PILOT Replication

October 2020

1 Introduction

To compare the results of the PILOT study and our replication, it is important understand the key similarities and differences between the both.

1.1 A few facts that are easy to deduce

1. Easier passwords are recovered faster through random guessing
2. RockYou dataset
   (a) **Our Paper:** 2.5 milion unique passwords
   (b) **PILOT:** 2.7 million unique passwords
3. **Lamondre:** It is the only password that we can recreate using the digraphs available to us
   (a) The random number of guesses required to get *lamondre* using our dataframe is 55,110 and the PILOT reports the number as 397,213. If I were to assume the best case scenario, i.e, they used 1296 digraphs for their model, then the metric $(randomguess)/(totalnumberofinstances)$ yields PILOT: 0.0609, Replication: 0.2046

1.2 Differences

The biggest difference between the replication and the original paper is the number of digraphs used. Since they haven't mentioned the total number of digraphs used, I'm assuming that they have 1296 digraphs.

2 PILOT metric comparison

Here are some of the metrics used in the paper.

1. **Avg** : average number of attempts to guess a password.
2. **Stdev:** standard deviation

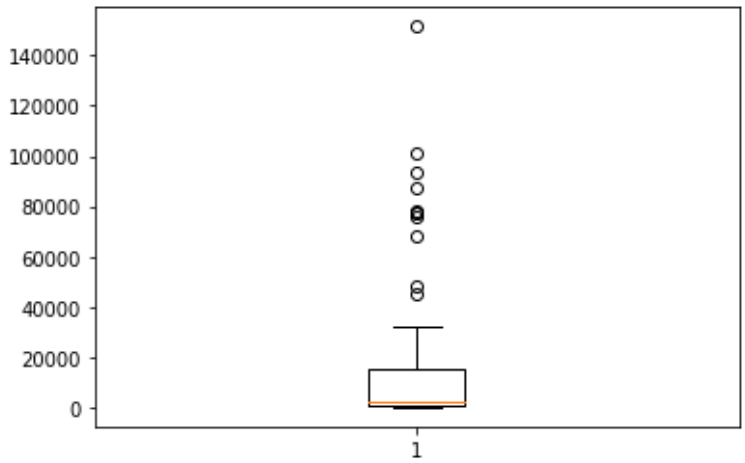

Figure 1: Boxplot representing the number of attempts required to guess lamondre

3. **Rnd:** number of attempts required for random guessing
4. **Best:** number of attempts of the best guess

Right off the bat, we can say that our model outperforms random guessing for an infrequent password like *lamondre*. I applied the ranking strategy discussed to predict the ranking of the password *lamondre* 100 times. Then, I compiled a list of the final rankings to obtain the metrics discussed in this section.

Here are a few basic measures of variability to gauge the differences between the two experiments (Disclaimer: These numbers are quite deceptive in terms of comparing the overall performance of the models)

Our Replication:

1. Mean: 14,374
2. Standard Deviation: 26,863
3. Median: 2,252

PILOT:

1. Mean: 301,906
2. Standard Deviation: 334,681
3. Median: 199,314

The average number of attempts to retrieve *lamondre* is about 14,374 in our replication and the PILOT paper reports the number as 301,906. If we were to do a brute force comparison between the results obtained:

Let A = average number of attempts to guess the password and N = total number of relevant passwords, then if we were to look at the metric $\frac{A}{N} * 100$ for both the papers, the result comes out to be 5.33 and 4.63 respectively.
NOTE: Here, I made an assumption that the total number of relevant passwords for the PILOT was equal to 6.5 million, i.e, the length of the RockYou dataset.

A similar metric can be used for comparing the best attempt. In some instances, the classifier ranked the password the highest in the dictionary. Although, before making bold claims I would like to verify those results with a few hundred additional iterations.

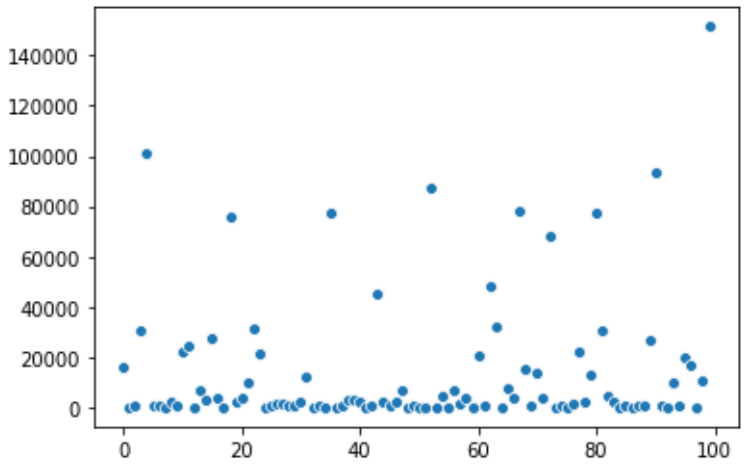

Figure 2: Scatterplot representing the number of attempts required to guess lamondre

3 Threshold

Adding threshold is distorting the average rank to about 147,554 (this needs some additional verification). In general, the threshold is able to distort the overall predicting process for *lamondre*

4 Conclusion

Overall, I believe that our model is on par with the model that they were using. However, a few things that I would like to discuss in this meeting are

1. In what other ways can we compare the models? I've tried measuring most metrics that were accounted in the paper.
2. This result could be biased because we tested it only for one password. In what other ways could we test the model's correctness
3. Should we really be comparing the results?