Background

There are situations in companies where malicious agents attempt to connect to computer networks. The agents perform various types of attacks in order to achieve various goals. The goal of an Intrusion Detection system is to monitor different activity in a computer network and classify the activity is either normal or malicious.

Task

The project is to build a classifier that is able to distinguish between malicious attacks and normal behaviour. The provided dataset contains features that have been engineered from millions of TCP connections that simulate a US AirForce LAN. Additional information can be found at http://kdd.ics.uci.edu/databases/kddcup99/task.html

The goal is to correctly identify all the malicious activities while minimizing the number of false positives. Too many false positives would result in a company being swamped with alerts they are unable to investigate.

Deliverable Items:

- Analysis Presentation with Recommendations
- Notebook (.ipynb) with markdown documentation and appropriate headers showcasing your Exploratory data analysis, experiments
- Productionized Code, meaning a cleaned up ipynb that just has your final model object and the code required to run the pipeline of load, extract, transform, predict, evaluate (File containing model object; code to read new independent variables of same array as original, prediction, writing predictions to csv file)
- link to your git hub project repo. I will be doing git hub pull requests to review your code here
- One pull request to each of your colleagues github repos for a fix or suggestion
- Power Point Presentation structure:
    - Title
    - Background
    - Problem Statement/What are we trying to solve/predict and what implications does it have
    - Data set profiling
    - feature selection
    - model selection
    - model evaluation
    - Best model and variables
    - What does mean, and why is it important?
    - Next Steps for the team now that we have built these variables and identified important features to monitor?

Notes on Power Points Presentation:

- Tailored for Intended audience (At the end of this brief)
- Should follow the structure above

Mandatory:

- Apply what you've learned so far about exploratory data visualization (EDA, visualization), feature correlation, feature selection, preprocessing (scaling, typing, mapping, dropping), feature engineering (if applicable)
- Modelling experiments should be recorded and output as a dataframe to organize your thoughts model planning
    - from sklearn.model_selection import GridSearchCV
    - hint: if using grid search use mygridsearchobject.cv_results_ → dataframe
- Use **all the models** that we have discussed in class so far **WITH PIPELINES** (appropriate transformation with respect to different data types) and tune with **GRID SEARCH** (appropriate parameter searching
- **You must state the null model (simply predicting the most popular class: what would be the accuracy, etc of such a model) and try to BEAT the null model**
- Explore your model coefficients (data visualization) and give some interpretation why your best model was selected (might be challenging)
- **IMPORTANT**: Your own personal insight into other interesting trends or patterns you noticed during your EDA. For every visualization, transformation, modelling selection decision, I want you to:
    - **EXPLAIN** what you are observing,
    - how this **CONNECTS** with the larger question and domain at large
    - **PROPOSE** further avenues for investigation in your dataset or if is beyond the scope of this dataset, how might the team better answer this problem with additional resources
    - **WHY** is it important that you are pursuing this line of investigation
- 3 user defined functions
- Create or use your designated project repo for this project:
    - README.md: make sure to fill this out to give a high level showcase about what the project is about, what domains does it cover from class, what you learned from your dataset/modelling, and what you learned from doing the project. This is to not only showcase your progress as you move along in the curriculum but also for anyone who visits (employers, etc) to see your abilities

Audience:

IT Security Manager and Team; HR Director