Econometric Theory and Method
Group Assignment (25 marks)

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The variable `Withdraw` in the dataset `ATM_sample.csv` is the total cash amount withdrawn per day from an ATM, recorded from the ATM network of a bank.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Withdraw	The total cash withdrawn a day (in 1000 local currency)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 sample data `ATM_sample.csv`.

1.1 Test error

The prediction accuracy will be computed using mean squared error (MSE) on the test data. Let $\hat{y}_i$ 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

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

where n_{test} is the number of observations in the test data. Your goal is to propose a model that produces the smallest test error. We try to mimic the real situation, when the actual data which will be used for prediction are not know to a data scientist when a model is created. Therefore, the test error is not observed to you.

2 Submission Instructions and Document requirements

1. Each group needs to submit 3 files (or more if necessary) via the Moodle site (to avoid multiple submissions, only one member of your group should submit):
   - 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>**.
   - Minutes of the group meeting(s), **Group_xxx_meeting.pdf** outlining who does what and when
2. **Group_xxx_document.pdf** needs to include
   - Introduction: summary of the task and the main findings (1 page or less)
   - Exploratory Data Analysis: data visualisation, correlation matrix, pairwise plots, etc. Tell some preliminary story behind the data.
   - Models and methods. Which models/methods are used and why, how the models are trained, etc. with sufficient justifications. You need to consider several models. The description should be detailed enough so that other data scientists are able to implement the task. All the numerical results are reported up to *four* decimal places. This part may have subsections for each model/sub-model and overall final model comparison/selection.
   - Discussion: discuss model interpretability (if any), cautionary notes, etc
   - References (I suggest at least 3)
   - Clearly and appropriately present any relevant graphs and tables to support your finding.
   - 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_sample.csv` and `ATM_test.csv`) are in **the same folder** as the Python file. If you wish to use deep learning models (but given that we cover them in the end this is not required), please, use **Keras (with Tensorflow backend)**.
   - 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 test_error
print(test_error)
```

     The idea is that, when a 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 (or in an unlikely event that it does not run, understand what may be an issue).

- For you to test whether your code runs smoothly, a small subset (5%) of the full test dataset is provided. This dataset has the same format as the full test data `ATM_test.csv`. You should not rely on computing test error from this data but rather use your sample to propose the best model (using appropriate model selection methods).
- 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 in a separate file **Group_xxx_meeting.pdf**). You may use the template provided for preparing meeting minutes. 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 15 marks, and the Python implementation contributes 10 marks. The marking is structured as follows:

1. Your report in **Group_xxx_document.pdf**: The maximum 15 marks are allocated based on
   - the appropriateness of the chosen forecasting method;
   - the details, discussion and explanation of your data analysis procedure;
   - written presentation.
2. Your implementation in **Group_xxx_implementation.ipynb**
   - Given that this file runs smoothly and a test error is produced, the 10 marks will be allocated based on clarity of the code and ability of the code to reproduce all results in the report.
   - If the marker cannot get **Group_xxx_implementation.ipynb** to run or a test error is not produced, some partial marks may be allocated based on the appropriateness and clarity of the code. Here your code comments will be even more useful.

Do not forget to keep the minutes during your group meeting(s) and attach the minutes as 2 marks will be deducted if they are not submitted.

To keep this as a contest, the group who will reach the smallest test error will get 2 extra **bonus marks**.