StumbleUpon

Exploring a dataset provided by Kaggle

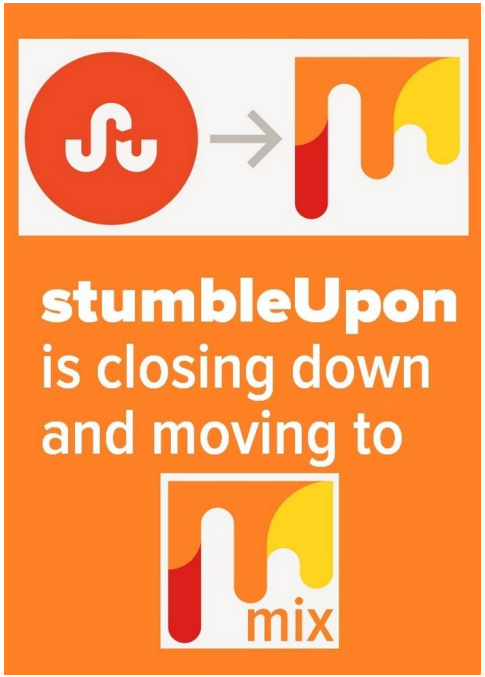
Erica Calvi
January 2022

What was StumbleUpon?

- Founded in 2002
- Ran for 16 years
- In 2018 transitioned into Mix.com

“A social network that helps you discover unique and interesting things across the Web”

https://smallbiztrends.com/2014/08/what-is-stumbleupon-how-do-i-use-it.html



Description of the Data Set

Descriptive Statistics of StumbleUpon Data

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
alchemy_category_score	5,053	0.6	0.2	0.1	0.5	0.8	1.0
avglinksize	7,395	2.8	8.6	0.0	1.6	2.6	363.0
commonlinkratio_1	7,395	0.5	0.2	0.0	0.3	0.6	1.0
commonlinkratio_2	7,395	0.2	0.1	0.0	0.1	0.3	1.0
commonlinkratio_3	7,395	0.1	0.1	0.0	0.02	0.1	1.0
commonlinkratio_4	7,395	0.05	0.1	0.0	0.0	0.1	1.0
compression_ratio	7,395	2.3	5.7	0.0	0.4	0.6	21.0
embed_ratio	7,395	-0.1	0.3	-1	0	0	0
frameTagRatio	7,395	0.1	0.04	0.0	0.03	0.1	0.4
html_ratio	7,395	0.2	0.1	0.05	0.2	0.3	0.7
image_ratio	7,395	0.3	1.9	-1.0	0.03	0.2	113.3
linkwordscore	7,395	30.1	20.4	0	14	43	100
non_markup_alphanum_characters	7,395	5,716.6	8,875.4	0	1,579	6,377	207,952
numberOfLinks	7,395	178.8	179.5	1	82	222	4,997
numwords_in_url	7,395	5.0	3.2	0	3	7	22
parametrizedLinkRatio	7,395	0.2	0.2	0.0	0.04	0.2	1.0
spelling_errors_ratio	7,395	0.1	0.1	0.0	0.1	0.1	1.0

Quick Data Facts:

Source: Kaggle

Created at least 8 years ago

7395 observations

27 variables

Label classification is split about 50/50

Pre-Processing

NULL Variables
(either char or unimportant):

- framebased
- url
- urlid
- boilerplate
- news_front_page

Converting Variables:

- alchemy_category - char to factor
- alchemy_category_score - char to num
- hasDomainLink - num to factor
- is_news -> char to factor
- Is_news -> converts all “?” to 0 - assumption
- lengthyLinkDomain -> num to factor
- label -> num to factor

Pre-Processing

Outliers

- Experimented with filling outliers with IQR - however the supervised models chosen are robust to outliers and produced the same results.

Missing Values

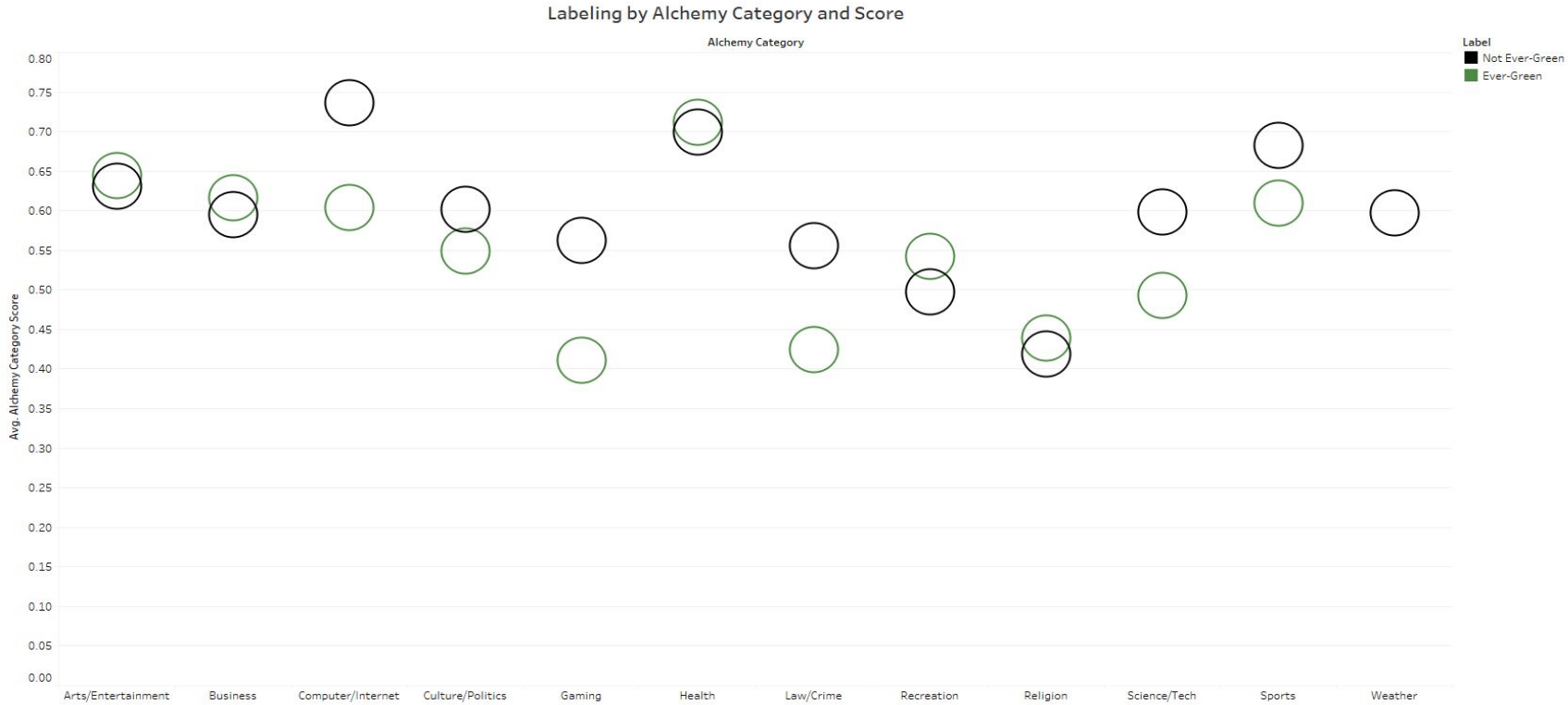
- Used Random Forest to predict missing values of alchemy_category_score
- Couldn’t use to fill factor categories such as alchemy_category

Exploratory Analysis

Questions Prior to Analysis:

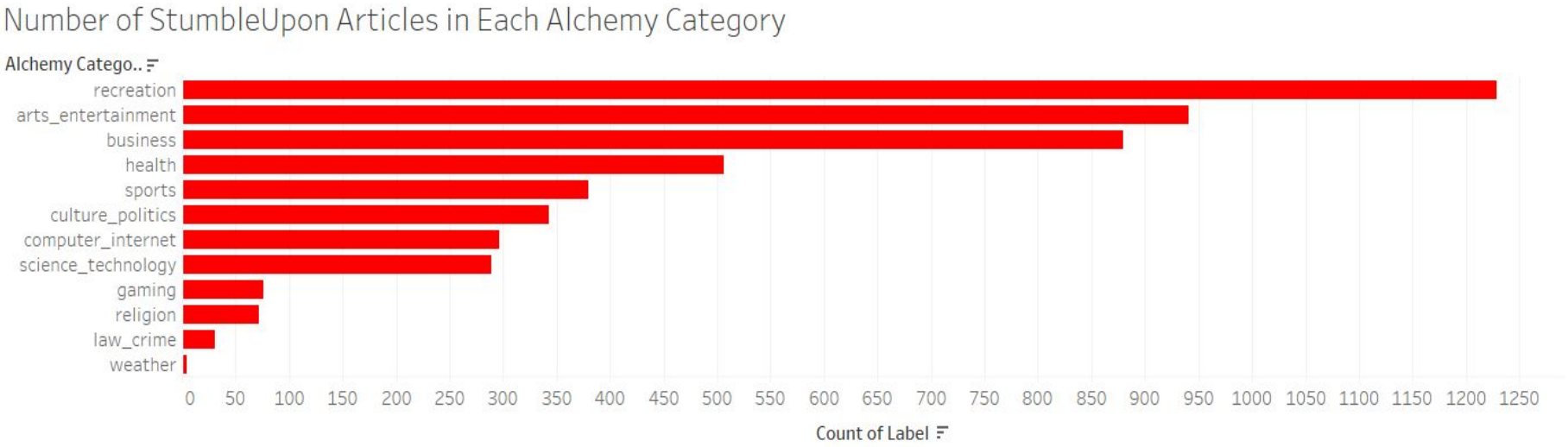
- Does a high alchemy score lead to an evergreen label?
- How is the data distributed amongst alchemy categories?
- Which variables play the biggest role in determining if the website is labeled evergreen?
- What are the most common used topics/words? Does common usage imply likelihood of being label evergreen?

Alchemy Score Chart



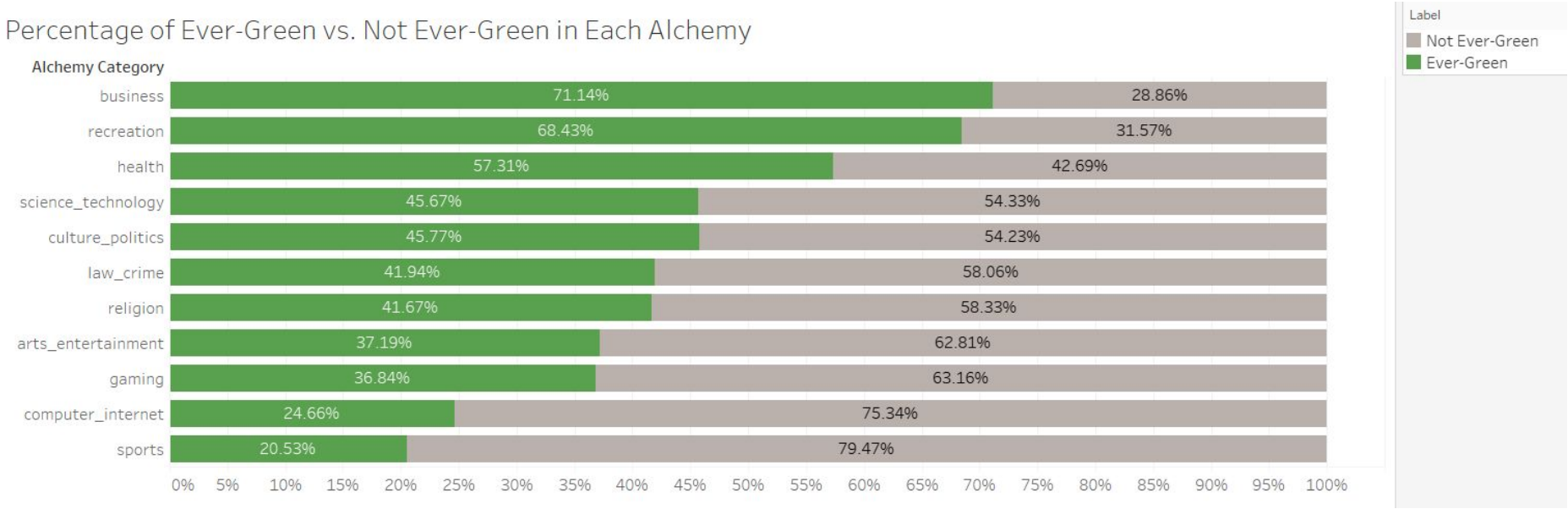
Category labels '?' and 'unknown' were removed for this visualization

Alchemy Categories



Category labels '?' and 'unknown' were removed for this visualization

Alchemy Categories - Evergreen or Not



Category labels '?' and 'unknown' were removed for this visualization

The Unsupervised Model

Bag-of-Words

Unsupervised ML Model - “Bag-of-Words”

- The goal was to uncover the relationship between the **key words** within the **Boilerplate** and the **Evergreen label**

- The Bag-of-Words technique identified all words within the boilerplate variable

- Used a filter to remove stop words and words used less than 500 times
  - Reduced the number of words from 92,642 to 602

Unsupervised ML Model - “Bag-of-Words”

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1	add	added	adding	age	ago	air	alcohol	amazing	american	amount	app	apple	april	art	article	august	awesome	baby	bacon	bad	bag	bake	baked	baking
2	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	1	0	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1

- Each row represents each url - which has a unique boilerplate
- Each column is one word that appeared at least 500 times in the boilerplate column of the original data set
- If the boilerplate for each url contained the word, then a “1” was assigned, if the word was not in the boilerplate it was assigned a “0”

“Bag-of-Words” Results

- These are the top 20 non-filler words used at least 500 times
- % Evergreen represents how often a word was contained in a URL/boilerplate that was labeled evergreen

Top 20 Words		
Word	Total Count	% Evergreen
Recipe	7161	91%
Cup	6476	88%
Time	6306	60%
Food	6053	70%
Chocolate	5596	90%
Minutes	5241	87%
Add	5207	87%
Recipes	4943	86%
Butter	4777	93%
Sugar	4467	88%
Cream	3890	89%
Top	3882	65%
People	3711	43%
Cheese	3709	92%
Water	3651	76%
Day	3648	60%
Health	3491	49%
Baking	3352	93%
Cake	3340	88%
Salt	3205	91%

What words are most likely to be Evergreen?

Creamy

Preheat

Sprinkle

Rack

Tsp

Crust

Mixture

What Words are Least Likely to be Evergreen?

Game - 23%

Technology - 17%

News - 20%

Fashion - 12%

Sports - 18%

The Supervised Models

Logistic Regression, Naive Bayes',
Classification Tree and Random Forest

Supervised ML Model - Logistic Regression

- Exploring how the independent variables affect the label

- Applied step function to model to weed out weak variables

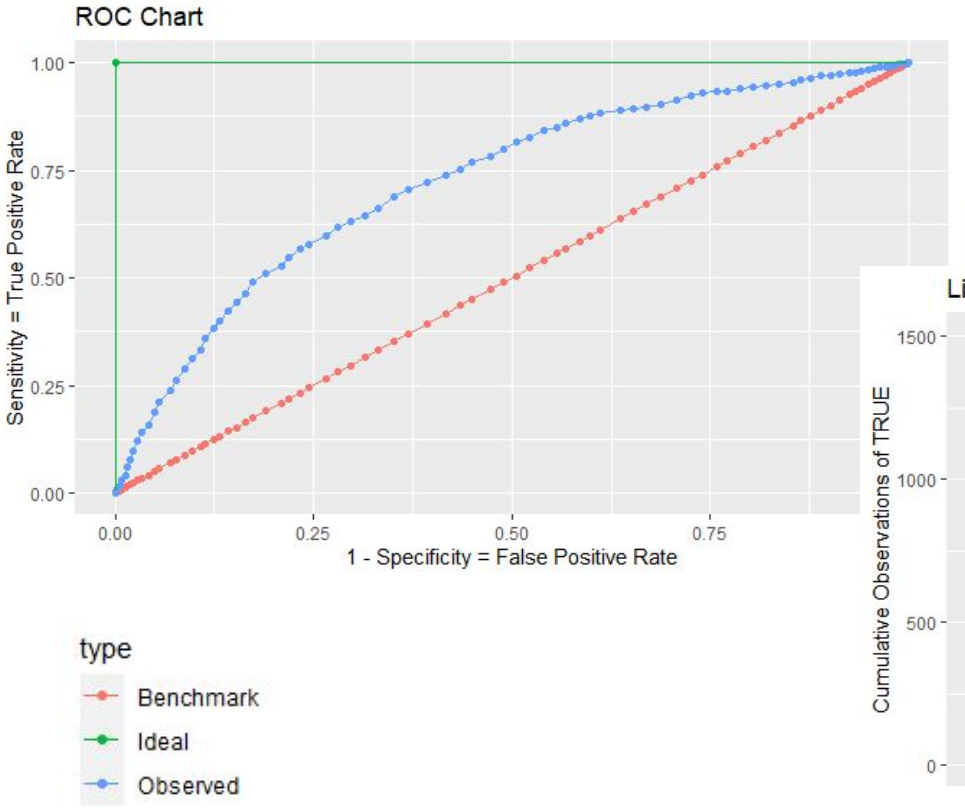
Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	1.100e+00	1.687e-01	6.523	6.90e-11	***
alchemy_category1	9.478e-01	1.163e-01	8.152	3.57e-16	***
alchemy_category10	-1.708e-01	1.649e-01	-1.036	0.300276	
alchemy_category11	-1.257e+00	1.836e-01	-6.846	7.62e-12	***
alchemy_category12	-1.267e+01	2.249e+02	-0.056	0.955059	
alchemy_category13	1.590e+00	1.146e+00	1.388	0.165241	
alchemy_category2	-4.500e-01	1.061e-01	-4.240	2.23e-05	***
alchemy_category3	-9.277e-01	1.902e-01	-4.877	1.08e-06	***
alchemy_category4	-6.079e-02	1.532e-01	-0.397	0.691510	
alchemy_category5	-7.636e-01	3.513e-01	-2.174	0.029713	*
alchemy_category6	4.534e-01	1.322e-01	3.429	0.000605	***
alchemy_category7	-1.157e-01	4.894e-01	-0.236	0.813063	
alchemy_category8	7.585e-01	1.000e-01	7.583	3.39e-14	***
alchemy_category9	-7.230e-01	3.876e-01	-1.865	0.062126	.
alchemy_category_score	-5.275e-01	1.612e-01	-3.273	0.001064	**
commonlinkratio_1	5.457e-01	2.230e-01	2.447	0.014417	*
commonlinkratio_3	3.517e+00	8.159e-01	4.311	1.63e-05	***
commonlinkratio_4	-2.847e+00	9.423e-01	-3.021	0.002516	**
frameTagRatio	-6.534e+00	1.026e+00	-6.370	1.90e-10	***
image_ratio	-8.040e-02	3.303e-02	-2.434	0.014927	*
linkwordscore	-2.448e-02	2.241e-03	-10.924	< 2e-16	***
non_markup_alphanum_characters	-2.019e-05	5.573e-06	-3.624	0.000291	***
numberOfLinks	9.542e-04	2.754e-04	3.464	0.000531	***
spelling_errors_ratio	-1.941e+00	4.544e-01	-4.272	1.94e-05	***

Alchemy_Category as Factors

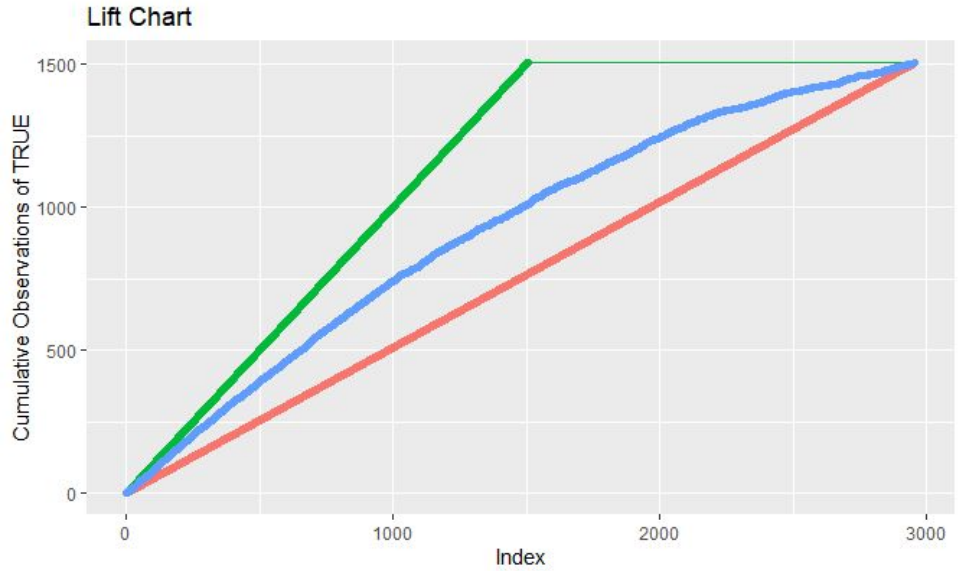
(1) Business, (2) Arts/Entertainment, (3) Computer/Internet, (4) Culture/Politics, (5) Gaming, (6) Health, (7) Law/Crime, (8) Recreation, (9) Religion, (10) Science/Technology, (11) Sports, (12) Weather, (13) Unknown

Supervised ML Model - Logistic Regression



Error Rate: 33%
Bench Error: 49%
AUC ROC: 0.72

Sensitivity: 70% correct trues (1 - evergreen)
Specificity: 62% correct falses (0-not evergreen)



Supervised ML Model - Naive Bayes'

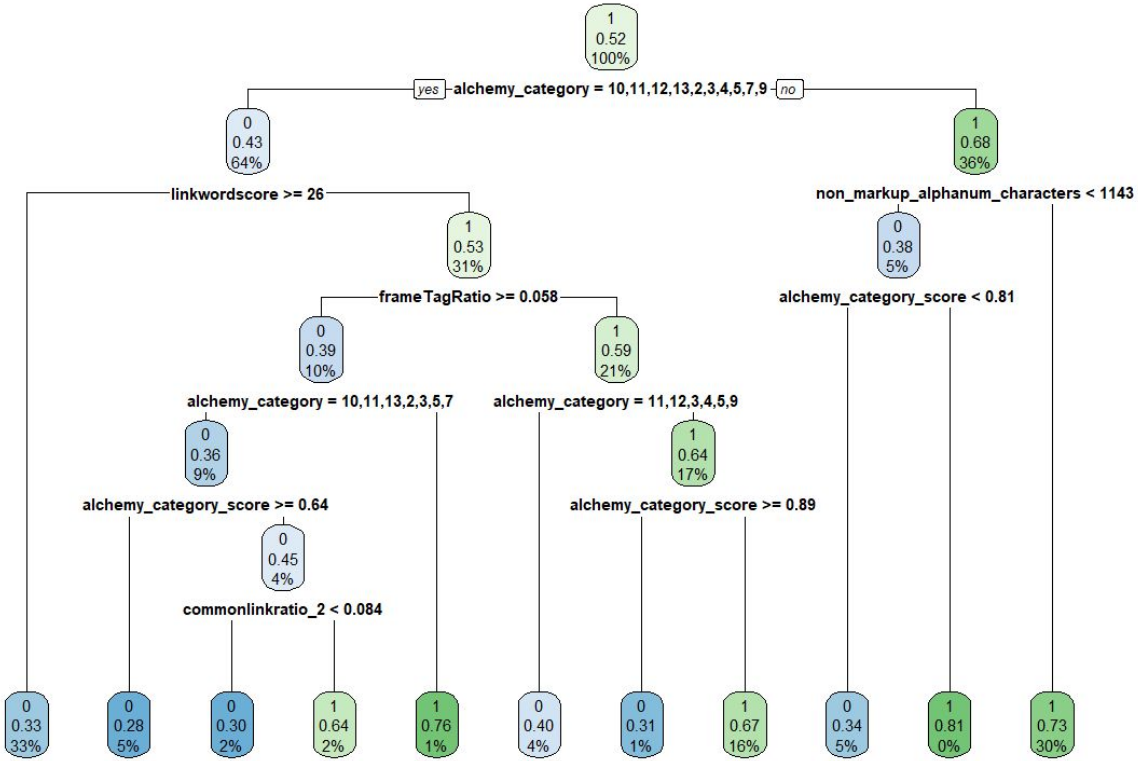
- Took bag of words created by unsupervised "bag-of-words" model and converted dataframe so that each word is a column and each row was 0 or 1, depending on if a data entry boilerplate contained that column word
- Helped determine the odds for a given word to be labeled evergreen

	Observations	
Predictions	0	1
0	662	308
1	75	434
Total	737	742

The odds that a boilerplate containing "nutella" is labeled evergreen is 17:1

Error Rate: 26%
Bench Error: 50%
Sensitivity: 58% correctly predicted trues (1 - evergreen)
Specificity: 90% correctly predicted falses (0- not evergreen)

Supervised ML Model - Classification Tree



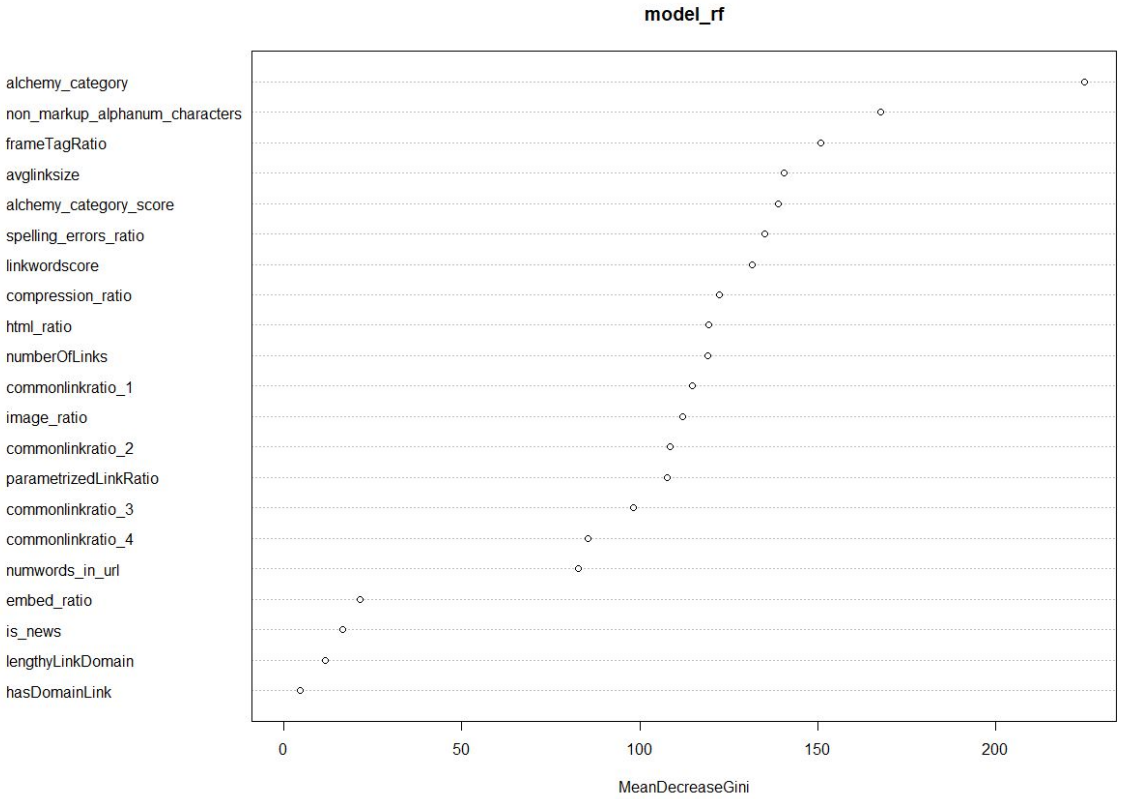
Error Rate: 34%

Bench Error: 50%

- Started with overfit model and then pruned

Helps outline which variables are more important in determining the label and other splits within.

Supervised ML Model - Random Forest



Error Rate: 29%

Bench Error: 50%

- Uses 500 trees

This plot also helps us visualize which variables are the most important for determining the label.

Conclusion

- Many of the numeric variables in the dataset that focus on the technical aspects of the website aren't very good at explaining what makes a website evergreen.
    - Alchemy Category appears to be the most important within those variables
- Using bag-of-words to explore the boilerplate variable was very helpful at revealing what topics were more likely to be labeled evergreen
- Topics about food appear to be more likely to be labeled evergreen, while topics such as news, fashion, and sports are not
    - Believed that new information about news,fashion, and sports are generated frequently, so old articles/webpages are more likely to be "ephemeral"

Questions?

Thank you!