Project: Introduction to Machine Learning Applications

First report Due 10/30/2023 11:59 pm via LMS.
Final presentations in-class – 11/30/2023; 12/04/2023; 12/07/2023
Final report due – 12/07/2023 11:59 pm via LMS

PROJECT OBJECTIVE
The goal is to develop an understanding of how individuals approach machine learning projects, following the entire process from exploratory data analysis to modeling and evaluation.

PROJECT SELECTION
Please look through the available Kaggle competitions at: https://www.kaggle.com/competitions and select one project. More details on Kaggle can be found here: https://www.kaggle.com/getting-started/44939

You should avoid image-based data, projects that only require visualizations and tutorial-style competitions. DO NOT consider tutorial datasets or the datasets that we used in the class during lectures or for homework. Ideally the project should be aligned with some type of application related to business (or the domain of your major). Please choose datasets that have at least 10,000 data points. If the data associated with a competition is large enough that your personal computer is unable to load it, please sample the data.

For 6000-level: it is highly encouraged that you consider submitting your final project report for a conference submission.

Make sure you include your Kaggle competition URL, your full name and RIN when submitting project reports.

First Report (30 points, due by October 30th, 2023):

**First report doesn't have any constraints on page spacing and document formatting. **

1. **Summary** (1 pages)

This should be a summary in your own words of the problem, data, preprocessing techniques you used and any initial observations. When you describe the data, please provide description of atleast 10 features; Description of class label -- please try to include tables instead of showing the code output.

2. **Benchmarking of Other Solutions** (2 pages)

Identify 3 other Kaggle *solutions* completed by others. The solution should include a score on the Kaggle prediction task. You can find them by selecting a project and then clicking on the

link to Kernels. Summarize the features, modeling approach, and performance in a table. Then do further research to comment on the approach and try to characterize what makes the kernel successful than others. At this stage, you are not expected to build your own model yet.

3. **Data description and Initial Processing** (3 pages)

This section should include basic characterization of data. You should run and report basic statistics on the data and generate at least 3 visualizations. You can review other kernels to understand some different approaches to the data, but in this section, you are required to generate all the analyses. In the preprocessing, state clearly what has been done to make sure data is ready to build a model – including important visualizations/tables. Please check if these visualizations are helping understand your data better. **Do not submit the report without any explanations for the visualizations included in the report**. Make sure the visualizations are clear and text in them are readable.

Final Report – including resubmission of sections2&3 (170 points, due by December 7th, 2023):

REPORT SECTIONS

1. **Executive Summary** (1 pages)

This should be a summary in your own words of the problem, data, and findings. When you describe the data, please provide description of at least 10 features; Description of class label -- please try to include tables instead of showing the code output.

2. **Benchmarking of Other Solutions** (2 pages)

Identify 3 other Kaggle *solutions* completed by others. The solution should include a score on the Kaggle prediction task. You can find it by selecting on the project and then clicking on the link to Kernels. Summarize the features, modeling approach, and performance in a table. Then do some further research to comment on the approach and try to characterize what makes the kernel more successful than others.

Notebook Name	Feature Approach	Model Approach	Train/Test Performance

3. **Data description and Initial Processing** (3 pages)

This section should include basic characterization of data. You should run and report basic statistics on the data and generate at least 3 visualizations. You can review other kernels to understand some different approaches to the data, but in this section you are required to

generate all analyses. In the preprocessing, state clearly what has been done to make sure data is ready to build a model – including important visualizations/tables. Please check if these visualizations are helping understand your data better.

4. **Modeling** (3 pages)

Modeling should examine the relevance of different independent variables (features) and different algorithms. You should examine at least 3 different models and be able to also explain the relevance of different independent variables. When you show the performance of the models, please use different ROC curve/precision-recall graphs or tables that include varied train-test splits, how the performance in terms of precision, recall, accuracy and F-score are changing, etc.

5. **Appendix**

Submit your well commented code.

Your proposed solution and how it is handling the drawbacks of the 3 solutions – atleast address basics of why your model is good and useful to solve this problem. An example scenario: the 3 solutions you chose when you modify the training data (in terms of train-test split), their accuracy is going low at a faster rate compared to your proposed solution -- this is something because of the way you are processing and handling your features.

6. **Formatting**

Please use the ACM conference style format when you submit your final project report.
.docx file:
https://www.acm.org/binaries/content/assets/publications/word_style/interim-template-style/interim-layout.docx
latex template:
https://www.acm.org/binaries/content/assets/publications/consolidated-tex-template/acmart-primary.zip
If you use overleaf, template can he found here:
https://www.overleaf.com/project/new/template/23074?brandVariationId=166&id=61153342&latexEngine=pdflatex&mainFile=sample-authordraft.tex&templateName=ACM+Conference+Proceedings+Primary+Article+Template&texImage=texlive-full%3A2021.1

NOTE: If you copy and paste from the Kaggle description that is plagiarism and you will be reported to the Associate Dean's office and receive a 0 on the project grade.

PROJECT EVALUATION

The description below describes an ideal project. Projects will be evaluated subjectively by the instructor according to this rubric.

- *Formatting (10 points).* The student presented the report in a format that indicated professionalism and care in the organization, writing, and presentation of the overall report while using the template provided.

- *Executive summary (20 points).* The student was able to present the results of modeling in a way that is rich and interesting as well. There is clear representation of key predictors and key algorithms used. There is a summary of the results and key findings.
- *Benchmarking of other solutions (30 points).* There is a clear insightful comparison of approaches, and he predictive characteristics of the different models are clearly compared in a table with appropriate conclusions. There are outside resources consulted in the description of specific algorithms if relevant.
- *Data description and initial processing (40 points).* The student was able to clearly present an overall picture of the data using techniques presented in the class. This includes basic structure field by field descriptions as well as visualization and basic statistics. Where necessary they have adequately used techniques for cleaning the data or generating new features.
- *Analysis of relevance of independent variables (25 points).* The student was able to clearly present justification of the value of different independent variables. Where possible, exploration of feature creation is provided.
- *Analysis of performance of different model types (25 points).* There are outside resources consulted in the description of specific algorithms if relevant. Outside sources may give clarity and there is evidence of some model tuning.
- *Commented Code (20 points, as needed).* Clearly commented code has been provided in the assigned Jupyter notebook. Please DO NOT submit data used towards this project but include only the code as a single Jupyter notebook.

PROJECT SUBMISSION

- *The project report is to be submitted to the LMS. Submission urls will be shared by the instructor.
- NOTE: If you copy and paste from the Kaggle description that is plagiarism and you will be reported to the Associate Dean's office and receive a 0 on the project grade.
- Final Code Submission must be done as a .zip file via LMS.
- **Grading is based on the quality of content you presented but not the quantity**

If you prefer choosing another project that is not a Kaggle competition, please contact me before you proceed forward All the above instructions still apply in case you choose to work on a dataset with minor modifications.

Project presentation will be graded separately for 30 points max.

Here is a checklist you should go through before you submit your final project:

- This is a business-focused report. No code should be included. Many students try to include large sections of code (you can submit code separately). You can screenshot visualizations. Otherwise try and put figures in a word document and then word to create a nice table.

- The EDA (and really all sections) should be a mixture of analyses and interpretations. You can't just paste plots you generated without any explanation.
- There should be a clear summary table that compares the performance of the proposed algorithms on a metric that is clearly labeled.
- Spend a bit of time on formatting. Do you follow the required format?
- Don't just give lists. Use paragraphs with headings to separate your report.
- It also isn't ideal to put long paragraphs of descriptions on tables.