Prueba de conocimiento analítico

Guía rápida de desarrollo de proyecto

José Max Barrios

1- Análisis de datos

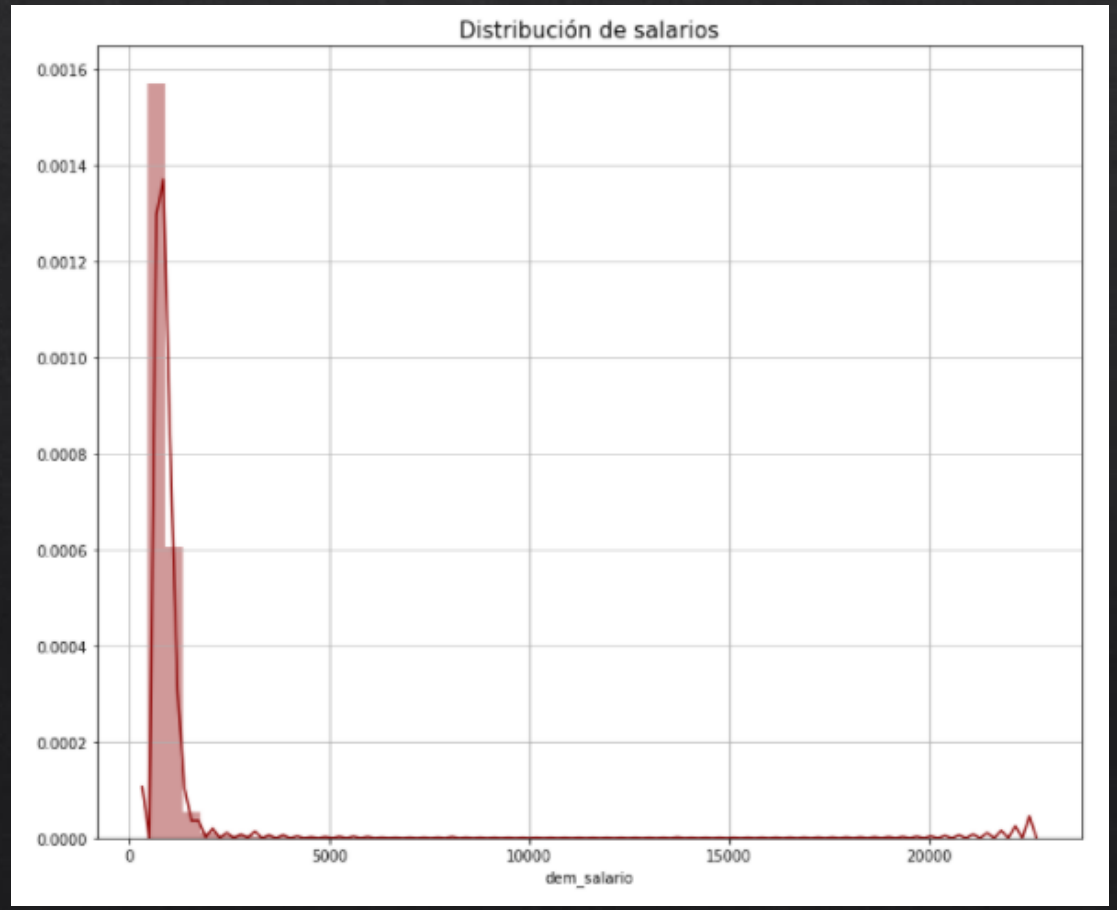
Carga de base de datos de entrenamiento y prueba.

- Identificación de valores nulos
- Substitución con media para valores nulos
- Identificación de tipos de datos.
- Eliminación de algunos datos NaN.
- Forma Final de entrenamiento 4988, 49

comer	pdcto_flag_garantizado	pdcto_flag_hipoteca	pdcto_flag_pp	pdcto_flag_seguros	pdcto_flag_tiene_tdc	pdcto_flag_tiene_tdd	pdcto_ivc_actual	dem_salario
0	0	1	0	0	0.0	0.0	2	1252.07
0	0	1	0	0	0.0	0.0	1	1033.50
0	0	0	0	1	1.0	1.0	2	1073.25
0	0	0	1	1	1.0	1.0	3	1020.00
0	0	0	0	0	0.0	1.0	1	766.40

```
llave_cod_cliente                0
admin_antiguedad_banco           0
admin_flag_gerenciado            0
buro_creditos_otros_bancos       0
buro_score_apc                 697
buro_wallet_share                8
comp_flag_atm                    4
comp_flag_bpi                    4
comp_flag_cnb                    4
comp_flag_pos                    4
comp_flag_suc                    4
comp_perc_atm                    0
comp_perc_canal_fisico           0
comp_perc_cnb                    0
comp_perc_bpi                    0
comp_perc_pos                    0
comp_perc_suc                    0
comp_score_digital               0
comp_txn_atm                     4
comp_txn_bpi                     4
comp_txn_cnb                     4
comp_txn_pos                     0
comp_txn_suc                     4
comp_usd_atm_prom                4
comp_usd_bpi_prom                4
comp_usd_cnb_prom                4
comp_usd_pos_prom                4
comp_usd_suc_prom                4
dem_edad                         0
dem_planilla                     0
finc_bal_act                     0
finc_bal_pas                     0
finc_perc_act_tc                 0
finc_perc_pas_tc                 0
finc_sva                       208
```

Análisis Exploratorio



Desde este momento ya que vemos la distribución de los salarios, se puede anticipar que nuestro modelo será capaz de predecir mejor salarios que esten por debajo de los 1000 dolares. En la celda de detalles estadisticos de igual manera se observa que 75% de los salarios esta por debajo de 937.00 dolares.

2-Análisis de Variables

Se Crearon 2 o 3 algoritmos por cada análisis de variables realizado. Estos son los análisis.

1- Feature selection por correlación

2- P-value menor de 0.05

3-Todas las variables

4- Intuitivo

Métrica de evaluación

MAPE-value	Accuracy of forecast
Less than 10%	Highly Accurate Forecast
11% to 20%	Good Forecast
21% to 50%	Reasonable Forecast
More than 51%	Inaccurate Forecast

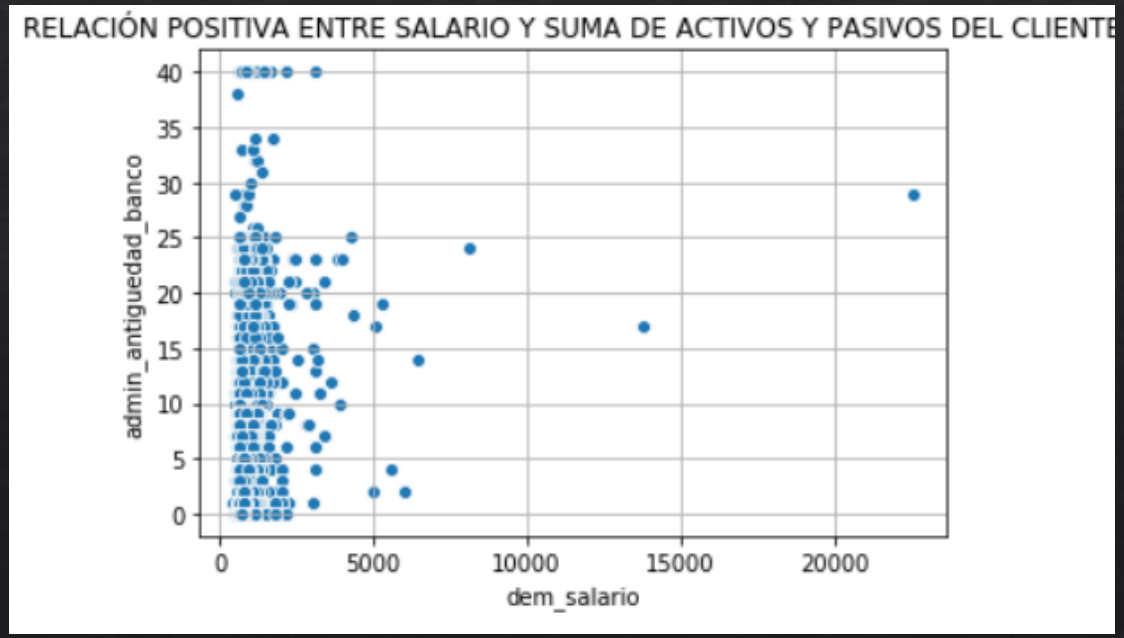
Source: Lewis, C.D., 1982.

3- Modelos

3.1 Feature selection por correlación

dem_salario
0.175915
0.161151
-0.130686
0.117816
0.109411
0.102152
0.125597
0.103169
0.122621
-0.115481
0.104222
0.143573
0.126665
0.108696
-0.105596
1.000000

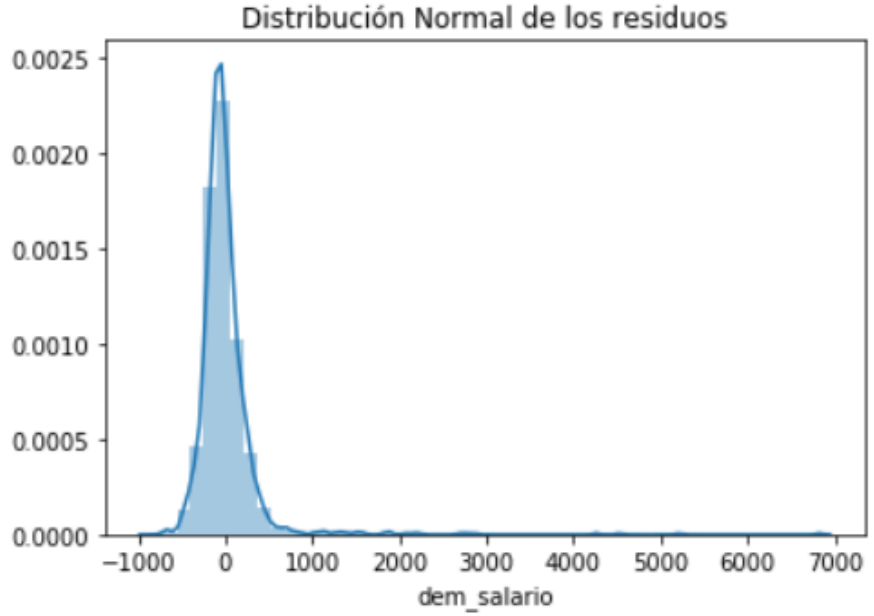
Se Extrajeron todos los atributos que estuviese en correlación nula (0.0), y se extrajeron los atributos que tenían un poco más de correlación aunque baja, ya que podrían servir como mejores predictores.



3.1 Feature selection por correlación

MAPE 19.792274797647185 %

Durbin-watson: 2.067799870062239



Algoritmo	MAPE
Linear Regression	19.59%
Artificial Neural Network	16.25%
Random Forest Regressor	18.30%

	f-statistic	p-value
dem_salario	2.201465e+17	0.000000e+00
admin_antiguedad_banco	1.592243e+02	5.895625e-36
buro_score_apc	1.329363e+02	2.259270e-30
finc_bal_pas	1.049403e+02	2.192455e-24
llave_cod_cliente	8.784697e+01	1.045706e-20
comp_perc_atm	8.663506e+01	1.909609e-20
finc_tamano_comercial	8.129986e+01	2.713598e-19
comp_usd_bpi_prom	7.991217e+01	5.415967e-19
dem_edad	7.611292e+01	3.598506e-18
comp_perc_bpi	7.018265e+01	6.952672e-17
finc_bal_act	5.475344e+01	1.594916e-13
comp_usd_pos_prom	5.364123e+01	2.791977e-13
comp_txn_bpi	5.257788e+01	4.770162e-13
finc_perc_pas_tc	3.038276e+01	3.725131e-08
comp_txn_pos	1.474333e+01	1.247096e-04
finc_sva	1.380410e+01	2.051138e-04
comp_usd_suc_prom	9.424878e+00	2.152131e-03
comp_txn_atm	7.038270e+00	8.003899e-03
comp_txn_suc	4.332707e+00	3.743764e-02
finc_perc_act_tc	4.199956e+00	4.047724e-02

3.2 P-valor significativo

Se extrajeron variables con valores con el p-valor < 0.05 con respecto a la variable de salario.

3.2 P-valor significativo

```
Epoch 1/7
418/418 [==============================] - 0s 839us/step - loss: 22.7030
Epoch 2/7
418/418 [==============================] - 0s 834us/step - loss: 16.5980
Epoch 3/7
418/418 [==============================] - 0s 872us/step - loss: 16.3012
Epoch 4/7
418/418 [==============================] - 0s 996us/step - loss: 16.10120
Epoch 5/7
418/418 [==============================] - 0s 970us/step - loss: 16.1072
Epoch 6/7
418/418 [==============================] - 0s 754us/step - loss: 16.0046
Epoch 7/7
418/418 [==============================] - 0s 750us/step - loss: 15.8659
MAPE: 17.149599423612365 %

VALORES ACTUALES VS PREDICCIONES
    dem_salario  salario_estimado
0        613.57        749.352478
1       1107.00        852.368164
2        964.70        866.640320
3        687.40        863.071899
4        661.07        713.512146
5        697.96        703.451172
```

Algoritmo	MAPE
Linear Regression	19.79%
Artificial Neural Network	17.14%
Random Forest Regressor	17.71%

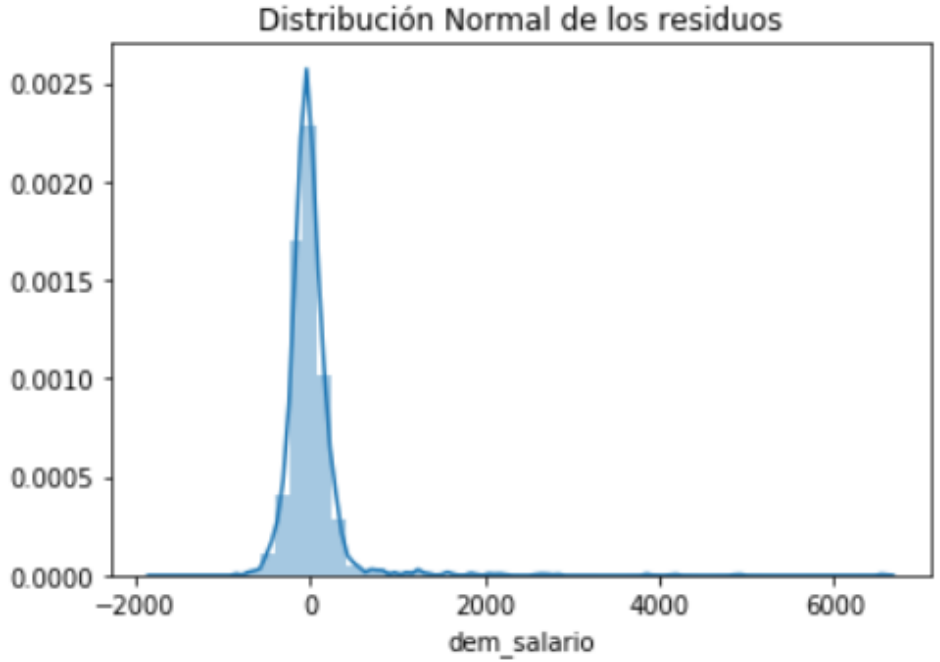
3.3 All Variables

Se utilizaron todas las variables ya aunque no tengan una correlación directa con la variable objetivo, probablemente realizaban buen complemento entre sí para predecir

3.3 All Variables

```
MAPE 18.794119397621117 %

Durbin-watson: 2.0641758380298327
```



Algoritmo	MAPE
Linear Regression	18.79%
Artificial Neural Network	15.73%
Random Forest Regressor	16.39%

3.4 Intuitivo

Vamos a utilizar las variables que muestran directamente el valor de activos o pasivos del cliente sin saber cuales estos son, ya que estos afectan directamente cualquier ingreso que tenga el cliente.

3.4 All Variables

```
Epoch 4/7
418/418 [==============================
Epoch 5/7
418/418 [==============================
Epoch 6/7
418/418 [==============================
Epoch 7/7
418/418 [==============================
MAPE: 17.38825516470161 %

VALORES ACTUALES VS PREDICCIONES
    dem_salario  salario_estimado
0        613.57        691.386108
1       1107.00        732.491150
2        964.70        691.386108
3        687.40        794.416138
4        661.07        691.552979
5        697.96        694.043152
6       1030.00        710.622437
```

Algoritmo	MAPE
Linear Regression	88673370.93 %
Artificial neural networks	17.38%

3.3 All Variables

Se utilizaron todas las variables ya aunque no tengan una correlación directa con la variable objetivo, probablemente realizaban buen complemento entre sí para predecir.

4. Predicciones en Base de datos de prueba.

Vamos a utilizar el modelo de regresión lineal con features selection.

Vista previa del archivo →

	llave_cod_cliente	dem_salario
0	484	975.149764
1	622	1036.510460
2	845	879.863613
3	884	881.502082
4	961	918.361135
5	1247	1000.352854
6	1382	1125.415205
7	1391	875.253912
8	1410	958.739611
9	1425	1005.197805

5- Conclusiones

La mayoría de nuestros modelos han entregado un MAPE entre 15% - 20% de error, el cual es bastante bueno. Sin embargo, los datos suministrados no son buenos para predecir la variable del salario, ya que no tienen ninguna correlación tan fuerte con la variable objetivo

En esta ocasión, no ha sido por modelos, sino por baja calidad de datos, que nuestro modelo no se puede mejorar.

Existen algunas técnicas de transformación como logaritmicas entre otras que siven para mejorar la correlación, pero en este caso al ser tan baja no contribuia o mejoraba demasiado el modelo.

Hemos elegido el modelo de regresión lineal con selección de atributos, ya que este contiene menor información de variables, y es casi tan bueno como los demás modelos realizados, sin variables que no aportan nada al modelo, como pudiese ser en el caso del modelo con todas las variables. O en caso de las redes neuronales ya que necesitan un poco más de poder computacional y su tiempo de entrenamiento es un poco más lento. Además que hemos realizado el test de normalidad de los residuos y el test de Durbin watson, y los residuos están normalmente distribuidos y el resultado del test ha sido de 2.06, lo que demuestra que no hay problemas de multicolinealidad (valores buenos entre 1.5 y 2.5).

Nuestro modelo es bastante bueno para predecir salarios debajo de los 1000 dolares, ya que la mayoría de los datos (75 porciento) de los datos de entrenamiento estaban por debajo de 975 dólares.

El modelo intuitivo realmente no es tan bueno, ya que a pesar de que los resultados son parecidos a los demás, estaban muy sesgados a predecir valores bajos de salarios.

La cantidad de datos tampoco era suficiente como para realizar un modelo más robusto-