Informe 3 Actividad 5

Data Science with Python

Build and Evaluate Models

Edwin Spencer

Definición de Framework y EDA

El marco de trabajo a utilizar será el BADIR ya que, según la necesidad del negocio, es la herramienta que más se ajusta para cumplir con los objetivos planteados

Objetivos del negocio

- Identificar la causa raíz del por qué va en aumento el número de clientes de Credit One que incumple el pago de sus préstamos.
- Creación de modelos analíticos para generar un scoring y poder así clasificar correctamente los clientes

Plan de análisis

Justamente este es uno de los puntos por el cual se definió utilizar el marco de trabajo BADIR, ya que con la información (DataSet) que se cuenta hasta el momento nos permite poder identificar claramente los objetivos del negocio, además de que con base a esto se puede definir una hipótesis y poder definir un plan de trabajo para poder cumplir con los objetivos planteados.

Colección de datos

Con respecto a los datos, se cuenta con datos históricos los cuales representan cantidad amplia de registros (30000) además de contener variables importantes que nos ayudarán a poder identificar la causa raíz que nos exige la organización, además de poder generar un Score que permita filtrar los mejores clientes, a partir de acá se procederá a realizar una depuración y validación de los datos, para obtener resultados más certeros.

Estructura del dataSet

- Monto del crédito
- Genero del cliente
- Nivel de escolaridad
- Estado civil
- Edad
- Pago a tiempo de los últimos 6 meses: Indica si se realizó el pago a tiempo, retrasado por un mes o mas
- Monto de la factura para los últimos 6 meses
- Monto de los pagos realizados para los últimos 6 meses
- Comportamiento del cliente

Gestión de los datos

Como parte de la administración o gestión de los datos, se debe realizar varias acciones importantes para poder aprovechar de la mejor manera la información contenida en el set de datos, las acciones a ejecutar a los datos serían las siguientes:

- Depuración
- Transformación
- Reducción
- Discretización

Set de datos

Parte del análisis inicial que se pudo realizar a los datos se puede identificar que existen datos socioeconómicos que podrían ser importante de incorporar y así poder obtener resultados más valiosos, por ejemplo, datos referentes a su empleabilidad, sus ingresos mensuales, deudas en otras entidades, etc son algunos de datos que podrían ayudar a seleccionar mejor a los clientes.

Preparación y exploración de los Datos

Acciones ejecutadas

- Entendimiento del dataset a través de los comandos Head(), describe(), info(), columns() y dtypes
- Validación de los datos nulos a través del comando isnull().any() detectando que no existían valores nulos.
- Además se realizó la verificación de si existen valores duplicados con el comando duplicated().any() sin encontrar ningún valor duplicado.
- Por otra parte, procedí a modificar el tipo de las siguientes variables a category:
    - SEX
    - EDUCATION
    - MARRIAGE
    - Las demás variables quedaron de igual forma.

Creación y evaluación de modelos

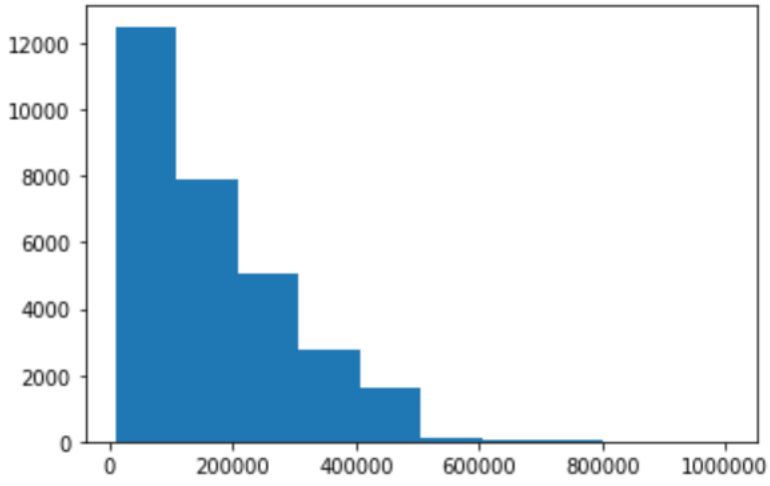
¿Qué atributos en los datos podemos considerar estadísticamente significativos para el problema en cuestión?

Se contempla como variable dependiente PAY_6 y las variables a contemplar para el entrenamiento del dataSet son las siguientes:

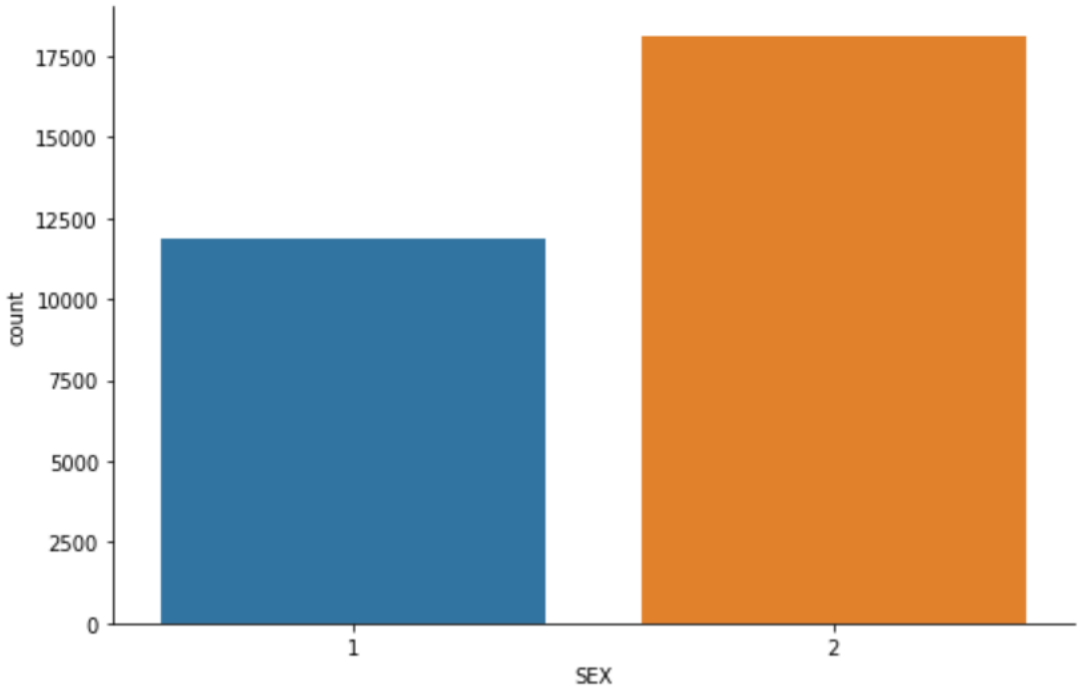
- ID
- PAY_0
- PAY_2
- PAY_3
- PAY_4
- PAY_5

¿Qué información concreta podemos derivar de los datos que tenemos?

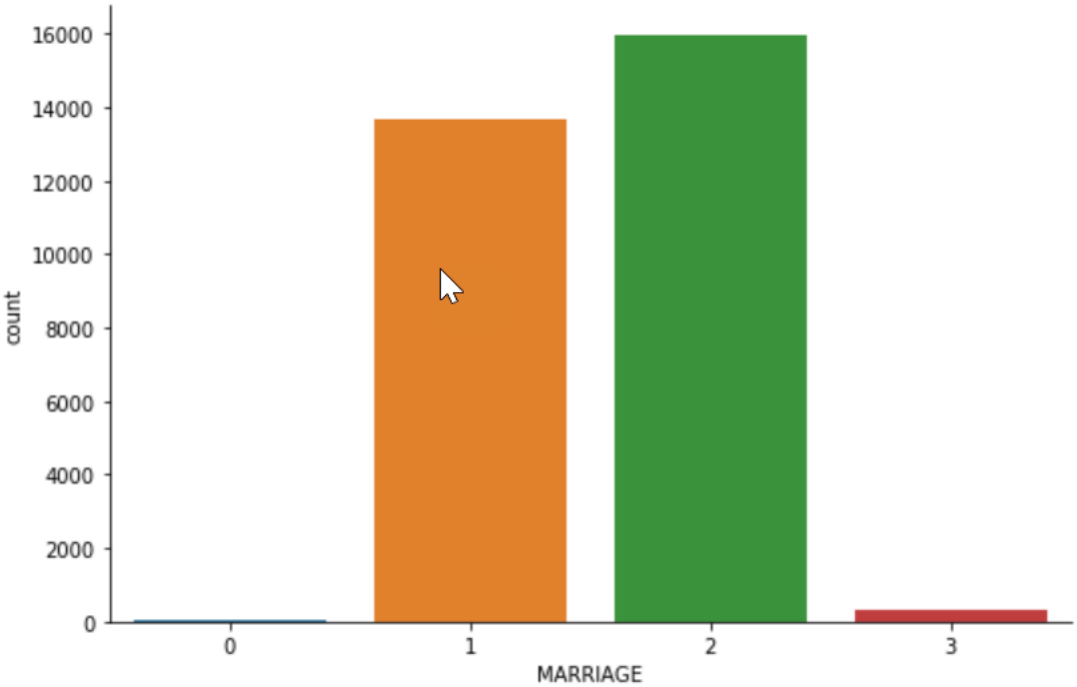
1. Más del 50% de los clientes tienen créditos por menos de 200 mil dólares.



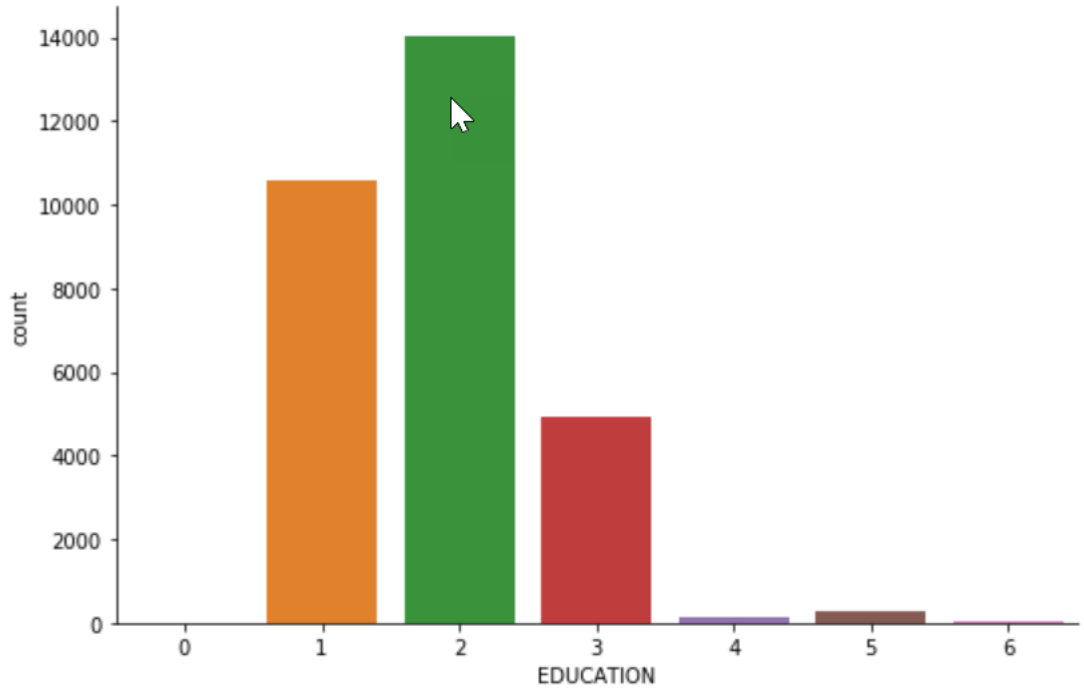
2. Del total de los clientes, más del 50% de los clientes son Mujeres.



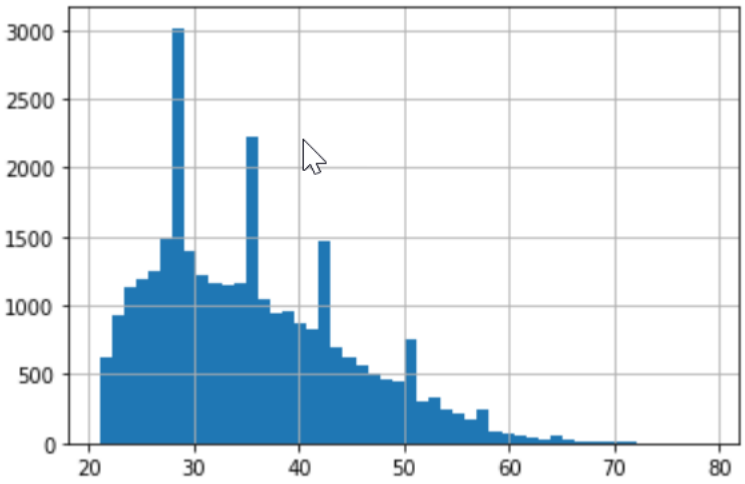
3. Más del 50% son solteros.



4. Un poco más del 40% son graduados de universidad.



5. Una mayor cantidad de clientes se encuentran entre los 20 y 30 años.



¿Qué métodos probados podemos usar para descubrir más información y por qué?

Como parte del proceso se evaluaron 3 modelos Random Forest, Regresión Linea y SVR los cuales, posterior a los entrenamientos nos da los siguientes resultados:

- Random Forest

```
In [51]: print(cross_val_score(modelRF, X_train, Y_train))
         [0.6257677  0.66437752 0.6535121 ]
```

- SVR

```
In [52]: print(cross_val_score(modelSVR, X_train, Y_train))
         [0.62133553 0.64600261 0.63568442]
```

- Regresión lineal

```
In [136]: print(cross_val_score(modelLR, X_train, Y_train))
          [0.59932269 0.63507676 0.62672925]
```

A partir de estos resultados, tomo la decisión de utilizar el modelo Random Forest para correr el set de pruebas.

```
In [137]: modelRF.score(X_train,Y_train)
Out[137]: 0.6916676685799448
```

```
In [138]: predictions = modelRF.predict(X_validation)
```

```
In [139]: rmse = sqrt(mean_squared_error(Y_validation, predictions))
```

```
In [140]: predRsquared = r2_score(Y_validation,predictions)
```

```
In [141]: print('R Squared: %.3f' % predRsquared)
          print('RMSE: %.3f' % rmse)
          R Squared: 0.656
          RMSE: 0.680
```

Presentando el siguiente resultado:

En este punto en particular tuve diferentes inconvenientes, que por más que investigue y utilicé diferentes formas de hacer las cosas no logre resolver el problema.

Adjunto evidencia de lo realizado y lo cual como indique antes por más que lo intenté no pude resolver, para efectos personales seguiré revisando el tema para buscar una solución al inconveniente.

```
In [151]: plt.scatter(Y_validation, predictions, color=("blue","red"), alpha = 0.5)
          plt.xlabel('Ground Truth')
          plt.ylabel('Predictions')
          plt.show();

---------------------------------------------------------------------------
ValueError                                Traceback (most recent call last)
~\AppData\Local\Continuum\anaconda3\lib\site-packages\matplotlib\axes\_axes.py in _parse_scatter_color_args(c, edgeco
lors, kwargs, xshape, yshape, get_next_color_func)
   4290                     valid_shape = False
-> 4291                     raise ValueError
   4292             except ValueError:

ValueError: 'c' argument has 2 elements, which is not acceptable for use with 'x' with size 7500, 'y' with size 7500.
```

```
In [149]: print(accuracy_score(Y_validation, predictions, normalize=False))

---------------------------------------------------------------------------
ValueError                                Traceback (most recent call last)
<ipython-input-149-42a8bc66133d> in <module>
----> 1 print(accuracy_score(Y_validation, predictions, normalize=False))

~\AppData\Local\Continuum\anaconda3\lib\site-packages\sklearn\metrics\classification.py in accuracy_score(y_true, y_p
red, normalize, sample_weight)
    174
    175     # Compute accuracy for each possible representation
--> 176     y_type, y_true, y_pred = _check_targets(y_true, y_pred)
    177     check_consistent_length(y_true, y_pred, sample_weight)
    178     if y_type.startswith('multilabel'):

ValueError: Classification metrics can't handle a mix of unknown and continuous targets
```

```
In [152]: print(confusion_matrix(Y_validation, predictions))

---------------------------------------------------------------------------
ValueError                                Traceback (most recent call last)
<ipython-input-152-0802aaa9ff56> in <module>
----> 1 print(confusion_matrix(Y_validation, predictions))

~\AppData\Local\Continuum\anaconda3\lib\site-packages\sklearn\metrics\classification.py in confusion_matrix(y_true, y
_pred, labels, sample_weight)
    251
    252     """
--> 253     y_type, y_true, y_pred = _check_targets(y_true, y_pred)
    254     if y_type not in ("binary", "multiclass"):
    255         raise ValueError("%s is not supported" % y_type)

ValueError: Classification metrics can't handle a mix of unknown and continuous targets
```

```
In [147]: print(classification_report(Y_validation, predictions))

---------------------------------------------------------------------------
ValueError                                Traceback (most recent call last)
<ipython-input-147-caa371b435c1> in <module>
----> 1 print(classification_report(Y_validation, predictions))

ValueError: Classification metrics can't handle a mix of unknown and continuous targets
```