Diplomado en Análisis de Datos |
Diplomado en Machine Learning

Curso: Aprendizaje Supervisado
Prof. Rolando de la Cruz

Tarea 1

Desarrollo obligatorio en grupos de 2, 3 o 4 integrantes.

Fecha de entrega: hasta las 23.59 hrs del 16/11/25 vía webcursos. Sólo 1 integrante del grupo debe subir la tarea.

DESCRIPCIÓN DEL PROBLEMA

Un banco quiere entender cómo los hábitos bancarios de los clientes contribuyen a los ingresos y a la rentabilidad. El banco tiene la edad del cliente y la información de la cuenta bancaria, por ejemplo, si el cliente tiene una cuenta de ahorros, si el cliente ha recibido préstamos bancarios y otros indicadores de la actividad de la cuenta. El conjunto de datos disponible en el archivo "BankRevenue.csv" contiene información de 16 variables sobre 7.420 clientes del banco:

1. **Rev_Total:** Total revenue generated by the customer over a 6-month period.
2. **Bal_Total:** Total of all account balances, across all accounts held by the customer.
3. **Offer:** An indicator of whether the customer has received a special promotional offer in the previous one-month period. Offer=1 if the offer was received, Offer=0 if it was not.
4. **AGE:** The customer's age.
5. **CHQ:** Indicator of debit card account activity. CHQ=0 is low (or zero) account activity, CHQ=1 is greater account activity.
6. **CARD:** Indicator of credit card account activity. CARD=0 is low or zero account activity, CARD=1 is greater account activity.
7. **SAV1:** Indicator of primary savings account activity. SAV1=0 is low or zero account activity, SAV1=1 is greater activity.
8. **LOAN:** Indicator of personal loan account activity. LOAN=0 is low or zero account activity, LOAN=1 is greater activity.
9. **MORT:** Indicator of mortgage account tier. MORT=0 is lower tier and less important to the bank's portfolio. MORT=1 is higher tier and indicates the account is more important to the bank's portfolio.
10. **INSUR:** Indicator of insurance account activity. INSUR=0 is low or zero account activity, INSUR=1 is greater activity.

11. **PENS:** Indicator or retirement savings (pension) account tier. PENS=0 is lower balance and less important to bank's portfolio. PENS=1 is higher tier and of more importance to the bank's portfolio.
12. **Check:** Indicator of checking account activity. Check=0 is low or zero account activity, Check=1 is greater activity.
13. **CD:** Indicator of certificate of deposit account tier. CD=0 is lower tier and of less importance to the bank's portfolio. CD=1 is higher tier and of more importance to the bank's portfolio.
14. **MM.** Indicator of money market account activity. MM=0 is low or zero account activity, MM=1 is greater activity.
15. **Savings:** Indicator of savings accounts (other than primary) activity. Savings=0 is low or zero account activity, Savings=1 is greater activity.
16. **AccountAge:** Number of years as a customer of the bank.

Se quiere construir un modelo que permita al banco predecir la rentabilidad para un cliente determinado. Un sustituto de la rentabilidad del cliente disponible en nuestro conjunto de datos es el ingreso total (**Rev_Total**) que un cliente genera a través de sus cuentas y transacciones. El modelo resultante se utilizará para predecir los ingresos del banco y guiarlo en futuras campañas de marketing.

Objetivo: Desarrollar un modelo predictivo de regresión lineal múltiple para predecir el ingresos total de un determinado cliente. El desarrollo del modelo predictivo deberá incluir los siguientes puntos:

1) Análisis exploratorio de las variables. Use gráficos y estadísticos apropiados para describir cada variable.
2) Análisis de correlaciones.
3) Especificación del modelo. Especifique correctamente la relación funcional entre la variable target y los atributos predictores.
4) Selección de variables. Use técnicas de selección de variables para determinar un modelo que haga la predicción requerida.
5) Use técnicas de regularización (Ridge, Lasso y Elastic Net) en el modelo especificado en el punto 3).
6) Medir la calidad predictiva de los modelos candidatos.
7) Elegir el mejor modelo predictivo.
8) Interpretar algunos coeficientes de interés.

Subir a la plataforma un Notebook de Python (en Colab) con la solución de la tarea. Al momento de seleccionar las muestras de entrenamiento y testing use una semilla (con la función apropiada en Python) y deje la semilla utilizada en su Notebook

Nota: No se aceptará como solución de la tarea solo el Notebook con los códigos/funciones de Python. Su desarrollo debe incorporar discusiones/interpretaciones, etc sobre los análisis realizados. Sus decisiones deben ser justificadas.