**Excerpts from Assignment Description (for info)**

**RT1: Model comparison**

Students should

• choose one dataset, but use this to artificially create a number of different synthetic datasets, each one with a different class imbalance.

• choose a small number of models (two or three). Models should be chosen to have different

fundamental properties that may influence performance on class imbalanced data, e.g. linear

versus non-parametric, different bias-variance trade-off.

~~• for each synthetic dataset, measure the performance of each model, and perhaps seek to understand the influence of key hyperparameters and their relationship to the class imbalance.~~

* CROSSED OUT PART IS NOT UNDERTAKEN WITH RT3, AND IS REPLACED WITH THE PART HIGHLIGHTED BELOW.

**RT3: Robustness to changes in class-imbalance**

Students should

• choose one dataset, but use this to artificially create a number of different synthetic datasets, each one with a different class imbalance.

• choose either a) two or three models as per RT1 or b) a single model and two or three sampling strategies as per RT2.

• evaluate training at under one degree of class imbalance and testing at another degree of class imbalance. Your aim is to understand how sensitive different methods are to differences in class imbalance between the training and testing sets.

This is the most challenging of the three research templates, but gives you the most freedom to explore choices and to stamp your own identity on the resulting investigation. An excellent investigation of research template 3 can lead to the highest marks on the project, but with the increased freedom comes a greater risk of producing a less coherent, weaker project too.

**1.3.1 Choosing a dataset**

You must choose **one dataset** which you may find online. Good sources for datasets are from Kaggle (https://www.kaggle.com/datasets) or UCI (<https://archive.ics.uci.edu/ml/index.php>). The dataset should ideally result in no more than 100000 datapoints after preprocessing (see Section 1.3.2), but you can always take a larger dataset and then take a subset for your experiments. However, you will need to make this subset available to the markers, so take the subset first save it down to file, work with that file, and submit that file as part of your code submission. Your final code should run on a desk computer (CPU only) in less than 30 minutes to obtain all the necessary results presented in your report from scratch.

**1.3.2 Preprocessing**

Preprocessing is the name we give to the manipulation, (re)formating, cleaning, subsampling and otherwise processing your data in advance to running your experiments. This can sometimes be a subject of research itself. For the purposes of this project you should think of most preprocessing as a predefined single collection of procedures that you decide on in advance. However, you will also need to somehow create class imbalanced subsets from your master dataset (this is discussed separately in Section 1.3.3 below). **You should not seek to investigate the effects of preprocessing.** This is something you do once to your data, and then use the same preprocessed data for all your experimental conditions.

For this reason, your preprocessing computation time does not have to be included in the 30 minute time allowance. Instead, you should run this on your data, then save the resulting data down to file and **submit this with your code archive**. However, **you must submit the code that you use for the preprocessing** and describe how to rerun this on the original dataset as part of your code submission.

**1.3.3 Derived datasets**

You must also decide how you are going to construct your derived datasets. Each derived dataset is going to be a subset of your master dataset, but each with a different class bias. You must decide the degree of class bias and how you will derive them. Derived dataset can share datapoints with one another (typically the case for RT1 and RT2), but you may decide that it is important for your experiments to keep them distinct (e.g. for RT3). You will also need to decide how you partition each derived dataset into train, validation and test data. There is no one way to do this, but you should explain and justify your approach in the report.

**Steps for preprocessing/ creating derived datasets and how it can fit in with the experimental design.**

1. **Start with original dataset** (approx. 91,000 data points, 83 columns. Split between classes is approx. 92 : 8).
2. **Get rid of null values and any duplicates to clean data** (use means of numerical data, modes for non-numerical, etc).
3. **Split this into 2 or 3 derived data sets of equal to or below 50% the size of the original dataset.** 
   1. At this stage we keep the initial ratio of imbalance (stratification). I.e. if we use 40,000 data points for a derived subset, we will keep the 92 : 8 split between classes. The selection of data points can be random.
   2. Assumption is that shared data between derived datasets is not
4. **Apply SMOTE to each derived dataset to create desired class imbalance:**
   1. Subset a can have an imbalance of (90:10) – similar to that of the OG dataset (only a small amount of extra minority class required).
   2. Setbset b can have an imbalance of (70:30).
   3. If needed, subset c can have the imbalance of (50:50).
      1. Note with this approach, the amount of majority class data stays the same, but the minority class increases. This means each derived dataset will have a different amount of data points to achieve the desired ratios. E.g. a derived dataset of 40,000 values will initially have 36,800 C0 and 3,200 C1. To achieve 70:30 split, with SMOTE, there will be approx. 52,600 values (36,800 C0 and now 15,800 C1)
      2. We could undersample the majority class for each derived dataset to reduce number of datapoints and achieve class imbalance, but we would lose even more data for analysis since we are already taking a subset of the master data.
      3. It will be difficult to take some splits (e.g. 70:30 and 50:50) directly from the master dataset, as there are not enough minority class data points to do this, this is why we need to bump up the number of minority class data points.
      4. SMOTE shouldn’t be applied to the master dataset because once we used it again with the derived data sets, the synthetic data would be used to create more synthetic data points, so analysis of the data could be impacted.
5. **Split each derived data set into training, validation and training sets.**
   1. A good split may be 70 training, 15 validation, 15 testing.
6. **Train each model on each training portion of the derived subset, and validate/tweak model hyperparameters based on the validation subset.** 
   1. “The validation set is a set of data, separate from the training set, that is used to validate our model performance during training.” - <https://www.v7labs.com/blog/train-validation-test-set>
   2. “This validation process gives information that helps us tune the model’s hyperparameters and configurations accordingly. It is like a critic telling us whether is moving in the right direction or not.” (same source as above).
7. **Use model trained on one type of imbalance on another imbalance** – possible combos include:

* Training on 90:10 – Tested on/or 70:30 and 50:50
* Training on 70:30 - Tested on/or 90:10 and 50:50
* Training on 50:50 - Tested on/or 90:10 and 70:30

This should allow us to answer the key part of RT3: “evaluate training at under one degree of class imbalance and testing at another degree of class imbalance. Your aim is to understand how sensitive different methods are to differences in class imbalance between the training and testing sets.”

**Questions for group**

* How much data to from the master dataset to use for each derived dataset?
  + We propose up to and below 50% to make sure that we don’t have over 100,000 data points in a derived dataset.
  + In process outlined above, derived datasets would share values. Can people doing the models let us know if this is an issue.
* Splits of class imbalances do we want (i.e. number of derived datasets)?
  + Happy to hear alternatives to this (e.g. from Andy’s message looks like he is proposing 80:20, 50:50, 60:40). Just depends on which splits we think will give us better results to discuss. The 90:10, 70:39 and 50:50 splits have larger differences and therefore it may be easier to see the impacts on class imbalance on the models.
* What split do we want for training, validation and test portions of each derived dataset?
  + We currently propose 70 training, 15 validation, 15 testing. We think validation and testing should be roughly equal so that there’s a good amount of data for validation and testing.