Trabajo Final

Predecir si un cliente se dará o no de baja de un producto/servicio es un dato fundamental para que las empresas puedan anticiparse a esta situación y tomar decisiones a tiempo. Este trabajo registra datos reales de individuos que poseen productos financieros en un importantísimo banco de Argentina. Se cuenta con 170.722 registros para entrenamiento y validación, con 103 variables que serán utilizadas para predecir si el cliente se dará de baja durante en el transcurso de los dos meses subsiguientes.

El archivo se denomina **Modelo_Clasificacion_Dataset.csv** y sus campos se indican en el excel adjunto denominado **Modelo_Clasificacion_Diccionario_de_Datos.xlsx**

Nota: hay varios datos faltantes en algunas variables. Estos valores faltantes pueden tratarse como un nivel particular de una variable categórica o procesarse mediante técnicas de eliminación o imputación.

Modelo de Predicción

Se deberá presentar los procesos y las estrategias implementadas, tanto de entrenamiento como de validación/selección de hiperparametros, a partir de las cuáles se arribe al modelo de mejor performance predictiva. Sólo para el modelo final seleccionado se deberá presentar el siguiente informe:

I. **Algoritmo utilizado para el modelo final:**
   a. Diseño del entrenamiento del modelo final
      i. Ingeniería de atributos: conversión de variables, procesamiento,normalización/estandarización, imputaciones, selección de variables, componentes ppales, etc.)
      ii. Determinación de Set de entrenamiento y Set de validación. Validación cruzada.
   b. Tipo de Modelo predictivo utilizado y sus parámetros e hiperparámetros
   c. Curva de selección utilizada para determinar los argumentos del pto b.

II. **Performance del modelo final:**
   a. Métrica ROC-AUC
   b. Justificación e interpretación de los resultados

Forma de entrega

1. Presentación del trabajo: archivo Word. Se deberá presentar un informe profesional en el que se expliquen los procesos llevados a cabo, su fundamentación y los resultados a los que se arribe.

2. Archivo completo de desarrollo del modelo : se considerará fundamental la presentación de las funciones y métodos, la definición de variables locales y globales utilizadas y la claridad en la documentación del código (comentarios utilizando el símbolo # antecediendo al texto explicativo)

3. Script o función de python que ejecute las predicciones sobre nuevos datos (a utilizarse para la prueba a realizarse en el siguiente punto, Calificación y "Competencia")

Calificación y "Competencia"

El trabajo será calificado en función de la presentación, justificación y resultados de los puntos I y II del apartado "Modelo de Predicción". A su vez se ha reservado un set de prueba externo de 18872 registros (extraído del dataset original) que no se incluyó en la entrega del material del presente trabajo, sobre el cual se correrá cada uno de los modelos presentados por los grupos, para luego determinar cuál de ellos es el de mejor performance. Este formato del tipo "competencia" es utilizado en la industria para el desarrollo comunitario de modelos, siendo su paradigma las competencias de la plataforma online "Kaggle", entorno en el que se han desarrollado, siguiendo esta metodología, variados reconocidos modelos de machine learning, destacándose el sistema de recomendación de películas de Netflix.

IMPORTANTE:
Tener en cuenta que el script final a entregar, deberá ser capaz de:

- Incorporar los datos de testeo final que están separados y el modelo nunca vió.
- Transformarlos con las mismas operaciones que fueron realizadas sobre los datos del entrenamiento.

La medida de performance utilizada para calificar y posicionar a los trabajos a los trabajos será: **ROC AUC.**