Evaluación de seguros de vida para Prudential

¿Puedes facilitar la compra de un seguro de vida?

Miembros del grupo

- Mauricio Aguiar Gil, CC 8162210 Ingeniería de Sistemas.
- Carlos Zapata Arango, CC 1044504411, Ingeniería de Sistemas.

1. **Describir el dataset que se va a utilizar**

   El dataset fue creado por la compañía Prudential y está enfocado en la evaluación de asignación de seguros de vida. Este conjunto de datos contiene más de 100 variables que describen atributos de los solicitantes de seguros de vida, y la variable objetivo a predecir es "Response", que es una medida ordinal de riesgo con 8 niveles, usada para decidir el nivel de aseguramiento del solicitante. El conjunto de datos está dividido en dos archivos: el archivo de entrenamiento "train.csv" y el archivo de prueba "test.csv".

   El archivo de entrenamiento contiene 128,655 instancias y 127 variables, mientras que el archivo de test contiene 55,362 instancias y 126 variables (la variable "Response" no está incluida en el archivo de test ya que es la variable a predecir). Además, el conjunto de entrenamiento incluye la variable objetivo "Response" y se puede utilizar para entrenar y validar modelos de aprendizaje automático.

   Es importante tener en cuenta que el conjunto de datos tiene una gran cantidad de variables categóricas, por lo que se requerirá la aplicación de técnicas de codificación de variables categóricas antes de poder entrenar los modelos. También es importante considerar que el conjunto de datos tiene una cantidad significativa de valores faltantes que deberán ser abordados antes de entrenar los modelos.

2. **Describir el problema predictivo a resolver**

   El problema es predecir la variable "Response" para cada solicitante en el conjunto de prueba. “Response” es una medida ordinal de riesgo que tiene 8 niveles; a partir de esto se puede concluir que este es un problema de clasificación multiclase.

3. **Mencionar las métricas de desempeño requeridas, tanto de machine learning como del negocio, si es el caso.**

- **Exactitud (accuracy**): mide la proporción de muestras clasificadas correctamente por el modelo.
- **F1-score:** es una medida de precisión y exhaustividad, que combina la precisión y el recall en una sola métrica.

- **Matriz de confusión**: es una tabla que muestra el número de muestras clasificadas correcta e incorrectamente por el modelo para cada clase, lo que permite identificar posibles desequilibrios de clases y errores de clasificación específicos.

4. **Mencione un primer criterio sobre cuál sería el desempeño deseable en producción**
   Dado que se trata de una compañía vendedora de seguros, está en su interés asegurarse de que no está asignando incorrectamente los niveles de riesgo a los solicitantes de seguros pues ello representaría grandes pérdidas materializadas en el pago de seguros a quienes no lo requieran. Se considera que una métrica de accuracy inferior al 10% es ideal. A grandes rasgos, se espera que el modelo tenga una precisión (accuracy) y un recall altos, indicativos de que se tiene un mínimo costo de errores de clasificación.