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 a 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).

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 4th at 19:00**.
- Please **do not** upload your dataset as part of the submitted zip file.
- 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.

- 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. MTU policy covering academic honestly and plagiarism can be found here. Any sources used should be clearly cited and referenced. Any plagiarism detected will result in a grade of 0% for the assignment.

- 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 [35%]**
- **Report – Methodology [10%]**
- **Report – Evaluation and Conclusions [35%]**
- **Project Code [15%]**

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

Dataset

You can select to use any of the datasets listed below. Please note that in order to download the datasets you will need to register for Kaggle.

1. Bank Marketing Dataset (Binary Classification Problem). The target is to predict the binary class target y - has the client subscribed to a term deposit.

2. Australia Weather Dataset: (Binary Classification Problem). The target is to predict RainTomorrow (each data instance contains weather related information about the current day, it also includes the target which is: does it rain the following day. Therefore, based on the weather-related information for the current day you want your model to predict if it will rain the following day).

3. Terrorist Dataset (Multiclass Classification Problem). The target here is to predict the column gname (the name of the responsible terrorist group. Here you can select the 10-15 most

prolific terrorist groups, that is extract all rows out of the dataset pertaining to these terrorists groups. There are a lot of feature columns in the dataset and there is a lot of missing data. You can remove any columns that have a significant amount of missing data or any columns that you deem irrelevant to the target.

4. Churn Modelling Dataset (Binary Classification Problem). This dataset contains bank customer details and the binary target is whether the customer left the bank (closed his/her account).

5. Adult Dataset: (Binary Classification Problem) This is an older dataset that has taken census data from the US. The features include information such as education, marital status, etc. You will notice one of the columns is if a person earns greater than 50K or less than 50k (this is the binary target).

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 and subsequently build and comprehensively evaluate a range of machine learning models. Appendix A gives an over the implementation workflow for the project.

Important: Because of the limited amount of time available for the assignment you can just use cross fold validation as the basis for your evaluation methodology (you don't need to use nested cross fold validation or even have a separate test set).

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 overall 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 (inclusive of graphs) 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 main results and overall findings.

In your introduction you should provide a description of your chosen dataset. You should clearly describe the features and the target classification value that you want to predict (along with any relevant observations you may have about the features and the target).

2. Report - Research [35%]

The section should outline the specific topic of research that you will incorporate into your study and why it is important and potentially relevant to the dataset. The objective of this section is that it allows you to select a specific pre-processing stage, research it in more depth and incorporate aspects of it into your methodology and results. It is important that you clearly describe the techniques you are using in the research section. You should demonstrate that you fully 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 can focus on:

- Outlier detection (Researching a range of techniques for performing outlier detection and investigating their impact).
- Dataset imbalance.
- Feature selection.
- Dealing with missing values.

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. You should also demonstrate a clear understanding of all techniques that you employ.

However, an essential component of this section is to 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 and demonstrate a detailed understanding of non-trivial techniques). 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.

3. Report - Methodology [10%]

Appendix A shows a typical high-level implementation workflow that you should adopt for this assignment.

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 and why these are being applied.

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 subsections (matching the 3 stages of Appendix A).

The results should be clearly interpreted and depicted (graphically where appropriate). 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 submitted 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 good practice for small datasets, please don't use it in this assignment due to the limited time available. 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.
2. Your full evaluation and results should be included in your final research report. A penalty will be incurred if you don't include your results in the report (if the results are not included in the report but are in included in an IPython notebook a penalty will apply).
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 not be penalized in any way for this. The grade you achieve is not at all dependent on the final performance of your model.
4. Hyper-parameter optimization can take a significant amount of processing time. Given that you have a very limited amount of time for completion of the project I strong recommend that you specify a small parameter search space.

5. Project Code [15%]

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 related to your research 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.

The following is the high level methodology that you should undertake for your assignment.

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)

    * Depending on the techniques you use in pre-processing, the above sequence may change. If you need to change this sequence that is perfectly fine.

2. Build a wide range of basic models [for example up to 5 different categories] (Use default parameters and build ML models)

3. Take the 2 best performing models from step 2 and perform hyper-parameter optimization to tune the model as best as possible. Remember hyper-parameter optimization can be time consuming so use a small parameter search space (you can always increase this if you have extra time).

Part 2: **Basic Experimentation**

Now that you have identified baseline models from Part 1. You should undertake some basic experimentation.

You should check to see if feature selection makes a difference to the performance of your two top models (notice we haven't included feature selection in the baseline process). Again, you will not be penalized at all if feature selection does improve the performance of the models.

You should clearly convey your understanding the feature selection technique you apply and why are applying it. You should also demonstrate that you can clearly evaluation the resulting performance.

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.