Desafío - Boosting Classifiers

- Para realizar este desafío debes haber estudiado previamente todo el material disponibilizado correspondiente a la unidad.
- Una vez terminado el desafío, comprime la carpeta que contiene el desarrollo de los requerimientos solicitados y sube el `.zip` en el LMS.
- Desarrollo desafío:
    - El desafío se debe desarrollar de manera Individual.
    - Para la realización del desafío necesitarás apoyarte del archivo *Apoyo Desafío - Boosting Classifiers*.

Requerimientos

Para esta sesión trabajaremos con una base de datos sobre rotación de clientes en una compañía de telecomunicaciones. El archivo contiene 3333 registros y 20 atributos. El vector objetivo a modelar es la tasa de rotación entre los clientes de una compañía de telecomunicaciones `churn`. Los atributos existentes hacen referencia a características de la cuenta de cada cliente.

Lista de atributos:

- **State:** Estado de Estados Unidos.
- **Account Length:** Tiempo en que la cuenta ha sido activada.
- **Area Code:** Código de área.
- **International plan:** Plan internacional activado.
- **Voice mail plan**: Plan de mensajes de voz activado.
- `number_vmail_messages`: Cantidad de mensajes de voz.
- `total_day_minutes`: Cantidad de minutos ocupados en la mañana.
- `total_day_calls`: Cantidad de llamadas realizadas en la mañana.
- `total_day_charge`: Cobros realizados en la mañana.
- `total_eve_minutes`: Cantidad de minutos ocupados en la tarde.
- `total_eve_calls`: Cantidad de llamadas realizadas en la tarde.
- `total_eve_charge`: Cobros realizados en la tarde.
- `total_night_calls`: Cantidad de llamadas realizadas en la noche.
- `tota_night_minutes`: Cantidad de minutos ocupados en la noche.
- `total_night_charge`: Cobros realizados en la noche.
- `total_intl_minutes`: Cantidad de minutos ocupados en llamadas internacionales.
- `total_intl_calls`: Cantidad de llamadas internacionales realizadas.
- `total_intl_charge`: Cobros realizados por llamadas internacionales.
- `churn`: 1 si el cliente se cambió de compañía, 0 de los contrario.

Los datos provienen del paquete `AppliedPreditiveModeling` de R.

Ejercicio 1: Preprocesamiento

- Grafique el comportamiento distributivo de los atributos y de la variable dependiente. Reporte brevemente el comportamiento de las variables.
- En base al comportamiento de los atributos, considere si es necesario implementar alguna recodificación o transformación de atributo. Algunas normas a seguir:
    - Para las variables categóricas, recodifíquelas en variables binarias.
    - Para aquellas variables numéricas que presenten alto sesgo, pueden transformarlas con su logaritmo.

Ejercicio 2: Comparación de AdaBoost y Gradient Boosting

- Entrene los clasificadores AdaBoost y Gradient Boosting para mejorar su capacidad predictiva en la medida de lo posible. Para ello, implemente una búsqueda de grilla con las siguientes especificaciones:

Modelo	**Grilla**
AdaBoostClassifier	{'learning_rate': [0.01, 0.1, 0.5], 'n_estimators': [50, 100, 500, 1000, 2000]}
GradientBoostingClassifier	{'learning_rate': [0.01, 0.1, 0.5],'n_estimators': [50, 100, 500, 1000, 2000], 'subsample': [0.1,0.5,0.9]}

- Si el tiempo de computación es alto, puede implementar la búsqueda con 1 validación cruzada.
- Reporte las métricas para los mejores modelos.

Ejercicio 3: Principales factores asociados

- Con el mejor modelo, reporte la importancia relativa de los atributos y comente cuáles son los que aumentan la probabilidad de fuga en los clientes.

    Tip: *Pueden implementar la función* `plot_importance` *que se encuentra en la lectura de Bagging y Random Forest.*

Ejercicio 4: Probabilidad de fuga

- El gerente general necesita saber en qué estados hay una mayor probabilidad de fuga de clientes. Para ello, identifique los tres estados con una mayor probabilidad de fuga.
- Implemente el modelo predictivo con el archivo `churn_test.csv`.
- Recuerde que para obtener la probabilidad de clase, debe utilizar la función `predict_proba` del modelo.