Practical ML – Assignment 2 (Scikit Learn)

Machine learning is an active area of research with a high level of impact on real-world problems.

The objective of this assignment is to allow you to explore an interesting and relevant machine learning dataset using Scikit-Learn. More specifically, you will be required to perform pre-processing, build and evaluate machine learning models and write a report on the results.

You will also be required to pick a specific area to research. This research should be integrated into your methodology and evaluation (more detail on this below).

Note: It is important that you do not start the assignment until I have validated your dataset. If you have not contacted me about your dataset yet please email me (ted.scully@cit.ie) as soon as possible.

Guidelines and Submission Instructions

- Please note you should upload all deliverable files (your python file and your report) into a single .zip file for submission. The submission deadline is **Sunday Dec 13th at 23:00**.
- Go to the "Assignment 2" unit on Canvas to upload your submission.
- It is your responsibility to make sure you upload the correct files.
- Please make sure you fully comment your code. You should clearly explain the operation of important lines of code.
- Please note that marks are awarded for code that is efficient, well structured and with minimum duplication.
- Late submissions will be penalized.
    - If you submit the assignment after the deadline but within 7 days, **10%** will be deducted from your final grade.
    - If you submit the assignment more than 7 days after the deadline but within 14 days, a **20%** penalty will be deducted.
    - A grade of **0%** will be given to any assignment submitted more than 14 days after the assignment deadline.

- Please reference any sources you use in your code or report.

- There is a zero tolerance policy with regards plagiarism. Software is used to detect any plagiarism that may be present in either your submitted document or in your code. CIT policy covering academic honestly and plagiarism can be found here.

- A discussion forum will be maintained where you can ask assignment related questions. It is very important that you do not share any code or refer in any way to the methodology you are using for solving the problems outlined below. If you are not fully clear on what is being asked in any part of the questions below you can look for clarification by submitting your question to the discussion forum.

Distribution of Marks

This project will account for **50%** of your overall module grade. The marks will be broken down as follows:

- **Report - Abstract and Introduction [5%]**
- **Report - Research [30%]**
- **Report – Methodology [10%]**
- **Report – Evaluation and Conclusions [35%]**
- **Project Code [20%]**

Each of the above components is described in more detail below.

Dataset

Your initial task will be to select an appropriate dataset. You should select either a regression or classification dataset. You should give an overview of your dataset (it's constituent features) and you should identify the column that will act as the target (either classification or regression target) for the model.

Project Overview

The project requires you to build machine learning models for your chosen dataset. You will need to perform pre-processing on your data. Follow the pre-processing steps outlined in the Scikit Learn lecture notes. You will need to build and comprehensively evaluate a range of machine learning models. The most promising models should then undergo hyper-parameter optimization.

You are also required to pick a specific topic to research and then incorporate the result of this research into your models and evaluate the impact. For example, if your dataset is imbalanced your research could focus on the techniques that are commonly used to address imbalance. You would then proceed to incorporate some of these into your evaluation and assess the impact on your results.

You should compose a research report detailing the work you have undertaken and the overall findings. You will find a template for the research paper in the assignment folder. This template adheres to the Springer paper specification and should be used for your report. The paper you submit should contain the following sections:

(i) Abstract
(ii) Introduction
(iii) Research
(iv) Methodology
(v) Evaluation
(vi) Conclusions and Future Work

I recommend that you do not exceed 8 pages for the research paper. I understand that some of you may have difficulty adhering to this limit. Please note that this is a recommended guideline, it is not a requirement and you will not be penalized if you exceed that page limit. More detail on each of these sections are provide below.

1. Report - Abstract and Introduction [5%]

Your abstract should provide a short summary of the work that you undertook as part of the project. It should primarily provide an account of the main objectives and a summary of the results.

In your introduction you should provide a description of your chosen dataset. You should clearly identify the target regression or classification value that you want to predict. Describe the motivation for building models for this dataset and the objectives of the study.

2. Report - Research [30%]

The section should outline the specific topic of research that you will incorporate into your study and why it is important. The objective of this section is that it allows you to select a particular stage of the pre-processing or model building process and research it in more depth and incorporate aspects of this into your methodology and results. It is also import that you clearly describe the techniques you are using in the research section. You should demonstrate that you understand the operation of the techniques you are going to employ.

There are a broad range of topics that you could consider for your research component. For example you could look at:

- Outlier Detection (Researching a range of techniques for performing outlier detection and investigating their impact).
- Dataset Imbalance
- Feature Encoding, etc
- Dealing with missing values etc

For many of the above areas I have demonstrated in the lecture notes a limited number of techniques. For example, with dataset imbalance we covered techniques such as random under and oversampling as well as SMOTE. If you were to select this topic of imbalance then you can start by comparing the impact of the techniques used in the lecture slides. You should demonstrate a clear understanding of all techniques that you employ.

However, **to grade well in the research section** you should undertake independent research. That is, you should demonstrate that you can research, understand and apply additional techniques not covered in the lecture notes (to grade very well here you need to undertake substantive research). You should clearly describe any techniques and integrate them into your process and evaluate the impact on your results. Please make sure to reference any sources you use.

[In the previous assignment (assignment 1) of this module you may have looked at the area of feature selection. This does not prohibit you from focusing on feature selection as a research topic here. However, it is very important that the techniques you cover for this assignment will differ from the feature selection technique(s) you selected for assignment 1].

3. Report - Methodology [10%]

Appendix A shows a typical high-level implementation workflow that you may undertake.

It is broken down into:

1. Part 1. Establishing a baseline
2. Part 2. Basic experimentation
3. Part 3. Research

This methodology section should outline the sequence of pre-processing steps that you undertook in order to prepare your data and the rationale for adopting these techniques (across both Part 1 (baseline) and Part 2 (basic experimentation)). You should demonstrate that you clearly understand any techniques you apply.

It should also describe the range of models you used in your initial model building phase. It should describe the hyper-parameter optimization technique that you employed and the range of parameters that you examined for each of the best performing models.

Notes:

1. There is no need to describe any aspects of Part 3: Research in this methodology section as that will be clearly described in the Research Section of your report.
2. You shouldn't include results in your methodology. That should be detailed in your evaluation section.

4. Evaluation and Conclusions [35%]

This section should contain a comprehensive evaluation of your results. You should report your results from building the initial baseline (Part 1 in Appendix A) including the initial model performance as well as the optimized results after hyper-parameter optimization. This section should also clearly communicate the impact of the basic experimentation (Part 2 in Appendix A). Also in this section you should clearly demonstrate the impact of your chosen research on the overall results (Part 3 in Appendix A). Please subdivide your evaluation into these three subsection.

The results should be clearly interpreted and depicted (graphically where possible). You should use a range of evaluation metrics. It is important you demonstrate a clear understanding of the evaluation metrics that you use.

Also please make sure you provide an intuitive method of cross-referencing between your code and the results in the evaluation. You could for example include the section number as a comment in your code. This will allow me to easily identify the code that generated each set of results.

This section should also include a conclusion, which outlines possible areas of future work.

Notes:

1. While nested cross fold validation is best practice, I don't expect you to use it in this assignment. Nested CV is very computationally expensive and may put you under pressure given the short time window available for the project. However, cross fold evaluation should be used where possible.
2. Your full evaluation and results should be included in your research report. A penalty will be incurred if you don't include your results in the report (even if they are in your IPython notebook).
3. If you end up getting poor performance results for your selected dataset (your models just don't perform well on your selected dataset) you will be penalized in any way for this. The grade you achieve is not at all dependent on the final performance of your model.

5. Project Code [20%]

All code should be completed using Python as the programming language. You should use Scikit Learn, NumPy and Pandas. You are free to use imported graphical libraries such as MatPlotLib or Seaborn (This is not a requirement. For example, you can also use tools such as Excel if generating graphs). You are also free to import Scikit-Learn contribution packages such as Imbalanced Learn. If you wish to use other external libraries please check with me in advance.

Your code should have a logical structure and a high level of readability and clarity. Please comment your code and put all code into functions. You code should be efficient and should avoid duplication.

Appendix A – Overview of Project Implementation Workflow.

Aa high level I would typically expect that you undertake the following workflow with your project implementation:

Part 1: **Establish a baseline**

1. Pre-processing your data
    a. Dealing with Outliers
    b. Dealing with Missing Values (if applicable)
    c. Handling Categorical Data (if applicable)
    d. Scaling Data
    e. Handling Imbalance (if applicable)

2. Build a wide range of basic models [for example up to 8 different categories] (Use default parameters and build ML models)

3. Take the best performing models from step 2 (for example the best 3 models) and perform hyper-parameter optimization to tune the model as best as possible. Remember hyper-parameter optimization can be time consuming so start with a small search space and increase if you have sufficient time available.

Part 2: **Basic Experimentation**

Now that you have some baseline models you should undertake some basic experimentation. For example, modify some of the pre-processing techniques you have undertaken. Check to see if feature selection makes a difference (notice we haven't included feature selection in the baseline process). The experimentation in this section is basic and will just be limited to feature selection and two other changes to the pre-processing pipeline. Your evaluation should demonstrate and interpret the performance with part 1 and part 2 separately.

Part 3: **Research**

Using the best models from Part 1 and 2 above you should now explore the topic for you research section and evaluation it's impact comprehensively on your model's performance.