MINERIA DE
DATOS

Andrei Alexandru Miu

Índice

Ejercicio sobre Minería de Datos y Aprendizaje Automático

Situación:

Una empresa de telecomunicaciones desea reducir la tasa de abandono de clientes. Para ello, decide aplicar técnicas de minería de datos y aprendizaje automático para predecir qué clientes tienen una alta probabilidad de cancelar su servicio en los próximos 3 meses.

Parte 1: Preparación de los Datos

1. ¿Cuáles son las principales fases del preprocesado de los datos? Menciona y explica al menos 3 técnicas utilizadas en esta etapa.

- **Limpieza de datos**: Eliminación de los valores nulos, duplicados o inconsistentes.

- **Normalización y escalado**: Implica tener rangos de valores muy diferentes, hay que aplicar la normalización Min-Max.

- **Codificación de las variables categóricas**: Hay que transformar las variables categóricas en numéricas, utilizando técnicas como one-hot encoding o label encoding.

2. ¿Por qué es importante el muestreo en el preprocesamiento de datos? Explica una situación en la que sería recomendable usar muestreo estratificado en este caso.

El muestreo es importante por que ayuda a reducir el costo computacional del entrenamiento del modelo.

En este caso, un muestreo estratificado sería recomendable si la cantidad de clientes que abandonan el servicio es mucho menor que la de clientes que se quedan.

Parte 2: Modelización

3. ¿Qué diferencia hay entre un modelo supervisado y uno no supervisado? ¿Cuál crees que es más adecuado para este problema? Justifica tu respuesta.

- **Modelo supervisado**: Tiene etiquetas (1/0).
- **Modelo no supervisado**: No tiene etiquetas (busca patrones).

Creo que el más adecuado sería el modelo supervisado, ya que disponemos de un histórico de clientes con su estado de abandono, lo que permite así entrenar los modelos para predecir la variable.

4. Considerando que queremos clasificar a los clientes en dos grupos (probable abandono o retención), ¿qué tipo de modelo supervisado recomendarías? Explica por qué.

El modelo de random forest creo que sería el más adecuado, ya que ofrece interpretabilidad y funciona bien para datos numéricos y categóricos.

5. En un árbol de decisión, ¿qué criterio se utiliza para dividir los datos en diferentes nodos? Explica brevemente el concepto de "pureza" en este contexto.

Los arboles de decisión funcionan y se dividen a través de una métrica.
La pureza se refiere a cual homogéneos sean los datos dentro de cada nodo.

Parte 3: Evaluación y Producción

6. Se ha entrenado un modelo de clasificación y se obtiene una matriz de confusión. ¿Cuáles son las métricas clave para evaluar su desempeño? Define **precisión**, **recall** y **exactitud**.

- **Precisión**: Indica cuantos clientes clasificados como “abandono” realmente lo son.

- **Recall**: Indica cuantos de los clientes que abandonan fueron identificados correctamente.

- **Exactitud**: Indica la proporción total de predicciones correctas sobre el total de los casos.

7. Una vez entrenado el modelo, se quiere poner en producción. ¿Qué método recomendarías para aplicar la inferencia en tiempo real en una plataforma web? Explica brevemente.

Se podría utilizar APIs, donde la plataforma envíe y reciba los datos en tiempo real, o usar un modelo desplegado en la nube como podría ser el de AWS.