Capstone Project

Modelling

The modelling will implement classification models from Sklearn and evaluate the prediction. Prior to implement model, I will decide the metrics I will used to evaluate the modelling process.

As in the previous Notebooks I prepared the dataset and created a personal transformer I will use it in the modelling pipeline with some other transformation such as "power transformation".

The modelling notebook will implement:

1. Decide metrics
2. Prepare configuration file for the personal transformer
3. Split the dataset
    a. Beginning -> 2018 training set
    b. 2019 -> validation set
    c. 2020 -> test set
4. Create pipelines
    a. Pre-processing steps
        i. Scaling
        ii. Adding PolynomialFeatures
        iii. OneHotEncoding
    b. Custom transformation
    c. Classifiers
5. Run the modelling step with pipelines
6. Select the best estimator
7. Fine tune the best estimator
8. Run the fine tunned estimator on the test set
9. Present and discuss the results

Everything will be implemented mostly with *scikit learn* (https://scikit-learn.org/stable/user_guide.html)

Metrics

I will work with 6 classification problems metrics to evaluate the performance of the models. Metrics are: Accuracy score, confusion matrix (3x3), Precision, Recall, F1 score and ROC Curve.

1. https://towardsdatascience.com/a-practical-guide-to-seven-essential-performance-metrics-for-classification-using-scikit-learn-2de0e0a8a040
2. https://towardsdatascience.com/6-useful-metrics-to-evaluate-binary-classification-models-55fd1fed6a20

ROC Curve

As the ROC Curve runs on a binary classification, I will apply it on the best estimator and use a scikit learn *OneVsRestClassifier* to prepare the data to compute and plot a ROC Curve.

1. https://stackoverflow.com/questions/45332410/roc-for-multiclass-classification
2. https://scikit-learn.org/stable/auto_examples/model_selection/plot_roc.html#sphx-glr-auto-examples-model-selection-plot-roc-py

3. https://medium.com/analytics-vidhya/understanding-roc-and-auc-metrics-in-classification-tasks-e5e7594cd6b
4. https://learn.extensionschool.ch/learn/programs/applied-data-science-machine-learning-v2/subjects/k-nearest-neighbors-v2/units/standardization-and-k-nn-v2

Configuration files for the custom transformer

I will prepare 4 configuration files. The 1st one will be a minimal file to evaluate a simple baseline. It will shift the net flows for 1, 2 and 3 periods.

The other configuration files will be more complicated. 2 will be based on the analysis and recommendation from the EDA Notebooks and the last one will a massive choice of features.

Split the dataset

I decided to split my dataset in a training set [< 2020], a validation set [=2020] and a test set [=2021]. I will verify the balance of the target data after splitting the dataset.

Pipelines

In the pipelines I will add pre-processing step, optional steps (as we did in the course) and transformation. I will use a grid search to evaluate multiple configurations of parameters.

1. https://medium.com/vickdata/a-simple-guide-to-scikit-learn-pipelines-4ac0d974bdcf

Models

- Logistic Regression (which is a linear model for classification problems).
- Support Vector Machine: (https://scikit-learn.org/stable/modules/svm.html)
- Nearest Neighbours Classification: (https://scikit-learn.org/stable/modules/neighbors.html#nearest-neighbors-classification)
- Decision Trees Classification: (https://scikit-learn.org/stable/modules/tree.html#classification)
- Neural Network Models (supervised) Classification: (https://scikit-learn.org/stable/modules/neural_networks_supervised.html#classification)

A meta estimator which fits several decision trees:

- Random Forest Classifier: (https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html?highlight=random%20forest#sklearn.ensemble.RandomForestClassifier)

I will run a grid search over multiple classifiers to elicit the best estimator.
https://stackoverflow.com/questions/23045318/scikit-grid-search-over-multiple-classifiers

Best estimator, fine tuning, and test set evaluation

At the end of the pipeline process, I will decide the best estimator, fine tune some parameters if needed and evaluate it against un unknown dataset which is the 2021 entry values (almost all Q1 data).