UNIVERSIDAD DE ANTIOQUIA
1 8 0 3

Entrega 1 – Introducción a la I.A

Brayan Daniel Oviedo Barreto, Julián Zaque Montoya, Lisset Andrea Zea
c.c 1037667170, 1152713576, 1001813095
Bioingeniería, Facultad de Ingeniería, Universidad de Antioquia
Marzo 01, 2023

- Descripción del problema predictivo

El problema predictivo que se va a resolver es predecir si un cliente de Santander realizará una transacción específica o no en función de su historial de transacciones pasadas y otros atributos relacionados.

- Dataset utilizado

El dataset a utilizar es el conjunto de datos de la competición "Santander Customer Transaction Prediction" alojado en Kaggle (Santander Customer Transaction Prediction | Kaggle). Este conjunto de datos contiene 200,000 instancias de entrenamiento y 200,000 instancias de prueba, y consta de 200 características numéricas (columnas) anónimas que representan los atributos de los clientes y su historial de transacciones.

El dataset proporcionado por la competición de Kaggle "Santander Customer Transaction Prediction" consta de dos archivos CSV:

1. ***train.csv:*** Este archivo contiene los datos de entrenamiento, que incluyen la variable objetivo "target" (0 o 1), que indica si el cliente ha realizado o no una transacción, así como otras 200 variables numéricas y categóricas. Este archivo contiene 200,000 filas y 202 columnas, incluyendo la columna ID que identifica de forma única a cada cliente.
2. ***test.csv:*** Este archivo contiene los datos de prueba, que incluyen las mismas 200 variables numéricas y categóricas que en el archivo de entrenamiento, pero no incluye la variable objetivo "target". Este archivo contiene 200,000 filas y 201 columnas, incluyendo la columna ID.

- Métricas de desempeño requeridas

Como métrica de machine learning, se utilizará el área bajo la curva ROC (AUC-ROC) para evaluar el rendimiento del modelo en el desarrollo del proyecto. Además, se requerirá una métrica de negocio que tenga en cuenta los costos y beneficios asociados con las predicciones del modelo. Dado que no se dispone de información específica sobre los costos y beneficios en esta competición, se utilizará la precisión (accuracy) como una métrica de negocio provisional.

La AUC-ROC mide la capacidad del modelo para distinguir entre las dos clases de la variable objetivo (clientes que han realizado una transacción y clientes que no han realizado una transacción). La curva ROC representa la tasa de verdaderos positivos (TPR) en función de la tasa de falsos positivos (FPR), mientras que el área bajo la curva (AUC) es una medida de la capacidad del modelo para discriminar entre las dos clases.

La fórmula para calcular la AUC-ROC es la siguiente:

$$AUC - ROC = \int_0^1 TPR(FPR) * dFPR$$

Donde TPR es la tasa de verdaderos positivos y FPR es la tasa de falsos positivos. La integral se calcula para los valores de FPR que van de 0 a 1.

- ## Desempeño deseable en producción

El desempeño deseable en producción se determinará en función de la métrica de negocio que se elija. Para esta competición, la precisión (accuracy) será utilizada como métrica provisional de negocio. En términos generales, se espera que el modelo tenga una precisión lo suficientemente alta como para justificar su uso en la toma de decisiones empresariales. Dado que no se dispone de información específica sobre el costo de los errores de predicción en esta competición, un primer criterio sobre cuál sería el desempeño deseable en producción podría ser alcanzar una precisión del 80%. No obstante, es importante tener en cuenta que esta es solo una cifra provisional y que la precisión óptima real dependerá de los costos y beneficios específicos asociados con las predicciones del modelo.

Referencias

- Santander Customer Transaction Prediction | Kaggle. (2023). Retrieved 12 March 2023, from https://www.kaggle.com/competitions/santander-customer-transaction-prediction/overview/description
- Receiver operating characteristic - Wikipedia. (2023). Retrieved 12 March 2023, from https://en.wikipedia.org/wiki/Receiver_operating_characteristic