The model was trained using a Random Forest Regressor, fit to the categories: number of comments, number of views, picture age, taken on date, and author id. This data was split, 70% training and 30% testing, and scores were 0.94 and 0.66 respectively.



Figure : Distribution of Posts by Region

Figure 1 to the left depicts a pie chart of the distribution of posts by region. This was taken from the “region” field of the original dataset. It can be seen that users of the site are relatively evenly distributed throughout all regions, meaning there is no noticeable trend in the amount of posts by region.



Figure : Distribution of Posts by Category

Figure 2, right, is a pie chart of the distribution of posts by category constructed from the “etitle” category in the dataset. From figure two, it is clear that ELEC, and DSL categories hold about half of the posts on the site.

The histogram in Figure 3 depicts the number of comments per post. The x-axis represents the number of comments, the y-axis represents the number of posts with that number of comments. While the graph seems to represent an exponential decay, since the x-axis is on a log scale, the data more closely resembles a negative linear trend.

Figure 4 is a histogram of the distribution of views per post. This graph follows a similar trend as that in Figure 3, with a local minimum in the minimum number of views bin, which will be discussed later in this report.

Figure 5 is a histogram of the number of upvotes per post, with number of votes on the x-axis and frequency on the y-axis. The distribution is a skewed right but normally distribution and centered at about 75 upvotes.

Shown in Figure 6, is a line graph that represents the activity of votes, views, comments, and posts with respect to time. Both posts and views have both gone through downturns in recent years, which may give insight into the local minimum in Figure 4. The posts with the lowest views are most likely the newer posts, because the older the post, the more time it has to accrue views. Given the downturn in recent posts, it stands to reason that this may be a cause for the local minimum.

The variance of the model from the actual number of votes a post received is depicted in the histogram in Figure 7. The distribution is centered about 0, which should be the case for an accurate model, and the data is also normally distributed.

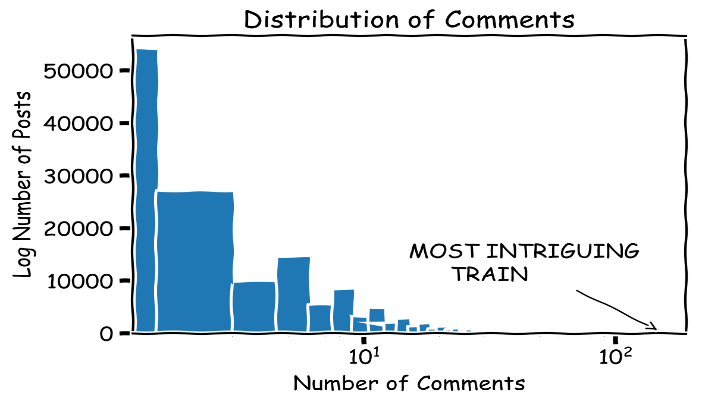


Figure : Distribution of comments

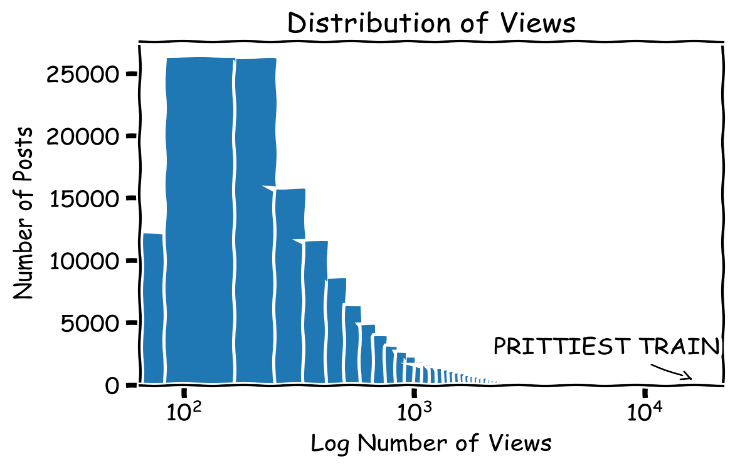


Figure 4: Distribution of Votes

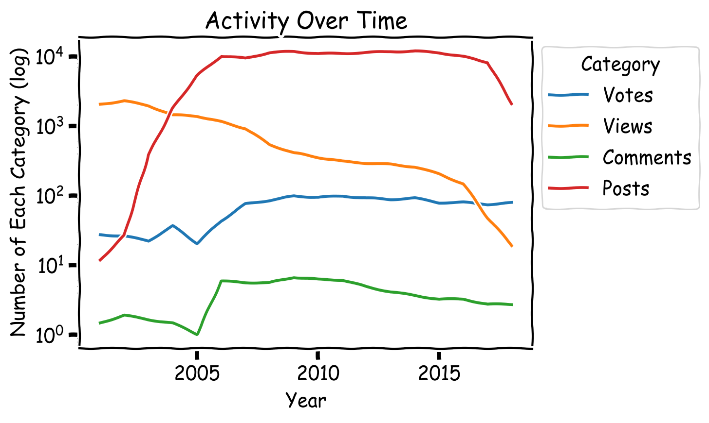


Figure : Site Activity Over Time

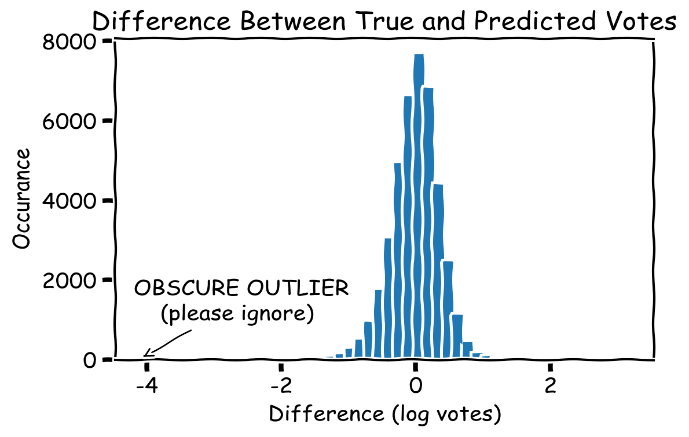


Figure : Difference Between Predicted and Actual Values

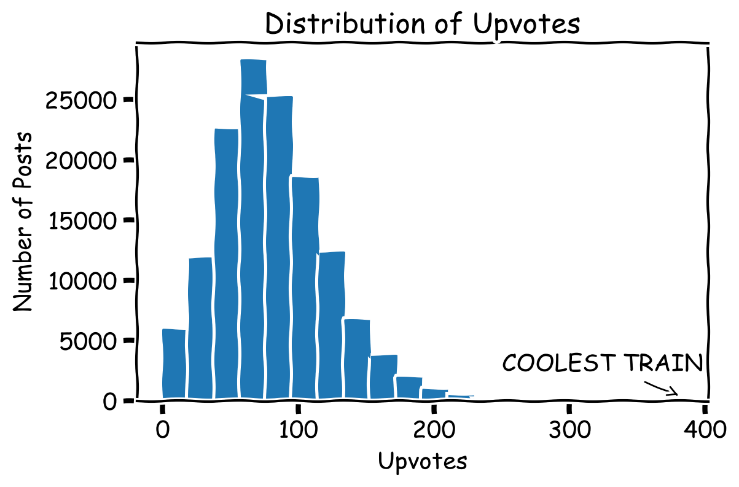


Figure 5: Distribution of Upvotes