18/12/2019
Jose Sebastian Estepa Andrade
Katherine Ramirez Cubillos
Francisco Perea De Zubiría
Introducción al machine learning para economistas

Datatón 2019

I. Variables:

Nuestro objetivo era que nuestras variables recogieran información acerca de las Características de la gestión de relaciones con el cliente (CRM), como proponen en San Pedro, J., Proserpio, D. y Oliver, N,(2015).

- **Valor total transferido**:
    - Para obtener el valor total transferido por usuario se sumó el valor transferido de cada una de sus transacciones financieras exitosas. Para ello se utilizó la función aggregate() en R. Posteriormente las añadimos a la base de datos del modelo utilizando el id como llave.
- **Número de transacciones exitosas:**
    - Primero filtramos con el objetivo de seleccionar únicamente las transacciones financieras exitosas. Posteriormente contamos el número de este tipo de transacciones, es decir el número de veces que el código de respuesta era cero.
- **Valor transferido promedio por por transacción:**
    - Se calculó el valor transferido promedio para cada usuario dividiendo su valor total transferido entre el número de transacciones exitosas que realizó.
- **Desposit:**
    - Se refiere al número de dispositivos por persona, lo cual realizamos con el comando n_distinct

Los datos fueron los proporcionados por Bancolombia para nuestro modelo empleamos las bases de formato lite. Una vez escogidas las variables se procedió a balancear la muestra según los valores de riesgo igual a 1, el método empleado fue balanceo por reducción con una muestra de 2735 datos.

Al realizar la prueba de significancia de las variables, se encontró que estas no eran significativas a un nivel del 5% sin embargo, decidimos analizar varios modelos.

II. Modelos

Para el ejercicio se realizaron siete modelos a comparar la predicción; Logistic, Naive Bayes, KNN, Decision Tree, Random Forest, Gradient Boosting y SVM. Después se procedió a comparar los modelos por las medidas Accuracy, Precision, F1 y ROC AUC. Como el criterio del concurso era últimamente la ROC AUC se le dió preferencia a este.

A la hora de graficar la curva ROC encontramos problemas con el modelo SVM que no nos dejaba realizar la gráfica por lo que se decidió eliminar este modelo.

Una vez realizando las comparaciones pertinentes encontramos que el modelo que generaba los mejores resultados fue Random Forest.

III. Bibliografía

Blumenstock, J., Cadamuro, G. & On, R. Predicting poverty and wealth from mobile phone metadata. Science 350, 1073–1076 (2015).

Blumenstock, J. E., Eagle, N. & Fafchamps, M. J. Airtime transfers and mobile communications: Evidence in the aftermath of natural disasters. Dev. Econ. 120, 157–181 (2016).

Björkegren, D. & Grissen, D. Behavior Revealed in Mobile Phone Usage Predicts Loan Repayment (SSRN, 2015).

San Pedro, J., Proserpio, D. & Oliver, N. MobiScore: Towards Universal Credit Scoring from Mobile Phone Data in User Modeling, Adaptation and Personalization (eds Ricci, F. et al.) 195-207 (Springer, 2015).