Universidad de la Ciudad de Aguascalientes

Guía Parcial II

Introducción a Ciencia de Datos
Rafael Dávila Bugarín

Teórico

1. Estudiar el quiz (posiblemente para esta parte no los deje usar computadora por lo sencillo del mismo).
2. Describa en un párrafo cuándo se una el método de máxima verosimilitud (*maximum likelihood estimator*) al momento de estimar los parámetros en la regresión logística.
3. Describa en un párrafo el proceso de un árbol de decisión.

Práctico

1. En un banco se desea saber qué hace que un cliente los abandone o no (churn rate). Se tiene la siguiente base de datos con las características del cliente:

Link:
https://www.kaggle.com/datasets/radheshyamkollipara/bank-customer-churn

1.1. Grafique un correlation heatmap[1] para ver cómo se relacionan las variables (RowNumber y CustomerID no deberían estar incluidas). ¿Qué variables a primera vista explican mejor la salida de los clientes?

1.2. A continuación, entrene 2 modelos usando solo las 2 variables que más se correlacionaron con la salida, uno de regresión logística y uno de árboles de decisión (Training set 80% en ambos casos, profundidad=2 para el árbol, y random_state=0 en ambos casos también). De acuerdo al accuracy score, cuál modelo fue mejor en el devset?

1.3. ¿Cuántos falsos positivos y falsos negativos hubo para cada modelo (usar la matriz de confusión)?

1.4. Sin ponerlo en código, ¿Qué le sucede al modelo si se tienen 4 niveles de profundidad en el árbol?

NOTA: Si tiran el dataset (como anteriormente he visto que sucede) aquí está la descripción de las variables y en el drive está el dataset.

RowNumber—corresponds to the record (row) number and has no effect on the output.
CustomerId—contains random values and has no effect on customer leaving the bank.
Surname—the surname of a customer has no impact on their decision to leave the bank.
CreditScore—can have an effect on customer churn, since a customer with a higher credit score is less likely to leave the bank.
Geography—a customer's location can affect their decision to leave the bank.

[1] Tome como base este código: https://medium.com/@szabo.bibor/how-to-create-a-seaborn-correlation-heatmap-in-python-834c0686b88e

Gender—it's interesting to explore whether gender plays a role in a customer leaving the bank.
Age—this is certainly relevant, since older customers are less likely to leave their bank than younger ones.
Tenure—refers to the number of years that the customer has been a client of the bank. Normally, older clients are more loyal and less likely to leave a bank.
Balance—also a very good indicator of customer churn, as people with a higher balance in their accounts are less likely to leave the bank compared to those with lower balances.
NumOfProducts—refers to the number of products that a customer has purchased through the bank.
HasCrCard—denotes whether or not a customer has a credit card. This column is also relevant, since people with a credit card are less likely to leave the bank.
IsActiveMember—active customers are less likely to leave the bank.
EstimatedSalary—as with balance, people with lower salaries are more likely to leave the bank compared to those with higher salaries.
Exited—whether or not the customer left the bank.
Complain—customer has complaint or not.
Satisfaction Score—Score provided by the customer for their complaint resolution.
Card Type—type of card hold by the customer.
Points Earned—the points earned by the customer for using credit card.

Acknowledgements

As we know, it is much more expensive to sign in a new client than keeping an existing one.

It is advantageous for banks to know what leads a client towards the decision to leave the company.

Churn prevention allows companies to develop loyalty programs and retention campaigns to keep as many customers as possible.