QTM2000 Name_____
Case Studies in Business Analytics
Professor Mathaisel
Fall 2021

Exam 2

Individual Assignment

The following pledge must be on your exam cover sheet and signed.

I pledge my honor that I have neither received nor provided unauthorized assistance during the completion of this work. Please Initial:_____

The objective of this project is to predict whether or not news is fake or real based on the text characteristics/attributes in the text. This type of prediction is known in analytics as classification. There are numerous techniques and models for classification, and the project will explore some of these techniques. For this project, you are being asked to write an "Executive Summary" to accomplish that goal using the dataset provided and its accompanying R script. Your summary should be brief, but it must include: an Introduction; Analysis; and Conclusion. The analysis must address three parts: 1. Data "Wrangling" and Visualizations; 2. Naïve Bayes; and, 3. Logistic Regression.

Rubric:

Part	1.	30 points
Part	2.	35 points
Part	3.	35 points
		100 points

Introduction:

Social media has become a channel to pass on information of what's happening. Often people perceive whatever conveyed in social media to be true, but news on social media can be both true and fake, and it can have a significant impact on the people, the government, and the economy. News can influence or shift these impact curves up or down depending on emotions and the political environment. Thus, it is important to identify the fake news from the true news. The authenticity of information has become a longstanding issue affecting business and society, both for printed and digital (social) media. On social networks, the reach and effects of information spread occur at such a fast pace that distorted, inaccurate, or false information can cause real-world impact. The sensationalism of not-so-accurate eye-catching and intriguing headlines aimed at retaining the attention of audiences to sell information has persisted all throughout history. But, the invention of social media outlets, like smart phones, has accentuated the problem.

Can analytics help distinguish between real and fake? Regardless of the context (in this case it is news), the notion of something being real or fake is a classification problem. So, let's use what we have learned about classification models to try to analyze this problem.

In the case at hand, we will perform an analysis on text, which means that we are dealing with Natural Language Processing (NLP). This course has not yet delved into NLP, but that does not mean that we cannot use the techniques that we have learned to analyze fake v. real news in an NLP environment. The only difference between the data that you have used in this course to date and the data for this project is that the data for this project is text, not numbers or categorical data. Text is really categorical data surrounded by other human language. The problem is that text also contains punctuation, "stopwords" ("for and but or so and yet") that are not significant, and other special characters (e.g., https://) that we don't need. So, we will be doing some data cleaning/wrangling to our text before we get to the classification routines. Learn a bit about NLP as you proceed through this project, and enjoy the trip!

Datasets:

There are two datasets: one from real (true) news outlets, like network news; the other from news outlets that are classified as fake news. Both contain text data related to the 2016 U.S. presidential election. This is your first introduction to text data, as opposed to numerical data, and how R handles text data. This brief introduction is your segway into NLP, which is a very interesting, popular, field of study in Data Science.[1]

The project contains two separate datasets of real (true) and fake news. FakenewsNet is a repository for an ongoing data collection project on fake news research at Arizona State University (ASU). The repository consists of a comprehensive dataset of Buzzfeed news and politifact. The FakenewsNet consists of multi-dimension information that not only provides signals for detecting fake news, but can also be used for research, such as understanding fake news propagation and fake news intervention. The Buzzfeed news dataset comprises a complete sample of news published in Facebook from nine news agencies over a week close to the 2016

[1] If you wish to learn more about Natural Language Processing, ask the instructor.

U.S. election from September 19 to 23 and September 26 and 27. Every post and the linked articles were fact-checked, claim-by-claim, by 5 BuzzFeed journalists. There are two datasets of Buzzfeed news: one dataset of fake news and another dataset of real news in the form of csv files.

Data Dictionary:

Both the Real (True) and Fake news datasets consist of the following main features/attributes:

- `title` : It refers to the headline that aims to catch the attention of readers and relates well to the major of the news topic.
- `text` : Text refers to the body of the article, it elaborates the details of news story. Usually there is a major claim which shaped the angle of the publisher and is specifically highlighted and elaborated upon.
- `subject`: It indicates the subject of the news article.
- 'date' : The date of the news.

Objective:

The objective of this project is to predict whether or not news is real (true) or fake based on the characteristics/attributes in the text. The project is divided into two main parts:

(1) Exploratory Data Analysis – to understand our data.

(2) Classification - to build a classification model that can detect fake news. We will use different classification models to classify documents into real/fake news categories.

Part 1. Data Visualization

Briefly and concisely tell a story of what you visualize from the graphs generated by the R script. Use as few words as possible, but convey your story.

Part 2. Naïve Bayes

Summarize the results of the Naïve Bayes model.

a. A brief discussion of what the model is trying to accomplish and how the technique works.
b. Discuss the resulting analysis.
c. An evaluation of the performance of the model.

Part 3. Logistic Regression

Summarize the results of the Logistic model.

a. A brief discussion of what the model is trying to accomplish and how the technique works.
b. Discuss the resulting analysis.
c. An evaluation of the performance of the model.