NEWS ARTICLES PROJECT

BY SIMRAN CELLINI

PRESENTATION OUTLINE

- Project motivation
- Project aim.
- Gathering and analyzing my data.
- Approach.
- Data Exploration.
- Data Exploration Part 2.
- Methodology.
- Naïve Bayes Result
- Support Vector Machine Results.
- Logistic Regression Results.
- Conclusion

PROJECT MOTIVATION

My Background

I studied TV production at university and I have a lot of previous working experience in the media industry in London and in Canada. I undertook placements at ITV, Toronto Film Festival, and more. I wanted to use my experience and knowledge from the media industry and analyze use it in a real data science project.

I chose this particular topic as there are different news articles being published both hourly and daily, meaning that there is a lot of data to be able to use and analyze with. In the near future I would potentially like to work as a junior data scientist working in the media industry.

PROJECT AIM

HOW TO PREDICT CATEGORIES FROM NEWS HEADLINES?

GATHERING & ANALYSING MY DATA

The first step was to gather and analyze my data. The original table showed 400.000 different news articles that were published by 10.000 different news publishers such as The Sun, Daily Mail and many more.

ID	TITLE	URL	PUBLISHER	CATEGORY	STORY	HOSTNAME	TIMESTAMP
1	Fed official says weak data caused by weather,...	http://www.latimes.com/business/money/la-fi-mo...	Los Angeles Times	b	ddUyU0VZz0BRneMioxUPQVP6sIxvM	www.latimes.com	1394470370698
2	Fed's Charles Plosser sees high bar for change...	http://www.livemint.com/Politics/H2EvwJSK2VE6O...	Livemint	b	ddUyU0VZz0BRneMioxUPQVP6sIxvM	www.livemint.com	1394470371207
3	US open: Stocks fall after Fed official hints ...	http://www.ifamagazine.com/news/us-open-stocks...	IFA Magazine	b	ddUyU0VZz0BRneMioxUPQVP6sIxvM	www.ifamagazine.com	1394470371550

ID: Number of news articles in the table.
URL: The web link to that particular article.
CATEGORY: Business, Entertainment, Technology and Medical.
HOSTNAME: The companies website.

TITLE: The different article headlines.
PUBLISHER: The company that published the article.
STORY: Different publishers writing about the same story.
TIMESTAMP: The number of seconds since January 1st 1970 GMT.

APPROACH

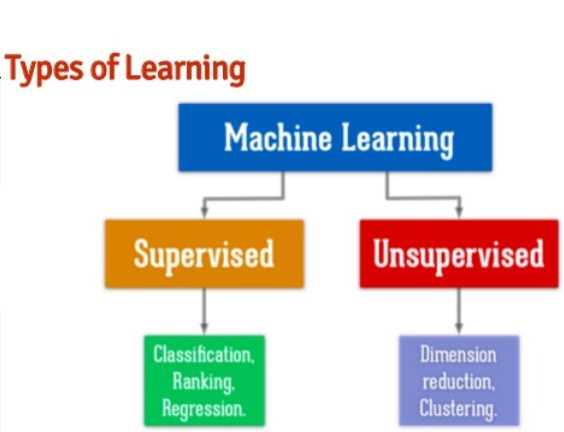
Inputs were taken from the table. They were the word counts for each title.

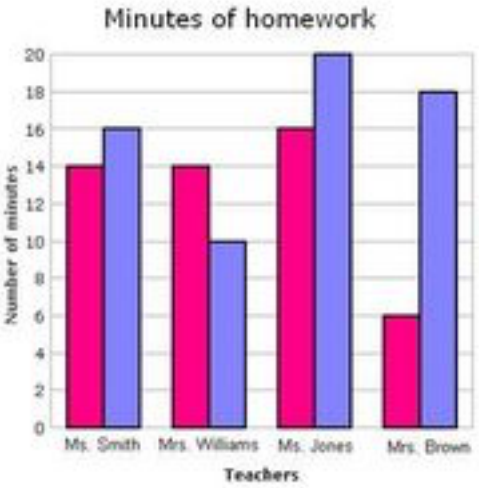
Model built on the training set to predict what category news headlines fall into when a new article is published.

Validate / Test the model on the test set . Explore data to understand which inputs are most important.

Analyse the results.

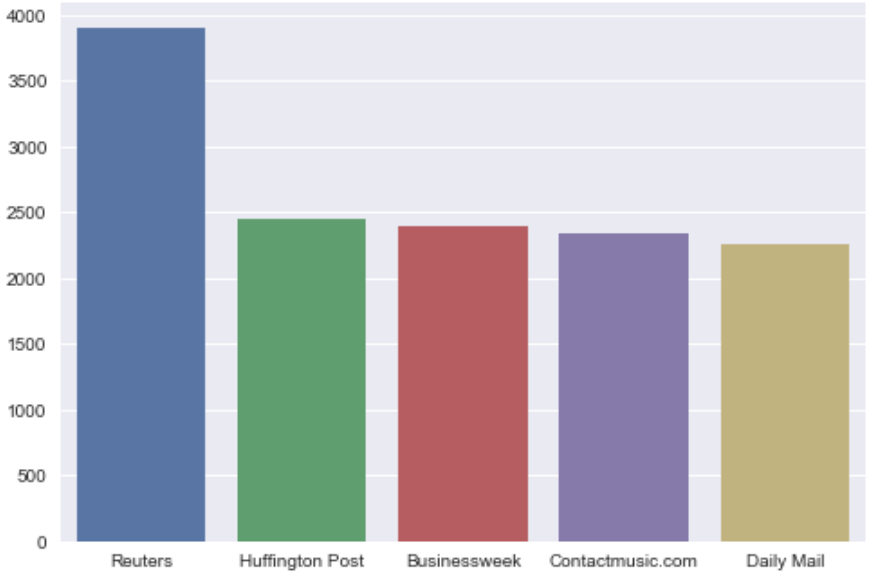
ID	TITLE	URL
33472	How Anita Baker Found Out There Was a Warrant ...	http://www.vintagevinylnews.com/2014/03/how-an...
329504	Frank Darabont could direct 'Snow White and th...	http://www.thenewage.co.za/129834-1022-53-Fran...
82591	How I Met Your Mother finale causes tears and ...	http://www.stuff.co.nz/entertainment/tv-radio/...



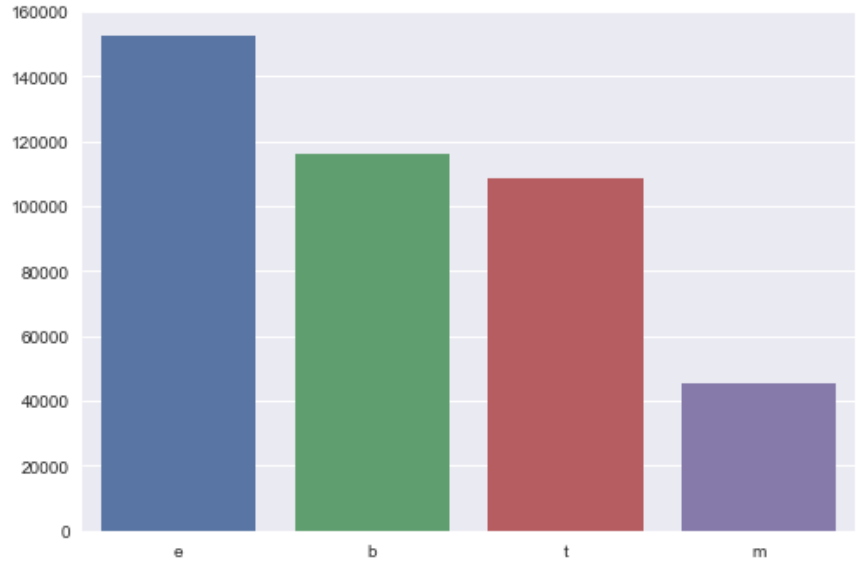


DATA EXPLORATION

- For **Bar Chart One** I first wanted to find out what my top 5 publishers were out of the 10.000 publishers. Who published the most news articles. **For Bar Chart Two** I wanted to find out what category published the most news article.



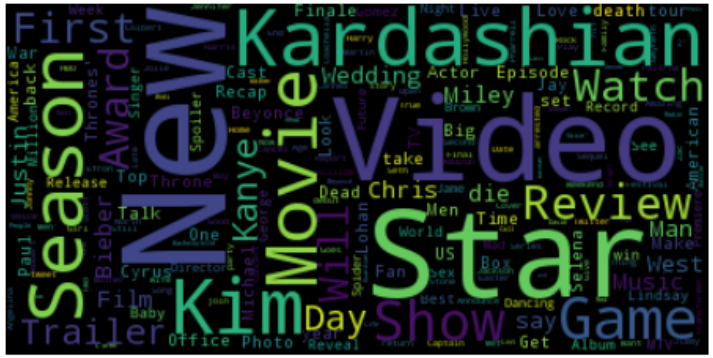
Bar Chart One



Bar Chart Two

DATA EXPLORATION PART 2

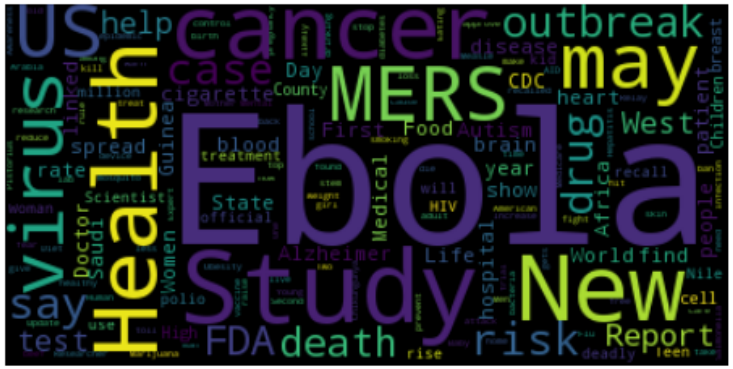
During my data exploration I also wanted to find out what the most popular words were in each category. I created four individual word clouds.



Entertainment Category Word Cloud

Business Category Word Cloud

Technology Category Word Cloud



Medical Category Word Cloud

METHODOLOGY

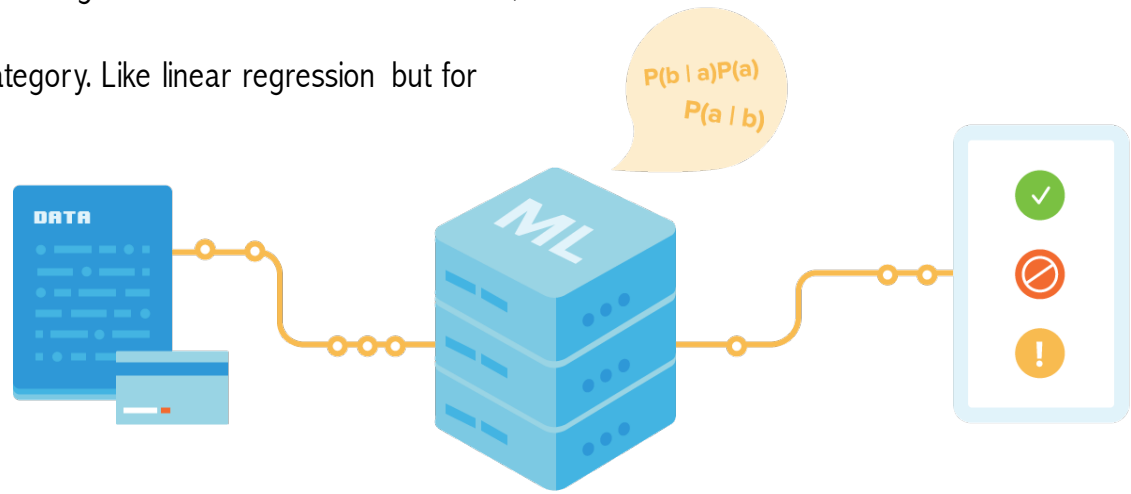
Below are the three machine learning models that I used.

Models

Naïve Bayes (Looks at each feature on its own, individual word count in a title, looks at the relationship)
Support Vector Machine (Splits the two classes by the inputs e.g. business or non business class, the largest separation the better)
Logistic Regression (Calculates the probability of each category. Like linear regression but for classification)

Evaluation

Confusion Matrix
Precision & Recall
Accuracy



NAÏVE BAYES RESULTS

CONFUSION MATRIX

Predicted

Actual		Business	Entertainment	Medical	Technology
	Business	31544	616	523	2026
	Entertainment	636	43917	446	573
	Medical	553	268	12546	312
	Technology	2329	867	272	29298

PRECISION & RECALL

	precision	recall	f1-score	support
b	0.91	0.90	0.90	35062
e	0.96	0.96	0.96	45668
m	0.92	0.91	0.91	13787
t	0.89	0.91	0.90	32209
avg / total	0.93	0.93	0.93	126726

ACCURACY

92%

SUPPORT VECTOR MACHINE RESULTS

CONFUSION MATRIX

Predicted

Actual	Business	Entertainment	Medical	Technology
Business	32527	503	446	1609
Entertainment	467	44571	227	382
Medical	304	190	12919	143
Technology	1764	404	195	30075

PRECISION & RECALL

	precision	recall	f1-score	support
b	0.93	0.93	0.93	35062
e	0.98	0.98	0.98	45668
m	0.95	0.94	0.94	13787
t	0.93	0.93	0.93	32209
avg / total	0.95	0.95	0.95	126726

ACCURACY

95%

LOGISTIC REGRESSION RESULTS

CONFUSION MATRIX

Predicted

Actual

	Business	Entertainment	Medical	Technology
Business	32542	504	268	1748
Entertainment	500	44707	111	350
Medical	515	287	12788	197
Technology	1573	456	105	30075

PRECISION & RECALL

	precision	recall	f1-score	support
b	0.93	0.93	0.93	35062
e	0.97	0.98	0.98	45668
m	0.96	0.93	0.95	13787
t	0.93	0.93	0.93	32209
avg / total	0.95	0.95	0.95	126726

ACCURACY

95%

CONCLUSION

- To sum up the two best models to use was either Support Vector Machine or Logistic Regression. It shows that by looking at news headlines we can predict what type of category the headline will fall into. Both of the models gave an accuracy of 95 %.

- I also found out that certain key words frequently appeared for a particular category. For example:

- Entertainment: Movie, Kardashian, New, Kim, Season.
- Business: US, Stock, China, New, Market, Billion.
- Technology: Google, Apple, Samsung, New, Android, Price.
- Medical: Health, New, Study, Virus, Ebola.

- As the word **NEW** appears in all of the categories it shows that it would not be a good predictor.

ANY QUESTIONS ?