UNIVERSIDAD ADOLFO IBÁÑEZ

Fundamentos de Ciencia de Datos – Prueba 01

Fecha de publicación: 09 de septiembre de 2021

Objetivo

El objetivo de esta prueba es realizar un análisis exploratorio de datos para un conjunto de datos reales, aplicando los conceptos básicos del análisis exploratorio de datos para un problema sencillo utilizando el lenguaje Python.

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.

Debe usar las técnicas de preparación de datos y EDA, vistas hasta el momento, para concretar los siguientes puntos:

a. **Descripción**: primero debes describir cada una de las variables de los datos adquiridos, excepto el identificador del cliente. Debes partir por señalar el tipo de variable, entre categórica nominal, categórica ordinal o numérica. Luego, debes analizar la distribución de cada una de estas variables para finalmente señalar cualquier aspecto relevante basado en la distribución observada. En caso que no detecte algún aspecto relevante, justifíquelo. **El puntaje de esta parte de la prueba es grupal y equivale a 2 puntos.**

b. **Búsqueda de relación entre variables**: al banco le interesa saber qué condiciones influyen en que un cliente los abandone. La idea es implementar una estrategia para detectar tempranamente a clientes que vayan a abandonar, con el fin de ofrecerles una atención especializada que permita que sigan su relación con el banco. **Por lo mismo, en esta parte debes hacer un análisis sobre qué variables son las que influyen sobre el abandono de los clientes**. Por ejemplo, aquí hay algunas preguntas que te pueden inspirar:

- ¿Cómo distribuye la edad de clientes que abandonan vs los que no abandonan?
- ¿La relación entre el cupo total y el cupo por pagar es distinta en clientes que abandonan vs clientes que no abandonan?
- ¿Cómo distribuyen los montos de las transacciones para clientas que abandonan vs clientes que no abandonan?

Se espera que hagas un análisis a fondo de estos puntos, apoyado en todas las técnicas vistas hasta ahora, principalmente en el análisis estadístico y en la visualización de datos. En el caso de que creas que no hay ninguna relación entre el abandono y alguna otra variable, debes justificarlo en profundidad. **El puntaje de esta parte de la prueba es individual y equivale a 2.5 puntos**.

c. **Visualización de las conclusiones**: basado en las conclusiones del punto **b.**, debes crear un gráfico explicativo que muestre la relación encontrada. **Este gráfico se deberá crear para una audiencia que no sabe nada de estadística**, pero de todas formas, deberá mostrar claramente la relación observada. **La nota de esta parte de la prueba es individual y equivale a 1.5 puntos**.

La fecha de entrega será el lunes 4 de octubre 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: análisis de los tipos de datos.
1.0 pto: análisis de la distribución de los datos.
0.5 pts: principios básicos de la visualización de datos.

Evaluación del punto **b.**:

1.0 pto: análisis para encontrar la relación entre los datos.
0.5 pts: consistencia entre la relación encontrada y lo que se muestra en el análisis y en los gráficos.
1.0 pts: principios básicos de la visualización de datos.

Evaluación del punto **c.**:

0.5 pts: generar un gráfico que efectivamente sea para público general.
0.5 pts: consistencia entre la conclusión y lo que se muestra en el gráfico.
0.5 pts: principios básicos de la visualización de datos.

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 o que sean difíciles de entender.

Formato de entrega

Se deberá subir un puro archivo comprimido por grupo con los siguientes archivos:

1. Archivo en formato .pdf que contenga el reporte correspondiente a la parte grupal.
2. Archivo en formato .pdf que contenga el reporte correspondiente a la parte individual.
3. Archivo en formato .zip que contenga los códigos con los que se generaron los análisis de la parte grupal.
4. Archivo en formato .zip que contenga los códigos con los que se generaron los análisis de la parte individual.