Logotipo, nombre de la empresa

Descripción generada automáticamente

Universidad de Montevideo – Introducción a la Ciencia de Datos

Examen 2023

30 de Enero, 2023

La directiva de la cadena de gimnasios más grande del país definió en el último trimestre disminuir la tasa de cancelación de sus usuarios. Como una de las primeras medidas contratan a un equipo de científicos de datos, entre los que se encuentra usted, para crear un modelo de machine learning para determinar la probabilidad de cancelación de sus afiliados y así dirigir accionables que permitan alcanzar el mencionado objetivo.

Al comenzar a trabajar le es entregado a usted las siguientes fuentes de datos:

A – Datos relativos a los usuarios conteniendo:

* Documento de identidad del usuario.
* Número identificatorio del usuario
* Fecha de nacimiento del usuario.
* Dirección de residencia del usuario.
* Fecha de inscripción del usuario.
* Variable booleana indicando si el usuario proviene de algún plan empresarial.
* Último nivel educativo del usuario.
* Forma de pago del usuario
* Plan al que el usuario está suscripto.

B – Datos de ingreso / egreso de las insttalaciones de la empresa conteniendo:

* Número identificatorio del usuario
* Fecha y hora del registro
* Variable booleana que indica si el registro se trata de un ingreso o egreso.

1- Dentro del contexto del campo de machine learning, ¿a que tipo de problema se está enfrentando? ¿Cuál es la variable objetivo?

2 – Seleccionar 5 variables de la fuente de datos A y describir si se trata de una variable nominal, ordinal, intervalo o ratio.

3-Crear, mediante *Feature engineering,* 4 nuevas variables a partir de las fuentes de datos A y B.

4 – Proponga un modelo para resolver el problema planteado y descríbalo.

5 – Luego de puesto en productivo el modelo este comienza a producir probabilidades de cancelación de forma mensual. La directiva de la empresa plantea realizar comunicaciones a los usuarios de mayor cancelación; una vez definido un medio y formato de comunicación este se prueba mediante un experimento.

Los resultados del experimento arrojaron lo siguiente:

* Tasa de cancelación en población contactada: 7%
* Tasa de cancelación en población control: 5%

Tomando los tamaños de la población control y comunicada se realizó un test de hipótesis que arrojó un **p-value** de **0.003.**

Reflexione sobre el resultado del experimento.

6 – TFIDF es una técnica estadística que se usa con frecuencia en dominios de aprendizaje automático, como en la clasificación de textos. TFIDF mide la relevancia de una palabra en un documento en comparación con todo el corpus de documentos. Dispone de un corpus (D) que contiene los siguientes documentos:

* 1. Documento 1 (d1): “A quick brown fox jumps over the lazy dog. What a fox!”
  2. Documento 2 (d2): “A quick brown fox jumps over the lazy fox. What a fox!”

Cual de las siguientes opciones considera correcta? Justifique:

1. Usando TFIDF, la palabra "fox" es igualmente relevante tanto para el documento d1 como para el documento d2.
2. Usando TFIDF, la palabra "fox" es más relevante para el documento d2 que para el documento d1.
3. Usando TFIDF, la palabra "fox" es más relevante para el documento d1 que para el documento d2.
4. Se ha proporcionado información insuficiente para calcular TFIDF.

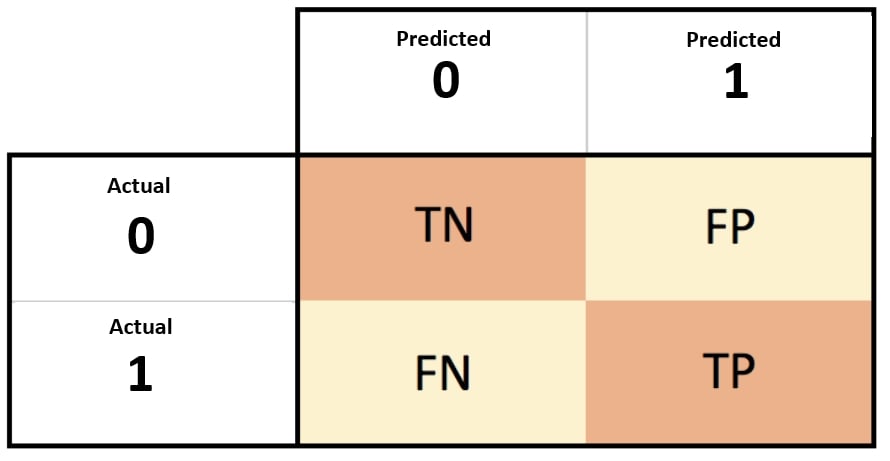
Para el IDF usar:

7 – Suponga que usted tiene un dataset tabular en el cual existe una columna categórica con valores faltantes y dos columnas continuas sin valores faltantes. Usted quiere imputar la columna con valores faltantes pero no puede imputarla utilizando mecanismos simples como la moda porque es imprescindible imputarla con la mayor precisión posible.

Cómo podría usted imputar la columna utilizando aprendizaje supervisado?

8 – Suponga que usted tiene un dataset de fraudes bancarios. Este dataset contiene 10.000 filas, de las cuales 9900 corresponden a casos donde no hubo fraude y 100 casos donde si hubo fraude. Usted entrena un model para clasificar fraudes y tras evaluarlo usted ve que el modelo tiene un *Accuracy = 99%* y un *Recall = 0%.*

Como interpreta estos resultados? Es un buen modelo?



9 – Que es la validación de resustitución? Suponga que usted entrenó un modelo utilizando este mecanismo de validación. Que cree que ocurrirá con la performance del modelo cuando se enfrente a datos reales? Qué otro mecanismo de validación conoce que reflejará mejor la performance real?

10 - Outliers:

* 1. ¿Qué formas de detectar outliers conoce?
  2. Una vez detectados, ¿qué tratamiento debe darles?