Course Project

Information security and machine learning

Complete a data analytics project in Information security

- The data analytics project stages will be distributed in the three milestones
- The paper will talk also about the team data analytics project
- You need to have a team from 1-3 students
- The team should pick one of the datasets in discussion board and reserve it making sure that the dataset is not used by any other team

---

1. Create an account in https://www.kaggle.com
2. See my sample report for assignment 1 to be used as a rubric for this assignment. As I demoed in the sample this dataset (https://www.kaggle.com/c/microsoft-malware-prediction/kernels), exclude it also from your selection.
3. Go over the sections: Data, Kernels, Discussion, Activity to understand the data analytic goals behind the dataset. Then summarize this in 1-2 pages.
4. Download one of the following data analytic tools (or else use Python or Java IDEs/packages if you have programming skills):

    a. RStudio : https://www.rstudio.com (requires R)

    b. RapidMiner: https://rapidminer.com/get-started

    c. Weka : https://www.cs.waikato.ac.nz/ml/weka

    d. Those who will be using programming languages/codes such as Python, Java, and R will be given a grading advantage (as it requires more effort to program than to use

ready to use tools). However, you are not required to do so.

5. Each milestone should include both report and code files

Guidelines for Grading/Assessment of Milestone 1

- Each Milestone has roughly a third of overall project grades. Both milestone presentation and report will contribute to the milestone grade.
- The team will be assessed based on the overall progress up to the milestone. There are minimum criteria (related to data science projects' tasks and also template components). In addition, top 10-20 % will be assessed based on relative contribution/achievement among all teams.
- Report content should reflect what you have completed so far.
- The final research report should follow research paper templates (You can see examples in the following links: http://open.lib.umn.edu/writingforsuccess/chapter/13-1-formatting-a-research-paper/, https://style.mla.org/formatting-papers/, http://www.aresearchguide.com/4format.html, https://explorable.com/research-paper-format, https://owl.english.purdue.edu/owl/resource/560/18/, https://www.ece.ucsb.edu/~parhami/rsrch_paper_gdlns.htm.
- I will also upload samples of "anonymous good previous submissions" for your preference
- You will submit in Milestone1: (1) Abstract, (2) Research Questions, (3) Introduction + ((4) analysis progress: e.g. data collection, and preprocessing activities), with size of no less than 3 pages (standard times new roman 12 fonts 1.5 space)
- Be prepared to present both your code progress as well as your report (Presentation is part of the grade for both code and report).

- For Milestones, you should prepare a power point presentation, but you will not be graded for the presentation part unless you present your work orally.
- You have to evaluate all other students presentations according to the assessment form template. Submit your assessment before you leave live session, Although your assessment to others will not be impacting your grade, yet lacking to submit them will cause your own presentation to be out of 70% instead.

Guidelines\Rubrics to deliver Project Deliverable 1

Report components

1. Overall Goals / Introduction (15 %)
2. Research Questions: Research Hypothesis (15 %)

3. Code, preprocessing (30%)
4. Code-Visualization (40%)

Guidelines\Rubrics to deliver Project Deliverable 2 and Final Deliverable

- In deliverables 2 and final, you add the new sections in addition to any modifications in earlier sections from previous deliverables.
- Make sure you submit report in addition to any supporting code

---------- Deliverable 2 components ----------

(1)(Previous/Related Contributions) (30 %)
[Kernel codes and also relevant literature]

As most of the selected projects use public datasets, no doubt there are different attempts/projects to analyze those datasets. 30 % of this deliverable is in your overall assessment of previous data analysis efforts. This effort should include:

- Evaluating existing source codes that they have (e.g. in Kernels and discussion sections) or any other refence. Make sure you try those codes and show their results
- In addition to the code, summarize most relevant literature or efforts to analyze the same dataset you have picked.
- If you have a new dataset with no or limited Kernel, survey literature not necessary on any work on this dataset in particular, but in the domain of the dataset (as you may have many other similar or relevant datasets)

(2)Features Selection / Engineering (40 %)

(See this link for content of the next section)

https://www.kaggle.com/WinningModelDocumentationGuidelines

- **We talked about feature selection methods and I uploaded several samples about that. I expect you for the least to reuse such code towards your dataset**
- You can use the following questions to guide you in what to include in this section (if you can answer all those questions with evidences/screenshots from your work, that is great)
- What were the most important features?
- We suggest you provide:
- a variable importance plot (an example here about halfway down the page), showing the 10-20 most important features and
- partial plots for the 3-5 most important features
  - If this is not possible, you should provide a list of the most important features.
  - How did you select features?
  - Did you make any important feature transformations?
  - Did you find any interesting interactions between features?
  - Did you use external data? (if permitted)

Simple Features and Methods (Partial, part of the final grading)

Many customers are happy to trade off model performance for simplicity. With this in mind:

- Is there a subset of features that would get 90-95% of your final performance? Which features? *
- What model that was most important? *
- What would the simplified model score?

* Try and restrict your simple model to fewer than 10 features and one training

method

(3)Training Method(s) (Partial, part of the final grading) 30 %

- Use the questions below to guide your effort in this section. (if you can answer all those questions with evidences/screenshots from your work, that is great)

- What training methods did you use?
- Did you ensemble the models?
- If you did ensemble, how did you weight the different models?
  A6. Interesting findings

- What was the most important trick you used?
- What do you think set you apart from others in the competition?
- Did you find any interesting relationships in the data that don't fit in the sections above?
  .

−−−−−Previous components should be complete by Deliverable 2, Next components can be partial in deliverable 2 to be completed in deliverable 3 (They will not be graded part of deliverable 2)

Model Execution Time (Partial, part of the final grading)

Many customers care about how long the winning models take to train and generate predictions:

- How long does it take to train your model?
- How long does it take to generate predictions using your model?
- How long does it take to train the simplified model (referenced in section A6)?
- How long does it take to generate predictions from the simplified model?

Ensemble and Deep learning methods (Partial, part of the final grading)

We will cover ensemble, DL methods in the last two weeks. Details, quality and thoroughness of evaluated Ensemble and DL methods are important factors in grading of the final deliverable

References (Partial, part of the final grading)

Citations to references, websites, blog posts, and external sources of information where appropriate.

Part of the final report

A comparison study

Compare results in your own work/project with results from previous or other contributions (data

and analysis comparison not literature review)

Summary

Summarize the most important aspects of your model and analysis, such as:

The training method(s) you used (Convolutional Neural Network, XGBoost)

The most important features

The tool(s) you used

How long it takes to train your model

Quality Criteria (10-20% of overall project):

1. **Thorough performance analysis**: Results in data analysis can be misleading. Without detail analysis of different performance metrics (e.g. accuracy, recall, ROC, AUC, etc.) one-side view of results can present incomplete and inaccurate findings. Presenting a thorough analysis for overall performance of your models will show that you did not ignore any factor in your model.
2. **Following standard project templates**: You can find through the Internet several standard templates for data science projects (How to structure your code, data, etc.). While following standard templates is not a must or required but will be considered as part of quality criteria. Here are examples of code templates for different programming environments:
    a. R and RStudio: http://projecttemplate.net/getting_started.html
    https://nicercode.github.io/blog/2013-04-05-projects/
    https://community.rstudio.com/t/data-science-project-template-for-r/3230/10
