

# **Profesional III Analítica**

-

**Departamento de  
Analítica y Gestión  
de Información**

Julian Felipe  
Baquero Mesa  
Julio de 2019

# Lectura y análisis de las bases de datos

*Búsqueda de variables con valores nulos y características distintas a las enunciadas en el planteamiento*

## Clientes

```
Int64Index: 10655 entries, 0 to 10692
Data columns (total 26 columns):
CONSECUTIVO      10655 non-null int64
COMPRA_CARTERA   10655 non-null int64
EDAD             10655 non-null int32
SCORE            10655 non-null float64
ING_EST          10655 non-null float64
ANT_DAV          10655 non-null float64
SALDO_DAV_ANT    10655 non-null float64
SALDO_DAV_ACT    10655 non-null float64
CUPO_DAV         10655 non-null float64
ANT_SF           10655 non-null float64
SALDO_SF_ANT     10655 non-null float64
SALDO_SF_ACT     10655 non-null float64
CUPO_SF          10655 non-null float64
NUM_PROD_P1      10655 non-null float64
NUM_PROD_P2      10655 non-null float64
NUM_PROD_P3      10655 non-null float64
NUM_PROD_P4      10655 non-null float64
SAL_OUT_DAV      10655 non-null float64
ACT_LAB_E        10655 non-null uint8
ACT_LAB_I        10655 non-null uint8
ACT_LAB_P        10655 non-null uint8
ENTIDAD_COMPRA_No aplica 10655 non-null uint8
ENTIDAD_COMPRA_P1 10655 non-null uint8
ENTIDAD_COMPRA_P2 10655 non-null uint8
ENTIDAD_COMPRA_P3 10655 non-null uint8
ENTIDAD_COMPRA_P4 10655 non-null uint8
dtypes: float64(15), int32(1), int64(2), uint8(8)
```

- Se deben imputar datos nulos en las variables de actividad laboral y entidad de compra de cartera
- En general las variables tenían varios valores nulos que se decidieron omitir. Por lo cual pasamos de un conjunto de 10693 líneas a uno de 10655
- Se crean variables dicotómicas con las variables categóricas para facilitar su uso en los modelos a plantear

## Datos mensuales

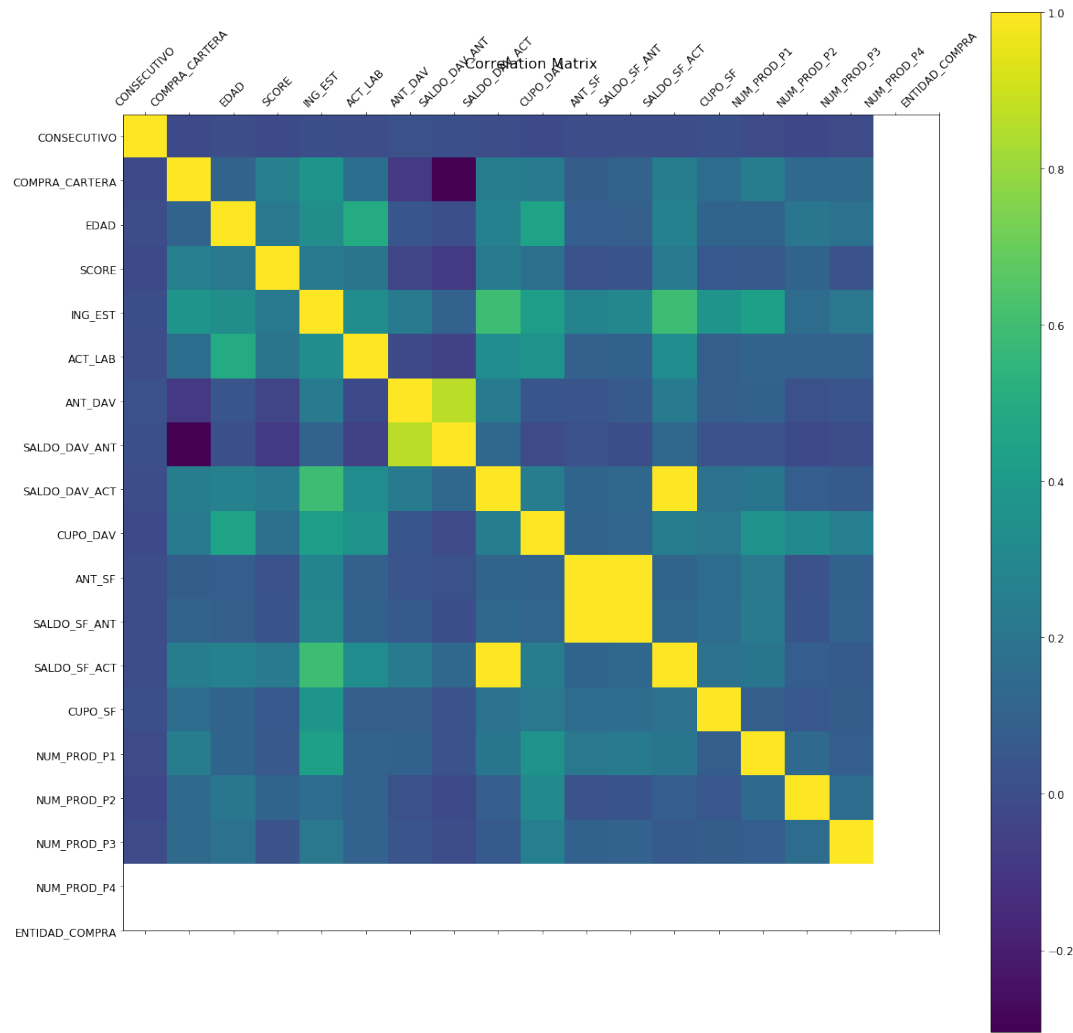
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 42772 entries, 0 to 42771
Data columns (total 5 columns):
CONSECUTIVO      42772 non-null int64
mes_lag          42772 non-null int64
MORA_DAV         42772 non-null float64
MORA_SF          42772 non-null float64
NUM_CONS         42582 non-null float64
dtypes: float64(3), int64(2)
memory usage: 1.6 MB
```

- Únicamente la variable que tiene valores nulos es NUM\_CONS
- Se requiere procesamiento extenso para lograr incluir la información que contienen los datos de esta base en la base de clientes

# Análisis descriptivo

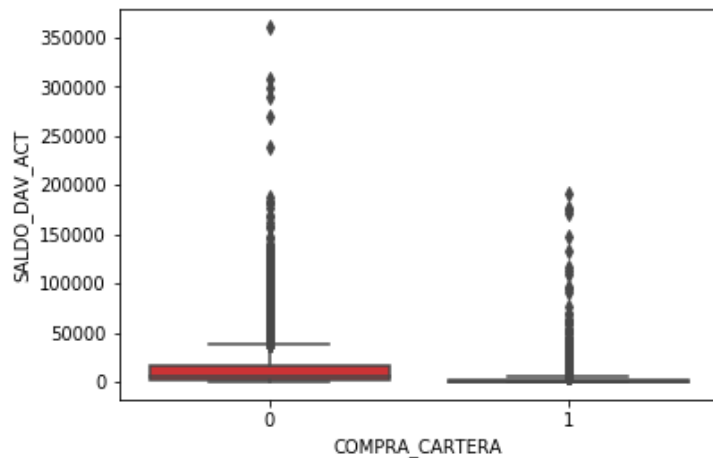
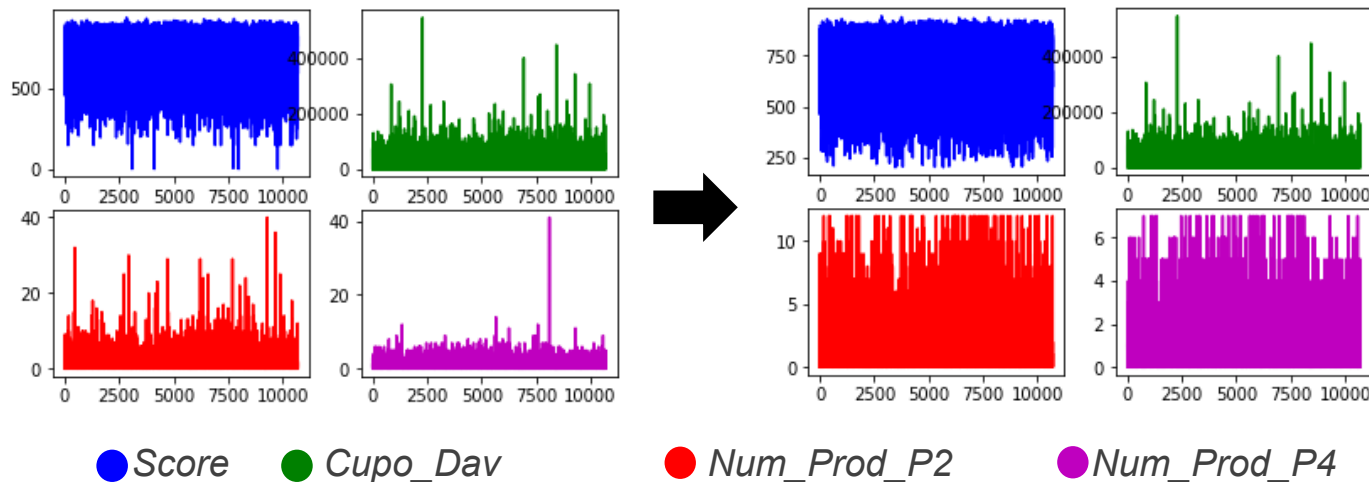
Variables *CONSECUTIVO* y *ENTIDAD\_C*  
*OMPRA* no presentan ninguna correlación  
con las demás variables.

Fuerte correlación negativa entre las variables *COMPRA\_CARTERA* y *SALDO\_DAV\_ACT*, lo cual nos lleva a plantear que los clientes más susceptibles a la compra de cartera son aquellos que presentan un bajo saldo con Davivienda.



Manejo de datos  
atípicos y  
comprobación gráfica

El código adjunto  
describe el tratamiento  
que recibieron para el  
manejo de los datos  
atípicos



No existe evidencia estadística suficiente para asegurar una correlación entre el saldo actual con el banco y la compra de cartera. Sin embargo, existen más clientes a los cuales no se les ha comprado cartera y estos clientes tienden a presentar un mayor saldo en Davivienda.

# Construcción del modelo y resultados

*Se evalúan tres modelos y tres métricas de desempeño para encontrar el modelo ideal para dar respuesta a nuestro caso*

**Modelos:** Regresión logística | *Árbol de Decisión* | Random forest

**Métricas:** Accuracy | *F1-score* | F-beta-score

$$\textit{Precision} = \frac{TP}{TP + FP}$$

$$\textit{accuracy} = \frac{TP + TN}{TP + FN + TN + FP}$$

$$\textit{Recall} = \frac{TP}{TP + FN}$$

$$F_{\beta} = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}}$$

$$F1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

*TP* = True positive    *FP* = False positive

*TN* = True negative    *FN* = False negative

# Construcción del modelo y resultados

## Modelo inicial

Metodo	Accuracy	F1-Score	F_Beta-Score
LogisticRegression	0.934685	0.922113	0.912504
DecisionTreeClassifier	1.000000	1.000000	1.000000
RandomForestClassifier	1.000000	1.000000	1.000000

- En este modelo se incluyeron únicamente las variables de la base de datos de clientes
- Se identificó un comportamiento extraño en los puntajes del árbol de decisión y random forest, por lo cual se decidió buscar las variables que producían estos resultados con un 100% de acierto

## Modelo final

Metodo	Accuracy	F1-Score	F_Beta-Score
LogisticRegression	0.933559	0.920734	0.910740
DecisionTreeClassifier	0.959084	0.952004	0.957375
RandomForestClassifier	0.958333	0.950645	0.946930

- En este modelo se eliminaron las variables sobre la entidad que compraban cartera, dado que estas variables traen implícita la información sobre la compra efectiva de cartera. Se incluyeron nuevas variables propuestas y las requeridas en el planteamiento
- Se obtuvo que el modelo con mejor desempeño en las métricas fue el árbol de decisión

# Conclusiones

- Se presenta satisfacción general con el árbol de decisión que permite clasificar correctamente a más del 95% de los clientes.
- Este modelo presenta un alto sesgo por parte de las variables del saldo, por lo tanto, se presenta la limitación de no contar con mayor información sobre los clientes o las entidades.
- Información que podría tener relevancia puede ser la demográfica para los clientes, junto con los productos que tiene con el banco o sobre los cuales se realizó la compra de cartera.
- El modelo puede alcanzar mejores resultado por medio de la optimización de sus parámetros. Sin embargo, esto no se realizó debido a que se obtendría un modelo más pesado y con el modelo final se alcanzan resultados que son bastante aceptables para el caso expuesto.
- Se evidencia la oportunidad de usar este modelo para identificar los clientes susceptibles a compra de cartera para evitar que otras entidades logren llevar a cabo esta compra
- Como respuesta a esto, el banco puede crear planes de prevención con los clientes identificados y buscar que ellos mantengan sus productos con Davivienda.
- Existen entidades como P2 que ya ha logrado comprar cartera de más de 2000 de los clientes del banco.

# Gracias

**Se puede consultar todo el material en el siguiente enlace:**

<https://github.com/julianfbaquero/Profesional3Analitica>

[julianfbaquero@gmail.com](mailto:julianfbaquero@gmail.com)