ECON3203/5403 Group Assignment (25 marks)

October 15, 2022

1 Background

Developing a predictive model for ATM cash demand is an important task for every bank. Suppose that you are employed by a bank, and your task is to optimise the bank's cash management by making smarter decisions about reloading its ATM network.

The variable `Withdraw` in the dataset `ATM_training.csv` is the total cash amount withdrawn per day from an ATM, recorded from the ATM network of a bank. The response variable and covariate variables are described in the following table.

Variable	Description
Withdraw	The total cash withdrawn a day (in 1000 local currency)
Shops	Number of shops/restaurants within a walkable distance (in 100)
ATMs	Number of other ATMs within a walkable distance (in 10)
Downtown	=1 if the ATM is in downtown, 0 if not
Weekday	= 1 if the day is weekday, 0 if not
Center	=1 if the ATM is located in a center (shopping, airport, etc), 0 if not
High	=1 if the ATM has a high cash demand in the last month, 0 if not

Your task is to develop a model for predicting the cash demand `Withdraw` based on the covariates.

The test dataset `ATM_test.csv` (not provided) has the same structure as the training data `ATM_training.csv`.

1.1 Test error

For the measure of prediction accuracy, please use mean squared error (MSE), computed on the test data. Let $\widehat{y}_i$ be the prediction of y_i where y_i is the i-th withdraw in the test data. The test error is computed as follows

$$\text{Test\_error} = \frac{1}{n_{\text{test}}} \sum_{y_i \in \text{test data}} (\widehat{y}_i - y_i)^2,$$

where n_{test} is the number of observations in the test data.

2 Submission Instructions

1. Each group needs to submit TWO files (or more if necessary) via the Moodle site (to avoid repeated submissions, one and only one member of your group should be responsible for submitting).

- A document file, named **Group_xxx_document.pdf**, that reports your data analysis procedure and results. You should replace the xxx in the file name with your group ID.
- A Python file, named **Group_xxx_implementation.ipynb** that implements your data analysis procedure and produces the test error. You might submit additional files that are needed for your implementation, the names of these files must follow the same format **Group_xxx_<name>**.

2. About your document file **Group_xxx_document.pdf**
   - Describe your data analysis procedure in detail: how the Exploratory Data Analysis (EDA) step is done, what and why models/methods are used, how the models are trained, etc. with sufficient justifications. The description should be detailed enough so that other data scientists, who are supposed to have background in your field, understand and are able to implement the task. All the numerical results are reported up to four decimal places.
   - Clearly and appropriately present any relevant graphs and tables.
   - The page limit is 20 pages including EVERYTHING: appendix, computer output, graphs, tables, etc.
3. The Python file is written using Jupyter Notebook, with the assumption that all the necessary data files (`ATM_training.csv` and `ATM_test.csv`) are in **the same folder** as the Python file. If you use deep learning models, then please assume that **Keras (with Tensorflow backend)** has been installed.
   - If the training of your model involves generating random numbers, the random seed in **Group_xxx_implementation.ipynb** must be fixed, e.g. `np.random.seed(0)`, so that the marker expects to have the same results as you had.
   - The Python file **Group_xxx_implementation.ipynb** must include the following code

```
import pandas as pd

ATM_test = pd.read_csv('ATM_test.csv')

# YOUR CODE HERE: code that produces the test error test_error

print(test_error)
```

     The idea is that, when the marker runs **Group_xxx_implementation.ipynb**, with the test data `ATM_test.csv` in **the same folder** as the Python file, he/she expects to see the same test error as you would if you were provided with the test data. The file should contain sufficient explanations so that the marker knows how to run your code.
   - In case you want to test your code to see if it runs smoothly and a test error is produced, a small subset (5%) of the full test dataset is provided. This data set has the same format as the full test data `ATM_test.csv`. You can use this subset of the full test data to see how the test error looks like, but note that this number will be different from the test error on the full test data.
   - You should ONLY use the methods covered in the lectures and tutorials in this assignment. You are free to use any Python libraries to implement your models as long as these libraries are be publicly available on the web.

4. Your group is required to submit meeting minutes (using the same submission link, but separated from **Group_xxx_document.pdf**). You may use the template provided for preparing meeting minutes. The more detailed the meeting minutes, the better (who does what next, what has been done by whom, etc.). In case of a dispute within a group, we will use the meeting minutes and/or request for more information to make adjustment to the individual marks. **Should a dispute occurs, please treat each other in a professional and respectful manner**.

3 Marking Criteria

This assignment weighs 25 marks in total. The content in **Group_xxx_document.pdf** contributes 10 marks, and the Python implementation contributes 15 marks. The marking is structured as follows.

1. The accuracy of your forecast: Your test error will be compared against the smallest test error (among all groups including the teaching team). The marker first runs **Group_xxx_implementation.ipynb**
   - Given that this file runs smoothly and a test error is produced, the 15 marks will be allocated based on your prediction accuracy, compared to the smallest MSE produced by the best group, and the appropriateness of your implementation.
   - If the marker cannot get **Group_xxx_implementation.ipynb** run or a test error isn't produced, some partial marks (maximum 5) will be allocated based on the appropriateness of **Group_xxx_implementation.ipynb**.
2. Your report described in **Group_xxx_document.pdf**: The maximum 10 marks are allocated based on
   - the appropriateness of the chosen forecasting method.
   - the details, discussion and explanation of your data analysis procedure.

4 Errors

If you believe there are errors with this assignment please contact the teaching team.