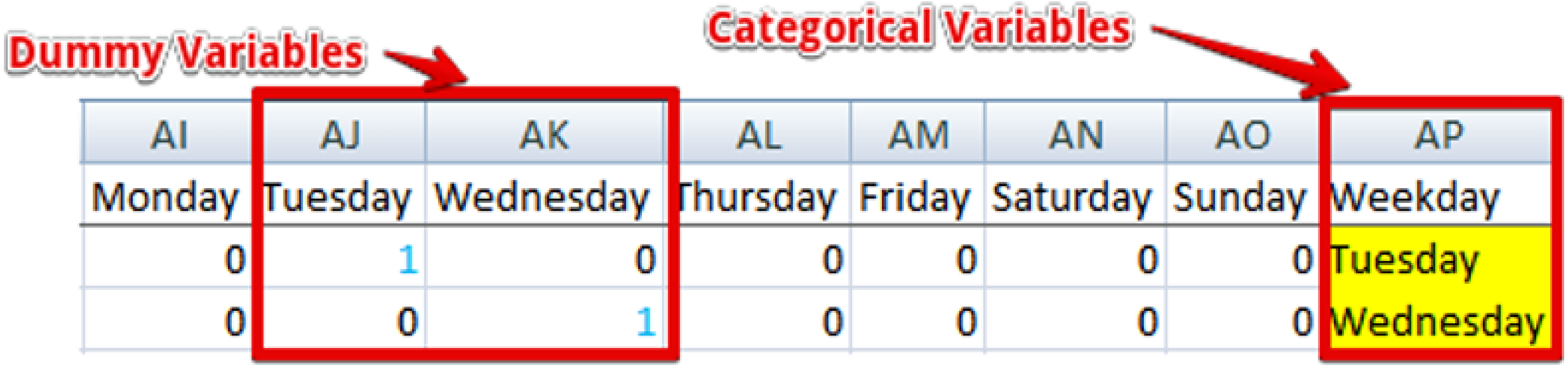
   We have used excel formula e.g. “=MID(A2,21,4)” to extract year and “=MID(A2,26,2)” to extract month from the “url” attribute.



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

DATA CLEANING-CREATE DUMMY VAR

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.



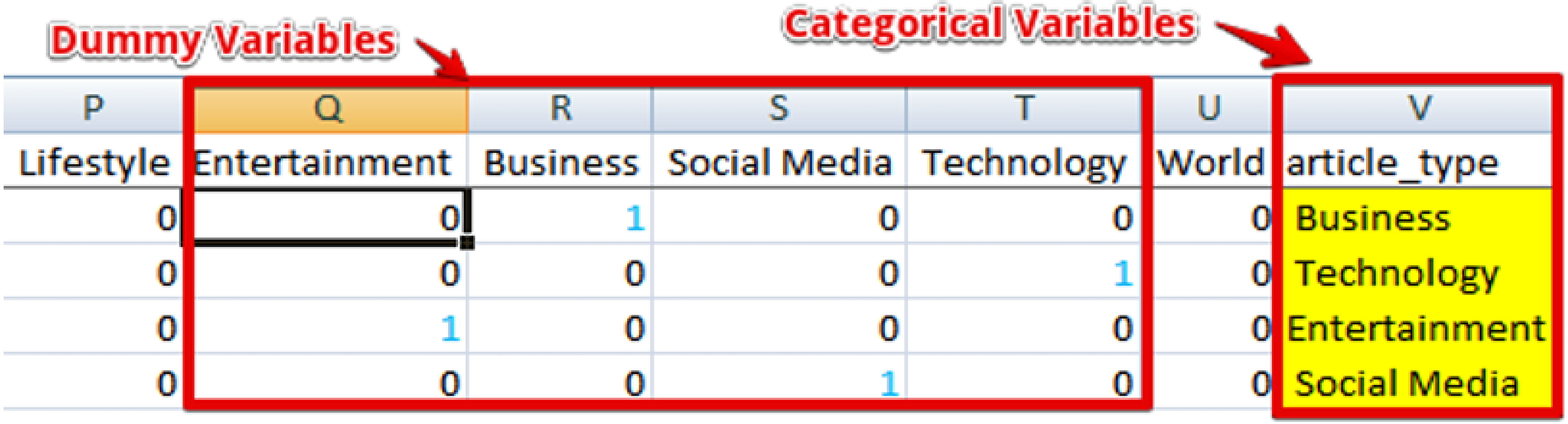
Dummy Variables

Categorical Variables

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

DATA CLEANING-CREATE DUMMY VAR

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.



Dummy Variables

Categorical Variables

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

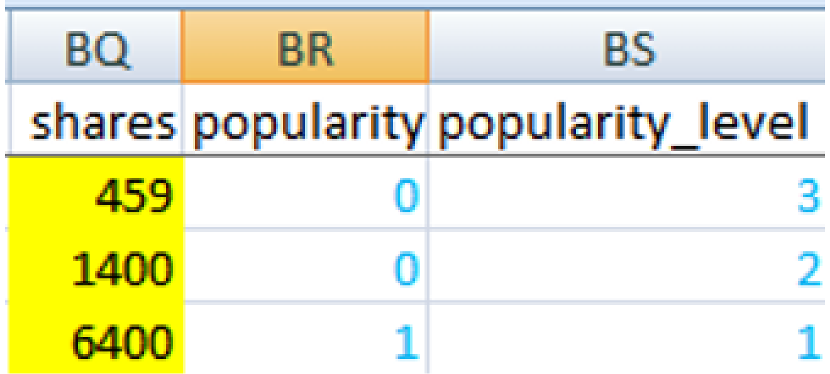
DATA CLEANING – CREATE POPULARITY LEVELS

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.

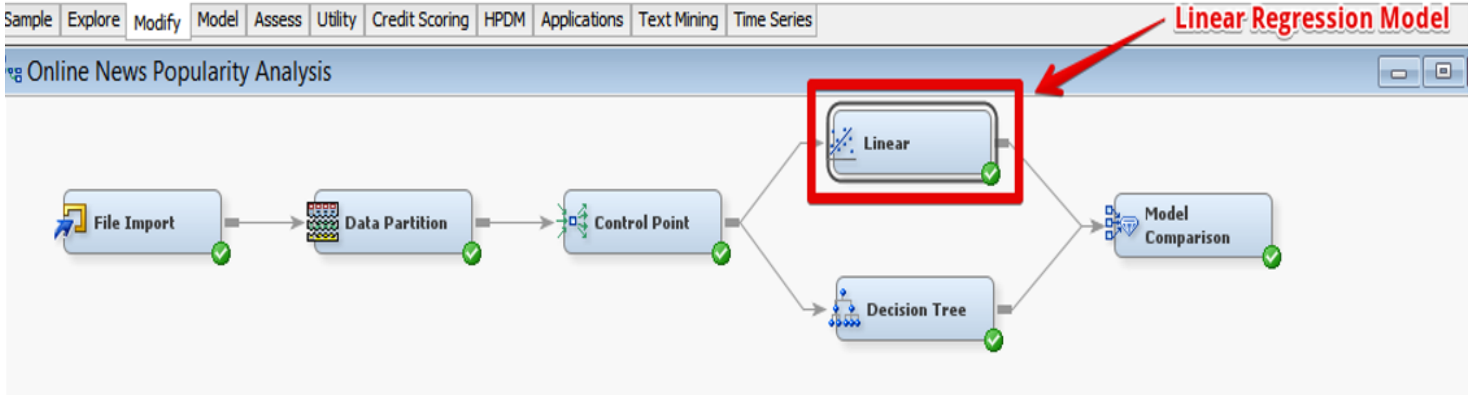
   (II) “=IF(BQ>2100,1,(IF(BQ=>1100 and <=2100),2,3))” – for ‘popularity’ where value = ‘1’ for shares >2100 and value = ‘2’ for shares between 1100 and 2100 and value = ‘3’ for shares < 1100.



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

OBJECTIVE 1 – PREDICT SHARES

Approach 1: Use Kitchen Sink Model on Linear Regression algorithm

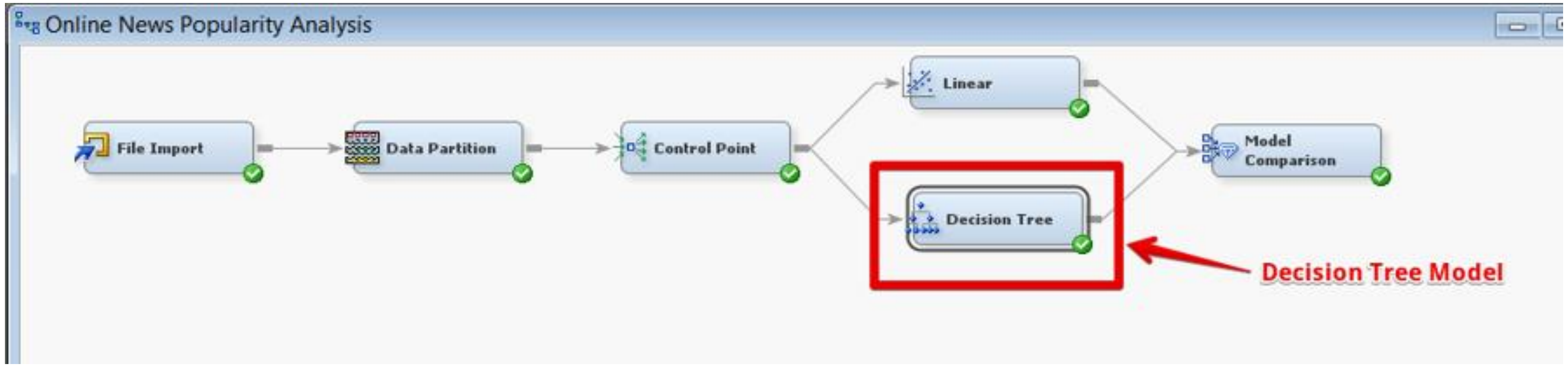


Results:

R-Squared	Adj R-Sq	Evaluation Criteria: MSE
0.0242	0.0199	51610069

OBJECTIVE 1 – PREDICT SHARES

Approach 2: Use Kitchen Sink Model on Decision Tree algorithm

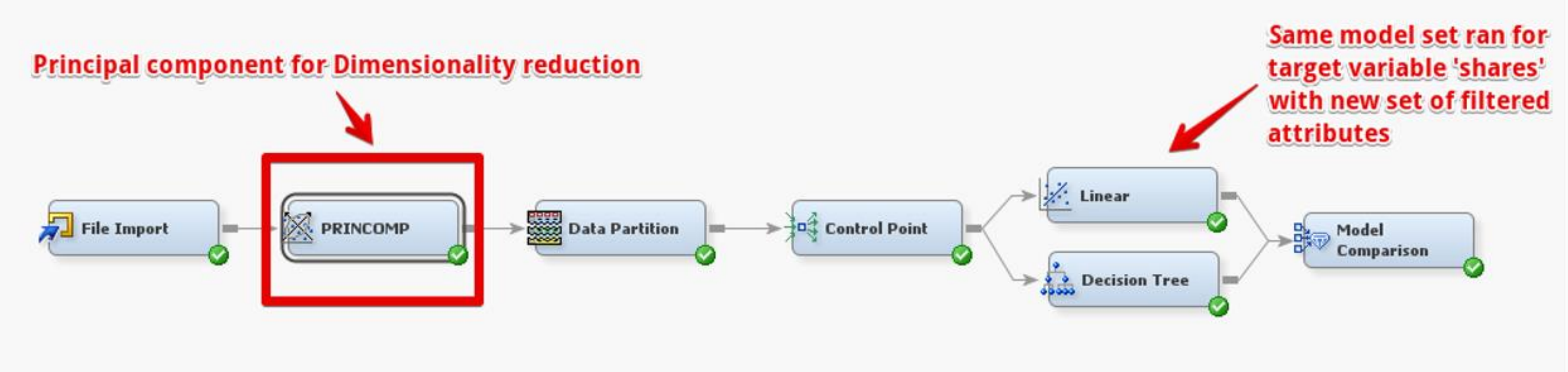


Parameters used and Results:

Depth	Leaf Size	No. of Rules	Interval Target Criteria	Evaluation Criteria: MSE
6	5	5	ProfF	51237007

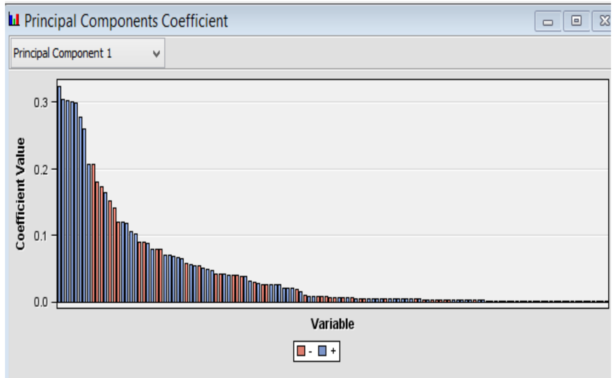
OBJECTIVE 1 – PREDICT SHARES

Approach 3: Use Principal Component Analysis and Linear Regression algorithm



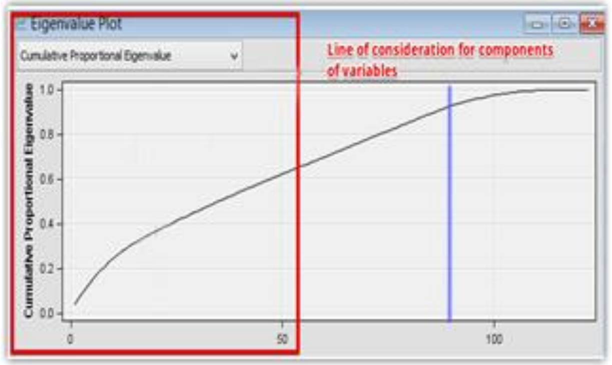
Principal component for Dimensionality reduction

Same model set ran for target variable 'shares' with new set of filtered attributes



Parameters used and Results:

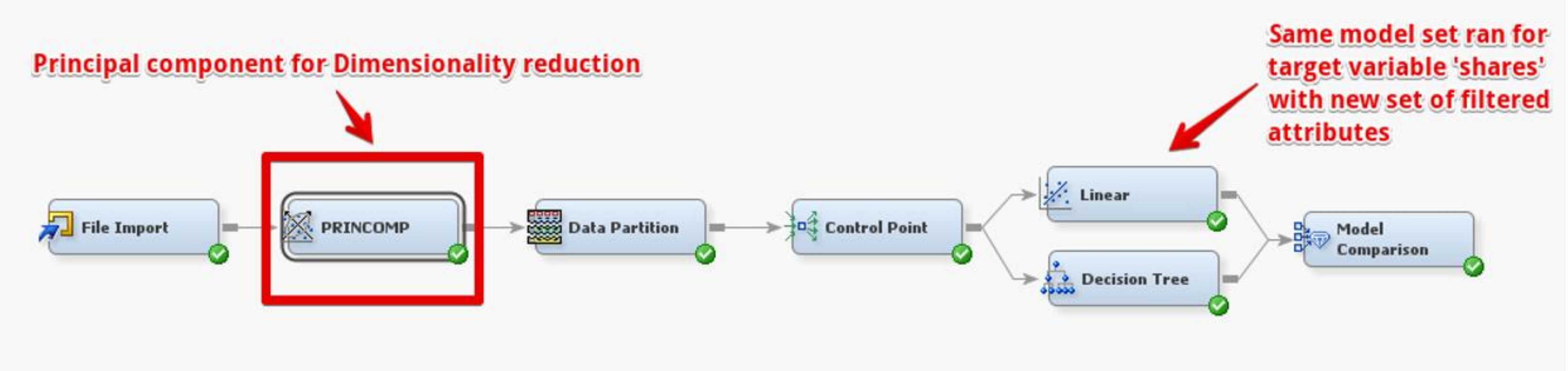
EigenValue	Cumulative Cut Off	Interval Target Criteria	R-Sq	Adj R-Sq	Evaluation Criteria: MSE
Correlation	0.8	ProfF	**0.0143**	**0.0136**	51482431



Line of consideration for components of variables

OBJECTIVE 1 – PREDICT SHARES

Approach 4: Use Principal Component Analysis and Decision Tree algorithm



Principal component for Dimensionality reduction

Same model set ran for target variable 'shares' with new set of filtered attributes

Parameters used and Results:

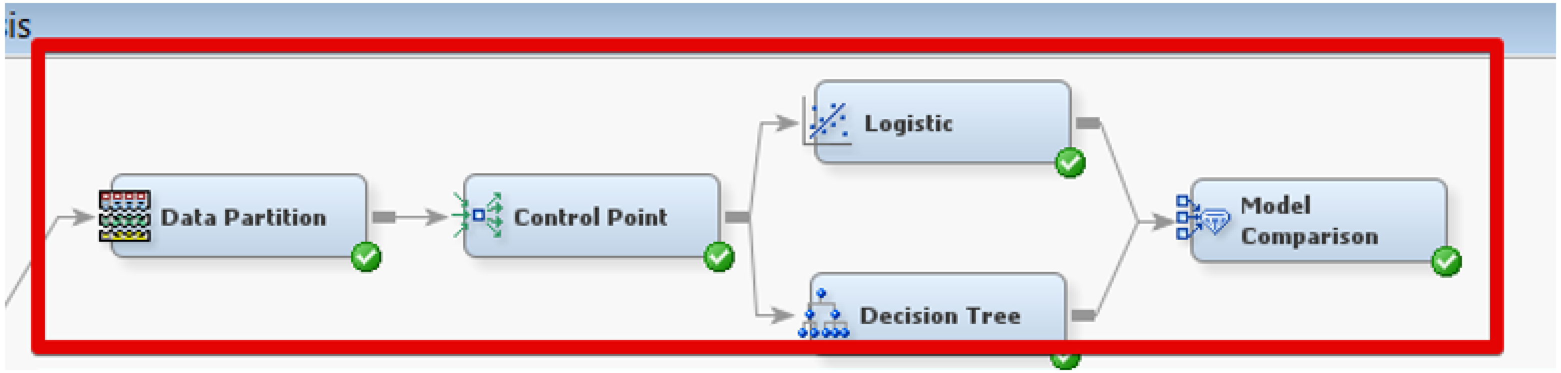
Depth	Leaf Size	No. of Rules	Interval Target Criteria	EigenValue	Cumulative Cut Off	Evaluation Criteria: MSE
6	5	5	ProfF	Correlation	0.8	52753115

OBJECTIVE 1 – CONCLUSION

- Adjusted R-Square is very low in all our approaches (Approx 2%)
- Only 2% of variance in target variable (‘shares’) can be explained which is too less to make predictions.
- 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 use a threshold value on continuous target variable and try to predict ‘SURGE’ or ‘PLUNGE’ in the stock price.
- Thus, we’ll predict popularity of Mashable article with a threshold
  - Popular [Shares > 1400]
  - Not Popular [Shares < 1400]

OBJECTIVE 2 – PREDICT POPULARITY

Approach 1: Use Kitchen Sink Model on Logistic Regression and Decision Tree algorithm



Logistic Regression Results:

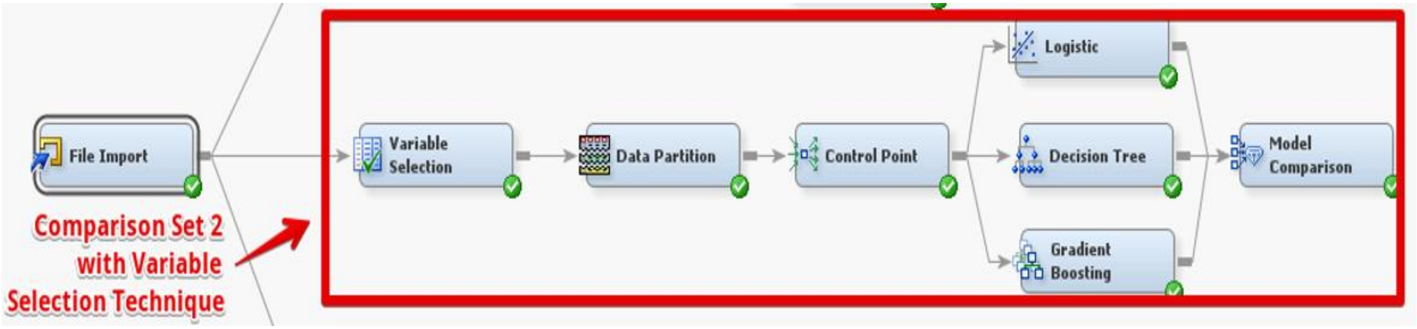
False -ve	False +ve	MISC_ Rate	Evaluation Criteria: Accuracy
1492	1271	0.35	65%

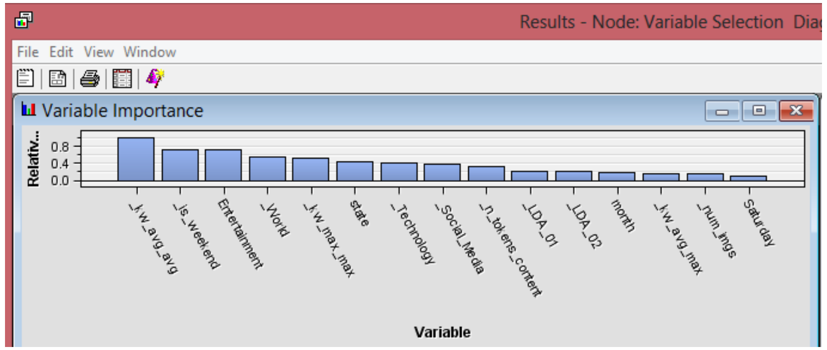
Decision Tree Results:

False -ve	False +ve	MISC_ Rate	Evaluation Criteria: Accuracy
1379	1471	0.36	64%

OBJECTIVE 2 – PREDICT POPULARITY

Approach 2: Use Variable Selection on Logistic Regression, Decision Tree & Gradient Boosting algorithm





Logistic Regression Results:

FN	FP	MISC_ Rate	Evaluation Criteria: Accuracy
1488	1441	0.37	63%

Decision Tree Results:

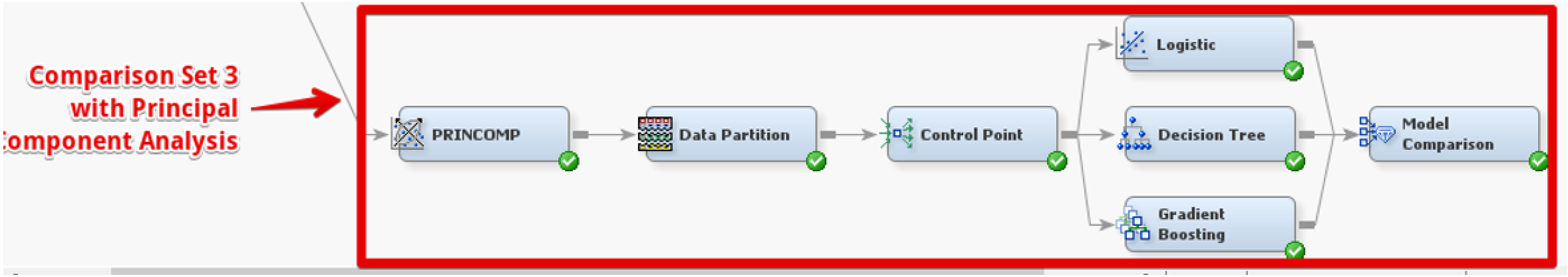
FN	FP	MISC_ Rate	Evaluation Criteria: Accuracy
1325	1530	0.36	64%

Gradient Boosting Results:

FN	FP	MISC_ Rate	Evaluation Criteria: Accuracy
1545	1451	0.38	62%

OBJECTIVE **2** – PREDICT POPULARITY

Approach 3: Use Principal Component Analysis on Logistic Regression, Decision Tree & Gradient Boosting algorithm



Logistic Regression Results:

FN	FP	MISC_ Rate	Evaluation Criteria: Accuracy
1533	1450	0.38	62%

Decision Tree Results:

FN	FP	MISC_ Rate	Evaluation Criteria: Accuracy
1235	1757	0.38	62%

Gradient Boosting Results:

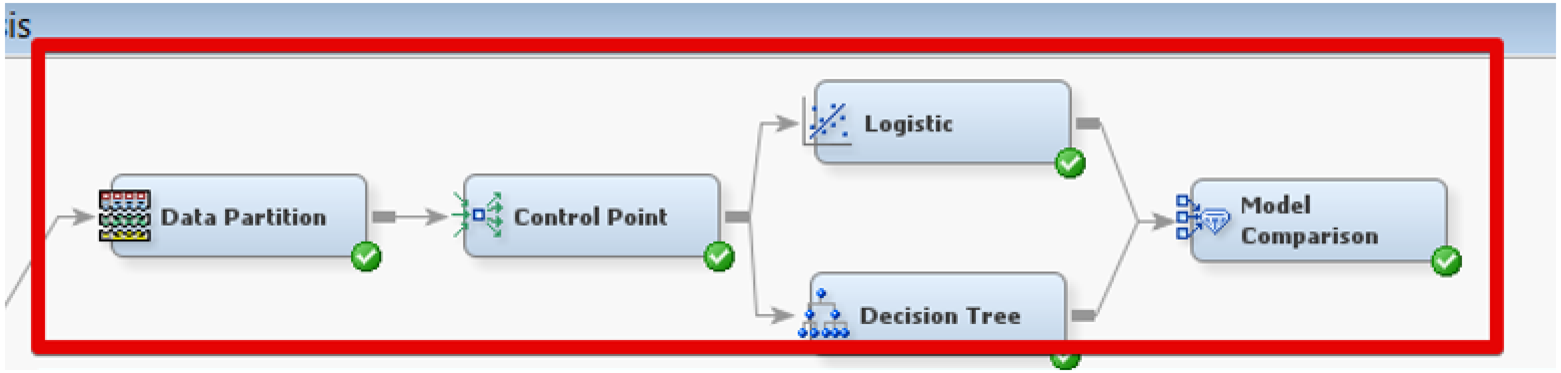
FN	FP	MISC_ Rate	Evaluation Criteria: Accuracy
1400	1545	0.37	63%

OBJECTIVE 2 – CONCLUSION

- Although with a kitchen sink model, we achieved ~65% accuracy using logistic regression, the model seems too complex.
- On compromising only ~1% accuracy, we built models using variable selection technique i.e. filtering input variables on R-Sq (for continuous) and Chi-Sq (for categorical). Thus, simplifying our model.
- Considering the fact that the value of False Negative is more alarming as compared to False Positive. Our selected model should have least FP value i.e. a cost-effective model for business strategy.
- As a result, we prefer **Decision Tree with Variable Selection** dimensionality reduction technique over any other model for prediction of binary target variable popularity.

OBJECTIVE **3** – PREDICT ORDINAL POPULARITY_LEVEL

Approach 1: Use Kitchen Sink Model on Logistic Regression and Decision Tree algorithm



Logistic Regression Results:

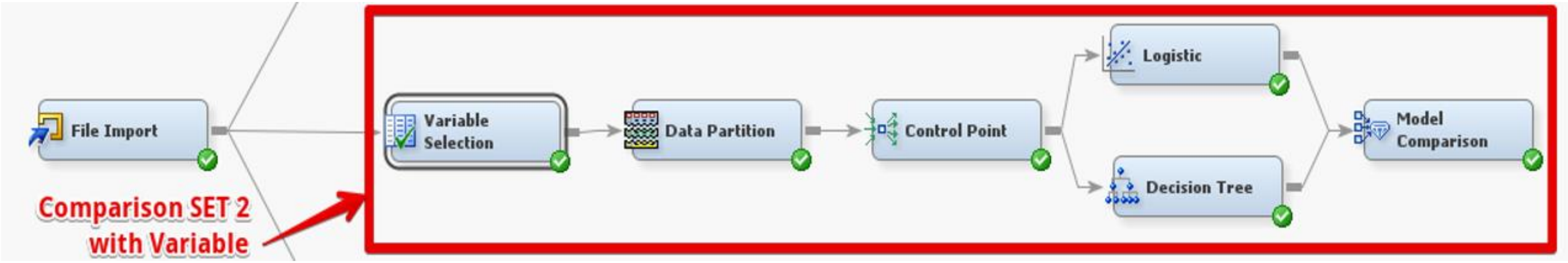
False -ve	False +ve	MISC_ Rate	Evaluation Criteria: Accuracy
757	2207	0.52	48%

Decision Tree Results:

False -ve	False +ve	MISC_ Rate	Evaluation Criteria: Accuracy
869	2164	0.53	47%

OBJECTIVE 3 – PREDICT ORDINAL POPULARITY_LEVEL

Approach 2: Use Variable Selection on Logistic Regression & Decision Tree algorithm



Logistic Regression Results:

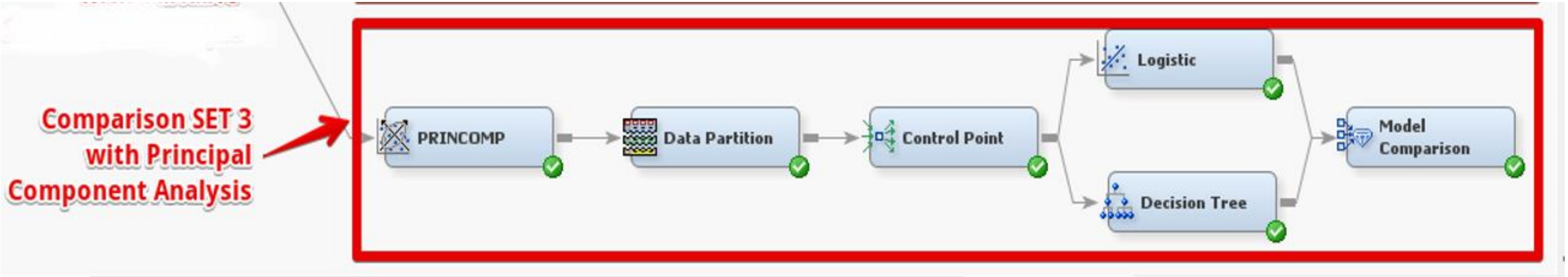
FN	FP	MISC_ Rate	Evaluation Criteria: Accuracy
748	2304	0.53	47%

Decision Tree Results:

FN	FP	MISC_ Rate	Evaluation Criteria: Accuracy
850	2234	0.53	47%

OBJECTIVE 3 – PREDICT ORDINAL POPULARITY_LEVEL

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



Logistic Regression Results:

FN	FP	MISC_ Rate	Evaluation Criteria: Accuracy
748	2433	0.54	46%

Decision Tree Results:

FN	FP	MISC_ Rate	Evaluation Criteria: Accuracy
811	2387	0.54	46%

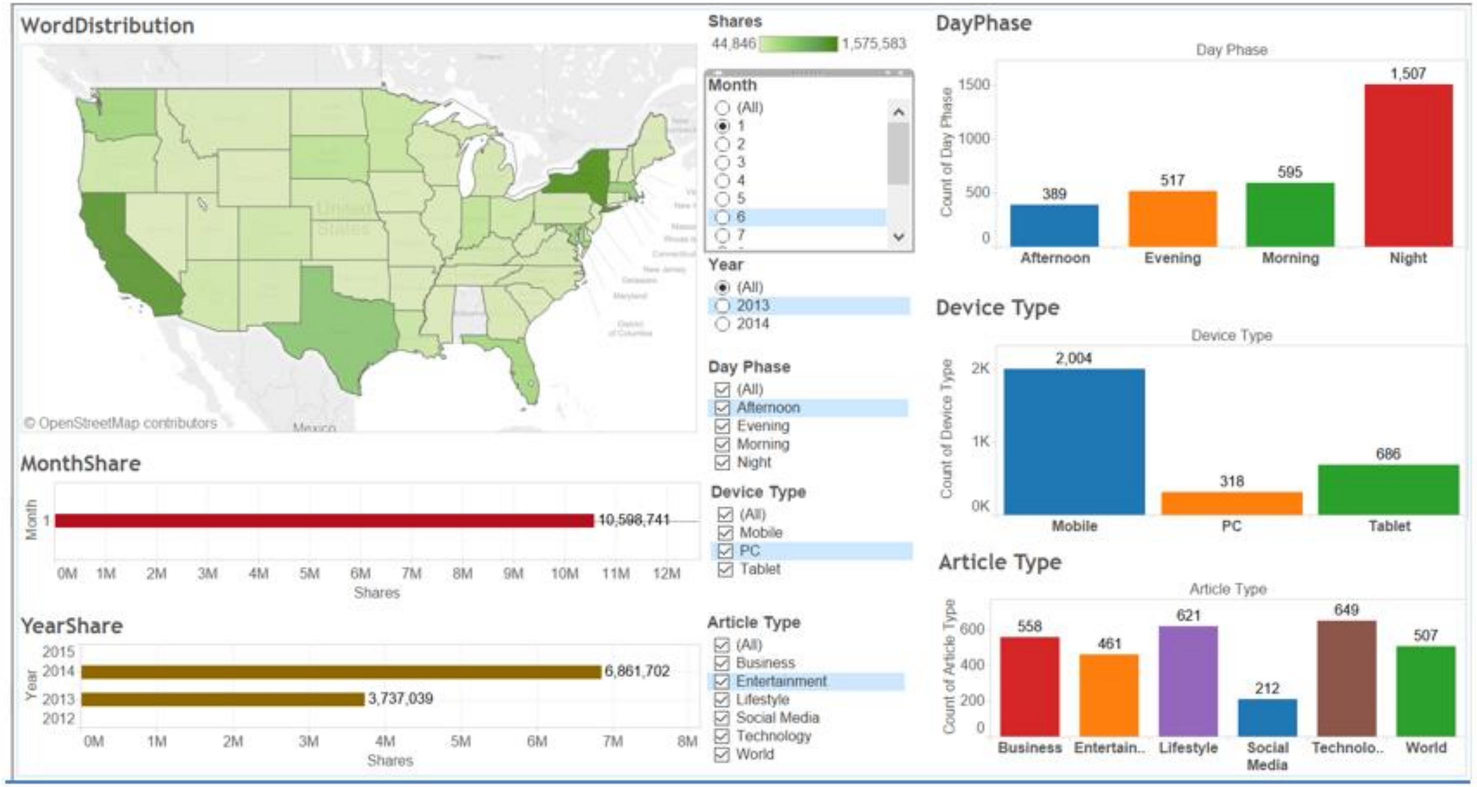
OBJECTIVE 3 – CONCLUSION

- Although with a kitchen sink model, we achieved ~48% accuracy using logistic regression, the model seems too complex.
- On compromising only ~1% accuracy, we built models using variable selection technique i.e. filtering input variables on R-Sq (for continuous) and Chi-Sq (for categorical). Thus, simplifying our model.
- Considering the fact that the value of False Negative is more alarming as compared to False Positive. This rules out the option of choosing Decision Tree over Logistic Regression.
- As a result, we prefer **Logistic Regression with Variable Selection** dimensionality reduction technique over any other model for prediction of Ordinal target variable popularity_level (High, Medium, Low)

TABLEAU VISUALIZATION - 1

Public Tableau: https://public.tableau.com/profile/jagpreet#!/vizhome/book1_10486/dashboard1

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



Month
○ (All)
● 1
○ 2
○ 3
○ 4
○ 5
○ 6
○ 7

Year
● (All)
○ 2013
○ 2014

Day Phase
☑ (All)
☑ Afternoon
☑ Evening
☑ Morning
☑ Night

Device Type
☑ (All)
☑ Mobile
☑ PC
☑ Tablet

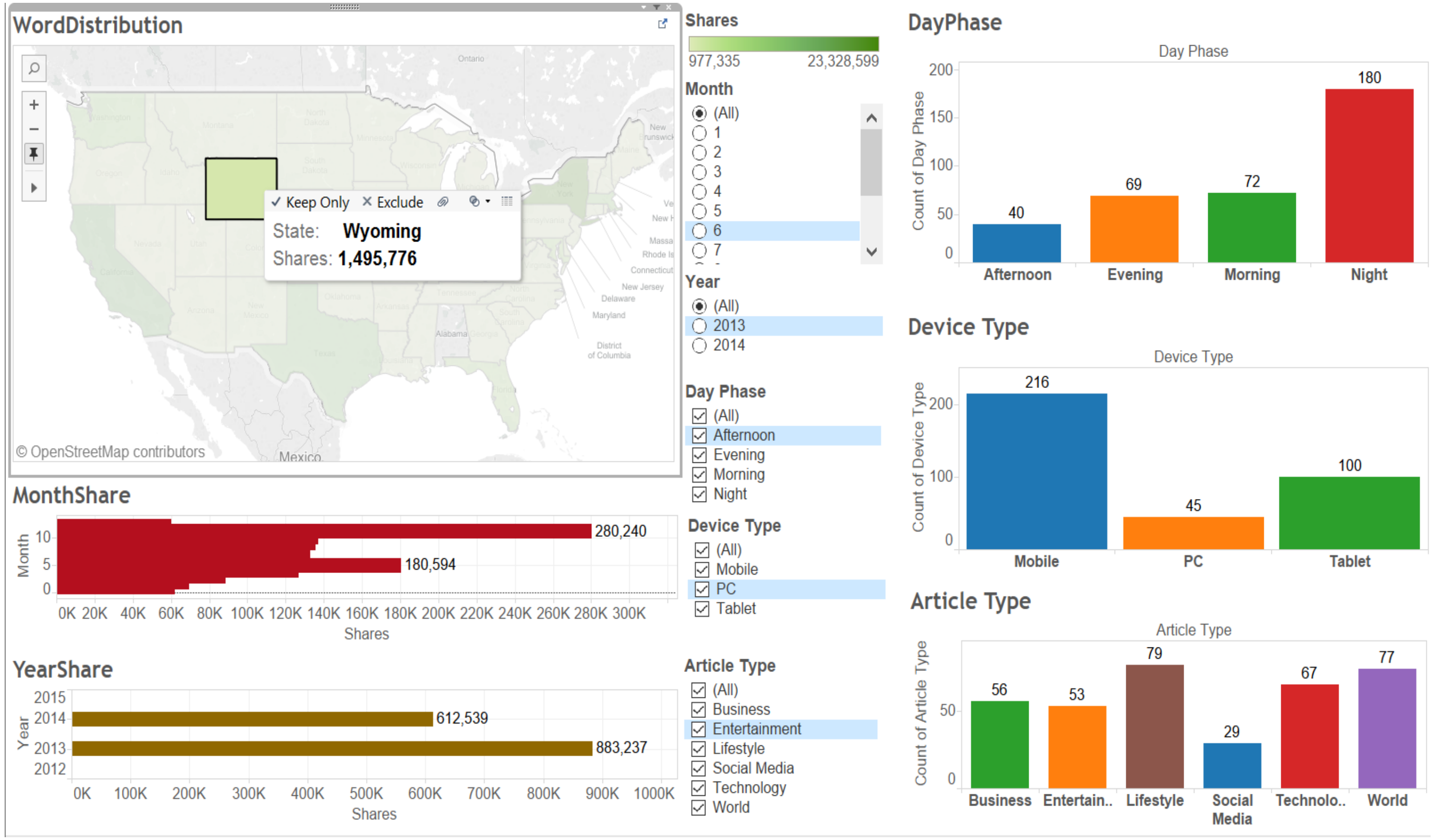
Article Type
☑ (All)
☑ Business
☑ Entertainment
☑ Lifestyle
☑ Social Media
☑ Technology
☑ World

- ❑ No. of shares in Jan 2014 is double than in Jan 2013.
- ❑ Mobile device is preferred to read Mashable articles.
- ❑ Cities - California, Texas, New York and Massachusetts has most of the authors.
- ❑ Most authors post during night hours.
- ❑ In Jan, people share maximum Technology related articles because the company CNET organizes CES product launch conference annually in the month of January.
- ❑ Henceforth, people stay active and share more articles on technology in Jan.

TABLEAU VISUALIZATION - 2

Public Tableau: https://public.tableau.com/profile/jagpreet#!/vizhome/book1_10486/dashboard1

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



State: Wyoming
Shares: 1,495,776

Shares
977,335 23,328,599

Month
● (All)
○ 1
○ 2
○ 3
○ 4
○ 5
○ 6
○ 7

Year
● (All)
○ 2013
○ 2014

Day Phase
☑ (All)
☑ Afternoon
☑ Evening
☑ Morning
☑ Night

Device Type
☑ (All)
☑ Mobile
☑ PC
☑ Tablet

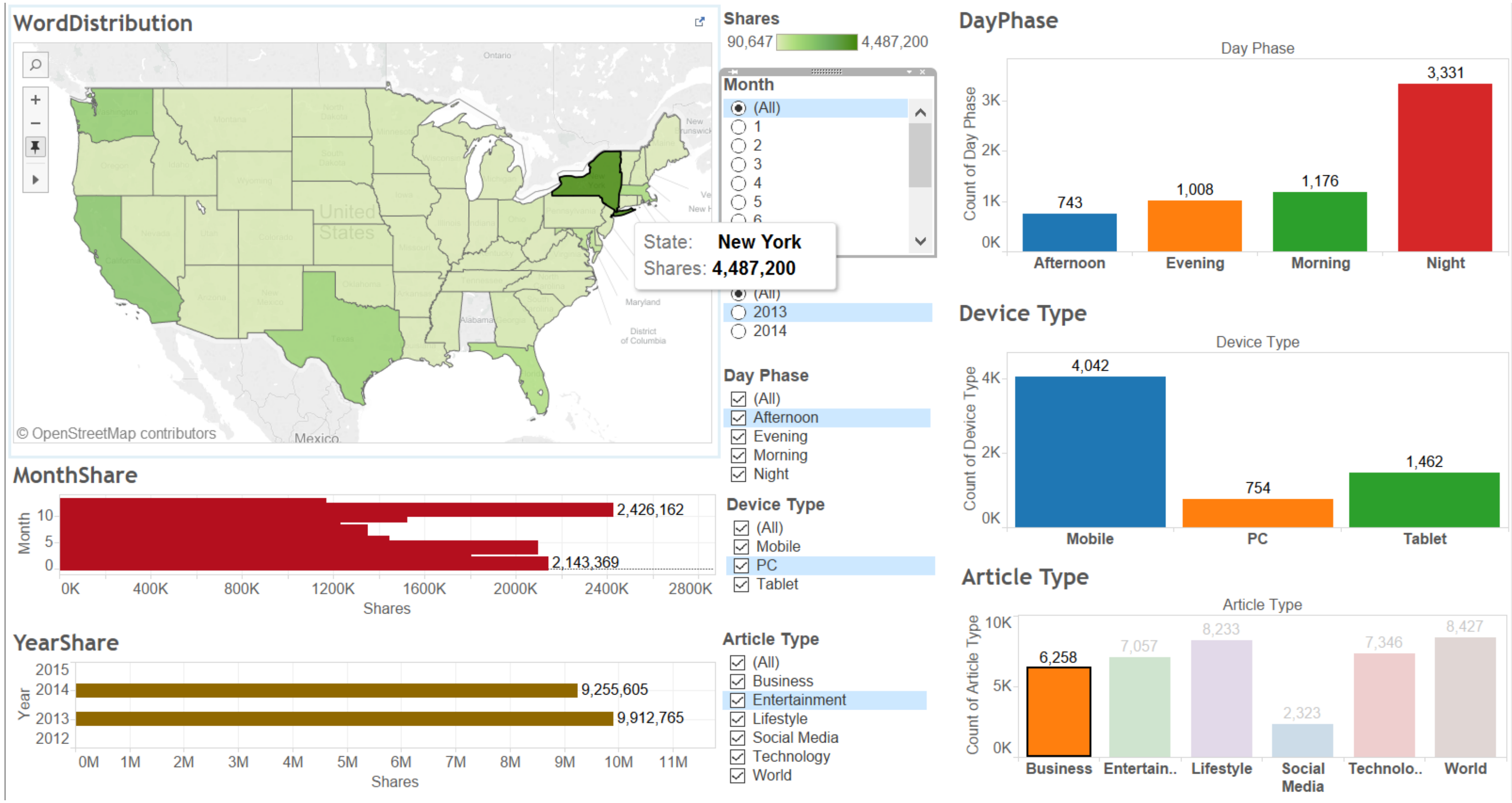
Article Type
☑ (All)
☑ Business
☑ Entertainment
☑ Lifestyle
☑ Social Media
☑ Technology
☑ World

- ❑ Authors in Wyoming publish more Lifestyle related articles.
- ❑ Such articles are shared more during last few months of a year.
- ❑ Festive like Christmas Holiday, Black Friday and Labor Day bring heavy discount on shopping, this makes people visit Lifestyle related Mashable articles even more.

TABLEAU VISUALIZATION - 3

Public Tableau: https://public.tableau.com/profile/jagpreet#!/vizhome/book1_10486/dashboard1

Insight 3: New York is the city with maximum business-minded crowd



Month
● (All)
○ 1
○ 2
○ 3
○ 4
○ 5
○ 6

● (All)
○ 2013
○ 2014

Day Phase
☑ (All)
☑ Afternoon
☑ Evening
☑ Morning
☑ Night

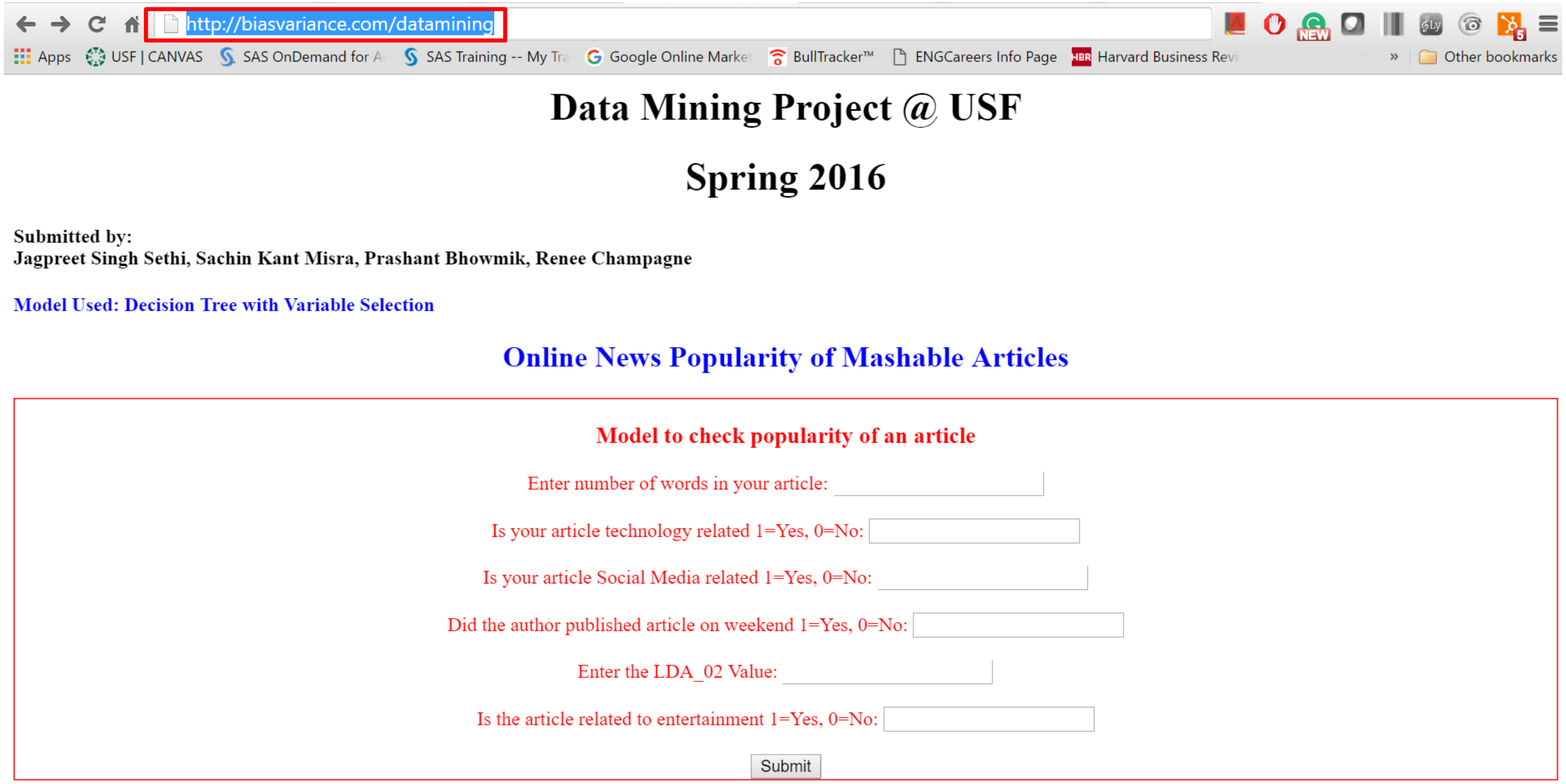
Device Type
☑ (All)
☑ Mobile
☑ PC
☑ Tablet

Article Type
☑ (All)
☑ Business
☑ Entertainment
☑ Lifestyle
☑ Social Media
☑ Technology
☑ World

- ❑ Most of the authors in New York publish articles on Mashable.com during night.
- ❑ People working in NY love to read and share "Business" centric Mashable articles,
- ❑ Henceforth, New York has maximum number of business-minded crowd.

MODEL IMPLEMENTATION

http://biasvariance.com/datamining (DEPLOYED ON AMAZON WEB SERVICES)



Data Mining Project @ USF

Spring 2016

Submitted by:
Jagpreet Singh Sethi, Sachin Kant Misra, Prashant Bhowmik, Renee Champagne

Model Used: Decision Tree with Variable Selection

Online News Popularity of Mashable Articles

Model to check popularity of an article

Enter number of words in your article:

Is your article technology related 1=Yes, 0=No:

Is your article Social Media related 1=Yes, 0=No:

Did the author published article on weekend 1=Yes, 0=No:

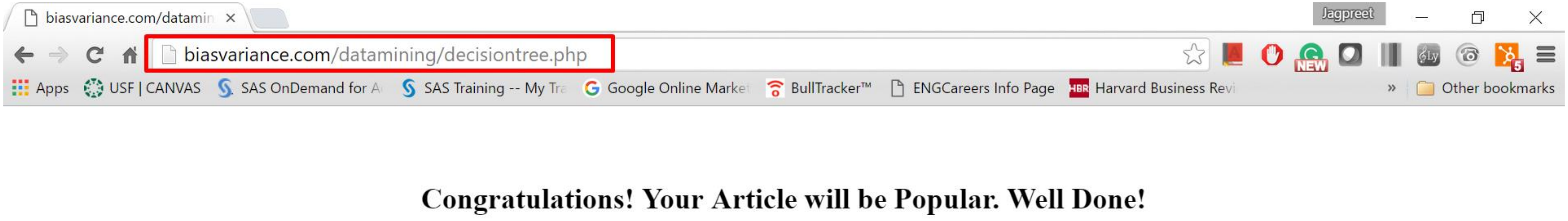
Enter the LDA_02 Value:

Is the article related to entertainment 1=Yes, 0=No:

Submit

MODEL IMPLEMENTATION

http://biasvariance.com/datamining (DEPLOYED ON AMAZON WEB SERVICES)



biasvariance.com/datamining/decisiontree.php

Congratulations! Your Article will be Popular. Well Done!

Thank you!