CSE 190
Assignment 2

Phat Huynh
A11733590

Nicholas Gibson
A11169423

1) Identify dataset

Reddit data. This dataset is chosen to study because as active users on Reddit, we'd like to know how a post become successful based on many categories such as: power users, posting time, titles of the post, certain subreddits (community), etc. We can apply this study not only for Reddit, but also other social networks with similar dataset such as Youtube, Facebook, Twitter, etc.

Reddit is a social networking, entertainment, and news website with a unique setting. Users can submit their contents in the form of texts or direct links. As of 2005, Reddit has 169 million unique visitors. The largest demographic are males from 18-29 years old. Registered users have the option to upvote or downvote the post, and this will determine the "karma" (final score) of the submission. This will determine the popularity of the post because when content receives too many downvotes in the first few minutes after being posted, it will not appear in the front page of Reddit (the top posts are on the front page and are visible to many users) and will be forever lost within other thousands of submissions. Most Reddit users are from the US. However, locations where submissions originated from do not play a big role on the success of the post. Instead, the amount of karma a user has dictates the popularity of submissions. Also, title length seems to play a role in the success of the post. Number of comments in each post is also important.

Here is some information about our dataset:

This data set is collected from http://snap.stanford.edu/data. The data was collected from July 2008 to January 2013.

There are 132,308 total submissions and 63,335 unique users. Only 16,736 of are unique images, which means the majority of the submissions are reposts. Images are reposted an average of 7.9 times.

Data fields in this dataset include:

- image id
- time of the submission
- title of submission
- total number of votes on this submission
- reddit user id of submission poster
- number of upvotes
- subreddit submission was posted to
- number of downvotes
- local time of the submission (can calculate location based on time difference)
- submission score
- number of comments
- reddit username of submission poster

2) Predictive task

With this dataset, we are able to research the submissions of images. As the main goal of our predictive task we want to predict the score of Reddit submissions. We decided we wanted to user link karma to help train our predictor. This information was not included in our data set, so we used the Reddit API to find the link karma of all unique users in the dataset. Out of the 63,335 unique users, 5,432 have deleted their account since the data was collected. For training our predictor we decided to only use accounts that have not been deleted. Out of our 132,308 total submissions, 32,573 are from users with deleted accounts. This left us with 99,735 submissions. We then split our data into training and test sets, 89,735 submissions in our training set and 10,000 submissions in our test set. To test our results we will calculate the mean squared error, absolute error, and fraction of variance unexplained on our test set using the predictor we will train with the training set.

We have several different hypotheses for this data. Our first hypothesis is that there is a positive correlation between user link karma and submission score. We will use linear regression to test this hypothesis, and this will be our baseline. Power users (users with high karma) know what makes a post popular, and are likely to get a high submission score. Conversely, users with low karma are inexperienced, do not know how to make popular submissions, and will receive low submission scores. Some high karma users receive celebrity status on Reddit, so their posts will likely receive a good score. Many users will look at the name of the original poster and notice that he or she is popular, assume that the post will be good, and proceed to upvote first without even read the content.

Our second hypothesis is that there exists a sweet spot for the title length of Reddit posts. This hypothesis was formed because we noticed that in our dataset, there are a variety of different lengths for all submission tittles, and they are all perform contrarily. A conclusion is drawn that titles that are too short are not informative, so they should receive a lower score. Titles that are too long take too long to read, and should also receive lower scores.

Our third and final hypothesis is that the submission score is positively correlated with the number of comments the submission received. On Reddit, the higher the post score the more prominently it is shown on the website. The top scoring submissions are shown on the front page, whereas lower scoring submissions are shown several pages after the front or there's a good chance that many will not see it. Another reason this may be true is that people are more inclined to post on popular submissions than less popular submissions, or that post has a really good content that catches people's attention quickly and engages them in a discussion. However, this last theory may not play a very big role because as inspected, many images are being reposted over time and only a small amount of them get a good score.

3) Literature

Himabindu Lakkaraju, Julian McAuley, and Jure Leskovec introduced the Reddit dataset to us in their study: *"What's in a name? Understanding the Interplay between Titles, Content, and Communities in Social Media."*

Lakkaraju, McAuley, and Leskovec mentions in their study that social network popularity problem has been researched many times in the past couple years. For example, predict the future success of a video on Youtube using its early view count done by Szabo and Huberman in 2010. Other authors like Artzi, Pantel, and Gamon used the language model to predict the popularity of Twitter posts.

A couple main tasks that Lakkaraju, McAuley, and Leskovec perform in their research:

- Developed a statistical model on four aspects: content of the submission, submission title, the community where the submission is posted, and the time when the submission is posted
- Two models: Community Model and Language Model
- Community Model: a submission's success is not related to its title, it depends on the subreddit (community) and the time of day of the submission: popular community and busy time will gain upvotes
- Language Model: the impact that the title has on the submission's success. Certain community favors some words over others. This is known as "good words" and "bad words"
- Their conclusion: it's hard to tell the success of a post using the content, title, community, and time. Therefore, need to develop models that separate these factors to study each feature.

How this study inspires us:

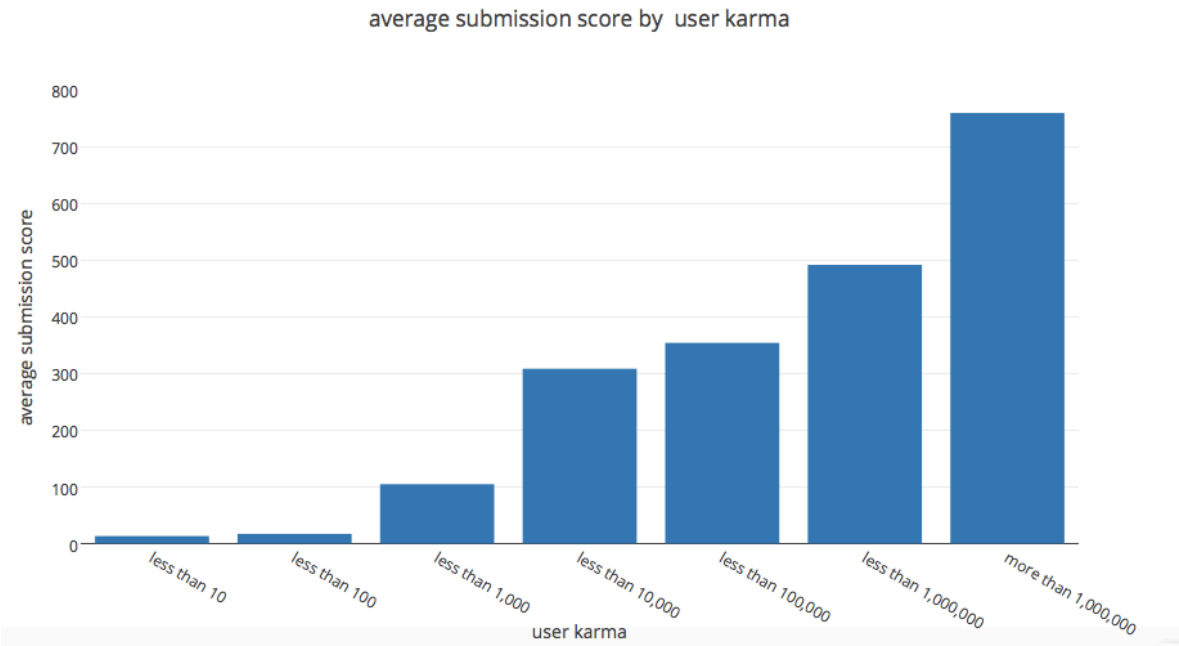
The four aspects of a successful submission: *content of the submission, submission title, the community where the submission is posted, and the time when the submission is posted* lead us to believe that users who have these good aspects practiced in their posts will receive higher scores. We decided to use different models than the ones presented because we'd like to find predicted tasks that aren't mentioned to possibly expand and improve this study. We draw a conclusion that when a user have more karma (positive score), the more likely they will get their content to the front page. We will then actively try to improve the result of the first linear regression model using users' karma. Next, instead of using good and bad words model like Lakkaraju, McAuley, and Leskovec, we inspect the title length and try to find the sweet spot that can gain a good score. The title length should not be too long or too short. And finally, we add comment as a feature of our linear regression predictor. Our purpose is not trying to mimic what has already been done before; we are using previous study as an inspiration to implement our own model to discover interesting finding

The conclusion from this work indeed supports and inspires our ideas for our study. However, since we're trying to implement different models, the results are not the same. The final result will be explained in part 4 below.

4) Results

Initially we tried using linear regression to calculate the submission score based on user karma. This gave us a weight of only 0.000117. We sorted the users by karma, and noticed that there is a big discrepancy between users. 38,274 users have more than 100,000 karma, and 19,620 users have less than 1000 karma. Our hypothesis is that users with very low karma will receive a very low submission score, and users with very high karma will receive a very high submission score. We decided to see how the linear regression predictor would perform as our baseline. After calculating the MSE and variance, we calculated the FVU to be 0.9948, which can be further improved. We realized that linear regression is not a perfect choice of predictor to use to test our hypothesis, so we decided to make a custom predictor.

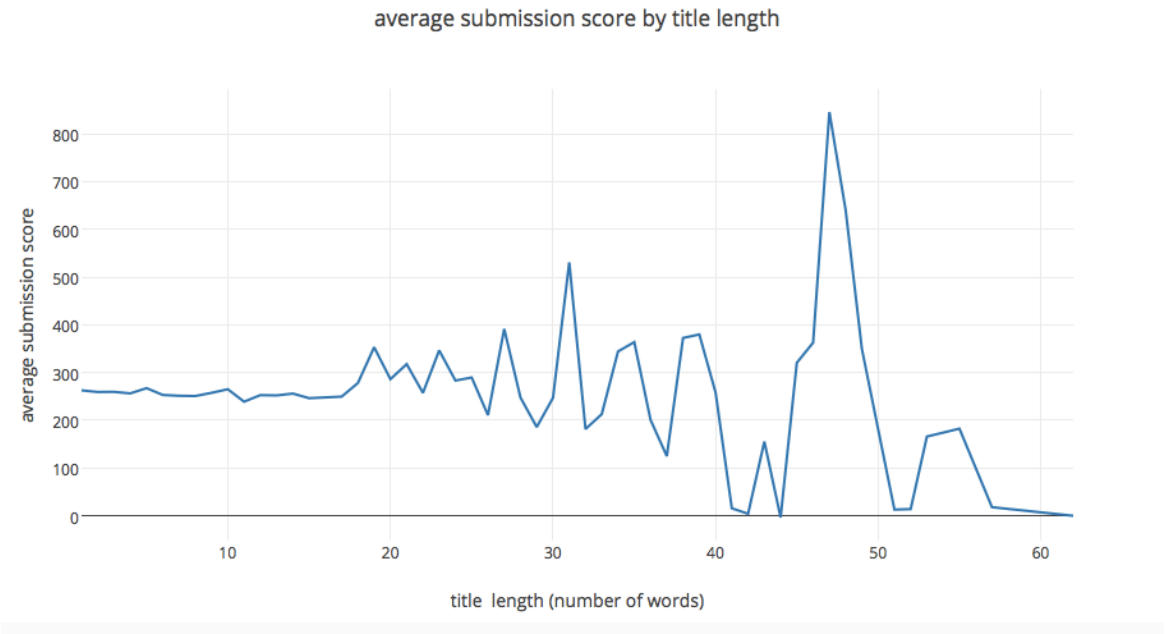
For our custom predictor, we calculated the average scores of users with less than 10 karma, 100 karma, etc., all the way up to 1,000,000 karma. The results confirmed our hypothesis; users with very low or very high karma are outliers in the data.



Our custom predictor showed promising results. It has an FVU of 0.9360, which is much better than the FVU of our linear regression predictor baseline. We decided to combine our custom predictor with linear regression, in order to more easily combine our other

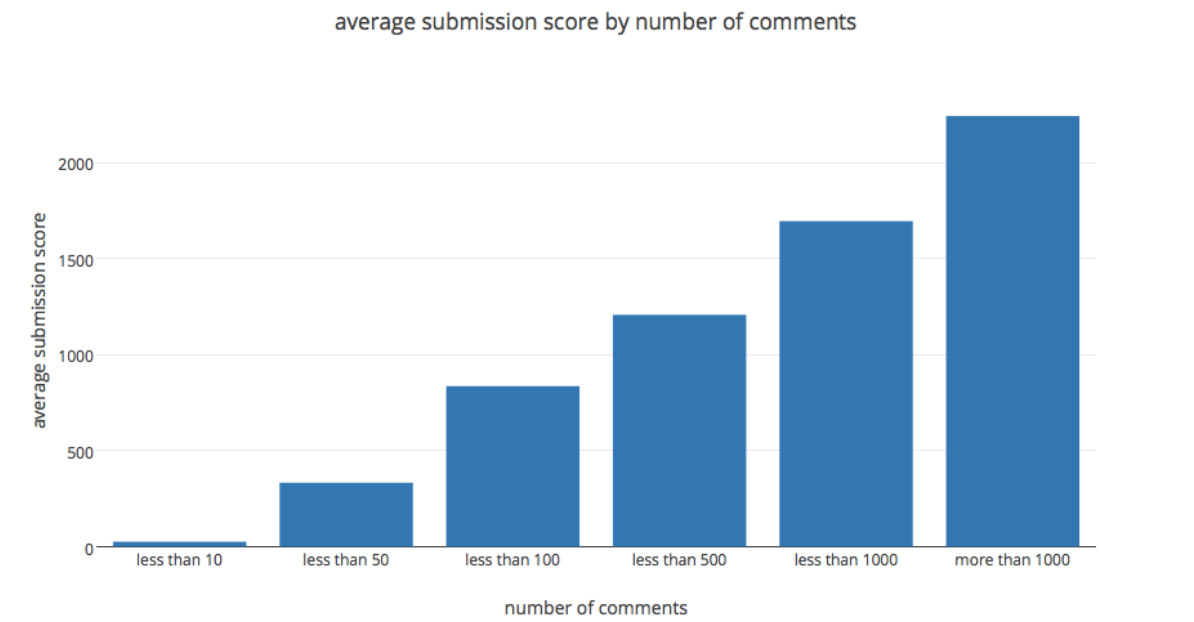
hypotheses. The FVU of this hybrid predictor was very similar to our first custom predictor.

The next technique we tried was finding a sweet spot for title length. Unfortunately, we discovered right away that our hypothesis was wrong, and there is no sweet spot for title length. As the below graph shows, the data clearly does not support our hypothesis.



This graph shows the average submission score based on title length. As the title length increases, the number of posts with that length decreases. That is why there are interesting outliers in the graph. Our findings conflict with Lakkaraju, McAuley, and Leskovec. Lakkaraju, McAuley et al. had conclusions similar to our hypothesis, but our findings indicate that there is no significant correlation between title length and submission score.

Adding number of comments in submission as a feature to our linear regression predictor had amazing results. Adding this feature improved our FVU to 0.5322. This is a large improvement from the baseline.



As you can see from the graph, the submission score is highly correlated with the number of comments. This correlation makes sense, as posts with higher scores are shown more prominently on Reddit, and the more people that see a post the more comments the post will have. McAuley et al. had similar results with their predictive analysis research. They used the number of comments to measure engagement, and found the number of comments was positively correlated with the submission score.

Conclusion

Our research yields a positive result on our first and most important hypothesis: a successful post on a social media site relies heavily on the popularity of the creator. In this study on Reddit dataset specifically, the user's popularity is measured by the amount of karma they have. We develop a model that confirms this hypothesis. Unfortunately, our strategy to find a sweet spot for title length does not work out as expected. The result is still recorded in the graph above. Finally, the amount of comments corresponds directly to the popularity of that post. All the models we built help us to predict whether a post made by a specific user and the comments that it has will be successful.

References

"40 Amazing Reddit Statistics (May 2015)." DMR. N.p., 26 Feb. 2014. Web. 31 May 2015.

H. Lakkaraju, J. J. McAuley, J. Leskovec. What's in a name? Understanding the interplay between titles, content, and communities in social media. ICWSM, 2013.