Slide 1

# Introduction

The dataset has been obtained from [www.kaggle.com](http://www.kaggle.com), originally provided by the London-based company PeerIndex. PeerIndex administers services in social media analytics, including assigning a score to online users to reflect their social capital. Companies are using social media analytics to improve their marketing techniques and to expand their client base, by observing online networking patterns for predictive analytics. Social media platforms, such as Twitter, have allowed many people who do not have conventional celebrity status still be able to spread messages and information to millions of others. Being social media savvy is a skill that not everyone can master; therefore, obtaining data on those who are and that have a lot of influence online can be very beneficial for improving a company’s marketing campaign. Not only is analyzing user influence on Twitter useful for understanding trends, by targeting those that are so-called “internet famous” is a cost-effective approach to advertising because you do not need to invest in the production of a commercial or need to pay a large sum of money to a celebrity ambassador.

The dataset I used contains multiple attributes describing the relationship between pairs of Twitter users, where each relationship is labelled according to a human’s judgement of who in the pair has more influence over their Twitter based social network. This type of data is intended for a classification problem to predict whether or not future users with particular characteristics would be considered influential or not on Twitter. Therefore the goal of my project was to compare the accuracy of two machine learning models: logistic regression and naïve Bayes. To see which one would correctly predict the most labels I used the statistical tool RStudio.

**Posts:** The main activity on Twitter is sending 140 character messages known as “posts” or “tweets”.

**Follower/Following:** When a Twitter user subscribes to another user’s posts it is called “following”. Likewise, if another user subscribes to your posts they are called your “follower”.

**Retweet:** If a Twitter user likes a post by one of the users they follow and they want to share it with their followers, they can repost it and this is called “retweeting”.

**Mentions:** Twitter users have a username that begins with an ‘at’ symbol. They take the format @username. You can use someone’s username to reply to one of their posts by including it in your post. This is called a “mention”.

**Lists:** A Twitter user can follow a group of users, known as a “list”, instead of just following an individual user.

Slide 2

# Literature Review

Here are the two main sources I used. First is “Measuring User Influence in Twitter: The Million Follower Fallacy”, by various researchers. They begin by discussing the role of influence in sociology and viral marketing, based on the theory that there are a minority of people that are more persuasive and these are the people that are the influencers in society. They talk about the value of studying these patterns for the purpose of marketing is that you get a better understanding of trends and innovations, as well as, by targeting the influentials you can use them to advertise at a low-cost but still have a far reach with your message. The specific purpose of this paper was to examine the influence of Twitter users of a variety of topics and their dynamics of time. They did an empirical analysis using the measurements of indegree (number of followers), number of mentions and number of retweets. I will not get into too much detail of their analysis, but some of their important conclusions are that having a high number of followers does not necessarily mean that you will have a high number of mentions and retweets. Also, the number of retweets is reflective of the value of the tweet, while the number of mentions is reflective of the value of the name recognition of the user, i.e. a celebrity would get a lot of mentions, whereas, a well-respected news organization would get a lot of retweets. A final conclusion I found interesting is that some users hold influence by being a respected voice on a variety of topics; I believe that this would be demonstrated by a user’s list count.

The second is “Preference Learning: An Introduction”. I think that this chapter begins by giving a good definition of a preference, which is: “a relaxed constraint which, if necessary, can be violated to a degree”. I like how non-committal this definition is because it mirrors the challenges in machine learning to be able to model and predict human behaviour. This reading describes the goal of preference learning as to build models using empirical data to predict an individual’s or a group of individuals’ preference, where a set of item with known preferences can be used to predict the preferences for a new set of items. Predicting preferences and the total order (ranking) of preferences is frequently used in e-commerce to provide personalized recommendations of products or services.

Slide 3

# Approach

The first step is to download the csv file from [www.kaggle.com](http://www.kaggle.com), then import it into RStudio using the read.csv() command. Then check for missing values and most importantly change the dependent variable from type integer to a factor.

The second step is to explore the dataset. First looking at the summary statistics used to describe the dataset, then to look at the behaviour of the data and visualize some of the attributes to have a more detailed understanding of the dataset.

The third step is to build the logistic regression model. Firstly, by creating the formula where ‘Choice’ is the dependent variable then split the data into 70% for training and 30% for testing. Use the glm() command in RStudio to build the model and evaluate its accuracy using the confusion matrix.

The fourth step is to build the naïve Bayes model, by again separating the data into 70% for training and 30% for testing. Use the naiveBayes() command in RStudio to build the model, then use a confusion matrix to evaluate the accuracy.

The fifth step is to compare the accuracy of the results and discuss which model is more appropriate and why. Then consider what could be improved.

The sixth step is to revisit what the original problem what and to provide concluding remarks. Comment on where or not the goal of the project was met and give insight for future research in social media analytics.

Slide 4

# Dataset

The dataset has 23 attributes and 5500 records. Each record describes the relationship between an individual A and an individual B. There are 11 pre-determined features, measured by PeerIndex, describing each user, they are: follower count, following count, listed count, mentions received, retweets received, mentions sent, retweets sent, posts, network feature 1, network feature 2, and network feature 3. Ignoring the first column, the first 11 attributes describe all the A users and the next 11 describe all the B users. The first column is the label assigned based on which user was judged as more influential. The first column is a categorical variable, where ‘1’ is for user A being more influential and ‘0’ is for user B being more influential, and the rest are of type ratio (integers and real). We can see from the standard deviation that the data has a wide spread and we know from the mean being larger than the median that the data is right skewed due to extreme values, in this case extremely large values.

Slide 5

# Results

The most important thing to take away from the summary statistics, as stated before, is that the data has a wide spread and is skewed right. This is initially indicated by the mean for each attribute being larger than the median and is confirmed by these plots. The density plot is the subset of the first 100 records of ‘A Mentions Received’ where ‘Choice’ equals ‘1’, meaning when the A users are viewed as more influential. We can see that most of the data is found below the value 25 000, while there are some extreme values. In the second plot, the data is also the first 100 records of ‘B Listed Count’ and ‘B Retweets Received’, but the data is split by labels ‘0’ and ‘1’. We can see when user B is considered more influential, we can see that a user’s list count has an exponential effect on the retweets they receive. At first there is a steady increase, the higher the list count the more retweets received and then there is a dramatic increase in retweets when the user is a part of a high number of lists. Another thing I looked at was the medians of each attribute when the data was separated into labels ‘1’ and labels ‘0’. I wanted to observe the difference between actions that were done by a user’s account and to a user’s account. As assumed, when the median action on to a user’s account was very high they were considered more influential; whereas, the median of their passive counterpart was less high. Here we have two examples of this. The first is the median retweets received and median retweets sent for users A and B when user A is considered more influential. There is a large difference between the retweets received by user A and by user B (in the first column) and the retweets sent by user A and by user B (in the second column).

Slide 6

# Results

In blue, we have the results from the logistic regression model. We can see that there are four attributes that are not statistically significant: A posts, B following count, B retweets received, and B posts. When we compare the accuracy, we have that label ‘1’ was correctly predicted 543 times and label ‘0’ was correctly predicted 650, which means that about 72% were correctly predicted. In yellow, we have the results of the naïve Bayes model, where label ‘1’ was correctly labelled 756 times and label ‘0’ was correctly labelled 129 times, which means that only about 53% of labels were correctly predicted. Based on these results we can say that the logistic regression model is more accurate when predicting labels. To see if we could improve the logistic regression model and prevent over-fitting the data, I removed the attributes “A posts” and “B posts” because they were the common none statistically significant features between A users and B users. The results improved to about 73% of labels being predicted correctly; however, now the following count for both A users and B users was considered not statistically significant. Therefore, I tried one more iteration by removing “A Following Count” and “B Following Count” but the accuracy did not change. Finally, I wanted to see if removing these attributes from the naïve Bayes model would improve it because the naïve Bayes algorithm is sensitive to complexity. It did improve the model, giving 65% of the labels being correctly predicted. While this is a decent improvement, the logistic regression model still preforms better.

Slide 7

# Conclusions and Discussion

The purpose of this project was to compare how a logistic regression and a naïve Bayes model preformed when predicting the labels of the given Twitter dataset. In regards to the original research question, the goals of the project were met. After executing the algorithms, we can see that the logistic regression model out-performed the naïve Bayes model with by correctly predicting 19% more of labels.

We know from looking at the mean and standard deviation we noticed that the data has a large spread and is right skewed by extreme values, which leaves the data vulnerable to over-fitting. Then when trying to improve our results we notice that by removing “A posts” and “B posts” from the logistic regression model we got 2% more accuracy. However, when we removed the variables “A following count” and “B following count” the results did not change significantly. To continue the comparison, we removed the same four variables from our naïve Bayes model and there was an 11% increase in accuracy. This is not surprising because the naïve Bayes algorithm is better for less complex data.

Our analysis is good starting point for marketing research and social media analytics because the data’s behaviour is consistent with the dynamics of Twitter in real life. One of the downsides of our dataset is that we did not curate it; therefore, there was a lack of understanding of all the variables. We had no understanding of how the “network features” were measured. Moreover, the attributes related to posts, retweets and mentions that were expressed as real numbers could have also been hiding important information. For example, if they were representing the rate at which users did those actions over a specific time period, having that information could have led to a wider range of analysis. Based on our results, we would recommend that future projects of this nature consider more closely the measurements of network interaction and the rates at which relevant Twitter actions are performed.