Daniel Gysi

Dr. Hallenbeck

CIS-241

December 5, 2020

Minecraft Data Mining Explanatory Report

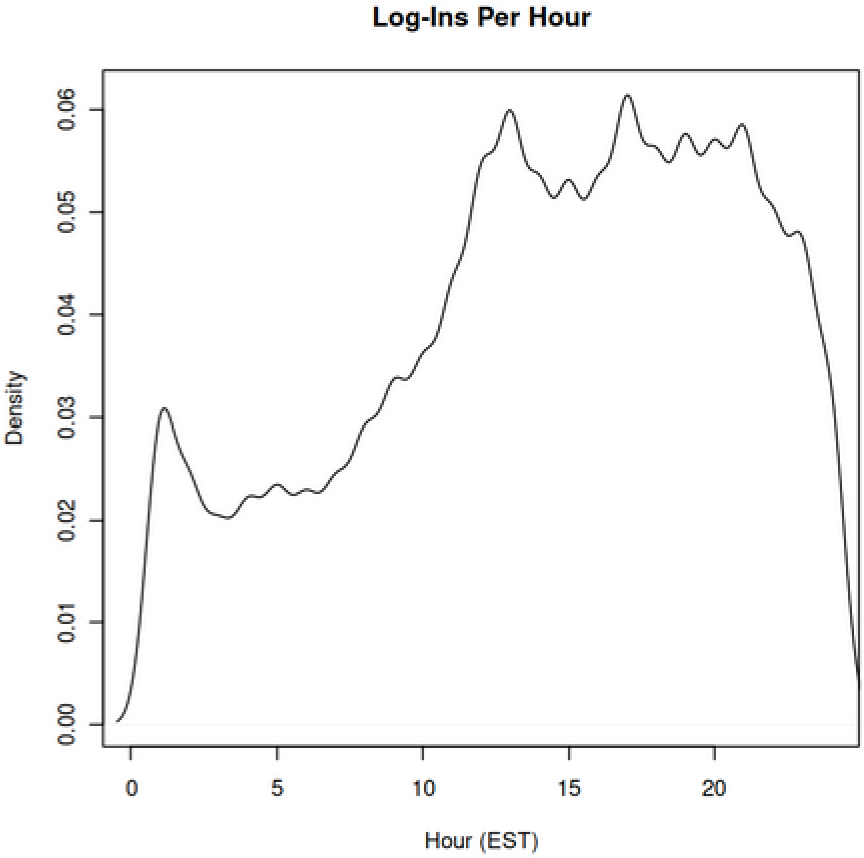
For our data set, we propose the following questions to be answered with explanatory data modeling. Can we predict the type of a block that was placed based on its location and user? Can we predict the type of a block that was broken based on its location and user and is it more accurate than placement? Can we predict whether a session action was a log-in or log-out? Given a user that logged onto the server, can we predict what hour they logged in?

Analysis

First, we look at whether we can predict what type of block was placed by a given user at a given location. This is useful because it indicates a potential pattern between these variables that may reveal more important information. Since we want to predict a categorical variable and we have substantial data but not substantial time, we use a naïve Bayes classifier. We trained the classifier with a training set consisting of 80% of our data randomly selected and tested it with a testing set composed of the remaining 20%. It takes every variable from the block placement data frame as input excluding the block type which it attempts to predict. There are 136 possible block types in the data set and our model is 19.5% accurate, meaning that it chooses the correct block type 19.5% of the time. Calculating other measures of accuracy is difficult here because of the large number of possible levels. However, other measures are not necessary because the model is of little practical utility anyway. 20% accuracy on 136 possibilities indicates that there is a pattern in block placement based on user and location which may be further identified through other means. We would expect one such pattern to be based on environmental blocks. For example, we expect endstone to be placed in the end dimension and by players who have made it to the end. The same goes for netherrack in the nether dimension. Perhaps we might expect certain wood types to be placed more in the areas where trees of that type grow. Further research is needed. To make a prediction on nonexistent data, we suppose that user SalC1 placed a block at (200, 128, 350) in the overworld at 9:50 PM EST on April 13, 2020. Our model predicts that this block is stationary lava. This makes sense as liquids are highly prevalent in our block placement data, and we would expect that a block placed at such a high height so near to spawn might be part of a "lava-cast," a structure created by pouring water on top of lava to create large cobblestone monoliths. We can only be 19.5% sure that this prediction is correct, but that is certainly better odds than guessing randomly.

Next, we apply the same question and methodology to block breaking. We would expect higher accuracy here due to the natural patterns of block generation (e.g. stone generates underground, wood and dirt near the surface, etc.). Sure enough, training our naïve Bayes classifier in the same fashion with block break data results in an accuracy of 31%, so we correctly guess the right block out of 136 possibilities 31% of the time. Again, the complexity of the prediction output makes it difficult to calculate and make sense of other accuracy measures, but this ~30% accuracy is enough to conclude that there is a clear pattern in the type of block broken based on other factors and that this pattern is easier to distinguish than the analogous pattern in block placements. When we test the model using the same manufactured data as the block placement test above, we find that the model selects lava. This makes sense for the same reasons as above. Testing the model on a more conventional break at (200, 35, 350) at the same time and date but by the user CactusDuper results in stone, as would be expected at that height in the overworld.

For our third question, we wonder if we can predict what hour a player will log onto the server based solely on that player’s username. This concept or similar concepts may be helpful to server operators attempting to predict and account for player activity or the activity of specific players that may have a high impact. Since we are predicting categorical data from categorical data with the smallest possible number of variables, we again start off with a naïve Bayes classifier. We also convert the time data into the hour of day to use for the model. We attempted to train the model on 80% of the log-in data as before so we could test it on 20%. Unfortunately, this failed, and the model did not produce any predictions at all. Hypothesizing that this may be due to the training set and testing set containing different usernames, we instead tried training the model on all the data from the log-in data frame, yet it still failed to create any predictions. Then, hypothesizing that this may be due to the complexity of the problem and with few other options, we attempted to train numerous neural networks with varying number of hidden layers, but the training process failed every time. Our failure is likely due to a flaw in methodology. Perhaps the data was not sufficient to produce meaningful patterns, or perhaps there was an issue with the data types or the small number of variables. Whatever the case, it was not feasible to work on this problem any longer. This leaves us with very few conclusions about the predictability of users’ log-in hours. It’s little consolation, but in the process, we were able to use some of the time conversion methods to create exploratory data on the number of logins per hour of day shown below.



Our final question is whether a model can predict whether a session action was a login or log-out action with greater than 50% accuracy. This would indicate that there is a perceivable difference between the activity of players when they’re logging in versus logging out which may be able to reveal more useful information. Since we’re predicting categorical data, we again used a naïve Bayes classifier. We used training and testing sets consisting of 80% and 20% of the total session data respectively and used all non-action variables as input to predict action. The resulting model was ~50% accurate. This indicates that there is no clear difference in the activity of players when they’re logging in versus logging out, at least as evidenced by the types of data recorded in these logs. Perhaps this is expected as for all but the first log-in and final log-out from a player, data has an identical partner of the opposite action. Further research may then pertain to whether there’s a difference between a player’s final log-out and any other log-outs they make, or whether the time that a player logs out is predictable. Information on either of these may prove useful to server operators trying to predict player activity.

Conclusions

This data set was perhaps not the best for the purpose of explanatory modeling. We found that there are clear and separate patterns in the type of block placed and the type of block broken based on location, user, and time. We failed to find whether there is a clear pattern between user and log-in hour. We found that there is no clear pattern between location, time, and user and session action (log-in or log-out). Our conclusions are not totally useful on their own, and perhaps this is why there appears to be no other research on explanatory modeling of this type of Minecraft data.