MACHINE LEARNING
EN FINANZAS
Problem Set 3

INTRODUCCIÓN

Has sido designado asesor de la Fed. Tu primera tarea asignada consiste en estimar el total de activos en riesgo del sistema financiero ante eventuales defaults de las entidades financieras. La Fed debe estar preparada para inyectar liquidez en tal magnitud.

Disponés de una base de datos de la FDIC con información contable relevante conjuntamente con el dato si la entidad fue a default o no 1 año después de la publicación de los mismo. Asimismo también contás con la misma información a hoy para las entidades actuales del sistema.

REQUISITOS

1. Implementar una clase `AnalistaDeRiesgo` que cumpla con las siguientes características:
   a. Aceptar en su constructor dos parámetros: **a)** una lista con tuplas (modelo, configuración para CV de los parámetros) que será utilizado para realizar la predicción, y **b)** el criterio de scoring a utilizarse.
   b. Poseer un método `load_data` que:
      i. Acepte dos parámetros: **a)** un `pandas.Dataframe` con la información histórica de las entidades, **b)** el nombre del campo indicador de default.
      ii. No retorne nada
   c. Poseer un método `get_report` que:
      i. Acepte dos parámetros: **a)** un `pandas.Dataframe` con la información de las entidades actuales, y **b)** una lista con los campos a utilizar como features.
      ii. Deberá retornar en la consola:

```
Entidades en riesgo de default = x.xx%
Total de activos del sistema (USD B): x,xxx,xxx.xx
Porcentaje de activos en riesgo de default: x.xx%
```

      iii. Este método debe entrenar por CV los modelos ingresados en el constructor, seleccionar el mejor modelo de acuerdo al scoring predefinido, y desarrollar la predicción implementando el workflow habitual en ML.
2. La rutina de testing del módulo deberá:
   a. Importar en `pandas.DataFrame` el dataset: 'central_bank_data.h5'. Los datos históricos están almacenados bajo el identificador 'bank_defaults_FDIC', y los datos de las entidades actuales en 'regulated_banks'
   b. Configurar nuestro `AnalistaDeRiesgo` con los modelos (a la derecha de la flecha se indica la configuración a usar para buscar el mejor):
      `KNN` -> `K` enteros entre 3 y 30
      `SVC` -> `C` debe ser 1, 10, 100, 500 o 1000; gamma será 'scale', y utilizar kernels 'linear' y 'Gaussian RBF'
      `Tree` -> cantidad mínima de observaciones por división: enteros entre 2 y 15.

NOTAS

1. Las variables a usar como regresores son: 'log_TA', 'NI_to_TA', 'Equity_to_TA', 'NPL_to_TL', 'REO_to_TA', 'ALLL_to_TL', 'core_deposits_to_TA', 'brokered_deposits_to_TA', 'liquid_assets_to_TA', 'loss_provision_to_TL', 'NIM' y 'assets_growth'
2. La columna que indica si está en default o no es 'defaulter'
3. Para el resto de la información requerida explorar el dataset.
4. Para cualquier corrida de CV utilizar 5 repeticiones.
5. Utilizar criterio de scoring el área debajo de la curva ROC.

ÉXITOS!