**Problem statement**

The goal of this project is to predict a Windows machine’s probability of getting infected by various families of malware, based on different properties of that machine.

|  |  |
| --- | --- |
| **1.context**  The malware industry is a well-organized, well-funded market dedicated to evading traditional security measures. Once a computer is infected by malware, criminals can hurt consumers and enterprises in many ways.  Microsoft owns billions of devices, the company wants to improve their security. As one part of it, company is asking to  develop techniques to predict if a machine will soon be hit with malware  **2.Criteria for success**  Performance metrics will be compared between models based on the area [under the ROC curve](http://en.wikipedia.org/wiki/Receiver_operating_characteristic) between the predicted probability and the observed label.  **3.Scope of solution space**  -Model is trained on 167 columns, with 5 gb of data  -To predict, HasDetection column, if the value is >0.5 then machine is infected | **4.Constraints within solution space**  The sampling methodology used to create this dataset was designed to meet certain business constraints, both in regard to user privacy as well as the time period during which the machine was running  **5. Stakeholders to provide key insights**   * Rahul Sagrolikar-Data science career track mentor, Springboard * Springboard-data science community platform   **6.Key data sources**   * data is sourced from Microsoft defender research which is in csv format from Kaggle competition * (<https://www.kaggle.com/c/microsoft-malware-prediction/data>) * training data is around 5gb with 167 columns. |