Capstone 2 Proposal - Marco Riva

What opportunities exist for Microsoft to build a machine learning model to predict the probability of a computer getting infected by a malware, with a performance 25% higher than random chance, by identifying the computer features which cause or prevent a malware infection, before the launch of the new Windows operating system in the Fall 2021?

The majority of individuals in our society own at least one computing device, may this be a smartphone, notepad, laptop, computer desktop, etc. In the last 30 years these devices, together with the Internet, have drastically revolutionized social interaction and enabled technological progress in various fields. Especially during the COVID-19 pandemic more and more of us have transformed our usual routine to a remote setting, exclusively. We pay bills, do shopping, order food and groceries, rent cars, uber, lyft, file taxes, take classes, watch movies on streaming service platforms, and, let’s not forget, even work from home, with our PC (or smartphone). However, this luxury comes with a great threat. Our computers and the internet know all there is to know about us: credit card number, bank accounts, social security number, emails, where we are and what we like - we post on instagram, don’t we? Though it may not be too much of a threat to have some of our data available to reputable companies, businesses and persons, so we can get personalized ads, things are quite different when it comes to sensitive and personal information. And it is no secret that hacking attacks are on the rise. Internet criminals, or hackers, and their vicious attacks are a serious issue which affect the government, organizations, companies, businesses and every single one of us and our daily lives. Even though systematic prevention is performed on electronic computing devices through the use of sophisticated anti-virus software, malware constitutes a serious threat to the privacy of billions of people, particularly because of how dynamically these threats evolve.

In the era of big data it is not too difficult, at least for a tech giant like Microsoft, to collect millions of instances of computers which were infected by malware, together with all the data that characterizes them. To find the correlation between computer features and malware infection is, instead, a more complicated but fascinating task. As Microsoft takes their customer’s security and privacy very seriously, they have provided an open source malware dataset that can be used to train machine learning algorithms to detect potential infections. In this project, I aim to leverage state-of-the-art machine learning techniques to develop a classification model which can predict the probability of a Windows machine getting infected by various families of malware, based on different properties of that machine. The metric of interest is the area under the curve (AUC) of the receiver operating characteristic (ROC) curve, with the particular goal of increasing the performance of the classifier model by 20-25% compared to random chance. The ROC AUC appears a valid metric for this case, as we also are interested in minimizing the false negative rate and maximizing the true positive rate.

Some of the constraints of this project include the sampling methodology chosen by Microsoft to generate the training and test dataset, since user privacy must be respected at all costs.

Furthermore, being the malware problem intrinsically a time series issue, due to the fact that computer operative systems and installed software are updated constantly, machines go online and offline, new computers are introduced and older are removed, etc., with the aforementioned constraint on privacy, made the sampling technique rather complicated. Even if the datasets have been roughly split by time, a disagreement between cross validation scores and test set scores is likely to arise.

A slide deck and project report summarizing the key findings for the data science team and Microsoft executives will be generated.

Finally, the datasets are in csv format and are available at:

https://www.kaggle.com/c/microsoft-malware-prediction/data