Proyecto de Introducción a la Inteligencia Artificial

Entrega 1

Cristian Camilo Julio Mejía

Universidad de Antioquia

Introducción a la Inteligencia Artificial

Raúl Ramos Pollan

Septiembre

2023

1. **Planteamiento del Problema**

   A la hora de contratar algún servicio público o privado, las empresas suelen adquirir información de las personas sobre algunos datos personales y residenciales. Es por ello, que en este proyecto se hablará sobre una compañía de telecomunicaciones ficticia, la cual ofrece en esta oportunidad servicio de internet. Se desea predecir si un cliente abandonará o no en función de su comportamiento e información demográfica. Cuando en la información suministrada se habla de La rotación (Churn), se refiere a cuando un cliente deja de utilizar los servicios de la empresa de telecomunicaciones.

2. **Descripción del DataSet**

   La base de dato a analizar fue adquirida a través de un Dataset alojado en Kaggle (**Telco customer churn: IBM dataset**), que tiene 7043 filas y 33 columnas. Este conjunto de datos contiene información sobre los clientes de una empresa de telecomunicaciones, incluida su información demográfica, los servicios a los que se han suscrito y si han abandonado o no.
   El documento que contiene los datos es un archivo en Excel.

   La información es la siguiente:

   CustomerID: ID único que identifica a cada cliente.
   Gender: El género del cliente: Masculino, Femenino
   Senior Citizen: Indica si el cliente tiene 65 años o más: Sí, No
   Tenure Months: Indica la cantidad total de meses que el cliente ha estado con la empresa al final del trimestre especificado anteriormente.
   Phone Service: Indica si el cliente contrata el servicio de telefonía residencial con la empresa: Sí, No
   Multiple Lines: Indica si el cliente contrata múltiples líneas telefónicas con la empresa: Sí, No
   Internet Service: Indica si el cliente contrata servicio de Internet con la empresa: No, DSL, Fibra Óptica, Cable.
   Contract: Indica el tipo de contrato actual del cliente: Mes a Mes, Un Año, Dos Años.
   Churn Label: Sí = el cliente dejó la empresa este trimestre. No = el cliente permaneció en la empresa. Directamente relacionado con el valor de abandono.

Churn Value: 1 = el cliente dejó la empresa este trimestre. 0 = el cliente permaneció en la empresa. Directamente relacionado con Churn Label.
Churn Score: un valor de 0 a 100 que se calcula utilizando la herramienta predictiva IBM SPSS Modeler. El modelo incorpora múltiples factores que se sabe que causan deserción. Cuanto mayor sea la puntuación, más probabilidades habrá de que el cliente abandone.
CLTV: Valor de vida del cliente. Un CLTV previsto se calcula utilizando fórmulas corporativas y datos existentes. Cuanto mayor sea el valor, más valioso será el cliente. Se debe monitorear la deserción de los clientes de alto valor.
Churn Reason: motivo específico de un cliente para abandonar la empresa. Directamente relacionado con la categoría de abandono.

Y otras características, como el lugar de residencia del cliente, la ciudad, el país, si tiene a alguien a cargo, el tipo de servicio elegido, etc.

3. Métricas

Hay varias métricas de rendimiento que podemos usar para evaluar nuestro modelo de aprendizaje automático, como exactitud, precisión, sensibilidad y puntuación F1. El modelo de predicción debería tener una tasa de aciertos del 0.7 o mayor; su tasa de precisión deberá ser superior al 0.6, y de este modo determinar si la predicción dada fue la correcta; el modelo a predecir ha de tener una sensibilidad de 0.8 o superior para determinar si en realidad identificó correctamente a los clientes que abandonarían la compañía, además, si el F1 Score es $\geq$ al 0.7 indicaría que hay un buen equilibrio entre la precisión y la sensibilidad lo que determinaría un buen desempeño en el modelo. Por otro lado, también podemos considerar métricas comerciales como el costo de los falsos positivos (es decir, predecir incorrectamente que un cliente abandonará) y el costo de los falsos negativos (es decir, predecir incorrectamente que un cliente no abandonará) Estos costos se pueden utilizar para calcular el valor esperado del modelo y determinar si vale la pena implementarlo en producción.

4. Desempeño

El rendimiento deseado en producción dependerá del contexto empresarial específico y de los costos asociados con los falsos positivos y falsos negativos. En general, nos gustaría que nuestro modelo tuviera alta exactitud, precisión y sensibilidad, al mismo tiempo que minimizara los costos asociados con predicciones incorrectas. Podemos utilizar técnicas como el aprendizaje sensible a los costos o el ajuste de umbrales para optimizar nuestro modelo para el rendimiento deseado en producción.

5. Referencia

IBM. (23 de Mayo de 2020). *Kaggle*. Obtenido de Kaggle: https://www.kaggle.com/datasets/yeanzc/telco-customer-churn-ibm-dataset