PMS: Predicting Malware Susceptibility

Manjot Singh Bilkhu
mbilkhu@ucsd.edu
A53280213

Tushar Dobhal
tdobhal@ucsd.edu
A53243953

1 Abstract

In this day and age of the internet, it is very tough to characterize the type of data we deal with. This naturally brings forward the need to be careful in understanding the type of data we deal with and whether it is safe for our machines or not. Malware attackers all over the world try to study specific user patterns and machine characteristics to target victims for their next attack. For example, an out-of-date linux kernel which does not have the latest security updates is a good enough opportunity for an attacker to fill in the gaps. Many big corporations have put forward the need for an *intelligent* data-driven solution, that can help characterize whether the particular user/machine is susceptible to malware or not.

2 Proposed Methodology

In particular, we plan to analyze two things:
1. Are there specific user patters/profiles which are more susceptible to malware attacks?
2. Are there specific machine characteristics which are more prone to a malware breach?
3. Can we use predictive modeling to propose a hypothesis that can predict malware attacks **before** they happen?

We plan to work with the Microsoft Malware Dataset [1], which has about 7.85 million samples for training and 7.85 million samples for testing. Each sample includes features like type of Operating System, which version of operating system, whether firewall is enabled or not, and so on. Since we have a lot of categorical features, we plan to experiment with Gradient Boosted Decision Trees like *LightGBM* [2] or *XGBoost* [3] for formulating our predictive model and validating our hypothesis.

References

[1] https://www.kaggle.com/c/microsoft-malware prediction/ Microsoft Malware Dataset, 2015.

[2] Guolin Ke et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Neural Information Processing Systems, 2017.

[3] Chen, Tianqi and Guestrin, Carlos. XGBoost: A Scalable Tree Boosting System. ACM, 2016.