FNCE30012: Foundations of Fintech - Assignment 5

This assignment builds on Lectures 10 to 12 and on Tutorials 8 to 11. You might want to consider using parts of the Python code discussed in those tutorials to answer some of the questions below.

Important: It is important that you *do not* change the type (markdown vs. code) of any cell, *nor* copy/paste/duplicate any cell! If the cell type is markdown, you are supposed to write text, not code, and vice versa. Provide your answer to each question in the allocated cell. *Do not* create additional cells. Answers provided in any other cell will **not** be marked. *Do not* rename the assignment files. All files in the assignment directory should be left as is.

Setting

Equifax Australia has provided us with synthetic loan application data from Australian proprietary companies. This data was generated to match the characteristics of *actual* lending proposals approved between February 2017 and March 2018. The Equifax data consists of two parts, which, to make it easier for you, we have merged together into one data set:

1. Company Business Trading History Data: This first part of the data set contains historical business trading data from 25,000 Australian proprietary companies who were granted a loan between February 2017 and March 2018.
2. Director Data: This second part of the data set contains information on up to four directors of each company. In case a company has more than one director, the corresponding data has been averaged across directors at the company level.

Since this is proprietary data that belongs to Equifax, we are not allowed to give you direct access to it. However, thanks to Jupyter Hub, you are able to access it remotely. In particular, using your knowledge from Tutorial 9, you are able to analyse it at an aggregate level and to use it for the estimation of credit scoring models.

The file called `Equifax_Data_Dictionary.xlsx` provides you with the dictionary for both company and director level data.

Helpful commands

The merged Equifax dataset will be referred to by the name `assignment5`. Please see Tutorial 9 for the details of how to run functions on remote data. We have implemented additional functions to help you with this assignment. Details of these functions are below:

1. `send_grouped_mean_request()`: This function takes inputs in the form of a dictionary containing names of two columns, and returns a Pandas dataframe that contains grouped means of a column with respect to another column. For example, using this function with input `{"data": "tutorial9.1","var": "age", "y": "SeriousDlqin2yrs"}` will return a dataframe where each entry has a value for `age` and the respective mean of `SeriousDlqin2yrs` for applicants' of that age (see Tutorial 10).
2. `send_glm_request()`: This function takes inputs in the form of a reference to the remote data ("data") and the dependent variable ("y"), i.e., `{"data": data, "y": y}`, and outputs the detailed results of a full-fledged logistic regression model without feature selection.
3. `send_logit_request()`: This function is similar to the `send_nn_request()` method, but it performs the remote estimation of a customised logistic regression. It takes inputs

in the form of a dictionary `{"data": data, "test": 0.2, "x": features, "y": y, "scale":"True"}` where:

```
4. a. "data": Reference to the remote data (see below)

5. b. "test": Fraction of the data used for testing

6. c. "x": List of features (independent variables) used by t
      he model

7. d. "y": Target variable (dependent variable) of the model

   e. "scale": Indicator ("True"/"False") for scaling
```

Predefined variables

For your convenience, we have predefined certain variables which you should reuse for this assignment:

1. `data`: The remote Equifax dataset that should be used in this assignment
2. `target`: The target variable (default indicator over 12 months)
3. `all_features`: The complete list of available features

Note: Please do not change the values of these variables.

In [51]:

```
# the merged Equifax dataset
data = "assignment5"

# the target (dependent) variable of interst (good/bad flag): 0 means no default within 12 months / 1 means default
target = 'Commercial_GBF_12m'

# the complete list of available features as of the loan approval date (see Equifax_Data_Dictionary.xlsx for details)
all_features = ['EFX_Comp_ID',
 'loanAmt',
 'External_Admin',
 'Petitions',
 'Writs_and_Summons',
 'Writs_and_Summons_Value',
 'Writs_and_Summons_LT_12M',
 'Writs_and_Summons_LT_12M_Value',
 'Writs_and_Summons_GT_12M',
 'Writs_and_Summons_GT_12M_Value',
 'Judgements',
 'Judgements_Value',
 'Judgements_LT_12M',
 'Judgements_LT_12M_Value',
 'Judgements_GT_12M',
 'Judgements_GT_12M_Value',
 'Directors',
 'Defaults',
 'Defaults_Value',
 'Defaults_12',
 'Defaults_12_Value',
 'Defaults_GT_12M',
 'Defaults_GT_12M_Value',
 'Telco_Defaults_LT_12M',
 'Telco_Defaults_LT_12M_Value',
 'Utility_Defaults_LT_12M',
 'Other_Defaults_LT_12M',
 'Other_Defaults_LT_12M_Value',
 'Defaults_Paid',
 'Defaults_Paid_Value',
 'Defaults_Unpaid',
 'Defaults_Unpaid_Value',
 'Credit_Enqry',
 'Credit_Enqry_Value',
 'Credit_Enqry_LT_12M',
 'Credit_Enqry_LT_12M_Value',
 'Credit_Enqry_GT_12M',
 'Credit_Enqry_GT_12M_Value',
 'Broker_Enqry',
 'Broker_Enqry_Value',
 'Broker_Enqry_LT_12M',
 'Broker_Enqry_LT_12M_Value',
 'Broker_Enqry_GT_12M',
 'Broker_Enqry_GT_12M_Value',
 'Mercantile_Enqry_LT_12M',
 'Mercantile_Enqry',
 'Mercantile_Enqry_GT_12M',
 'Mercantile_Enqry_GT_12M_Value',
 'ny7513_df_6m',
 'ny7514_df_12m',
 'ny7516_df_60_84m',
 'ny7517_df_tcut_6m',
 'ny7518_df_tcut_12m',
 'ny7520_df_tcut_60_84m',
 'ny7568_df_sts_unpd_60_84m',
 'ny7585_df_origamt_tcut_60_84m',
 'ny7586_df_origamt_60_84m',
 'ny7587_df_latamt_12m',
 'ny7588_df_latamt_60_84m',
 'ny7589_df_time_1',
 'ny7591_df_s_1_60_84m',
 'ny7601_adv_48_84m',
 'ny7999_enq_7d',
 'ny8000_enq_1m',
 'ny8001_enq_3m',
 'ny8002_enq_6m',
 'ny8003_enq_12m',
 'ny8006_enq_60m',
 'ny8028_enq_rm_1m',
 'ny8029_enq_rm_3m',
 'ny8030_enq_rm_6m',
 'ny8031_enq_rm_12m',
 'ny8034_enq_rm_60m',
 'ny8042_enq_tcut_1m',
 'ny8043_enq_tcut_3m',
 'ny8044_enq_tcut_6m',
 'ny8045_enq_tcut_12m',
 'ny8048_enq_tcut_60m',
 'ny8049_enq_own_3m',
 'ny8050_enq_own_12m',
 'ny8056_enq_amt_1',
 'ny8057_enq_amt_2',
 'ny8059_enq_time_1',
 'ny8060_enq_time_2',
 'ny8062_enq_amt_3m',
 'ny8063_enq_amt_60m',
 'np7504_dj_60m',
 'np7505_dj_time_1',
 'np7506_dj_out_amt_60m',
 'np7508_dj_out_60m',
 'np7509_wr_48m',
 'np7510_wr_out_48m',
 'np7511_wr_60m',
 'np7512_wr_out_60m',
 'np8500_dr_cur',
 'np8501_dr_time_max_cur',
 'np8502_dr_prev_60m',
 'np8508_pr_cur',
 'np8509_dr_prev_120m_ever',
 'na8905_ntb_flg',
 'na8902_age_fle_max',
 'na8904_age_fle',
 'na8908_age_ind',
 'na8920_em_time_1',
 'na8921_ad_time_1',
 'ny8056_enq_amt_1_is_1']
```

Please run the following cell to import the required libraries.

In [50]:

```
# Usual imports
import warnings
warnings.simplefilter(action='ignore', category=FutureWarning)

# Essential libraries for this assignment
from finml import *
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
%matplotlib inline
import nest_asyncio
from tfclient import send_nn_request, send_logit_request, send_glm_request, send_grouped_mean_request
nest_asyncio.apply()

# Suppress warnings for deprecated methods from TensorFlow
import tensorflow as tf
tf.compat.v1.logging.set_verbosity(tf.compat.v1.logging.ERROR)
```

In your assessment, please address the following questions.

Question 1 (2 marks)

Write a Python code that creates two bar plots of average default rates (`Commercial_GBF_12m`) depending on (i) whether a company was under external administration (`External_Admin`) or (ii) had filed petitions (`Petitions`). Make sure your plots' axes are appropriately labelled.

Answer 1

In [62]:

```
send_grouped_mean_request(data='Equifax_Data_Dictionary.xlsx',var=['External_Admin','Petitions'], y= 'Commercial_GBF_12m')
```

```
---------------------------------------------------------------------------
TypeError                                 Traceback (most recent call last)
<ipython-input-62-d2ecb1bb8591> in <module>()
----> 1 send_grouped_mean_request(data='Equifax_Data_Dictionary.xlsx',var=['External_Admin','Petitions'], y= 'Commercial_GBF_12m')
```

```
TypeError: send_grouped_mean_request() got an unexpected keyword argument '
data'
```

Question 2 (2 marks)

Write a Python code that creates two plots of average default rates (`Commercial_GBF_12m`) as a function of (i) the number of months since a director's last commercial default (`ny7589_df_time_1`) and (ii) the frequency of adverse commercial events over four years 48 months prior to application (`ny7601_adv_48_84m`). Make sure your plots' axes are appropriately labelled.

Answer 2

In [48]:

```
target=data['Commercial_GBF_12m']
```

```
---------------------------------------------------------------------------
TypeError                                 Traceback (most recent call last)
<ipython-input-48-8a5fc36e2e38> in <module>()
----> 1 target=data['Commercial_GBF_12m']

TypeError: string indices must be integers
```

Question 3 (1 mark)

How do you interpret the above plots from Questions 1 and 2? What is your conclusion?

Answer 3

YOUR ANSWER HERE

Question 4 (2 marks)

Run a full-fledged logistic regression model without any ex-ante feature selection. Based on the estimation output, select and report all features that are significant at the 5%-level (or below).

Note: To increase the stability of the estimation, Python will automatically omit certain variables.

Answer 4 - Code

In [ ]:

```
"""Write your code in this cell"""
# YOUR CODE HERE
pass
```

Answer 4 - Text

YOUR ANSWER HERE

Question 5 (3 marks)

Run a logit model using the function `send_logit_request()` and applying the following specifications:

1. Relative size of test data: 20%
2. Only use the features from Question 4 with a significance level below 5%
3. Scaling: "True"

Evaluate the testing performance of your logit model.

Answer 5 - Code

In [ ]:

```
"""Write your code in this cell"""
# YOUR CODE HERE
pass
```

Answer 5 - Text

YOUR ANSWER HERE

Question 6 (4 marks)

Write a Python code that estimates a series of full-fledged neural networks with the following specifications:

1. Number of layers: 1
2. Number of units: 2, 4, 16, 64, 256
3. Relative size of test data: 20%
4. Scaling: "True"

Generate one plot that shows each model's ROC ("roc"), both for testing and training. What is your conclusion?

Answer 6 - Code

In [ ]:

```
"""Write your code in this cell"""
# YOUR CODE HERE
pass
```

Answer 6 - Text

YOUR ANSWER HERE

Question 7 (2 marks)

Based on the testing performance of the above five neural network model, which one would you pick and why? Rerun the estimation of your chosen model.

Answer 7 - Text

YOUR ANSWER HERE

Answer 7 - Code

In [ ]:

```
"""Write your code in this cell"""
# YOUR CODE HERE
pass
```

Question 8 (3 marks)

Conduct an in-depth comparison between the "simple" logit model (Question 5) and your preferred neural network (Question 7). What are their respective potential advantages and disadvantages? If you were to run a credit scoring agency, which type of model you think your clients would prefer?

Answer 8

YOUR ANSWER HERE

Question 9 (3 marks)

The average loan amount across the Equifax sample is $75,000. Furthermore, let us make the following simplifying assumptions:

1. The interest rate charged for each loan under the simple logit model (Question 5) is 5% p.a.
2. Each loan has a duration of one year
3. If a loan defaults, the total amount is lost (zero recovery) and no interest payments occur
4. A loan application only gets granted, if the respective model predicts no default
5. Each granted loan generates administrative costs of 1% p.a.

You are running a business that lends loans of $75,000 to small companies. Based on the above **testing data**, what is the interest rate implied by your chosen neural network (Question 7), such that you will generate the same net income as under the simple logit model?

Note: For your calculations, you can neglect any time value of money effects.

Answer 9

YOUR ANSWER HERE

Question 10 (1 mark)

When you go through the dictionary provided in the file `Equifax_Data_Dictionary.xlsx`, you will notice that Equifax uses primarily legal data rather than accounting data to predict defaults. Why do you think that is?

Answer 10

YOUR ANSWER HERE

Question 11 (2 marks)

Discuss the pros and cons of using deep learning, i.e., hierarchical machine learning applied to big data, in the context of credit scoring.

Answer 11

YOUR ANSWER HERE