

Obligatorio Machine Learning - LOGIT

Descripción del problema:

Un banco desea saber la tasa de cancelación de clientes del próximo mes, de tal manera de poder llevar adelante una campaña con el objetivo de recuperar al cliente antes de que comunique su decisión de darse de baja

Para ello se busca crear el mejor modelo predictivo posible para la variable “Attrition_Flag” utilizando el dataset que se encuentra en el archivo “BankChurners.csv”.

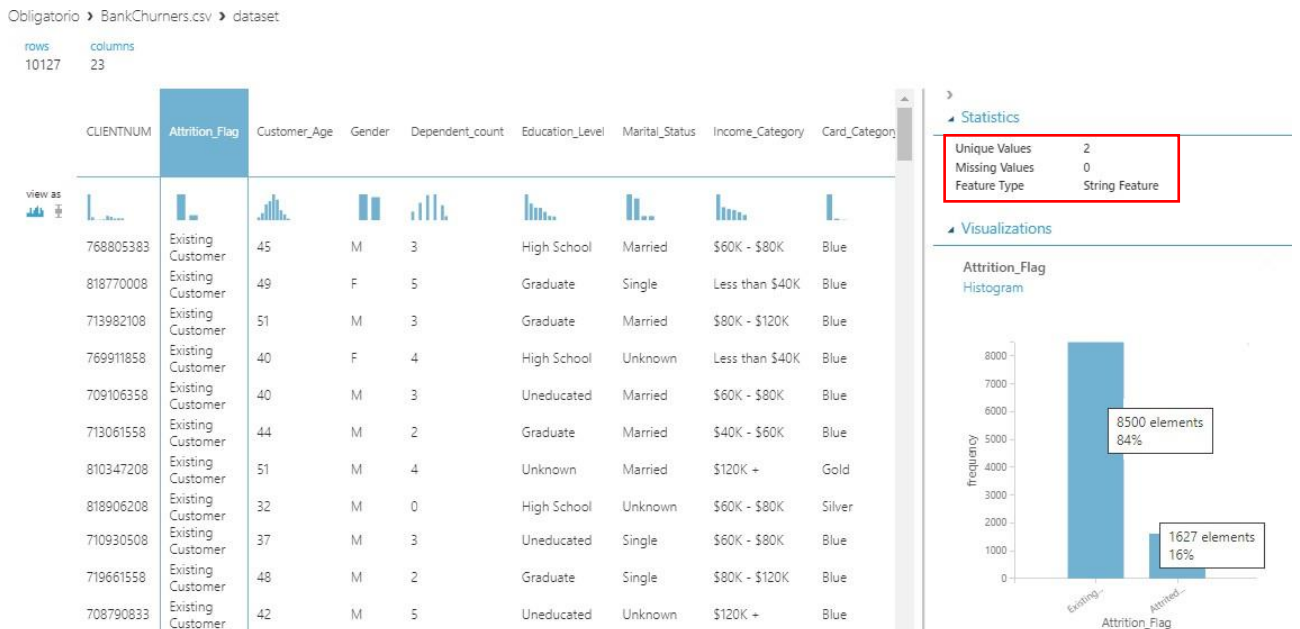
Referencias de variables:

- **CLIENTNUM:** Número de cliente → variable cualitativa.
- **Attrition_Flag:** Status de la cuenta al mes siguiente → variable cualitativa.
- **Customer_Age:** Edad del cliente → variable cuantitativa.
- **Gender:** Género del cliente → variable cualitativa.
- **Dependent_count:** Número de personas a cargo → variable cuantitativa.
- **Education_Level:** Nivel educativo → variable cualitativa.
- **Marital_Status:** Estado civil → variable cualitativa.
- **Income_Category:** Categoría de ingresos del cliente → variable cualitativa.
- **Card_Category:** Tipo de tarjeta → variable cualitativa.
- **Months_on_book:** Antigüedad de la cuenta → variable cuantitativa.
- **Total_Relationship_Count:** Cantidad de productos del cliente (cuentas y tarjetas) → variable cuantitativa.
- **Months_Inactive_12_mon:** N° de meses inactivo en los últimos 12 meses → variable cuantitativa.
- **Contacts_Count_12_mon:** N° de contactos en los últimos 12 meses (consultas o reclamos al banco) → variable cuantitativa.
- **Credit_Limit:** Límite de crédito → variable cuantitativa.
- **Total_Revolving_Bal:** Saldo no cubierto de la tarjeta (sería lo que el cliente lleva usado del monto en su tarjeta, es la diferencia entre Credit_Limit y Avg_Open_To_Buy) → variable cuantitativa.
- **Avg_Open_To_Buy:** Disponible de la tarjeta → variable cuantitativa.
- **Total_Amt_Chng_Q4_Q1:** Cambio porcentual de monto de consumos → variable cuantitativa.
- **Total_Trans_Amt:** Monto de consumos en los últimos 12 meses → variable cuantitativa.
- **Total_Trans_Ct:** Cantidad de transacciones en los últimos 12 meses → variable cuantitativa.
- **Total_Ct_Chng_Q4_Q1:** Cambio porcentual de cantidad de consumos → variable cuantitativa.
- **Avg_Utilization_Ratio:** Ratio de utilización de la tarjeta (es el resultado de hacer Total_Revolving_Bal sobre Credit_Limit) → variable cuantitativa.

Experimento en Azure:

<https://gallery.cortanaintelligence.com/Experiment/Obligatorio-Correa-Lopez-Mosco>

Análisis de la variable objetivo (y):

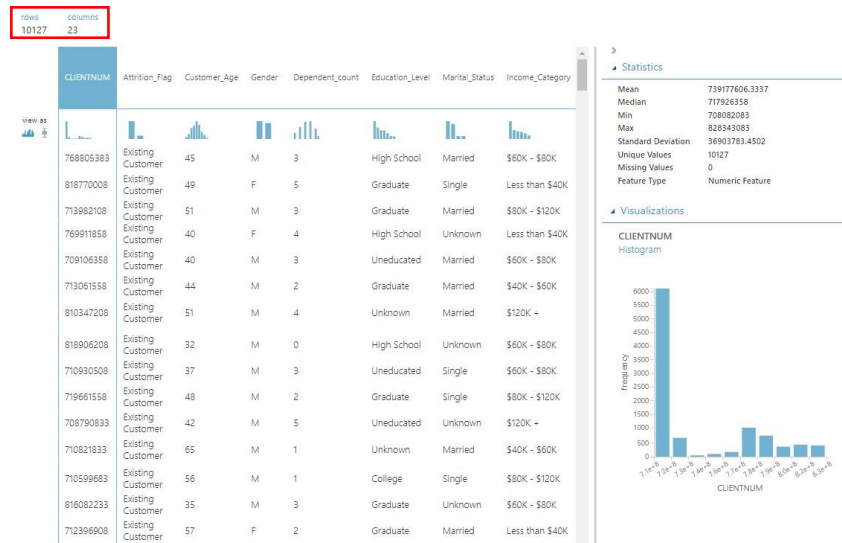


Nuestra variable objetivo es “Attrition_Flag”, la cual toma dos valores:

- **Existing Customer:** Es el cliente que va a seguir en el banco.
- **Attrited Customer:** Es el cliente que quiere dar de baja el servicio.

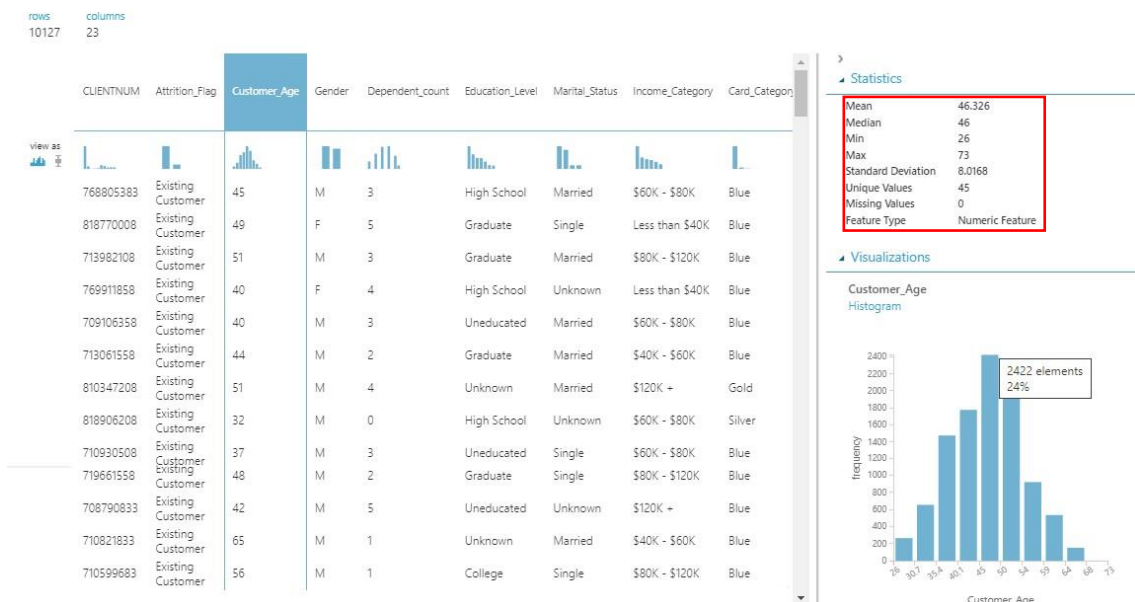
Observamos que la variable “Attrition_Flag” no tiene datos faltantes y es del tipo “String Feature”, esto indica que es una variable categórica. A partir de esta variable se van a crear las variables dicotómicas o también conocidas como “dummies”. Tiene 84% de clientes existentes y 16% de clientes que quieren darse de baja.

Breve análisis exploratorio del dataset (EDA):



El dataset consta de 10.127 filas, donde cada una de ellas representa a un cliente del banco. Las 23 columnas representan a 22 variables predictoras¹, además de la variable dependiente o a predecir.

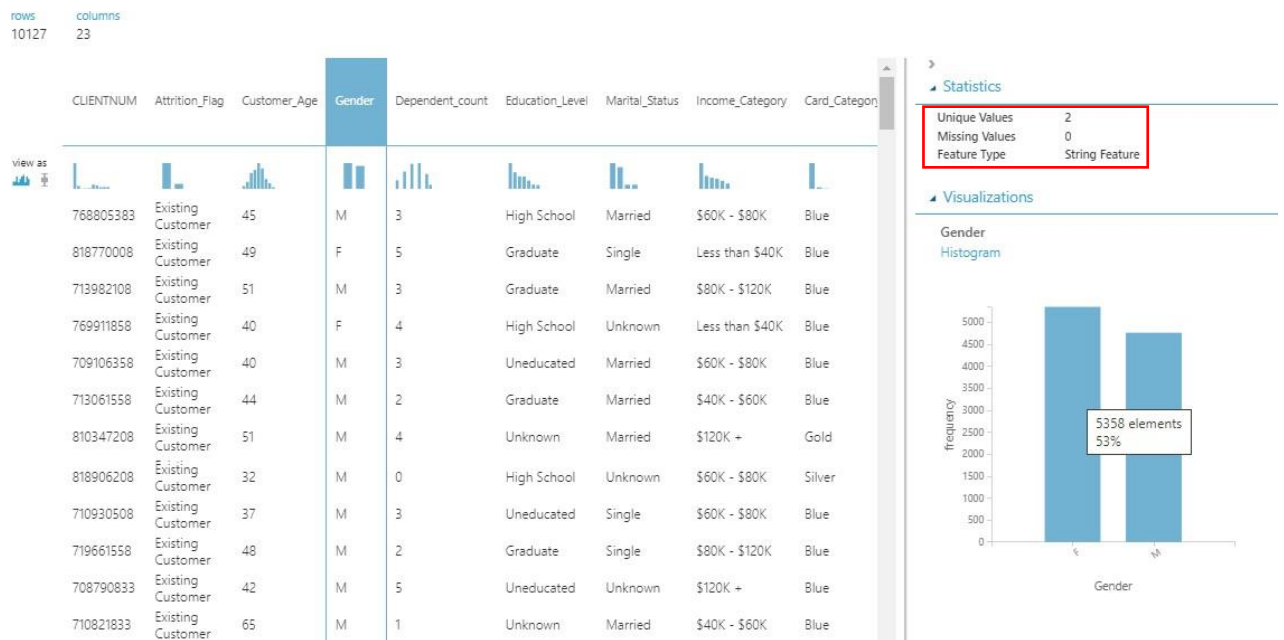
CLIENTNUM: Corresponde al número de cliente o “id”, esta variable no nos sirve porque al modelar ese dato no aporta nada ya que no significa nada en sí mismo, por lo tanto va a ser excluida.



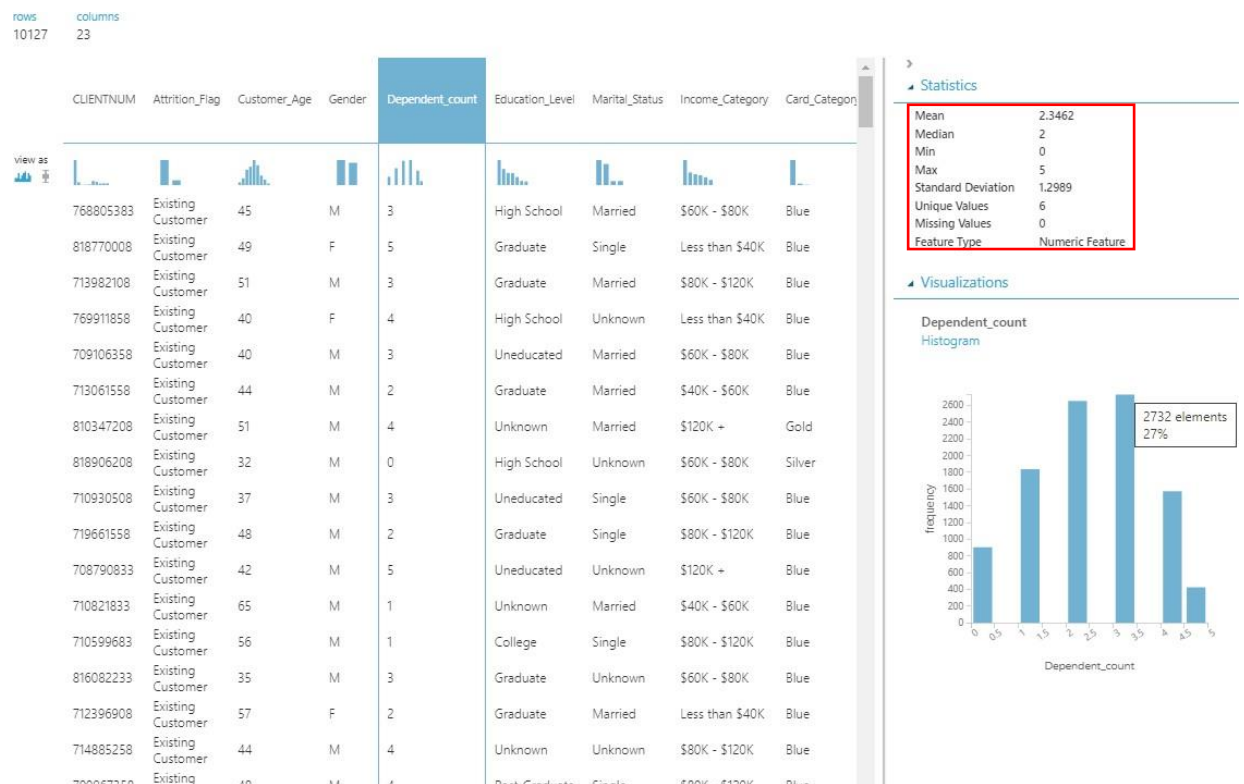
Customer_Age: Observamos que la variable correspondiente a la edad del cliente no tiene datos faltantes, la media es de 46,3 años, la edad máxima es de 73 años mientras que la mínima es de 26.

Si vemos la distribución, la mayoría de los clientes se concentran entre 45 y 50 años (24%).

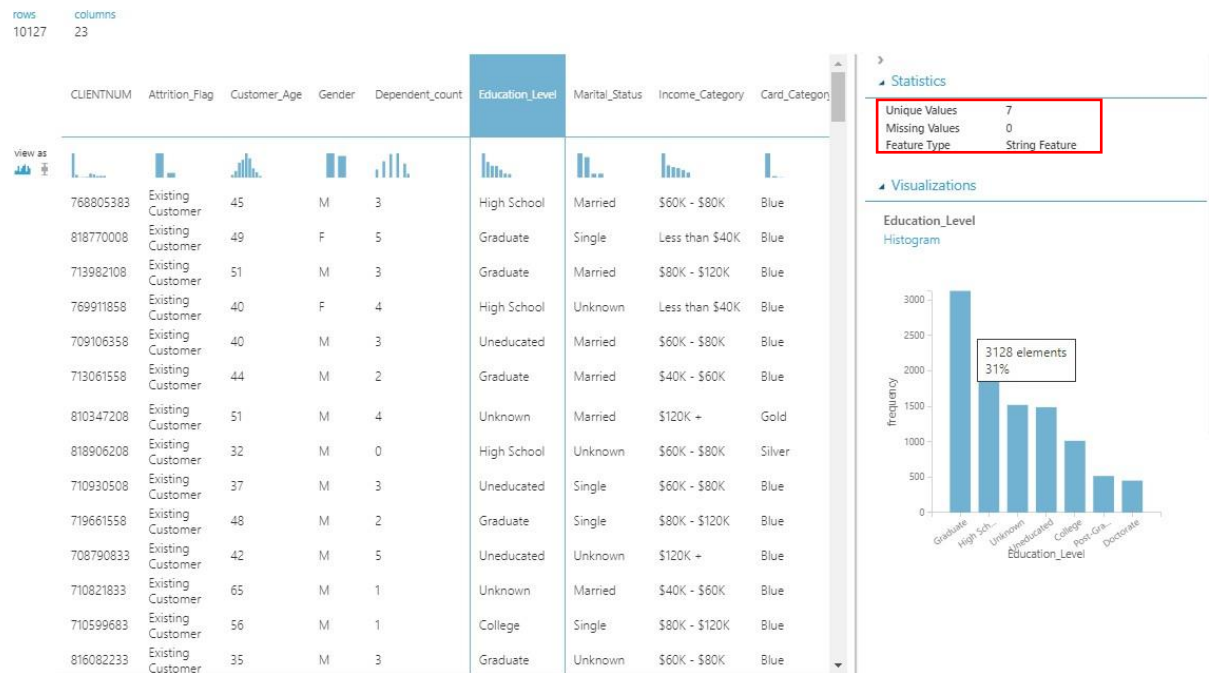
¹ Una variable predictora es aquella que se utiliza para predecir alguna otra variable o resultado.



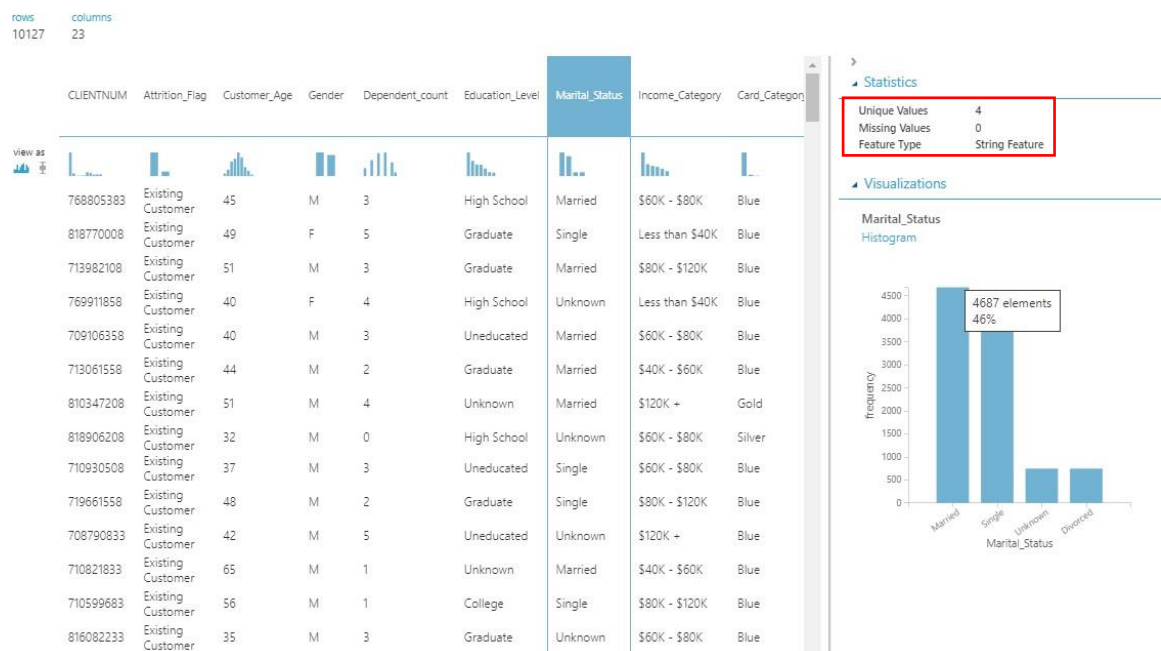
Gender: Observamos que es una variable categórica, toma dos valores distintos (femenino y masculino) y no tiene valores faltantes. Si vemos la distribución, la mayoría de los clientes son del sexo femenino (53%).



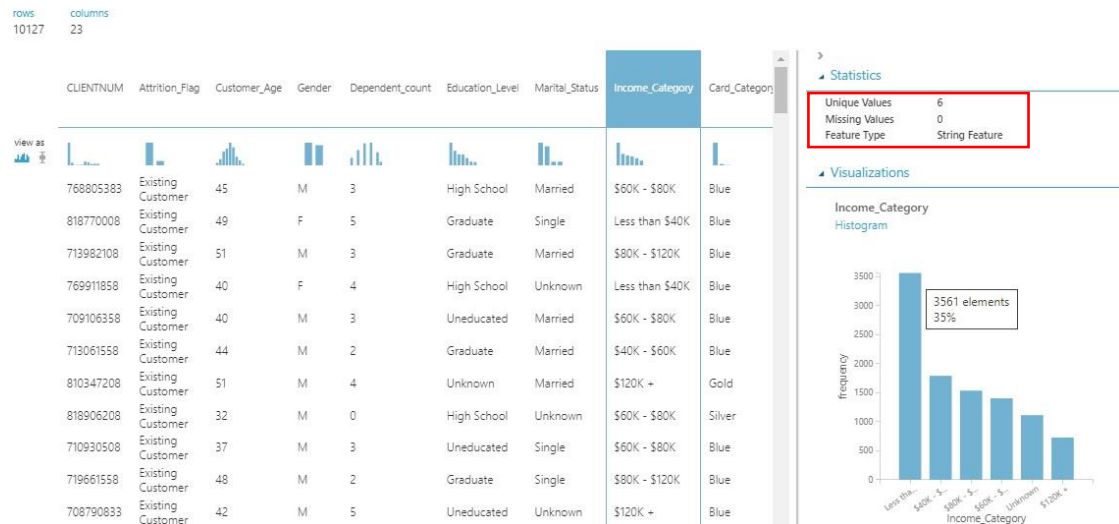
Dependent_count: Observamos que la cantidad de personas a cargo no tiene datos faltantes, la media es de 2 personas, el máximo es de 5 personas y el mínimo es de 0. Si vemos la distribución, la mayoría de los clientes se concentran en 3 personas a cargo (27%).



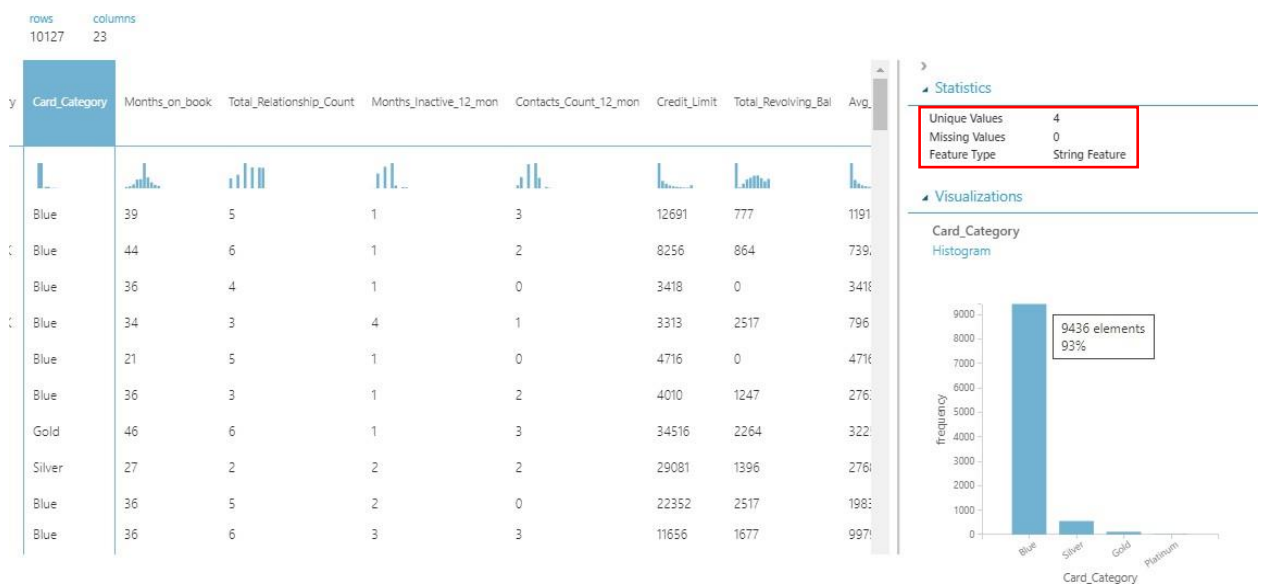
Education_Level: Observamos que es una variable categórica, toma siete valores distintos: Graduate (título de grado), High School (Bachiller), Unknown (el cliente no brindó el dato), Uneducated (no terminó el liceo), College (cursó o está cursando estudios universitarios), Post-Graduate (máster), Doctorate (doctorado) y no tiene valores faltantes. Si vemos la distribución, la mayoría de los clientes son graduados (31%), entonces podemos suponer que no van a querer darse de baja.



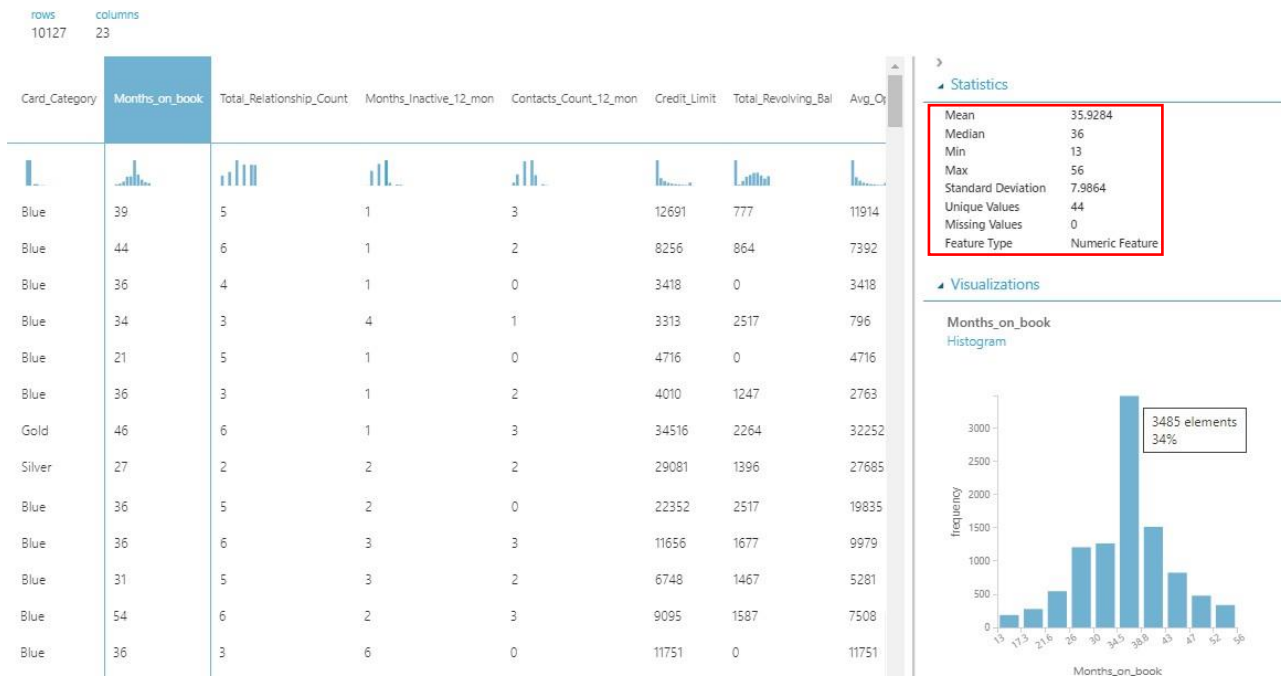
Marital_Status: Observamos que es una variable categórica, toma cuatro valores distintos: Married (casado), Single (soltero), Unknown (el cliente no brindó el dato o es viudo), Divorced (divorciado) y no tiene valores faltantes. Si vemos la distribución, la mayoría de los clientes son casados (46%), entonces podemos suponer que son los que tienen más estabilidad económica por lo que no van a querer darse de baja.



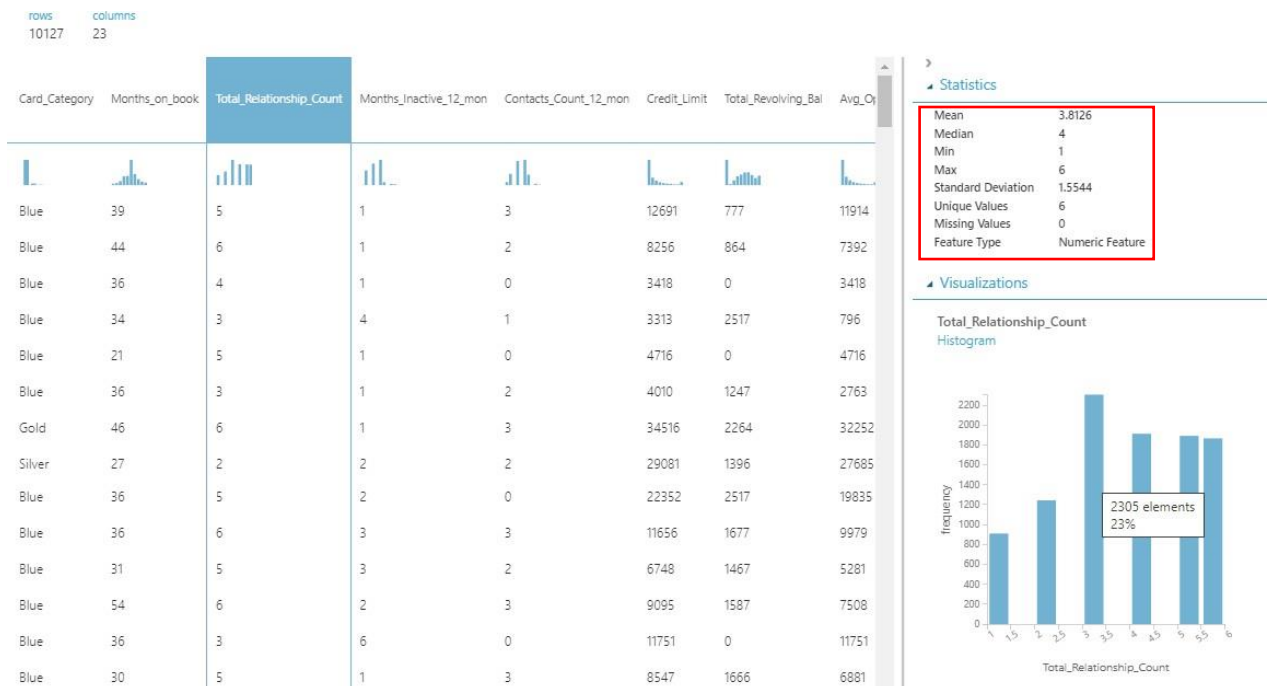
Income_Category: Observamos que es una variable categórica, toma seis valores distintos y no tiene valores faltantes. Si vemos la distribución, la mayoría de los clientes ganan menos de \$40K al año (35%), entonces podemos suponer que es posible que quieran darse de baja.



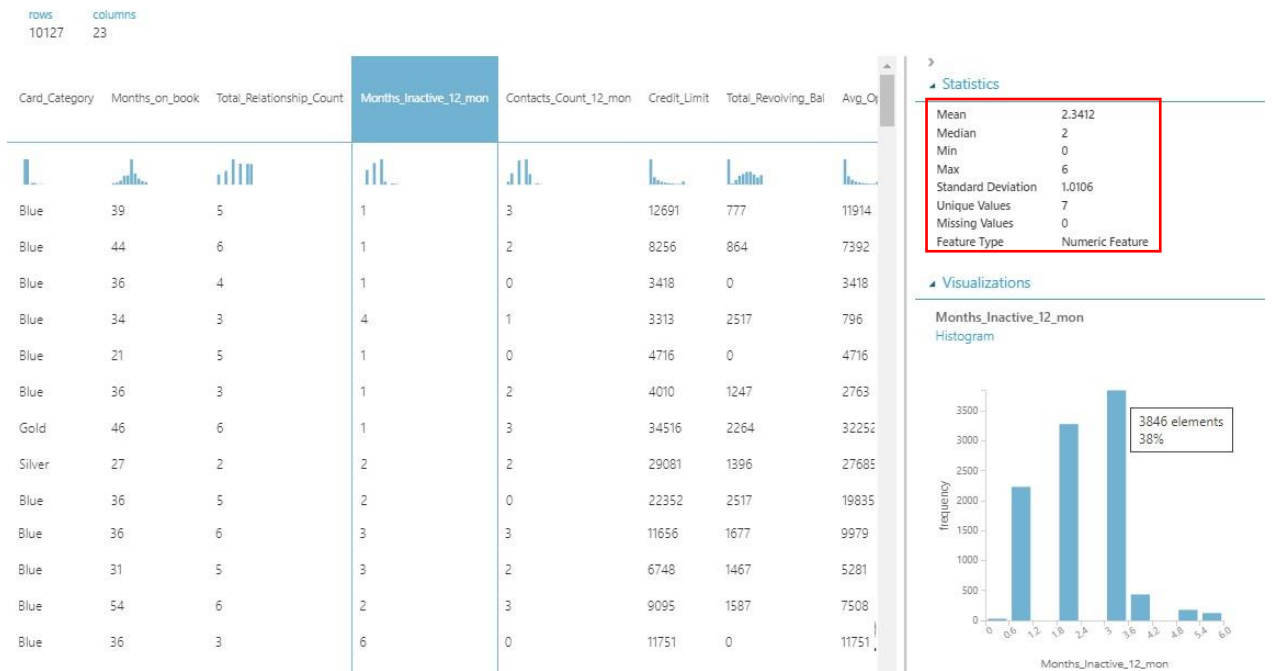
Card_Category: Observamos que es una variable categórica, toma cuatro valores distintos: Blue, Silver, Gold, Platinum y no tiene valores faltantes. Si vemos la distribución, la gran mayoría de los clientes tienen una tarjeta Blue (93%).



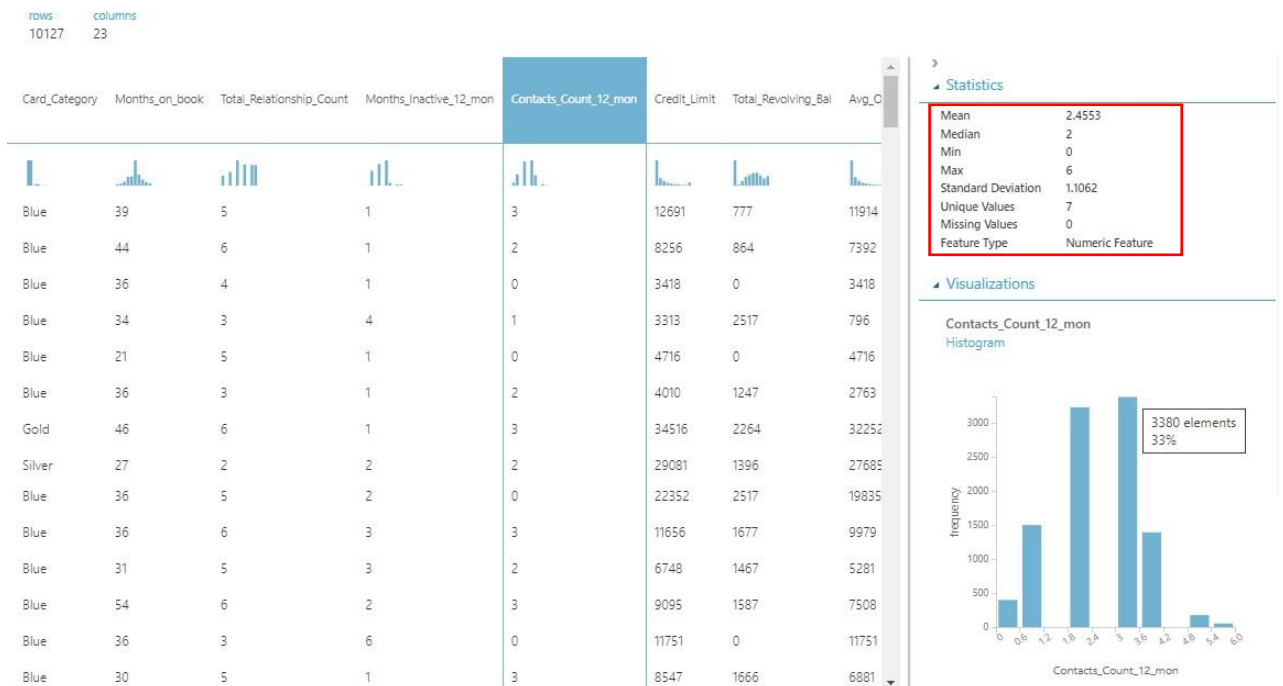
Months_on_book: Observamos que la antigüedad de las cuentas de los clientes no tiene datos faltantes, la media es de 35,9 meses, el máximo es 56 meses y el mínimo es 13. Si vemos la distribución, la mayoría de los clientes se concentran en una antigüedad entre 34.5 y 38.8 meses (34%).



Total_Relationship_Count: Observamos que la cantidad de productos de los clientes no tiene datos faltantes, la media es de 4 productos, el máximo es 6 y el mínimo es 1. Si vemos la distribución, la mayoría de los clientes se concentran en una 3 productos (23%).



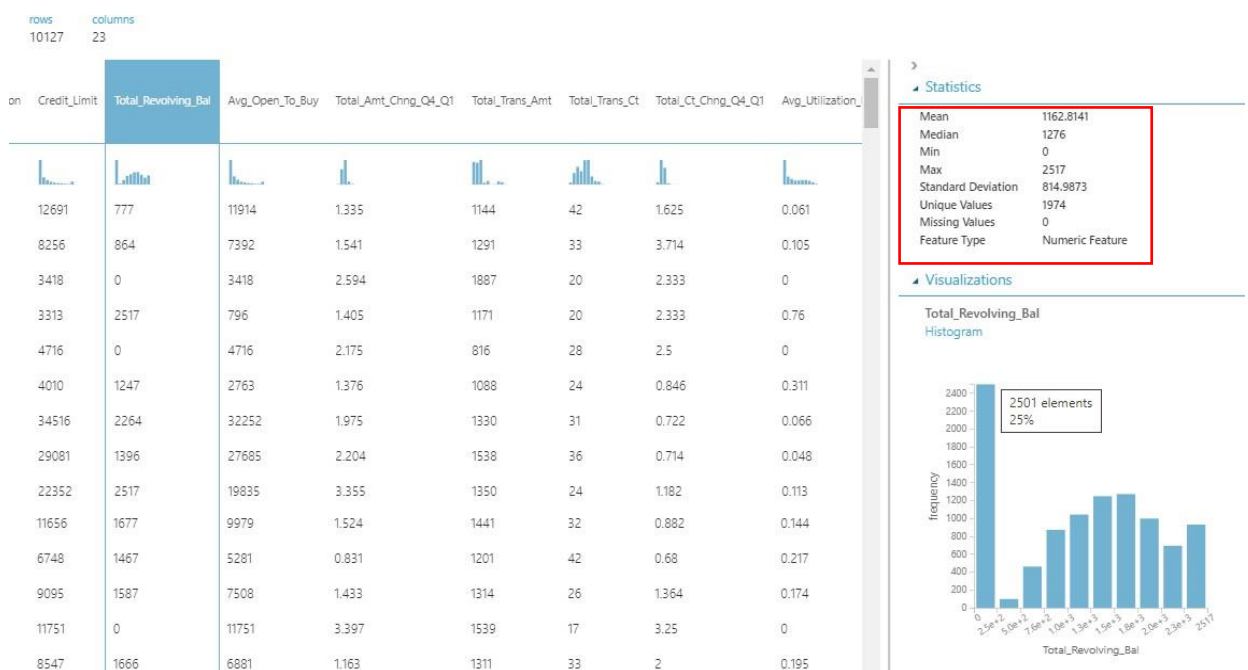
Months_Inactive_12_mon: Observamos que la cantidad de meses que los clientes estuvieron inactivos en el último año no tiene datos faltantes, la media es de 2.3 meses, el máximo es 6 y el mínimo es 0. Si vemos la distribución, la mayoría de los clientes se concentran entre 3 y 3.6 meses inactivos (38%).



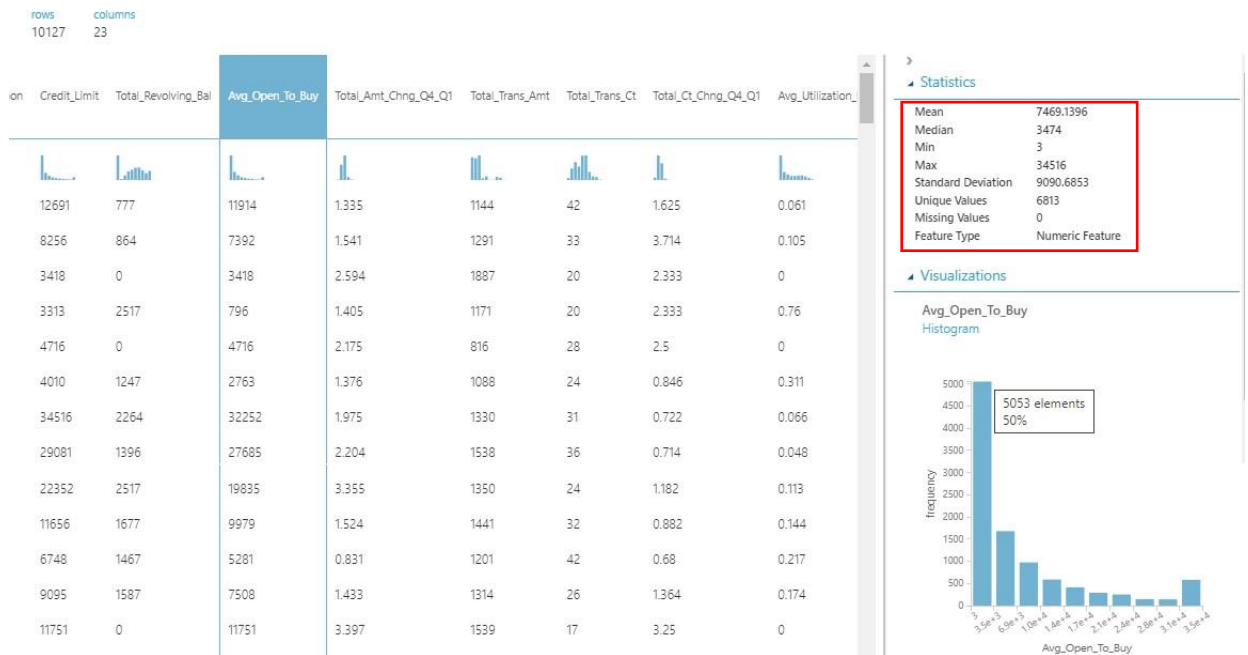
Contacts_Count_12_mon: Observamos que la cantidad de contactos que los clientes tuvieron con el banco en el último año no tiene datos faltantes, la media es de 3 contactos, el máximo es 6 y el mínimo es 0. Si vemos la distribución, la mayoría de los clientes se concentran entre 3 y 3.6 contactos (33%).



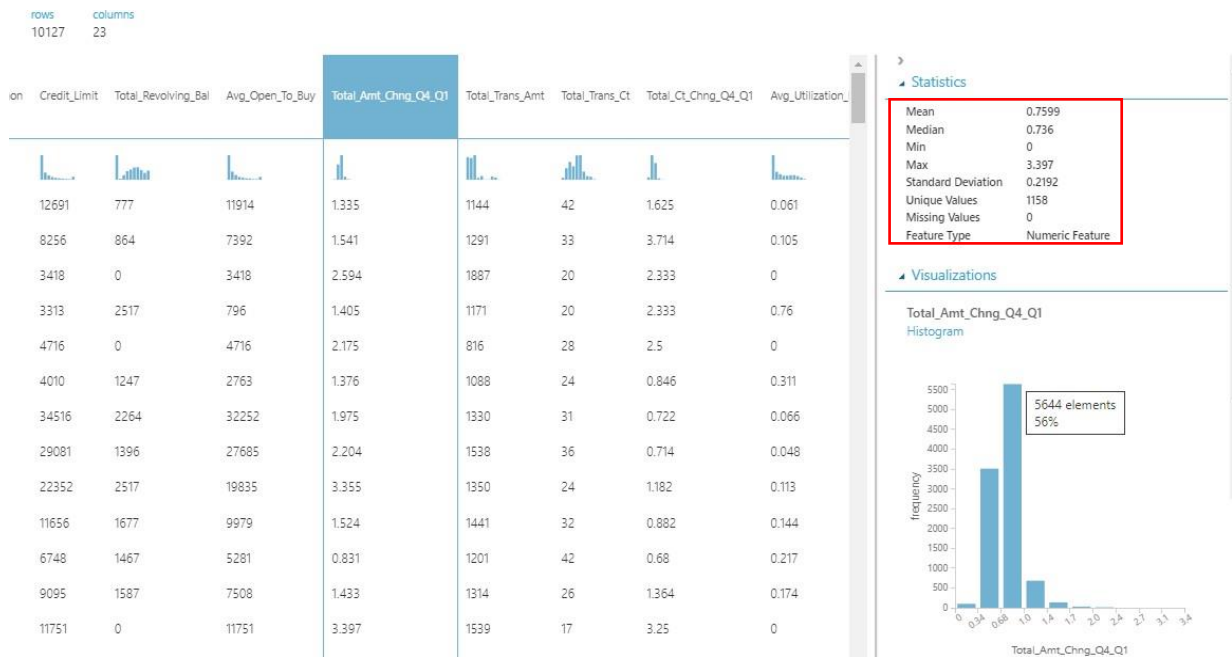
Credit_Limit: Observamos que la cantidad límite de crédito de los clientes no tiene datos faltantes, la media es de 8.631,9 unidades monetarias, el máximo es 34.516 unidades monetarias y el mínimo es 1438,3. Si vemos la distribución, la mayoría de los clientes se concentran entre 1.400 y 4.700 unidades monetarias (51%).



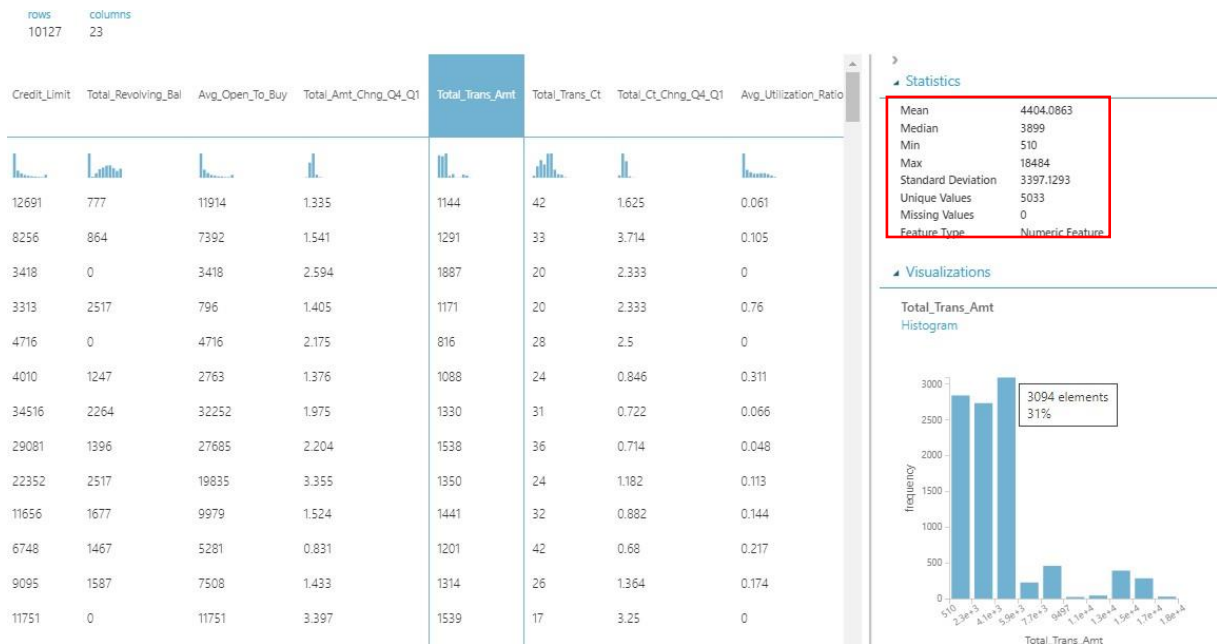
Total_Revolving_Bal: Observamos que el saldo no cubierto de la tarjeta no tiene datos faltantes, la media es de 1.162,8 unidades monetarias, el máximo es 2.517 unidades monetarias y el mínimo es 0. Si vemos la distribución, la mayoría de los clientes se concentran entre 0 y 250 unidades monetarias (25%).



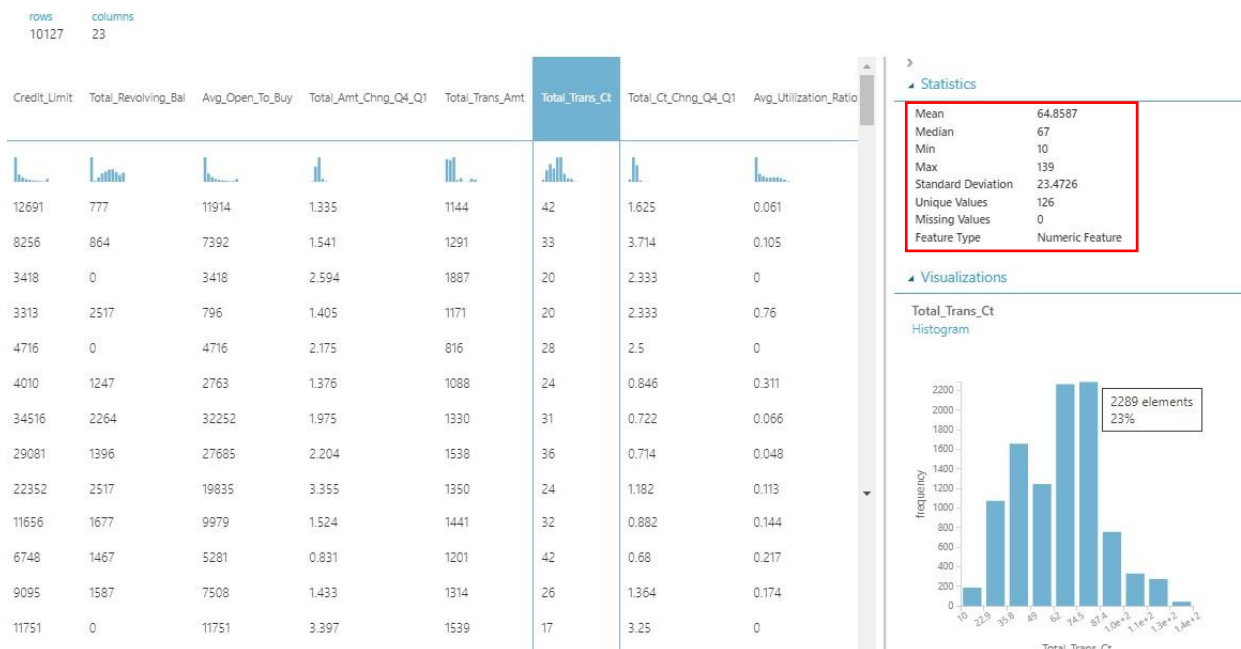
Avg_Open_To_Buy: Observamos que el saldo disponible de la tarjeta no tiene datos faltantes, la media es de 7.469,1 unidades monetarias, el máximo es 34.516 unidades monetarias y el mínimo es 3. Si vemos la distribución, la mayoría de los clientes se concentran entre 0 y 3.500 unidades monetarias (50%).



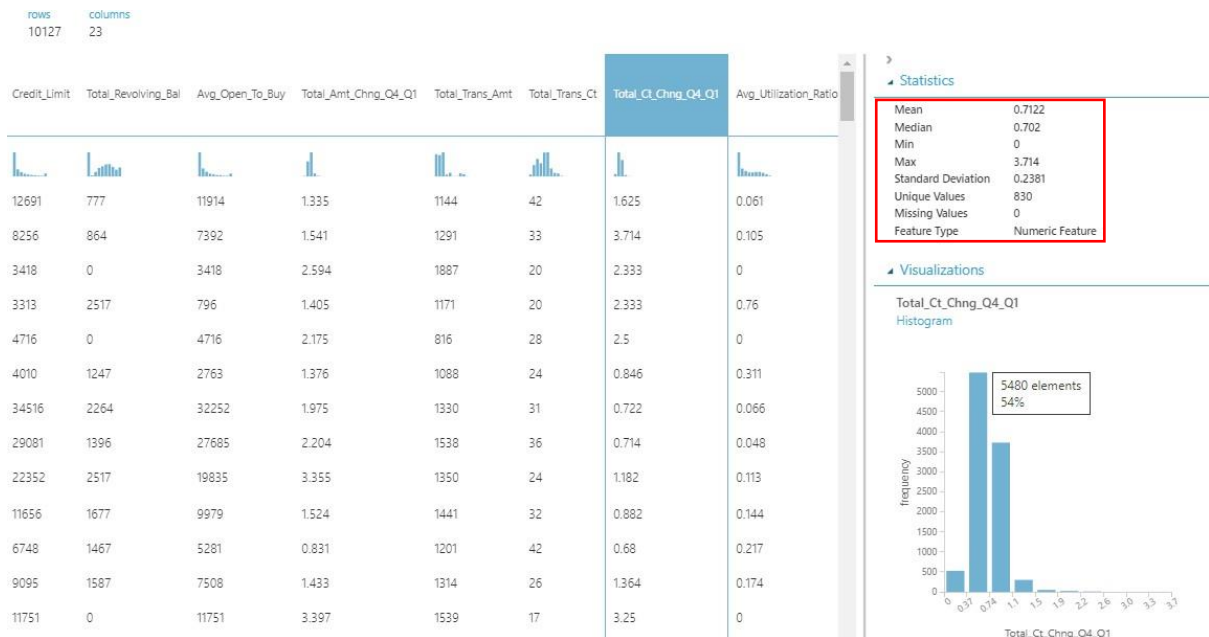
Total_Amt_Chng_Q4_Q1: Observamos que el cambio porcentual de monto de consumos no tiene datos faltantes, la media es de 0.76 por ciento, el máximo es 3,4% y el mínimo es 0. Si vemos la distribución, la mayoría de los clientes se concentran entre 0,68 y 1 por ciento (56%).



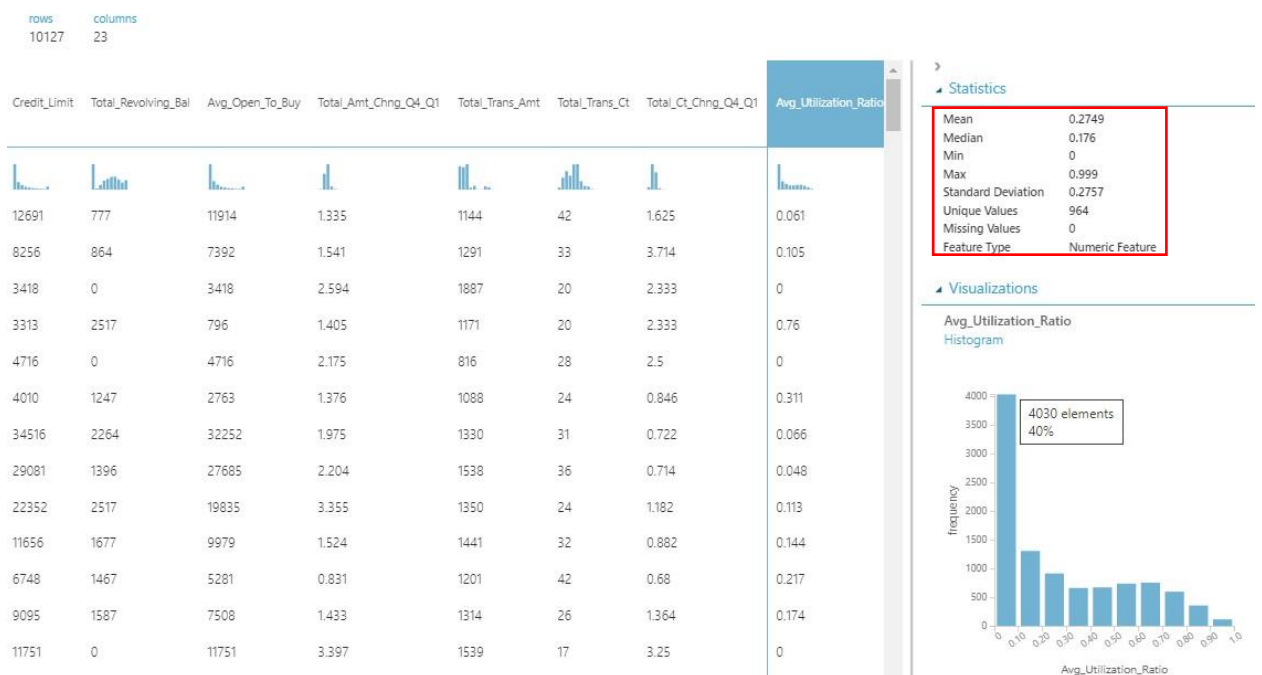
Total_Trans_Amt: Observamos que el monto de consumos en el último año no tiene datos faltantes, la media es de 4.404 unidades monetarias, el máximo 18.484 es y el mínimo es 510. Si vemos la distribución, la mayoría de los clientes se concentran entre 510 y 2.300 unidades monetarias (31%).



Total_Trans_Ct: Observamos que la cantidad de transacciones en el último año no tiene datos faltantes, la media es de 65, el máximo 139 es y el mínimo es 10. Si vemos la distribución, la mayoría de los clientes se concentran entre 10 y 229 transacciones (23%).



Total_Ct_Chng_Q4_Q1: Observamos que el cambio porcentual de cantidad de consumos no tiene datos faltantes, la media es de 0,71 por ciento, el máximo 3,71 es y el mínimo es 0. Si vemos la distribución, la mayoría de los clientes se concentran entre 0 y 0,37 por ciento (54%).

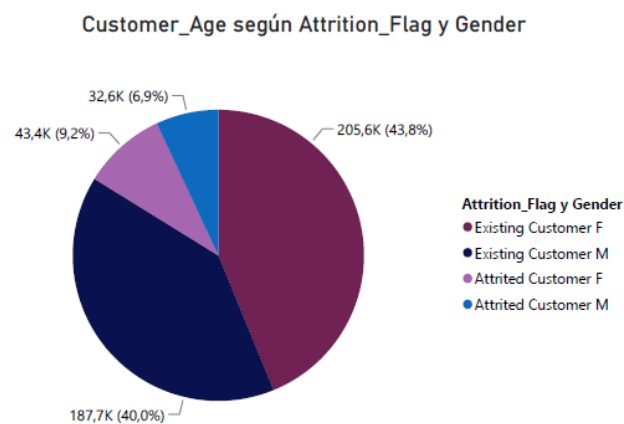


Avg_Utilization_Ratio: Observamos que el ratio de utilización de la tarjeta no tiene datos faltantes, la media es de 0,27 por ciento, el máximo 0,99 es y el mínimo es 0. Si vemos la distribución, la mayoría de los clientes se concentran entre 0 y 0,10 por ciento (40%).

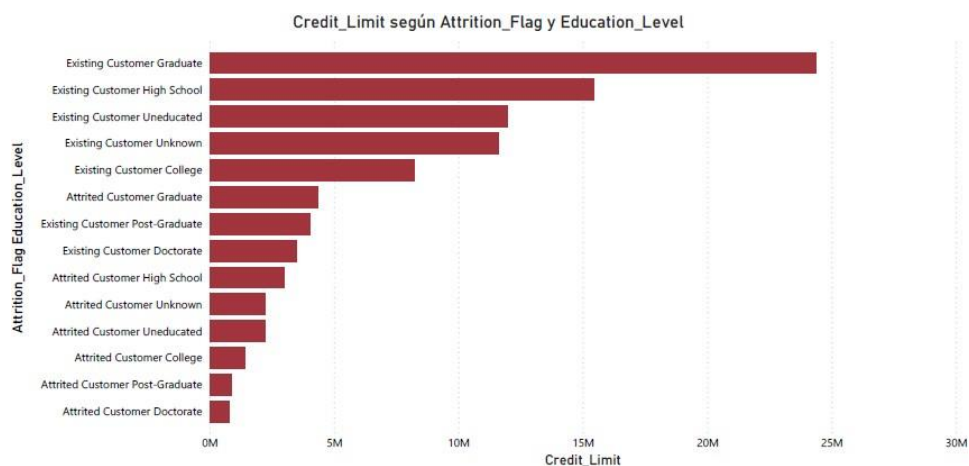
Las variables predictoras son todas excepto la variable objetivo “Attrition_Flag”, que es la que se va a predecir. Conviene agregarlas todas al modelo con excepción de las siguientes:

- **Naive_Bayes_Classifier_Attrition_Flag_Card_Category_Contacts_Count_12_mon_Dependent_count_Education_Level_Months_Inactive_12_mon_1:** No tenemos información sobre esta variable.
- **Naive_Bayes_Classifier_Attrition_Flag_Card_Category_Contacts_Count_12_mon_Dependent_count_Education_Level_Months_Inactive_12_mon_2:** No tenemos información sobre esta variable.

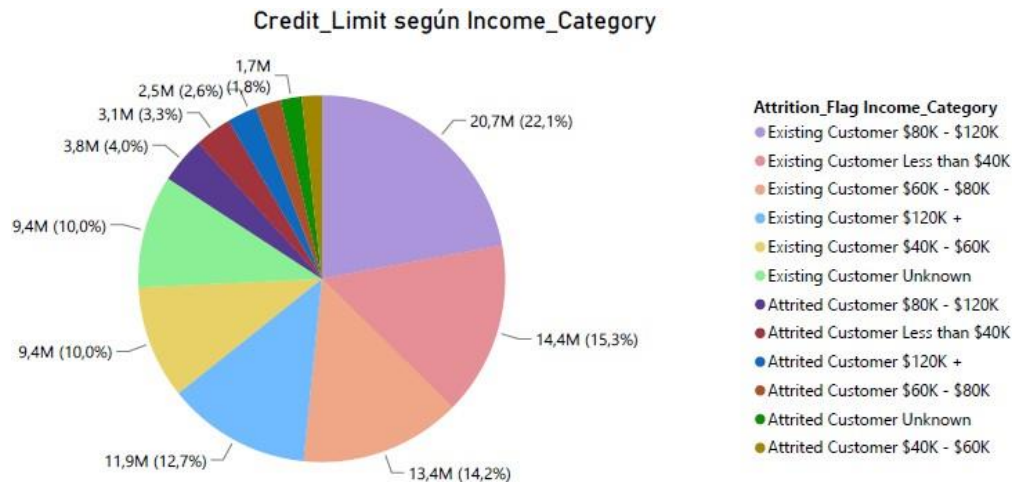
Realizamos un breve análisis exploratorio en Power BI:



Aquí observamos cómo se distribuyen los clientes según su edad, género y categoría de cliente. La mayoría de los clientes son mujeres, 43,8% quiere continuar siendo clientes del banco mientras que 9,2% quiere darse de baja.



Si hacemos el análisis de límite de crédito según el nivel educativo y la categoría de cliente, observamos que los clientes existentes graduados son los que tienen el mayor límite crediticio.



Observamos la representación del límite de crédito según la categoría de ingresos y la categoría de cliente. El mayor límite crediticio lo tienen los clientes que ganan entre \$80K - \$120K al año y quieren seguir en el banco.

Análisis de los resultados obtenidos para los siguientes modelos:

	AUC	True Positive	False Negative	False Positive	True Negative	Acuracy	Precision	Recall	F1 Score
Modelo base, con regularización	0,935	1646	52	126	201	0,912	0,929	0,969	0,949
Modelo 1: Sólo variables numéricas, con regularización	0,928	1641	57	132	195	0,907	0,926	0,966	0,946
Modelo 2: Variables numéricas + FE, con regularización	0,943	1650	48	120	207	0,917	0,932	0,972	0,952
Modelo 3: Variables numéricas + FE + Filter based, con regularización	0,944	1647	51	121	206	0,915	0,932	0,97	0,95

Cuanto mayor es el área bajo la curva (AUC) significa que el modelo es mejor para predecir el 0 como 0 y 1 como 1 → **a mayor AUC, mejor es el modelo** → como para este modelo base el AUC nos da un valor de 0.935 quiere decir que como primera aproximación al modelado está haciendo una predicción correcta.

Las funciones que les aplicamos a las variables crudas numéricas son las siguientes:

```
select * ,  
1- Avg_Open_To_Buy as inv_OpenToBuy,  
power (Credit_Limit,2) as sq_CreditLimit,  
power (Total_Revolving_Bal,2) as sq_TotalRevolving,  
power (Avg_Open_To_Buy,2) as sq_OpenToBuy,  
power (Credit_Limit,3) as cu_CreditLimit,  
power (Total_Revolving_Bal,3) as cu_TotalRevolving,  
power (Avg_Open_To_Buy,3) as cu_OpenToBuy  
from t1;
```

Usamos el módulo “Filter Based Feature Selection” para visualizar las variables que tengan muy poca relación con la variable objetivo y así excluirlas.



De forma arbitraria quitamos del modelo todas las variables con un score menor a 0.01 → Credit_Limit, sq_Credit_Limit, cu_Credit_Limit, Customer_Age, Months_on_book, Dependent_count.

Observamos que el AUC del modelo 3 da casi igual que el AUC del modelo 2, por lo que podemos concluir que las variables que quitamos no influyen en la variable objetivo.

Modelo 4: Variables categóricas




Las variables categóricas son:

- **Attrition_Flag:** Es la variable objetivo.
- **Gender:** Indica las categorías de género de una persona.
- **Education_Level:** Indica las categorías del nivel de educación de una persona.
- **Marital_Status:** Indica las categorías maritales de una persona.
- **Income_Category:** Indica las categorías de ingresos de una persona.
- **Card_Category:** Indica las categorías de tipo de tarjetas que puede tener una persona.

Realizamos el análisis bivariado para cada variable categórica y así poder conocer cómo es la relación de cada uno de los valores de las categorías con la variable objetivo.

a. Variable Education_Level:

```
select Education_Level,
       sum(case when Attrition_Flag = 'Existing Customer' then 1 else 0 end)*1.0/count(*) as Percentage_Existing_Customer,
       count(*) as Total
from t1
group by Education_Level
Order by Percentage_Existing_Customer;
```

Education_Level	Percentage_Existing_Customer	Total
		
Doctorate	0.789357	451
Post-Graduate	0.821705	516
Unknown	0.831468	1519
Uneducated	0.840619	1487
Graduate	0.844309	3128
College	0.847976	1013
High School	0.847988	2013

Percentage_Existing_Customer representa si los clientes van a continuar con su suscripción al banco.

- La categoría “Doctorate” corresponde a los clientes que cuentan con un doctorado y es la que tiene una tasa de “clientes existentes” (Percentage_Existing_Customer) más baja (0.789357) → está indicando que este tipo de “Education_Level” es el que cae en más clientes que no quieren continuar con los servicios del banco.
- La categoría “Post-Graduate” corresponde a los clientes que cuentan con un nivel educativo superior al título de grado, “Unknown” contiene los valores faltantes (en blanco) o no revelados, “Uneducated” corresponde a los clientes que no tienen educación, “Graduate” significa que se graduó de la universidad, “College” indica que fue a la universidad y “High School” indica que hizo el liceo → está indicando que estos tipos de “Education_Level” tienen clientes que quieren continuar en el banco en mayor medida.

Education_Level	Percentage_Existing_Customer	Total
Doctorate	0,789357	451
Post-Graduate	0,821705	516
Unknown	0,831468	1519
Uneducated	0,840619	1487
Graduate	0,844309	3128
College	0,847976	1013
High School	0,847988	2013

Las agrupaciones que hicimos fueron las siguientes:

- “High School”, “College” y “Graduate” se pueden agrupar porque tienen una tasa de “clientes existentes” muy similar → su comportamiento debe ser similar también (en promedio).
- “Uneducated”, “Unknown” y “Post-Graduate” se pueden agrupar porque tienen una tasa de “clientes existentes” también muy similar → su comportamiento debe ser similar también (en promedio).
- “Doctorate” lo dejamos solo porque es el que tiene la menor tasa de “clientes existentes”.

b. Variable Gender:

```
select Gender,
       sum(case when Attrition_Flag = 'Existing Customer' then 1 else 0 end)*1.0/count(*) as Percentage_Existing_Customer,
       count(*) as Total
from t1
group by Gender
Order by Percentage_Existing_Customer;
```

Gender	Percentage_Existing_Customer	Total
F	0.826428	5358
M	0.853848	4769

- La categoría “F” corresponde a los clientes de sexo femenino y es la que tiene una tasa de “clientes existentes” más baja (0.826428) → está indicando que este tipo de “Gender” es el que cae en más clientes que no quieren continuar con los servicios del banco.
- La categoría “M” corresponde a los clientes de sexo masculino → está indicando que este tipo de “Gender” tiene clientes que quieren continuar en el banco en mayor medida.

A partir de esta información podríamos crear una variable dummy que indique si la persona tiene registro de no querer pertenecer más al banco.

c. Variable Marital_Status:

```
select Marital_Status,
       sum(case when Attrition_Flag = 'Existing Customer' then 1 else 0 end)*1.0/count(*) as Percentage_Existing_Customer,
       count(*) as Total
from t1
group by Marital_Status
Order by Percentage_Existing_Customer;
```

Marital_Status	Percentage_Existing_Customer	Total
Unknown	0.82777	749
Single	0.830586	3943
Divorced	0.838235	748
Married	0.848731	4687

- La categoría “Unknown” contiene los valores faltantes (en blanco) o no revelados y es la que tiene una tasa de “clientes existentes” más baja (0.82777) → está indicando que este tipo de “Marital_Status” es el que cae en más clientes que no quieren continuar con los servicios del banco.
- La categoría “Single” corresponde a los clientes que son solteros, “Divorced” a los clientes que son divorciados y “Married” a los clientes que están casados → está indicando que estos tipos de “Marital_Status” tienen clientes que quieren continuar en el banco en mayor medida.


Marital_Status	Percentage_Existing_Customer	Total
Unknown	0,82777	749
Single	0,830586	3943
Divorced	0,838235	748
Married	0,848731	4687

Las agrupaciones que hicimos fueron las siguientes:

- “Married” lo dejamos solo porque es el que tiene la mayor tasa de “clientes existentes”.
- “Single” y “Divorced” se pueden agrupar porque tienen una tasa de “clientes existentes” muy similar → su comportamiento debe ser similar también (en promedio).
- “Unknown” lo dejamos solo porque es el que tiene la menor tasa de “clientes existentes”.

d. Variable Income_Category:

```
select Income_Category,
       sum(case when Attrition_Flag = 'Existing Customer' then 1 else 0 end)*1.0/count(*) as Percentage_Existing_Customer,
       count(*) as Total
from t1
group by Income_Category
Order by Percentage_Existing_Customer;
```

Income_Category	Percentage_Existing_Customer	Total
		
\$120K +	0.826685	727
Less than \$40K	0.828138	3561
Unknown	0.831835	1112
\$80K - \$120K	0.842345	1535
\$40K - \$60K	0.848603	1790
\$60K - \$80K	0.865193	1402

- La categoría “\$120K+” corresponde a los clientes que ganan 120 mil dólares o más al año y es la que tiene una tasa de “clientes existentes” más baja (0.826685) → está indicando que este tipo de “Income_Category” es el que cae en más clientes que no quieren continuar con los servicios del banco.
- La categoría “Less than \$40K” corresponde a los clientes que ganan menos de 40 mil dólares al año y tiene una tasa de “clientes existentes” baja (0.828138), muy similar a la anterior.
- Las demás categorías tienen una tasa de “clientes existentes” más alta.




Income_Category	Percentage_Existing_Customer	Total
\$120K +	0,826685	727
Less than \$40K	0,828138	3561
Unknown	0,831835	1112
\$80K - \$120K	0,842345	1535
\$40K - \$60K	0,848603	1790
\$60K - \$80K	0,865193	1402

Las agrupaciones que hicimos fueron las siguientes:

- “\$120K+”, “Less than \$40K” y “Unknown” se pueden agrupar porque tienen una tasa de “clientes existentes” muy similar → su comportamiento debe ser similar también (en promedio).
- “\$80K - \$120K” y “\$40K - \$60K” se pueden agrupar porque tienen una tasa de “clientes existentes” muy similar → su comportamiento debe ser similar también (en promedio).
- “\$60K - \$80K” lo dejamos solo porque es el que tiene la mayor tasa de “clientes existentes”.

e. Variable Card_Category:

```
select Card_Category,
       sum(case when Attrition_Flag = 'Existing Customer' then 1 else 0 end)*1.0/count(*) as Percentage_Existing_Customer,
       count(*) as Total
from t1
group by Card_Category
Order by Percentage_Existing_Customer;
```

Card_Category	Percentage_Existing_Customer	Total
		
Platinum	0.75	20
Gold	0.818966	116
Blue	0.839021	9436
Silver	0.852252	555

- La categoría “Platinum” corresponde a los clientes que tienen una tarjeta de ese color y es la que tiene una tasa de “clientes existentes” más baja (0.75) → está indicando que este tipo de “Income_Category” es el que cae en más clientes que no quieren continuar con los servicios del banco. Además observamos que muy pocas personas acceden a ella (20), entonces deberíamos agruparla con la siguiente categoría. Esta misma situación ocurre con la categoría “Gold” ya que solo tiene 116 observaciones → la vamos a agrupar con la categoría “Platinum”.
- Las demás categorías tienen una tasa de “clientes existentes” más alta y muchas más observaciones.

Card_Category	Percentage_Existing_Customer	Total
Platinum	0,75	20
Gold	0,818966	116
Blue	0,839021	9436
Silver	0,852252	555

Las agrupaciones que hicimos fueron las siguientes:

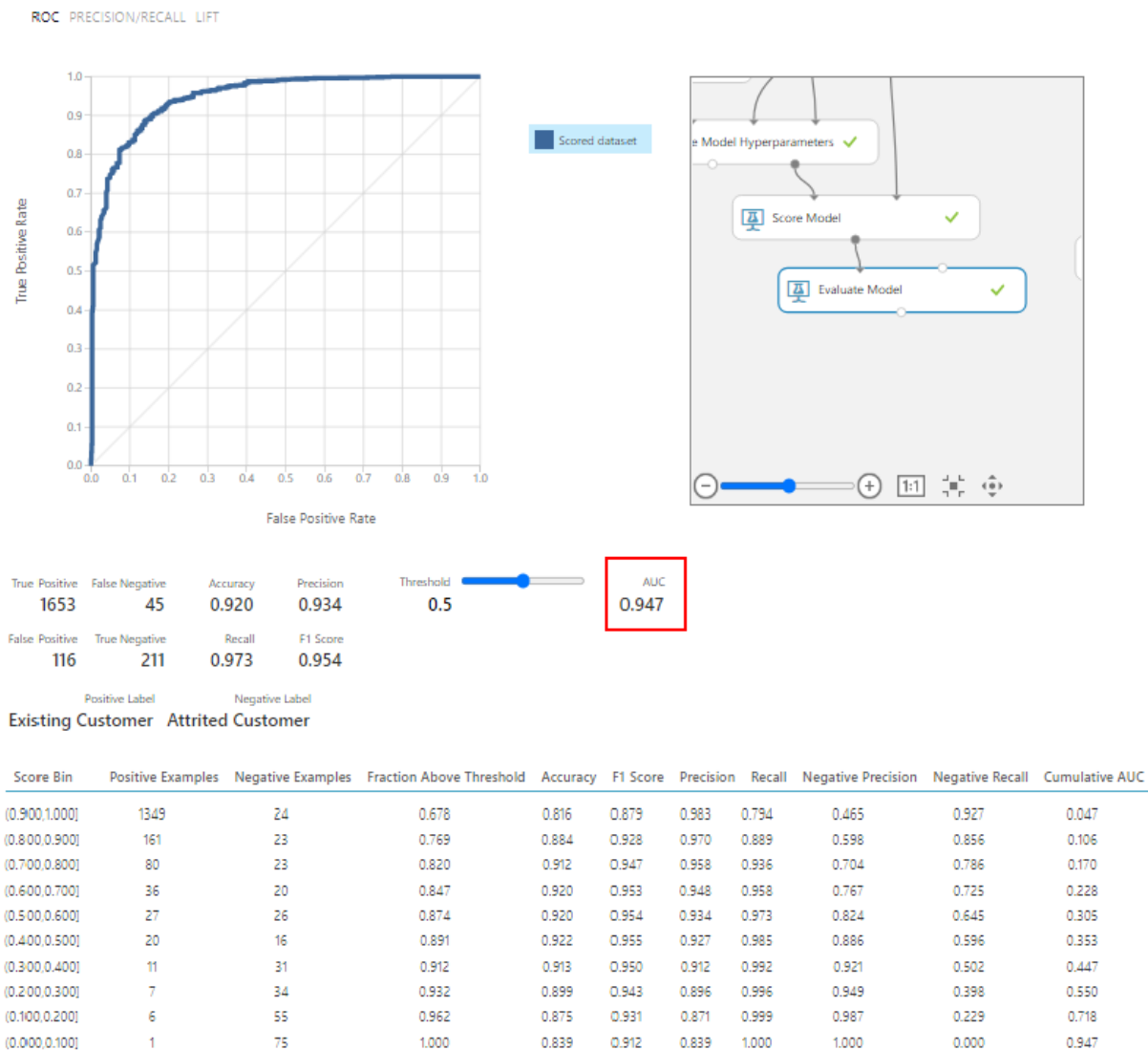
- “Platinum” y “Gold” se pueden agrupar por lo que definimos anteriormente.
- “Blue” y “Silver” los dejamos solos porque, si bien son los que tienen la mayor tasa de “clientes existentes”, difieren un poco entre ellos.

```

select *,
  1 - Avg_Open_To_Buy as inv_OpenToBuy,
  power (Credit_Limit,2) as sq_CreditLimit,
  power (Total_Revolving_Bal,2) as sq_TotalRevolving,
  power (Avg_Open_To_Buy,2) as sq_OpenToBuy,
  power (Credit_Limit,3) as cu_CreditLimit,
  power (Total_Revolving_Bal,3) as cu_TotalRevolving,
  power (Avg_Open_To_Buy,3) as cu_OpenToBuy,
  case when Education_Level in ('High School', 'College','Graduate') then 1 else 0 end as EL_sch_col_grad,
  case when Education_Level in ('Uneducted', 'Unknown', 'Post-Graduate') then 1 else 0 end as EL_uned_unk_postgrad,
  case when Education_Level = 'Doctorate' then 1 else 0 end as EL_doc,
  case when Gender = 'M' then 1 else 0 end as G_male,
  case when Marital_Status = 'Married' then 1 else 0 end as MS_mar,
  case when Marital_Status in ('Single', 'Divorced') then 1 else 0 end as MS_sing_div,
  case when Marital_Status = 'Unknown' then 1 else 0 end as MS_unk,
  case when Income_Category in ('$120K +', 'Less than $40K', 'Unknown') then 1 else 0 end as IC_120_less40_unk,
  case when Income_Category in ('$80K - $120K', '$40K - $60K') then 1 else 0 end as IC_unk_80_120_40_60,
  case when Income_Category = '$60K - $80K' then 1 else 0 end as IC_60_80,
  case when Card_Category = 'Blue' then 1 else 0 end as CC_blue,
  case when Card_Category = 'Silver' then 1 else 0 end as CC_silver,
  case when Card_Category in ('Platinum', 'Gold') then 1 else 0 end as CC_plat_gold
from t1;

```

Creamos las variables dummies para cada variable categórica usando las agrupaciones anteriores.

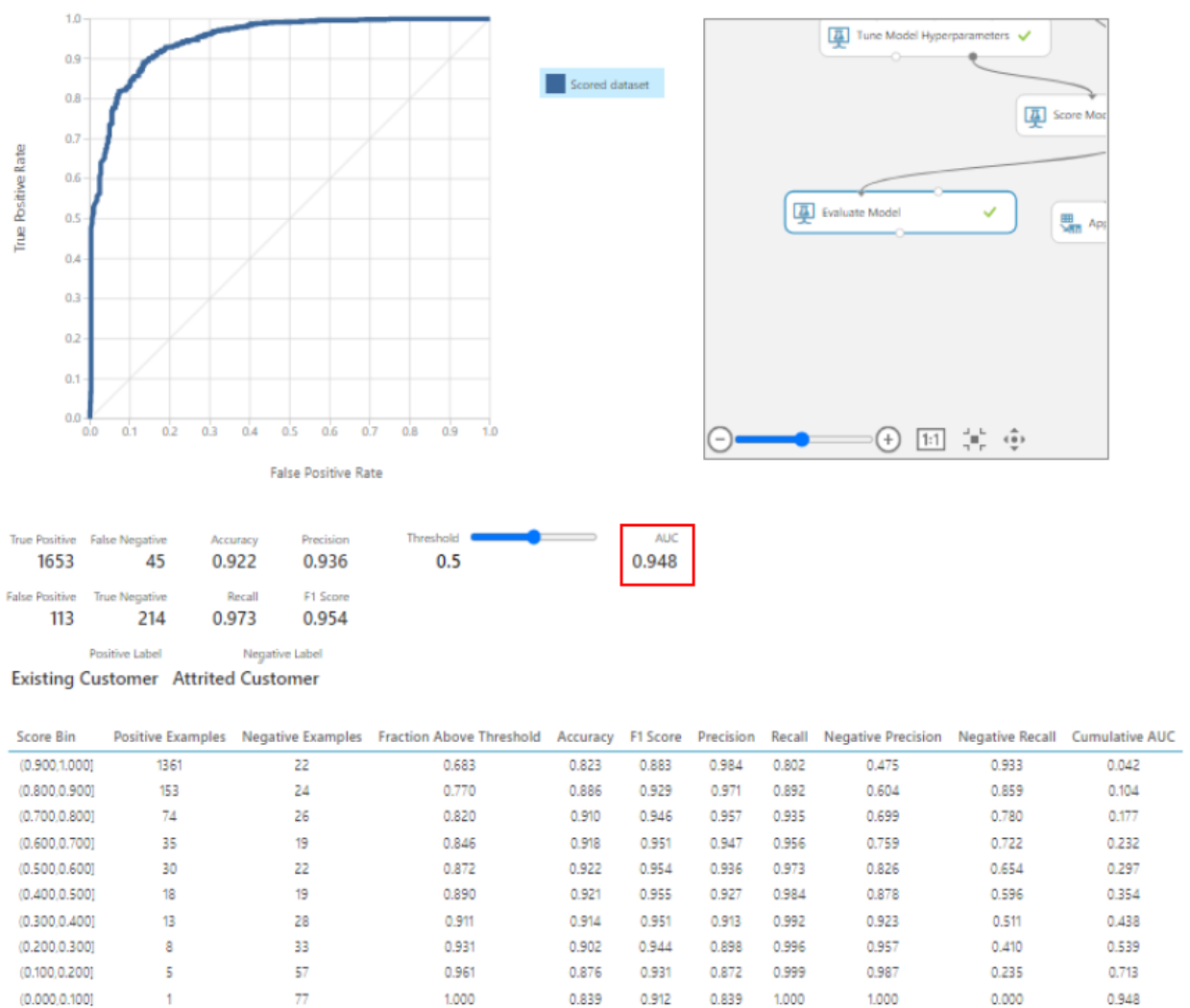


Observamos que el modelo 4 mejora respecto al anterior ya que el AUC aumenta.

Modelo 5: Variables interactivas

```
select *,
    EL_sch_col_grad * Avg_Open_to_Buy as Avg_Open_to_Buy_EL_sch_col_grad,
    EL_uned_unk_postgrad * Avg_Open_to_Buy as Avg_Open_to_Buy_EL_uned_unk_postgrad,
    G_male * Avg_Open_to_Buy as Avg_Open_to_Buy_G_male,
    MS_mar * Avg_Open_to_Buy as Avg_Open_to_Buy_MS_mar,
    MS_sing_div * Avg_Open_to_Buy as Avg_Open_to_Buy_MS_sing_div,
    IC_80_120_40_60 * Avg_Open_to_Buy as Avg_Open_to_Buy_IC_80_120_40_60,
    CC_blue * Avg_Open_to_Buy as Avg_Open_to_Buy_CC_blue,
    CC_silver * Avg_Open_to_Buy as Avg_Open_to_Buy_CC_silver
from t1;
```

Aquí estudiamos las interacciones del saldo disponible de la tarjeta de los clientes con las variables dummies que creamos anteriormente.



El AUC se incrementa ligeramente al agregar variables interactivas.

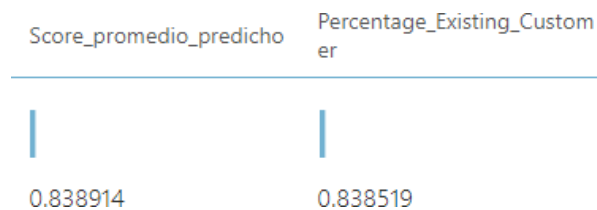
Conclusiones del análisis:

	AUC	True Positive	False Negative	False Positive	True Negative	Accuracy	Precision	Recall	F1 Score
Modelo base, con regularización	0,935	1646	52	126	201	0,912	0,929	0,969	0,949
Modelo 1: Sólo variables numéricas, con regularización	0,928	1641	57	132	195	0,907	0,926	0,966	0,946
Modelo 2: Variables numéricas + FE, con regularización	0,943	1650	48	120	207	0,917	0,932	0,972	0,952
Modelo 3: Variables numéricas + FE + Filter based, con regularización	0,944	1647	51	121	206	0,915	0,932	0,97	0,95
Modelo 4: Variables numéricas + FE + Filter Based y dummies, con regularización	0,947	1653	45	116	211	0,92	0,934	0,973	0,954
Modelo 5: Variables numéricas + FE + Filter Based + dummies e interactivas, con regularización	0,948	1653	45	113	214	0,922	0,936	0,973	0,954

El modelo que seleccionaríamos es el número 5, ya que es el que tiene mayor valor de AUC.

Comparamos el Score Promedio Predicho versus el Porcentaje de “clientes existentes” que hay en la muestra (valor real):

```
select
  avg([Scored Probabilities]) as Score_promedio_predicho,
  sum(case when Attrition_Flag = 'Existing Customer' then 1 else 0 end)*1.0/count(*) as Percentage_Existing_Customer
from t1;
```







En términos globales, el valor predicho (0.838914) y el valor real (0.838519) dan valores muy similares entonces podemos decir que el modelo es aceptable → 0,04%.

Una diferencia entre el valor predicho y el valor real de hasta un $\pm 5\%$ es lo que vamos a aceptar.

Ahora hacemos el análisis por deciles y averiguar si también es buena la predicción por tramos:

```
select Attrition_Flag,
  [Scored Probabilities],
  (Select count(*) from t1 as t1bis where t1bis.[Scored Probabilities] <= t1.[Scored Probabilities]) as Orden
from t1
order by [Scored Probabilities];
```

```
select Orden/200 as Grupo,
  avg(case when Attrition_Flag = 'Existing Customer' then 1 else 0 end) as Frecuencia_real,
  avg([Scored Probabilities]) as Frecuencia_estimada,
  count(*)
from t1
group by Grupo;
```

	Grupo	Frecuencia_real	Frecuencia_estimada	count(*)
Estos serían los deciles malos porque tienen la menor cantidad de "clientes existentes" si nos fijamos en la frecuencia_real. El límite lo marcamos en 0.60.				
	0	0.150754	0.172375	199
	1	0.565	0.572046	200
	2	0.81	0.808062	200
	3	0.92	0.908227	200
	4	0.97	0.954651	200
Estos serían los deciles buenos porque tienen la mayor cantidad de "clientes existentes" si nos fijamos en la frecuencia_real.	5	0.955	0.975036	200
	6	0.995	0.986513	200
	7	1	0.992878	200
	8	1	0.996436	200
	9	0.995	0.998683	200
	10	1	0.999709	26

Observamos que la frecuencia estimada se condice con la frecuencia real, entonces tramo por tramo vemos que la frecuencia_estimada también ordena de menor a mayor como la frecuencia_real y que además la frecuencia_estimada es bastante similar a la frecuencia_real.