UNIVERSIDAD ADOLFO IBÁÑEZ

Fundamentos de Ciencia de Datos – Prueba 02

Fecha de publicación: 11 de noviembre de 2021

Objetivo

El objetivo de esta prueba es construir y validar un modelo predictor sobre un conjunto de datos reales, aplicando los conceptos básicos del área de Machine Learning utilizando el lenguaje Python.

Resumen

Esta prueba es una continuación de la prueba anterior, en la que tuviste que encontrar relaciones en los dantos mediante técnicas de visualización y exploración de datos. Ahora tienes que entrenar y validar un algoritmo de *Machine Learning*. A grandes rasgos, vas a tener que:

- Pre-procesar los datos para que un algoritmo los pueda consumir.
- Escoger las mejores variables para entrenar un modelo de *Machine Learning*.
- Encontrar relaciones mediante técnicas de reducción de dimensionalidad.
- Entrenar y validar un modelo de *Machine Learning*.
- Explicar los resultados

Descripción

El *dataset* utilizado en esta tarea corresponde a los registros de los clientes de un banco. Este banco está preocupado porque muchos de sus clientes han cerrado sus cuentas, y por lo mismo, buscan analizar los datos para entender qué factores influyen en la fuga de clientes. En esta primera prueba, vas a explorar las columnas del *dataset* y vas a encontrar relaciones entre las mismas.

Los datos están en el archivo **BankData.csv**. Cada fila contiene información sobre un cliente en particular, como por ejemplo el identificador del cliente, si sigue siendo cliente del banco o ya cerró su cuenta, un resumen de las interacciones que ha tenido con el banco, entre otros datos. En concreto, las columnas son las siguientes:

- **CLIENTNUM**: identificador único del cliente.

- **Attrition_Flag**: si el cliente sigue en el banco (**Existing Customer**) o abandonó el banco (**Attrited Customer**).
- **Customer_Age**: edad del cliente.
- **Gender**: género del cliente.
- **Dependent_count**: número de personas que dependen financieramente del cliente.
- **Education_Level**: nivel de educación del cliente.
- **Marital_Status**: estado civil del cliente.
- **Income_Category**: rango del salario del cliente.
- **Card_Category**: tipo de la tarjeta de crédito.
- **Months_on_book**: número de meses que el cliente lleva en el banco.
- **Total_Relationship_Count**: número de productos que tiene el cliente en el banco.
- **Months_Inactive_12_mon**: número de meses inactivos durante los últimos 12 meses.
- **Credit_Limit**: cupo total de crédito de la cuenta del cliente.
- **Total_Revolving_Bal**: cupo utilizado aún no pagado por el cliente.
- **Total_Trans_Amt**: monto total de todas las transacciones del cliente en los últimos 12 meses.
- **Total_Trans_Ct**: número de transacciones total realizadas por el cliente en los últimos 12 meses.

Para esta entrega vas a tener que utilizar el análisis de los datos realizados en tu entrega anterior. Tomando como base ese análisis, se pide lo siguiente:

a. **Procesamiento de los datos**: en esta entrega vas a tener que procesar los datos. Esto implica hacer transformaciones de las columnas (por ejemplo, logarítmica), estandarización o normalización de las columnas, etc. Además, vas a tener que transformar las columnas categóricas nominales con **one-hot encoding**, y también las variables categóricas ordinales en columnas numéricas. Debes indicar todas las transformaciones y procesamiento que hagas sobre tus dantos, junto de una breve justificación de por qué hiciste dicha transformación/procesamiento. **El puntaje de esta parte de la prueba es grupal y equivale a 1.5 puntos; además, esto puede que lo necesites hacer más de una vez y no solo al principio de la prueba, por lo que están autorizados a hacer cualquier procesamiento de los datos en grupo.**
b. **Proceso de selección de variables**: también vas a tener que discutir cuáles crees que son las columnas más prometedores para ser utilizadas en tu modelo. Si bien el *dataset* dispone de varias columnas, no siempre es buena idea usarlas todas, sino que en general es mejor las columnas más significativas. Por lo mismo, en base al análisis de la entrega anterior, **discute y justifica** cuáles son las columnas

más relevantes en este caso. Para las columnas transformadas con one-hot encoding solo debes considerar las columnas originales. **El puntaje de esta parte de la prueba es grupal y equivale a 0.5 puntos.**

c. **Reducción de dimensionalidad**: algo habitual al hacer un proyecto del área de Machine Learning es reducir los datos a dos dimensiones para encontrar *clusters* o relaciones de algún tipo. Por lo mismo, se pide que para todas las variables numéricas en el *dataset* original hagas una reducción a dos dimensiones con PCA y con T-SNE; en el caso de PCA también se espera que discutas el porcentaje de varianza acumulada por componente. Luego, haz una visualización que permita encontrar alguna relación entre los clientes que abandonan y los que no (e.g. un scatterplot). Comenta si existe alguna relación que puedes apreciar gracias a la reducción de dimensionalidad. **El puntaje de esta parte de la prueba es individual y equivale a 1 punto; además, recuerda que es frecuente hacer pre-procesamiento (e.g. estandarización) antes de reducir dimensiones. Este pre-procesamiento cuenta como puntaje para el apartado a.**

d. **Entrenando (más de una vez) un modelo**: el siguiente paso es entrenar un modelo de *Machine Learning* con al menos dos subconjuntos de las variables del *dataset* que hagan sentido según lo que explicaste con tu grupo en el punto **b**. Para esta prueba vas a tener que usar el modelo **Random Forest** de la librería Scikit Learn. Además, vas a tener que probar con al menos dos configuraciones de los parámetros (por ejemplo, la profundidad máxima). **La nota de esta parte de la prueba es individual y equivale a 1 punto**.

e. **Validando los resultados**: para cada una de las configuraciones de tu modelo (en cuanto a columnas seleccionadas y parámetros del modelo), debes validar los resultados. Se espera que pruebes con todas las métricas vistas en clases, y además en cada caso, debes discutir los resultados de forma crítica. También, queda a tu criterio cómo divides el *dataset* en entrenamiento y prueba, pero debes justificar tu elección. Finalmente, debes discutir cuál es la mejor configuración para este problema en base a las métricas. **La nota de esta parte de la prueba es individual y equivale a 1 punto**.

f. **Explicando los resultados**: para finalizar, debes hacer un resumen de a lo más una página donde expliques a alto nivel el proceso que seguiste, los resultados y tus conclusiones. Tienes que pensar que esto lo van a leer en el banco varias personas, incluido gente que no es experta en Machine Learning. Si quieres, puedes anexar visualizaciones que hagan más entendible tu explicación. **La nota de esta parte de la prueba es individual y equivale a 1 punto**.

La fecha de entrega será el martes 7 de diciembre a las 23:59 a través de Webcursos.

Tienes que entregar un informe con todos tus análisis (ver formato detallado más abajo) de la forma más profesional posible. Piensa que esto lo va a leer un ejecutivo del banco que realmente necesita que lo ayudes a entender por qué se están fugando los clientes.

Para asegurar la participación de cada uno de los estudiantes, la nota de la entrega tiene una parte grupal y una individual. Por lo mismo, en las partes individuales del trabajo, se espera que el análisis y las respuestas sean originales en cada uno de los alumnos del curso. **De esta forma, cualquier informe que infrinja el código de honor de la universidad recibirá las sanciones correspondientes**.

Evaluación del punto **a.**:

0.5 pts: justificación de las transformaciones y del procesamiento.
0.5 pts: transformación de las columnas.
0.5 pts: procesamiento correcto de columnas categóricas.

Evaluación del punto **b.**:

0.5 pts: justificación de la elección de las variables más importantes.

Evaluación del punto **c.**:

0.5 pts: aplica correctamente algoritmos de reducción de dimensionalidad.
0.5 pts: analiza correctamente el resultado de los algoritmos.

Evaluación del punto **d.**:

0.5 pts: el entrenamiento del modelo es consistente con las columnas señaladas anteriormente.
0.5 pts: entrena varias configuraciones del modelo (máxima profundidad, distintos subconjuntos de columnas).

Evaluación del punto **e.**:

0.5 pts: calcula correctamente las medidas de desempeño.
0.5 pts: el análisis de las métricas es consistente.

Evaluación del punto **f.**:

0.5 pts: el reporte de los resultados es consistente con el resto del trabajo.
0.5 pts: el reporte de los resultados es para “público general”.

En caso de que un alumno no entregue su parte, ese alumno tendrá nota 1.0 y no afectará la nota final del resto del grupo. Además, el equipo docente se reserva el derecho a hacer descuentos por informes que estén mal presentados, mal redactados, que sean difíciles de entender o que no sigan las instrucciones.

Formato de entrega

Habrán dos buzones en Webcursos, uno para la parte grupal y otro para la parte individual. A cada buzón debe ir un informe en formato **.pdf** con el reporte correspondiente y un **.zip** que contenga los **códigos** asociados a cada parte (grupal vs individual).