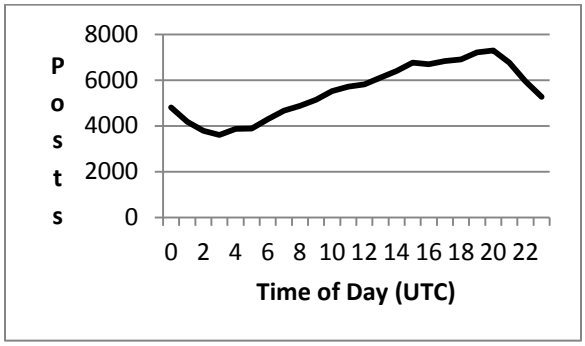
Matthew Hale
A10942651

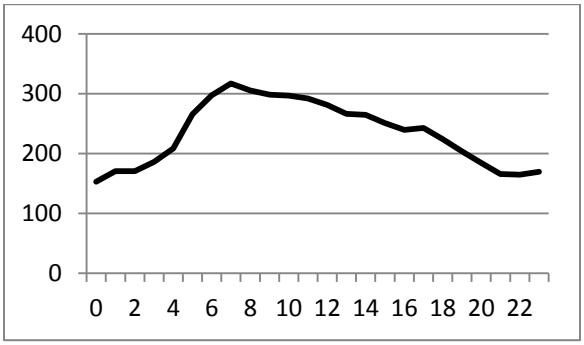
The dataset to be analyzed is the dataset of Reddit submissions suggested in lecture. (http://snap.stanford.edu/data/web-Reddit.html). In this dataset, there are 132,308 entries for analysis. Of these entries, however, there are only 16,736 unique images, and on average, each was resubmitted approximately 7.9 times. The timespan across which these Reddit submissions were made is July 2008 to January 2013. This dataset will be used to yield any possible information concerning the relation between a post's success, the number of posts concurrently being submitted to Reddit, and the time of day. Specifically, which feature makes a better predictor, in comparison to other plausible prediction features. Additionally, another goal is to discover a correlation between number of hourly posts and average score of all posts within the hour. Below is a graph plotting the amount of posts for each hour of the day, according to the datapoints. This chart will be referenced in greater detail later.

Figure 1: Distribution of Posts by Hour



This graph shows a clear relation between the time of day and the activity within Reddit. If there exists a relation between the frequency of concurrent posting and the success of a given post, then the above graph implies that there exists an optimal posting time for Reddit . Unfortunately, this graph alone does not provide many answers. One could hypothesize that the best time to post could be 20:00 UTC, because that hour experiences the greatest amount of posts, and therefore there must be a large number of users browsing Reddit. However, this hypothesis assumes that the given post's score will not be drowned out by the myriad of other posts being made. Then, the only two other plausible times according to Figure 1: The optimal time could be on the local minimum with the least amount of other competing posts being made, or the point of inflection, where the post frequency is most rapidly increasing. This next graph provides some answers. The direct relation between score average and time of day:

Figure 2: Average Score by Hour

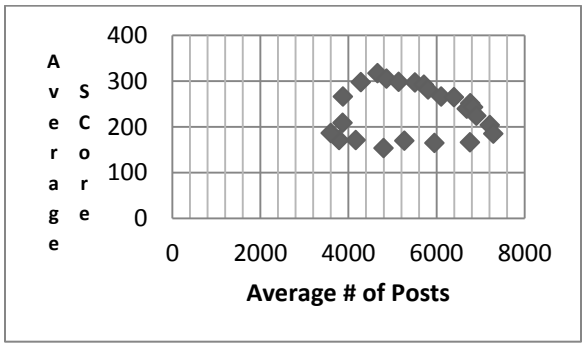


There are two things that this graph shows. It shows that the best time to post content would presumably be roughly 07:00 UTC, which according to Figure 1 is a time of low but increasing post frequency. It also shows that there may exist a relation between score and concurrent post frequency. Both graphs demonstrate a clear correspondence between Reddit activity and time of day, mainly due to their smooth curves which indicate stable but clear changes throughout the day.

As previously mentioned, the predictive task at hand is to predict a post's success based on the time of day in which it was posted, as well as based on how many other posted are

currently being made, and compare predictions. The predictions from the models for each feature (time of day vs. hourly post frequency) are to be analyzed for any similarities. If the prediction sets turn out to be similar, then there is reason to believe that there exists some sort of causality between time of day and Reddit traffic, and between Reddit traffic and the score of a post. This would provide an answer to the secondary goal of discovering a correlation.

Figure 3: Avg. Number of Posts made vs. Avg. score of posts, for each hour



As Figure 3 shows, there is no immediately clear trend between how much users post to Reddit and how highly each post scores. There are no signs of low variance or clusters for dimensionality reduction into metadata. This means that K-means clustering and Principal Component analysis cannot be used to help find a more efficient way of solving the predictive task. Judging by the crude sinusoidal patterns of Figures 1 and 2, there may be some sort of daily cycle in which post scores increase when post frequency decreases, and vice versa. The following predictive task aims to find more answers as to how these parameters affect one another.

Two least-squares regression models must be used to evaluate the impact of each feature, and the results of each model shall be compared with one another. Both of these models, however, will attempt to predict a post's score as a function of time. Instead of only predicting whether or not posts will be more successful before or after a given time, it will instead predict a post's popularity based on when it was posted throughout the course of the day, and calculate the significance of the post being made during a time of frequent posting. This rules out classification models, such as support vector machines and naïve Bayes. The only instance in models such as those could be used is if there was some element of classification involved in the predictive task, such as there being a "threshold" that lies between successful and unsuccessful posts. Hence, least-squares regression is the most viable model candidate. If a model has a lower MSE than the standalone baseline model, then it can be deemed more accurate, and in turn the feature with which it trained its predictor from is at least somewhat impactful in a post's score. As previously mentioned, all of the models will implement least-squares regression, and various training sets will be experimented with as well for each of the two main models. The MSEs of each of the baselines will be compared with the MSEs of the two main models.

One training set will contain only submissions of content that is among the top 8,371 highest scoring content, resulting in a training set that is approximate to half the total amount of unique content (16,736). Additionally, with an average repost rate of 7.9, we can expect the training set size to be roughly 66,130, which is very close to half the size of the entire set (66,154), and is therefore an appropriately sized training set. The reason this particular training set is selected is because the higher scores in the training data will guarantee larger data values are being plugged into the

regression equation. Posts that consistently see thousands of upvotes will have more pronounced fluctuations in score than do posts with only a couple dozen. The main drawback is the bias presented from training using only posts that have consistently scored highly.

Another training set will consist of the four highest scoring posts for each individual image, resulting in a training set of size 66,944, which is close to half the size of the entire dataset (66,154) and therefore makes for an appropriately sized training set. The reason for this set being included is to compare each post's best moments, and train our predictor based on the maximum potential of each post. The main drawback of using this training set is its lack of coverage across all hours of the day. According to the above chart, most of the elements in this training set can be expected to be between 06:00 and 08:00 UTC. The reduced range of the training data may lead to a skewed regression equation.

The third and final training set for the models model will simply pick a random training subset half the size of the total dataset. The randomness of this approach helps eliminate possible sources of bias that the other two models may have. The drawback of this set is that not all posts have been reposted the same amount of times, causing there to be potential for some outlier content that gets reposted hundreds of times to produce skew in the final result, assuming that more recycled content depreciates in score.

The standalone baseline model, also referred to as the control baseline, will be a basic regression model that will predict a post's score based on the average of the scores of its past submissions. This model serves as the most naïve method of prediction, and should be overtaken by all other models, provided the other models' prediction methods are decent enough. There will be two additional baseline models apart from the main models and control baseline which predict scores using features unrelated to the time of day. One will be a repost frequency feature model will take a training set at random of half the size of the total set. However, each element will include an additional parameter which represents how many times the image has previously been posted, according to the dataset. This aim of this parameter is to penalize posting stale content, and is to be used in conjugation with the other two baselines. The other baseline will use subreddit size as its predictor. Additionally, the Jaccard similarity of the top scoring predictions from the time of day model and the frequency model will be calculated, which may present a correlation that is sought after in the secondary goal of the study mentioned in the introduction. Although Jaccard similarities are more commonly used for recommender systems, a high Jaccard similarity would indicate that the if one datum is predicted to have a higher score based on the time of day it was posted, it will likely be predicted to have a high score based on the amount of other posts being made when the datum was submitted. Therefore, the frequency of posts would be directly correlated to the average score of the same posts. If both main models are shown to predict more accurately than the control, then an additional model using time of day paired with post frequency shall be developed and tested. Specifically, the additional model will be a combination of the two main models: One for time of day, and another for amount of concurrent posting.

Reddit has been the subject of multiple studies, presumably due to its accessibility and anonymity. The dataset used in this study is the same one that was originally gathered for an older Reddit case study conducted by Professor

McCauley and his colleagues at Stanford University[1]. The original study also aimed to predict the score of a Reddit post using linear regression. However, the original study used a wider variety of models and features. One such model examined the score differential of posts across different subreddit communities, and predicted what each community would score a given post. Another examined the language within the post title to predict a post's score. Using various predictive components, including the aforementioned models, the final predictions concluded that one could not predict the success of a post on its title alone. There are multiple factors at play, and the best title depends on the subreddit the post is being made in, the amount of times it gets posted, how recently it was last posted, and the image itself, among other possible factors. The prediction methods used in this study also consisted of linear regression, as well as other smaller methods used to make up the terms in the custom regression equation. Some of the other methods used were Jaccard's similarity, hierarchial classification, topic models, and other specialty models. Instead of attempting to mimic such a complex equation in this particular study, deriving predictions using a least-squares equation will suffice. No ideas were borrowed from this study's model for the sake of simplicity. Although there are many other studies that investigate success on social media predicted by lexical analysis of posts, few analyze the time of day at which the posts were made. However, one such study conducted by Lerman[2] studied the relationship between 'visibility' and 'interestingness' on content on Digg and Twitter. When one considers how the amount of posts being made is related to the 'visibility' of each post, a connection between this study and Lerman's can be made.

After testing the previously mentioned models, it quickly became clear that a linear regression model would not perform as well as a curved fit model. Thus the sine regression equations

$$y = \theta \sin(X)$$

$$y = \beta_1 \sin(X_{time}) + \beta_2 \cos(X_{freq})$$

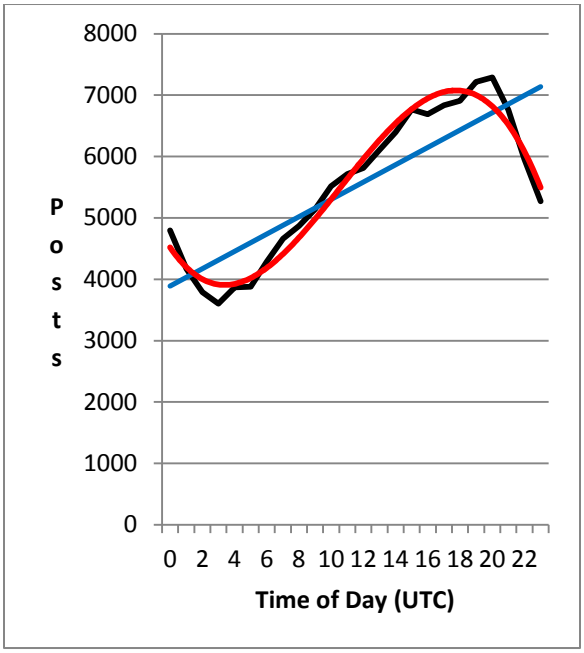
are better fits than least-squares linear regression equations, given the roughly sinusoidal pattern of the above data plotted over time. These equations was made to look as similar as possible to the more well-known linear regression, for clerical purposes. In the single parameter equation, theta still represents the unknowns vector, and X and y still represent features and labels, respectively. For the second equation, cosine was used to represent concurrent post frequency in hopes of being a better fit, as Figure 1 suggests. These equations were implemented in Python by being plugged in as function parameters to the scipy library's curve_fit function. These equations were found to dramatically reduce the mean squared error across all models models (except the baselines).

Feature X	Control MSE	Linear Reg MSE	Sine Reg MSE
X = Time of Day	215,103	214,145	108,302
X = Post Frequency	215,103	217,584	213,013
X = [Time of day, Post Frequency]	215,103	215,506	109,774
X = Number of times reposted	215,103	215,201	-
X = Relative Subreddit Size	215,103	221,179	-

Matthew Hale
A10942651

The above chart lists the MSEs for each model. For the top three models, the training sets were selected at random, as it ultimately proved to be the most accurate representation of training data based on lowest residual. The top 8,371 set performed second best, although its MSE was roughly 10% higher than the random set. The best of 4 set was the worst training set, with an MSE over 240,000 for the linear regression.

These feature representations could be improved by adding components similar to those used in Professor McCauley's original case study. Regardless, the best performing model was the time of day model.



An illustration of a linear trendline vs. a sinusoidal trendline.

One such reason the time of day model outperformed the frequency model is because the amount of posting traffic on Reddit is likely related to the time of day. Whereas the time of day model predicted with discrete hour values from 0 to 23 with 1-hour increments, the frequency predicted using feature values which fluctuated much less discretely over time, ranging from 4,000 to 7,000 with variable increments, adding an extra degree of indirectness to the frequency model.

As previously stated, the sine regression model provided the best predictions, confirming the notion of a period or cycle that repeats itself daily which was first suggested by the data in Figure 3. The model also confirms an clear association with post score and time of day, and a much more loose association between post score and amount of other posts being submitted. The sets of the top 10,000 posts predicted by the time of day model and the frequency model were compared by deriving the sets' Jaccard similarity. The aim of this task was previously mentioned as to determine if more posting hurts the average score of the posts. However, The Jaccard similarity between the two sets was:

$$P = \frac{ToD \cap F}{ToD \cup F} = 0.19423$$

This weak similarity between the predicted sets implies that there is little correlation between the two features, confirming that no such correlation exists, as was suspected by the secondary goal in the introduction. In conclusion, these calculations show that posts made during high-volume hours of the day are not necessarily expected to score lower. Additionally, both time of day and overall post frequency are valid predictors of a post's score, although time of day is much more accurate. In contrast, repost count and subreddit sizes are poor predictors for a post's score. However, due to the large time differential between training time and test time for the subreddit size model, another dataset could be gathered for more accurate prediction training.

Matthew Hale
A10942651

Works Cited

1. Lakkaraju, Himabindu, Julian J. McAuley, and Jure Leskovec. "What's in a Name? Understanding the Interplay between Titles, Content, and Communities in Social Media." *ICWSM*. 2013.
2. Lerman, Kristina, and Rumi Ghosh. "Information Contagion: An Empirical Study of the Spread of News on Digg and Twitter Social Networks." *ICWSM* 10 (2010): 90-97.