

Machine Learning Project - LOGIT

Description of the problem:

A bank wants to know the customer cancellation rate for the next month, in such way that it can carry out a campaign with the aim of recovering the customer before they communicate their decision to unsubscribe.

To do this, we seek to create the best possible predictive model for the variable "Attrition_Flag" using the dataset found in the "BankChurners.csv" file.

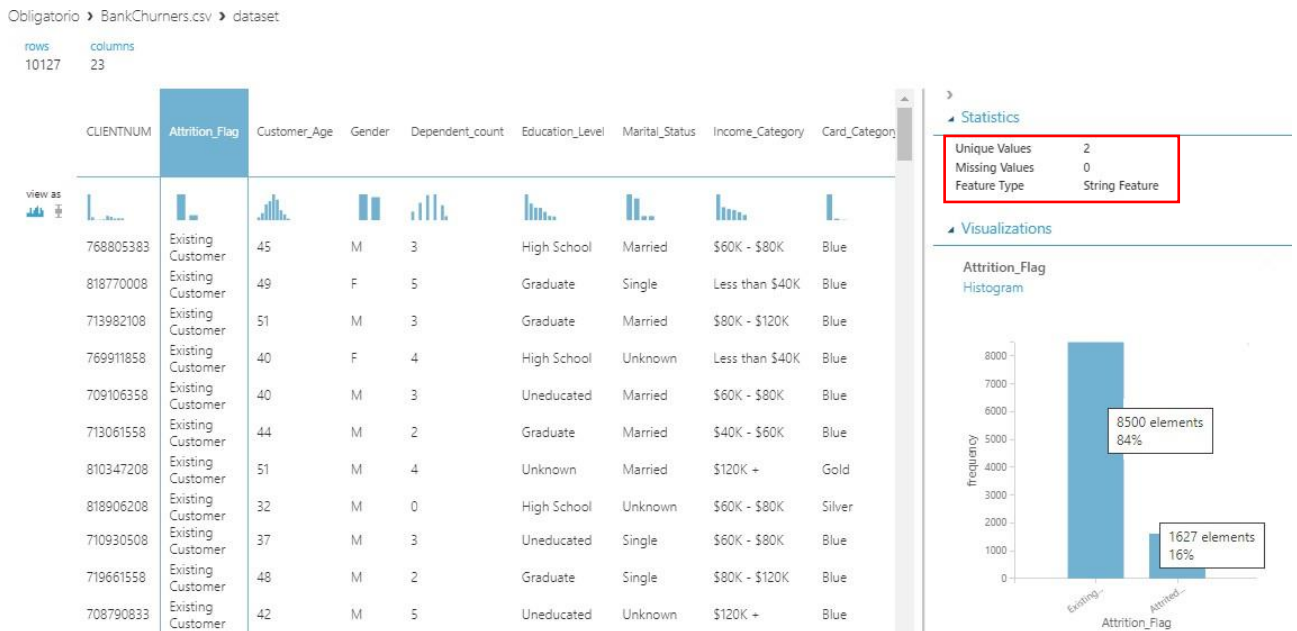
References to variables:

- **CLIENTNUM:** Number of clients → quantitative variable.
- **Attrition_Flag:** Account status the following month → qualitative variable.
- **Customer_Age:** quantitative variable.
- **Gender:** qualitative variable.
- **Dependent_count:** Number of people who are responsible for others → quantitative variable.
- **Education_Level:** qualitative variable.
- **Marital_Status:** qualitative variable.
- **Income_Category:** qualitative variable.
- **Card_Category:** qualitative variable.
- **Months_on_book:** Account's age → qualitative variable.
- **Total_Relationship_Count:** Number of customer's products (accounts and cards) → quantitative variable.
- **Months_Inactive_12_mon:** Number of months inactive in the last 12 months → quantitative variable.
- **Contacts_Count_12_mon:** Number of contacts in the last 12 months (inquiries/claims to the bank) → quantitative variable.
- **Credit_Limit:** quantitative variable.
- **Total_Revolving_Bal:** Uncovered balance of the card (it would be what the client has used of the amount on his card, it is the difference between Credit_Limit y Avg_Open_To_Buy) → quantitative variable.
- **Avg_Open_To_Buy:** Available amount in the card → quantitative variable.
- **Total_Amt_Chng_Q4_Q1:** Percentage change in the amount of consumption → quantitative variable.
- **Total_Trans_Amt:** Amount of consumption in the last 12 months → quantitative variable.
- **Total_Trans_Ct:** Number of transactions in the last 12 months → quantitative variable.
- **Total_Ct_Chng_Q4_Q1:** Percentage change in amount of consumption → quantitative variable.
- **Avg_Utilization_Ratio:** Card utilization ratio (it is the result of doing Total_Revolving_Bal divided by Credit_Limit) → quantitative variable.

Experiment in Azure:

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

Analysis of the target variable (y):

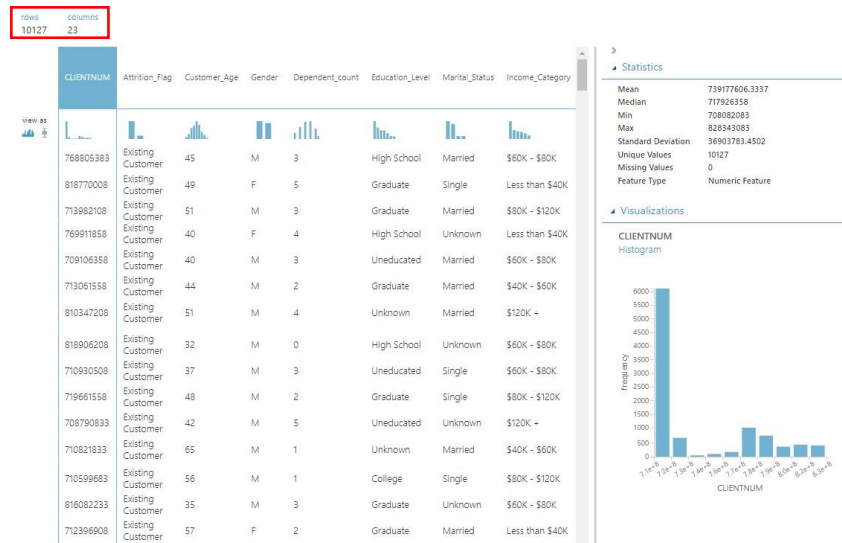


Our target variable is "Attrition_Flag", which takes two values:

- **Existing Customer:** It is the client who will continue in the bank.
- **Attrited Customer:** It is the client who wants to cancel the service.

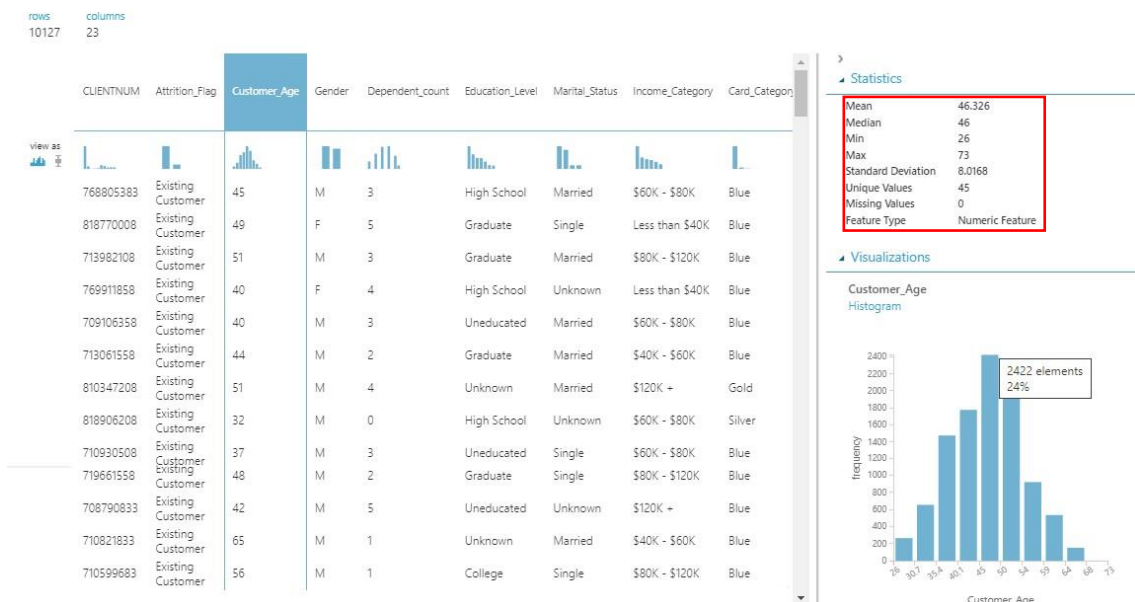
We observe that the variable "Attrition_Flag" has no missing data and is of type "String Feature", this indicates that it is a categorical variable. From this variable, the dichotomous variables or also known as "dummies" will be created. It has 84% of existing customers and 16% of customers who want to unsubscribe.

Brief Exploratory Data Analysis (EDA):

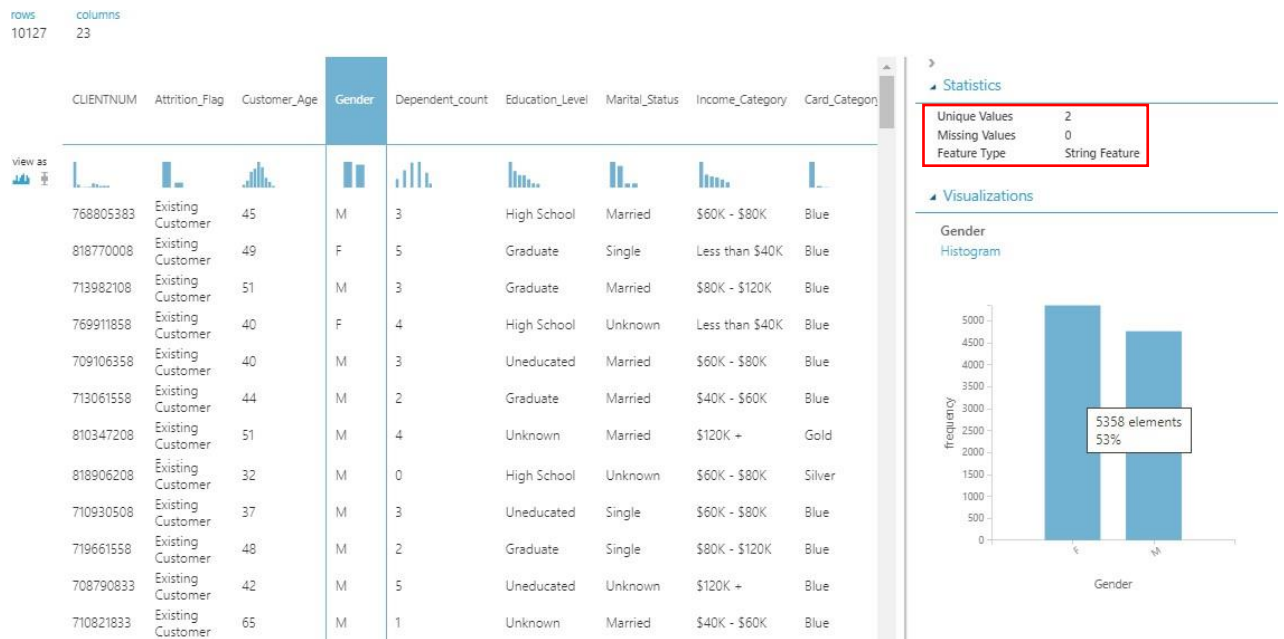


The dataset consists of 10,127 rows, each of which represents a bank customer. The 23 columns represent 22 predictor variables, in addition to the dependent variable (also called “to be predicted”).

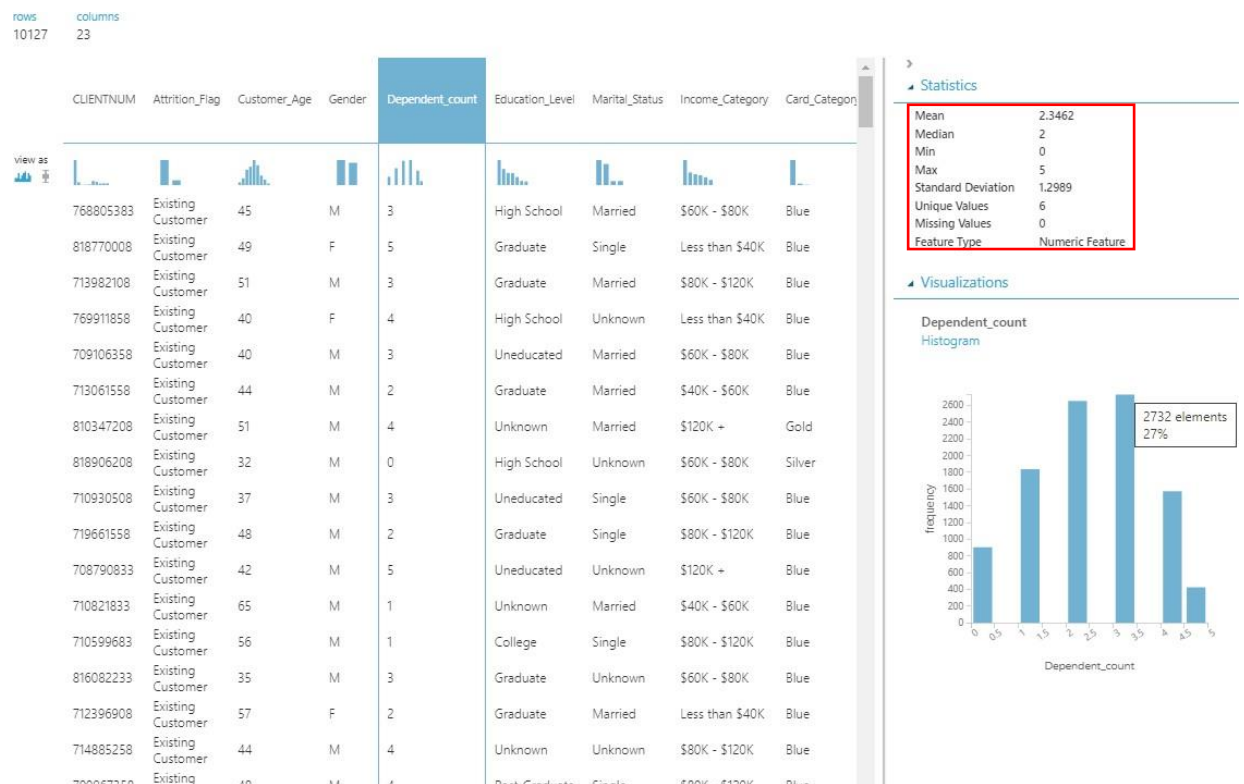
CLIENTNUM: It is not going to give us any type of information when running our model because it is a unique identifier of the customer and of the purchase and we are not going to need it to train our model.



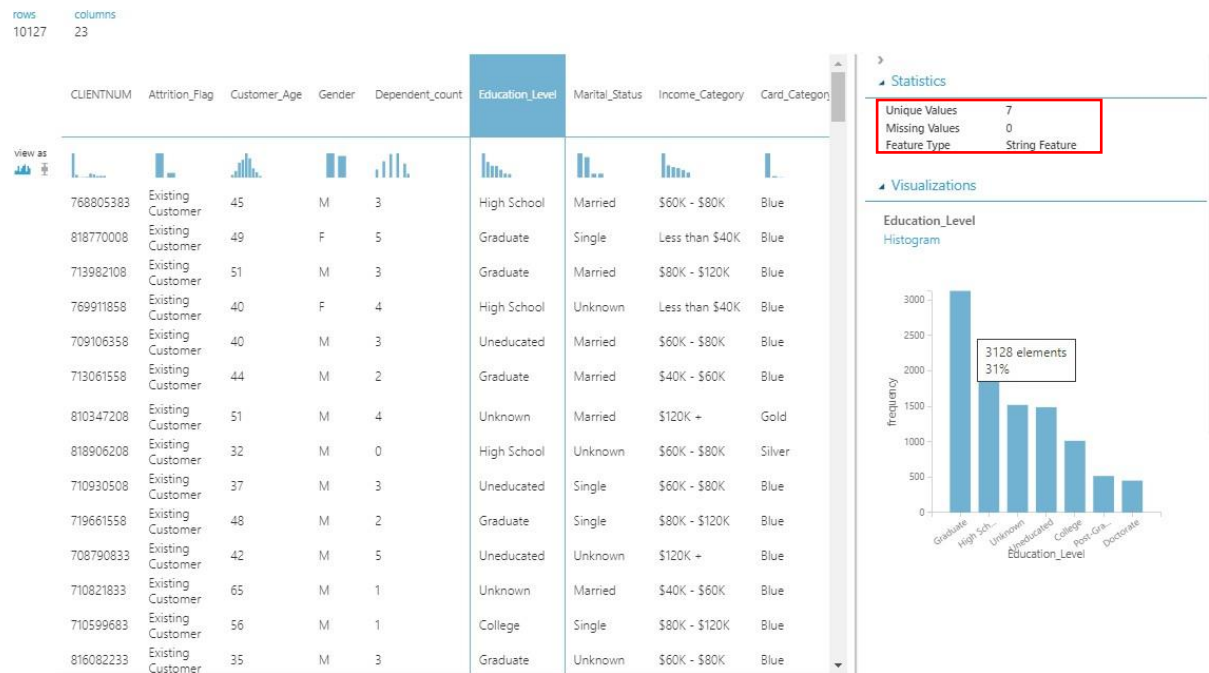
Customer_Age: We observe that the variable corresponding to the client's age has no missing data, the mean is 46.3 years, the maximum age is 73 years while the minimum is 26. If we look at the distribution, the majority of clients are between 45 and 50 years old (24%).



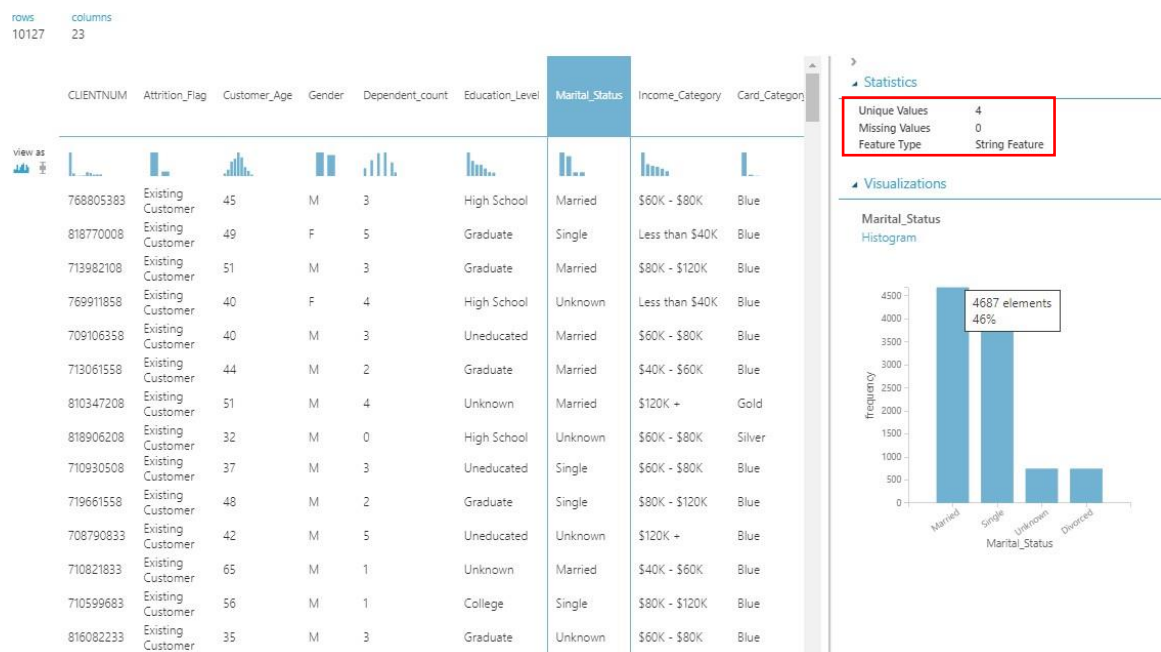
Gender: We note that it is a categorical variable, it takes two different values (female and male) and has no missing values. If we look at the distribution, the majority of clients are female (53%).



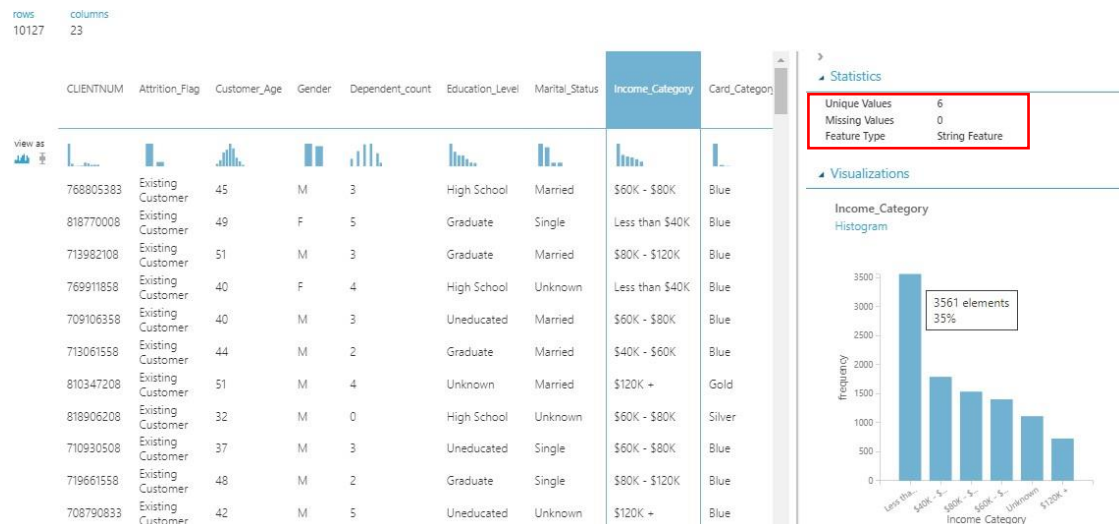
Dependent_count: We observe that the number of people who are responsible for others has no missing data, the mean is 2 people, the maximum is 5 people, and the minimum is 0. If we see the distribution, most clients are concentrated in 3 people in charge (27%).



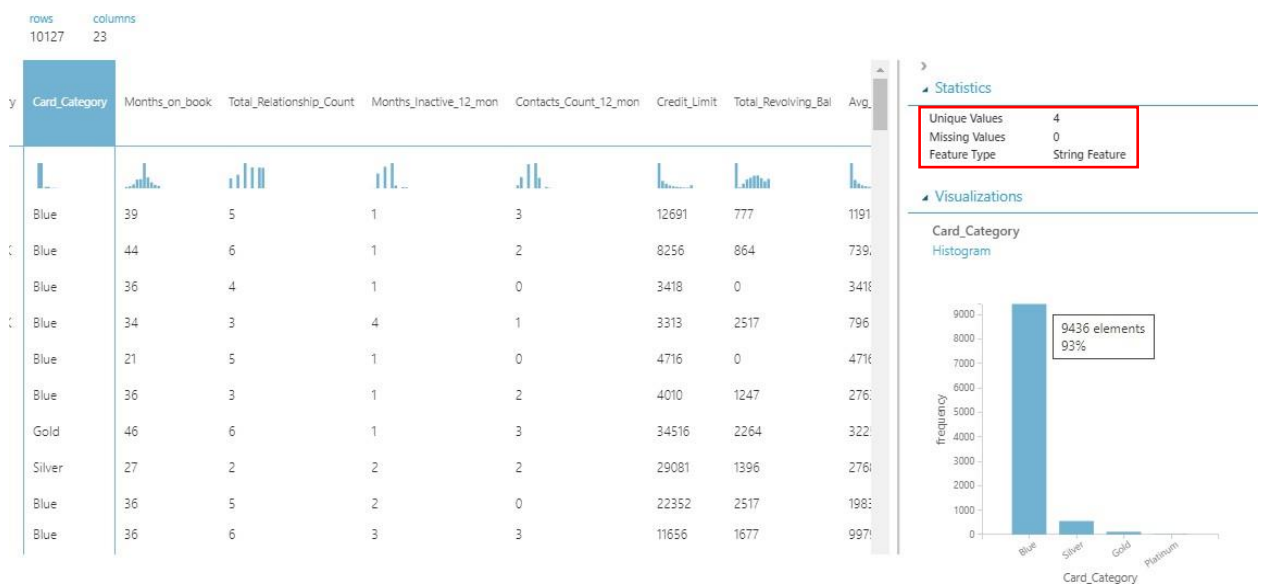
Education_Level: We observe that it is a categorical variable, it takes seven different values: Graduate, High School, Unknown (the client did not provide information), Uneducated (did not finish high school), College (he/she studied or is studying in college) Post-Graduate, Doctorate, and has no missing values. If we look at the distribution, most of the clients are graduates (31%), so we can assume that they will not want to unsubscribe.



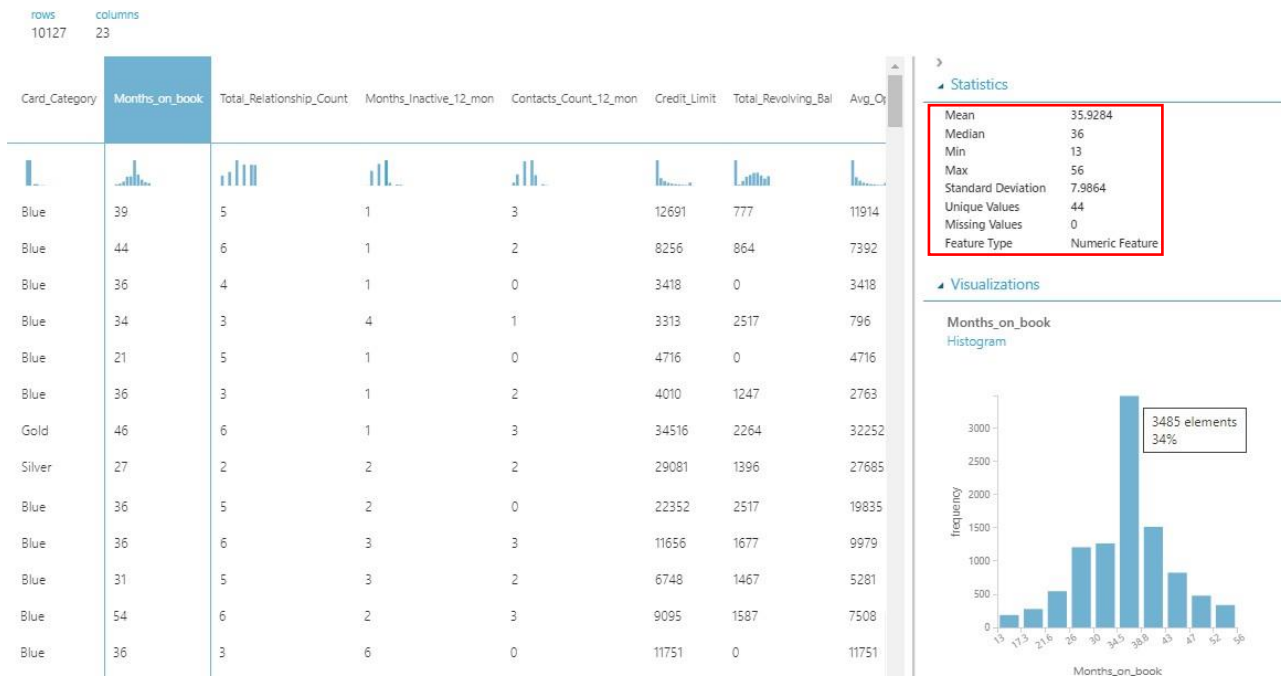
Marital_Status: We observe that it is a categorical variable, it takes four different values: Married, Single, Unknown (the client did not provide information or is a widow), Divorced and has no missing values. If we see the distribution, most of the clients are married (46%), so we can assume that they are the ones with the most economic stability, so they will not want to unsubscribe.



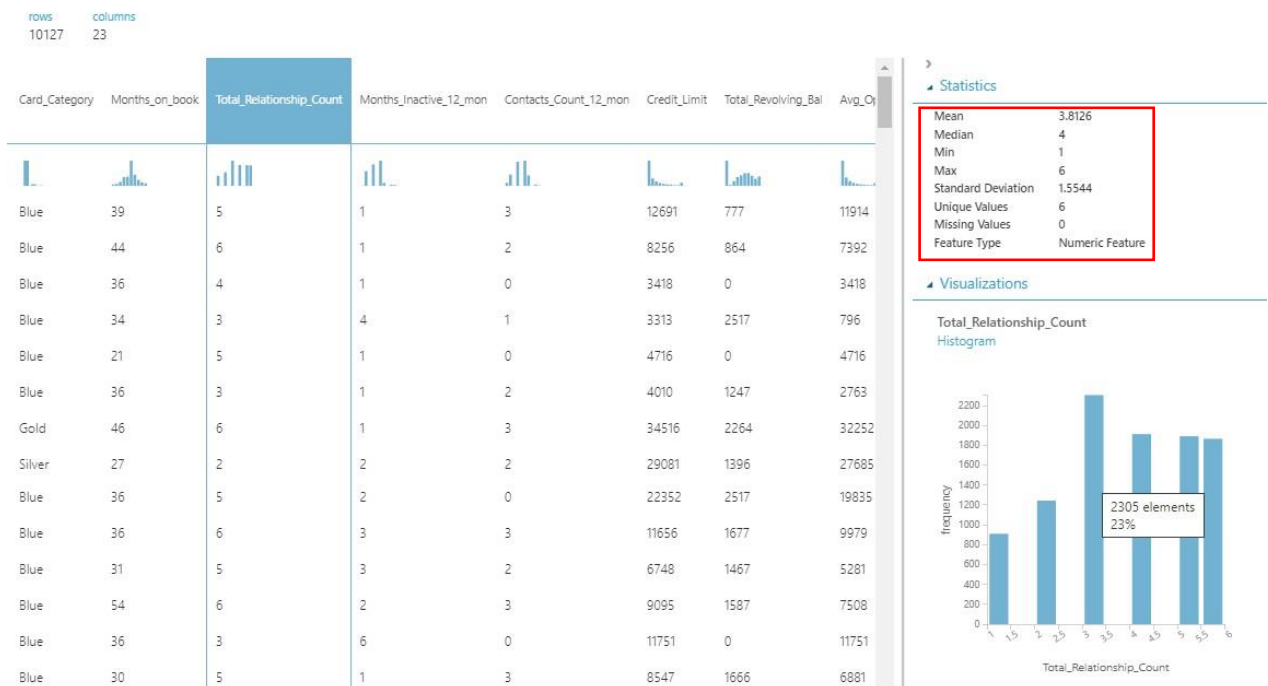
Income_Category: We note that it is a categorical variable, it takes six different values, and it has no missing values. If we look at the distribution, the majority of clients earn less than \$40K a year (35%), then we can assume that they may want to leave.



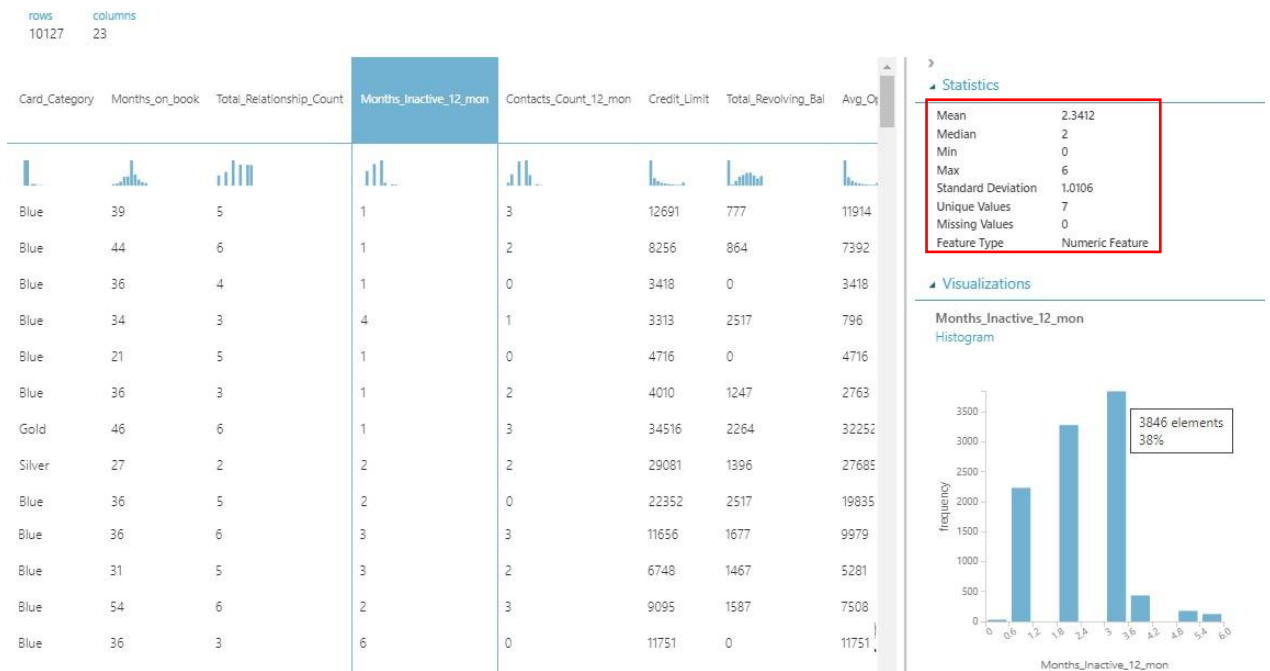
Card_Category: We note that it is a categorical variable, it takes four different values: Blue, Silver, Gold, Platinum and has no missing values. If we look at the distribution, the vast majority of customers have a Blue card (93%).



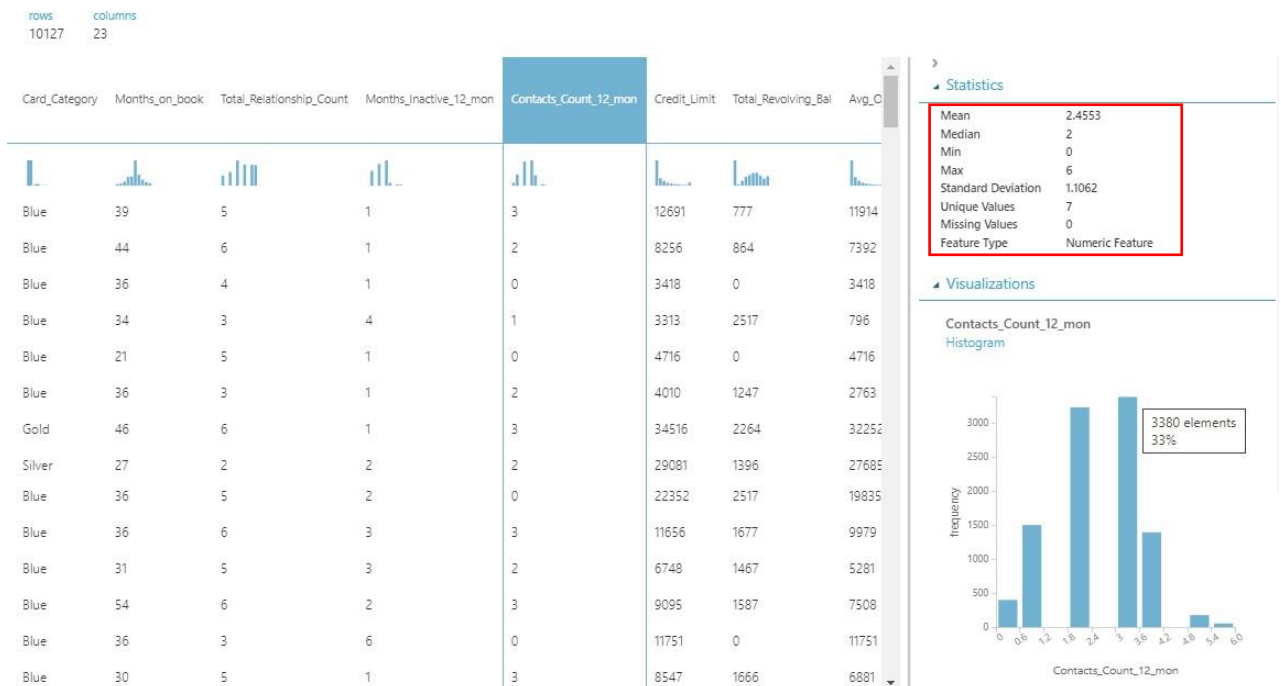
Months_on_book: We observe that account's age has no missing data, the mean is 35.9 months, the maximum is 56 months, and the minimum is 13. If we look at the distribution, most customers are concentrated in one age between 34.5 and 38.8 months (34%).



Total_Relationship_Count: We observe that the number of products of the clients have no missing data, the mean is 4 products, the maximum is 6 and the minimum is 1. If we see the distribution, the majority of the clients are concentrated in 3 products (23%).



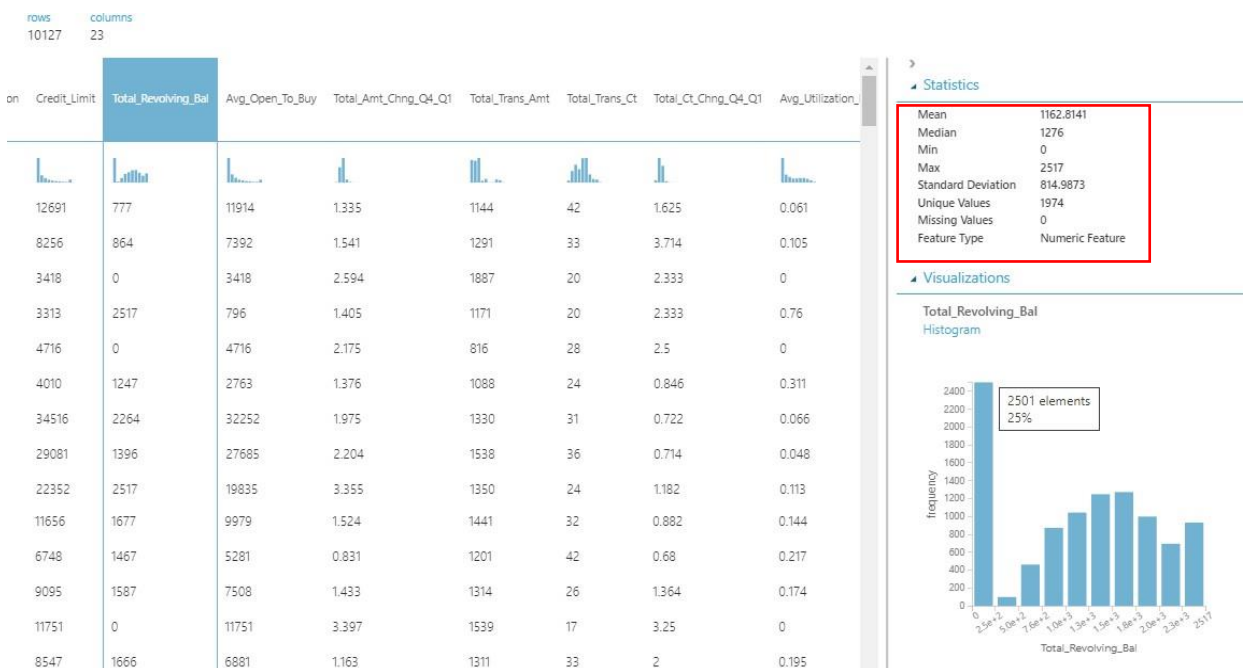
Months_Inactive_12_mon: We observe that the number of months that customers were inactive in the last year has no missing data, the mean is 2.3 months, the maximum is 6 and the minimum is 0. If we see the distribution, most of the customers are concentrated between 3 and 3.6 months inactive (38%).



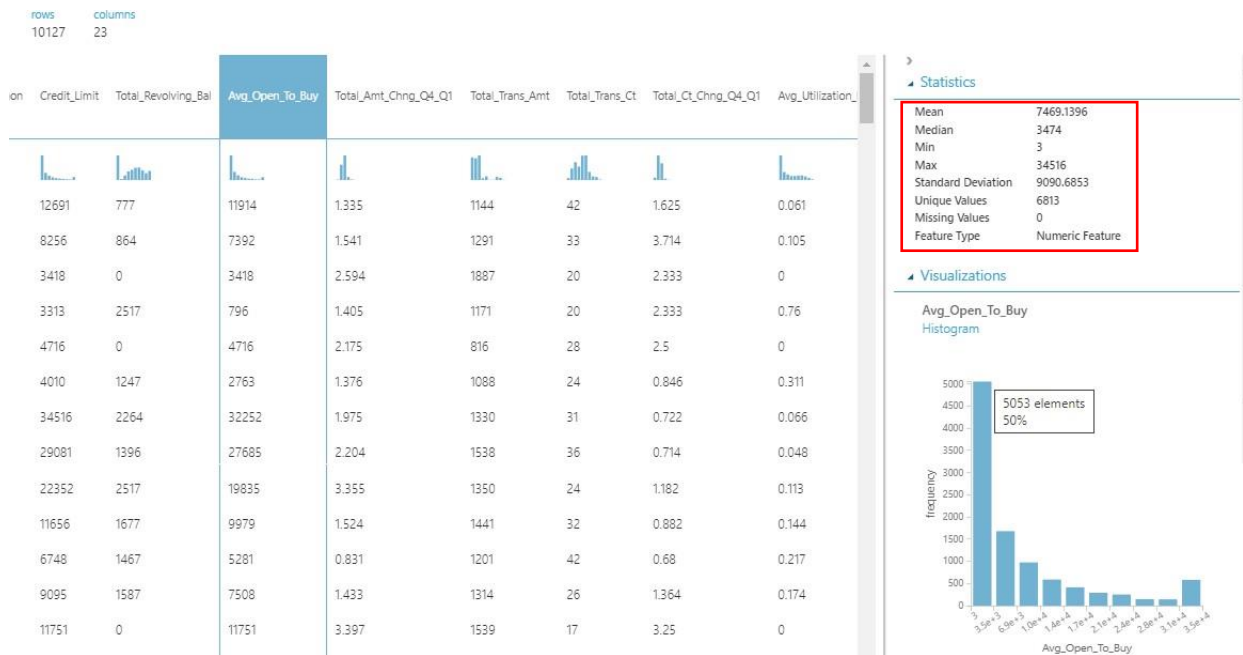
Contacts_Count_12_mon: We observe that the number of contacts that the clients had with the bank in the last year has no missing data, the mean is 3 contacts, the maximum is 6 and the minimum is 0. If we see the distribution, the majority of the clients are concentrated between 3 and 3.6 contacts (33%).



Credit_Limit: We observe that the client's credit limit amount has no missing data, the mean is 8,631.9 currency units, the maximum is 34,516 currency units, and the minimum is 1,438.3. If we look at the distribution, the majority of clients are concentrated between 1,400 and 4,700 monetary units (51%).



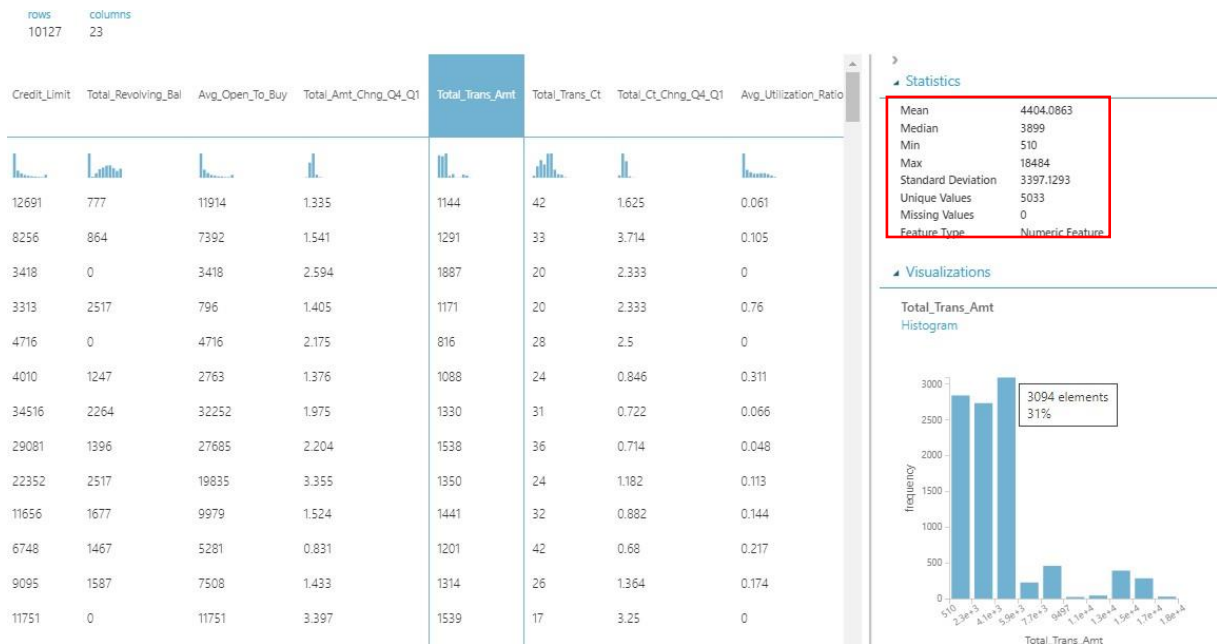
Total_Revolving_Bal: We observe that the uncovered balance of the card has no missing data, the mean is 1,162.8 monetary units, the maximum is 2,517 monetary units and the minimum is 0. If we look at the distribution, most customers are concentrated between 0 and 250 monetary units (25%).



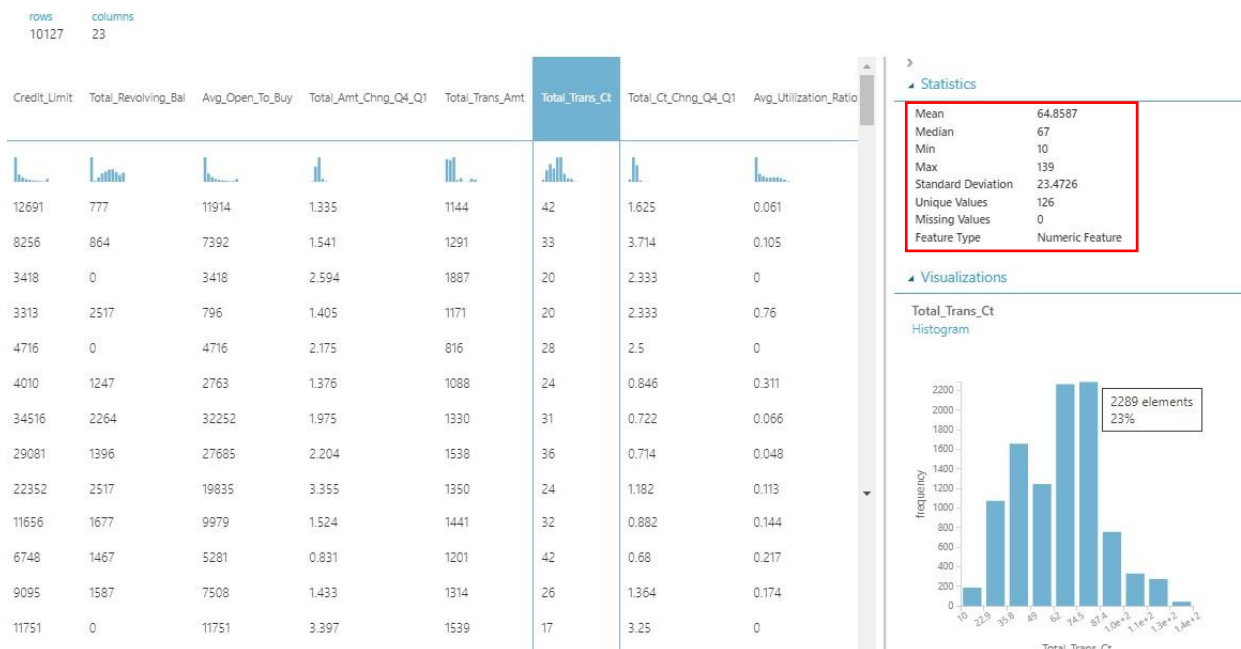
Avg_Open_To_Buy: We observe that the available balance of the card has no missing data, the mean is 7,469.1 monetary units, the maximum is 34,516 monetary units and the minimum is 3. If we see the distribution, most of the clients are concentrated between 0 and 3,500 monetary units (50%).



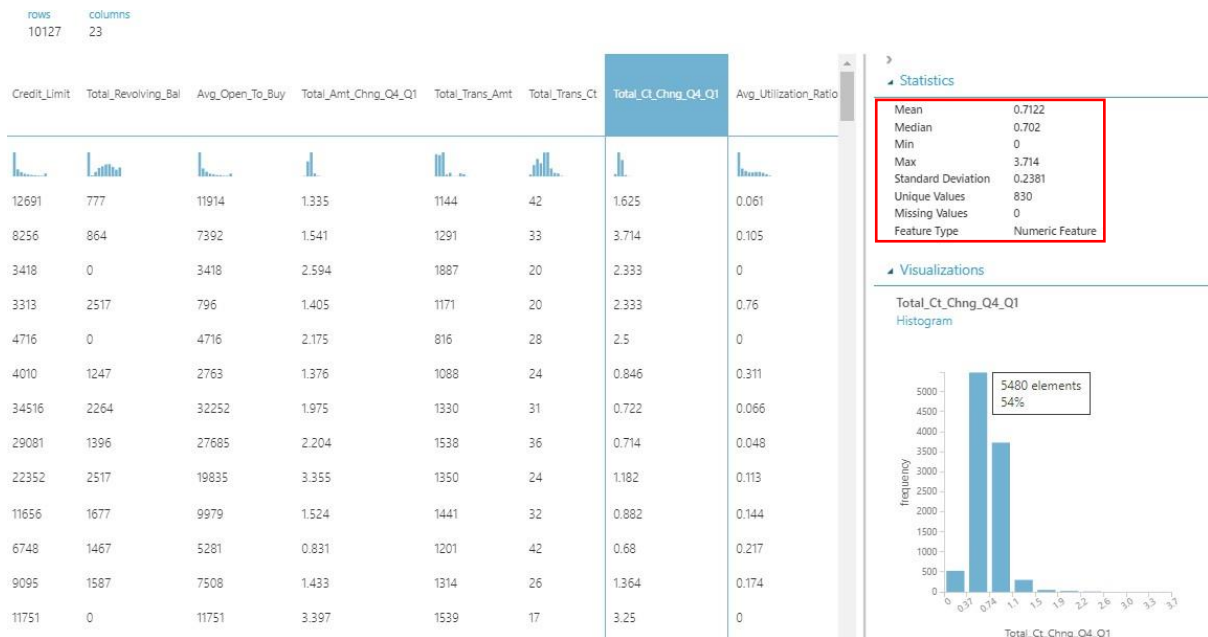
Total_Amt_Chng_Q4_Q1: We observe that the percentage change in the amount of consumption has no missing data, the mean is 0.76 percent, the maximum is 3.4% and the minimum is 0. If we see the distribution, most of the clients are concentrated between 0.68 and 1 percent (56%).



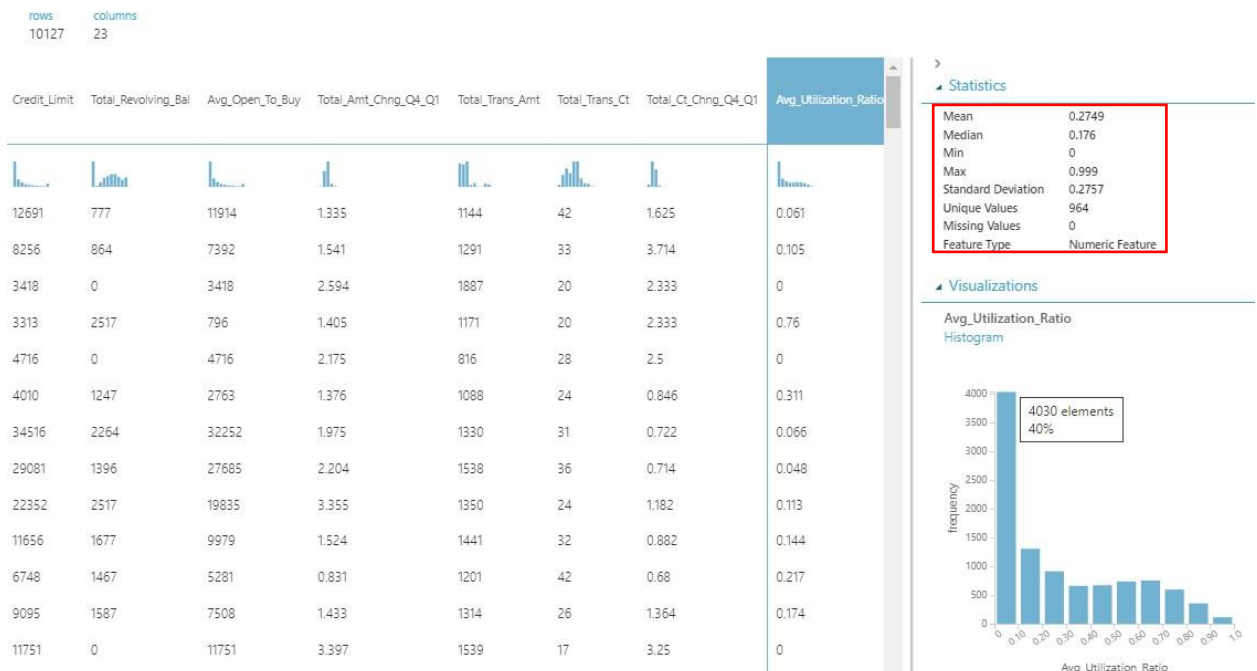
Total_Trans_Amt: We observe that the amount of consumption in the last year has no missing data, the mean is 4,404 monetary units, the maximum is 18,484 and the minimum is 510. If we see the distribution, most customers are concentrated between 510 and 2,300 units monetary (31%).



Total_Trans_Ct: We observe that the number of transactions in the last year has no missing data, the mean is 65, the maximum is 139 and the minimum is 10. If we see the distribution, the majority of clients are concentrated between 10 and 229 transactions (23%).



Total_Ct_Chng_Q4_Q1: We observe that the percentage change in quantity of consumption has no missing data, the mean is 0.71 percent, the maximum is 3.71 and the minimum is 0. If we see the distribution, most of the clients are concentrated between 0 and 0.37 percent (54%).

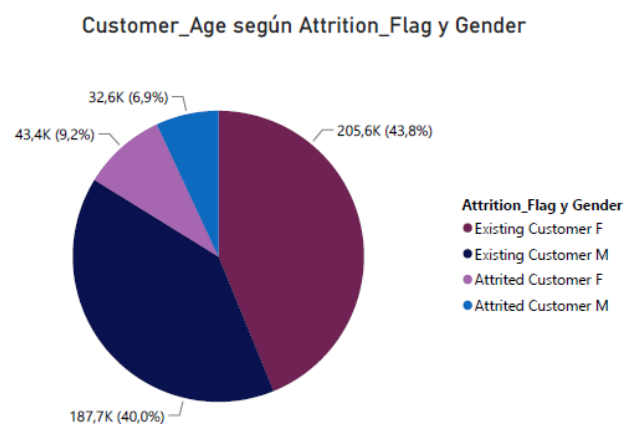


Avg_Utilization_Ratio: We observe that the card usage ratio has no missing data, the mean is 0.27 percent, the maximum is 0.99 and the minimum is 0. If we see the distribution, most customers are concentrated between 0 and 0.10 percent (40%).

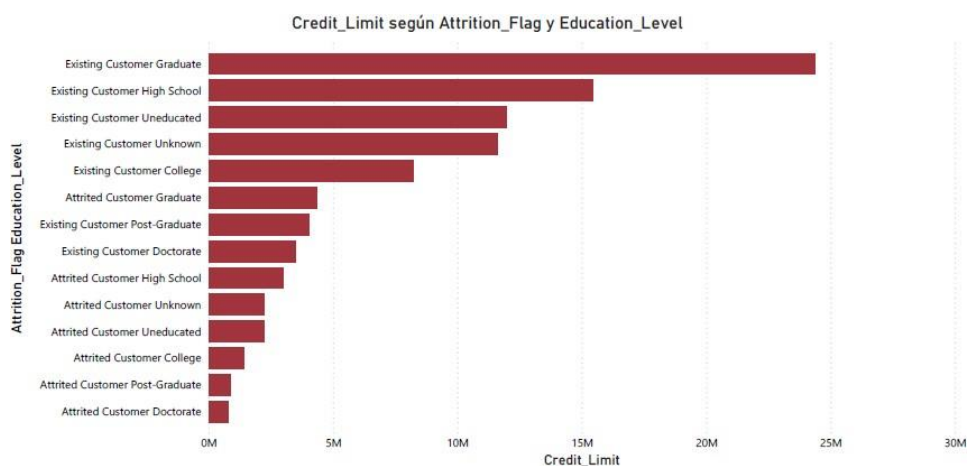
The predictor variables are all except the target variable "Attrition_Flag", which is the one that will be predicted. It is convenient to add all of them to the model with the exception of the following:

- **Naive_Bayes_Classifier_Attrition_Flag_Card_Category_Contacts_Count_12_mon_Dependent_count_Education_Level_Months_Inactive_12_mon_1:** We have no information on this variable.
- **Naive_Bayes_Classifier_Attrition_Flag_Card_Category_Contacts_Count_12_mon_Dependent_count_Education_Level_Months_Inactive_12_mon_2:** We have no information on this variable.

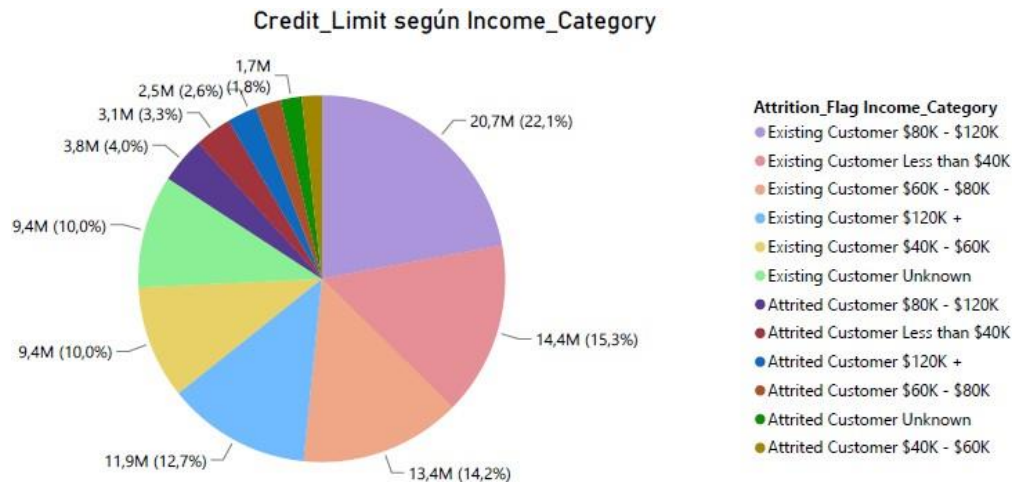
Brief Exploratory Analysis in Power BI Power BI:



Here we can see how customers are distributed according to their age, gender, and customer category. Most of the clients are women, 43.8% want to continue being clients of the bank while 9.2% want to unsubscribe.



If we do the credit limit analysis according to the educational level and the category of client, we observe that the existing graduated clients are the ones with the highest credit limit.



We see the representation of the credit limit according to income category and customer category. The highest credit limit is for customers who earn between \$80K - \$120K a year and want to stay with the bank.

Results obtained análisis for the following models:

	AUC	True Positive	False Negative	False Positive	True Negative	Accuracy	Precision	Recall	F1 Score
Base model	0,935	1646	52	126	201	0,912	0,929	0,969	0,949
Model 1: Only numerical variables	0,928	1641	57	132	195	0,907	0,926	0,966	0,946
Model 2: Numerical variables + Feature Engineering	0,943	1650	48	120	207	0,917	0,932	0,972	0,952
Model 3: Numerical variables + Feature Engineering + Filter based	0,944	1647	51	121	206	0,915	0,932	0,97	0,95

The higher the area under the curve (AUC) is, it means that the model is better at predicting 0 as 0 and 1 as 1 → **the higher the AUC, the better the model** → since for this base model the AUC gives us a value of 0.935, it means that as a first approximation to the modeling it is making a correct prediction.

The transformations that we apply to the numerical raw variables are the following:

```
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;
```

We use the “Filter Based Feature Selection” module to visualize the variables that have very little relationship with the target variable and thus exclude them.



Arbitrarily we remove from the model all variables with a score less than 0.01 → Credit_Limit, sq_Credit_Limit, cu_Credit_Limit, Customer_Age, Months_on_book, Dependent_count.

We observe that the AUC of model 3 is almost the same as the AUC of model 2, so we can conclude that the variables that we removed do not influence the target variable.

Model 4: Categorical variables

The categorical variables are:

- **Attrition_Flag:** It is the target variable.
- **Gender:** Indicates the gender categories of a person.
- **Education_Level:** Indicates the categories of a person's level of education.
- **Marital_Status:** Indicates the marital categories of a person.
- **Income_Category:** Indicates a person's income categories.
- **Card_Category:** Indicates the categories of type of cards that a person can have.

We perform the bivariate analysis for each categorical variable and, that way, be able to know how the relationship of each of the values of the categories with the objective variable is.

a. Education_Level Variable:

```
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 represents whether customers are going to continue with their subscription to the bank.

- The "Doctorate" category corresponds to customers who have a doctorate and is the one with the lowest rate of "existing customers" (Percentage_Existing_Customer) (0.789357)→ is indicating that this type of "Education_Level" is the one that falls into more clients who do not want to continue with the bank's services.
- The category "Post-Graduate" corresponds to clients who have an educational level higher than the bachelor's degree, "Unknown" contains the missing (blank) or undisclosed values, "Uneducated" corresponds to clients who have no education , "Graduate" means that they graduated from university, "College" indicates that he went to university and "High School" indicates that he attended high school → is indicating that these types of "Education_Level" have clients who want to continue in the bank to a greater extent.

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

The groups made were the following:

- “High School”, “College” and “Graduate” can be grouped together because they have a very similar “existing customer” rate → their behavior should be similar too (on average).
- “Uneducated”, “Unknown” and “Post-Graduate” can be grouped together because they also have a very similar “existing customer” rate → their behavior should be similar too (on average).
- We do not group “Doctorate” because it is the one with the lowest rate of “existing customers”.

b. Gender Variable:

```
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

- “F” category corresponds to female clients and is the one with the lowest rate of “existing clients” (0.826428). → indicates that this type of “Gender” is the one that falls into more clients who do not want to continue with the bank's services.
- “M” category corresponds to male clients → indicates that this type of “Gender” has clients who want to continue in the bank to a greater extent.

From this information we could create a dummy variable that indicates if the person has record of not wanting to belong to the bank anymore.

c. Marital_Status Variable:

```
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

- The “Unknown” category contains the missing (blank) or undisclosed values and has the lowest “existing customers” rate (0.82777) → indicates that this type of “Marital_Status” is the one that falls into more clients who do not want to continue with the bank's services.
- The “Single” category corresponds to clients who are single, “Divorced” to clients who are divorced and “Married” to clients who are married → indicates that these types of “Marital_Status” have clients who want to continue in the bank to a greater extent.

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

The groups made were the following:


- We do not group “Married” because it is the one with the highest rate of “existing customers”.
- “Single” and “Divorced” can be grouped together because they have a very similar rate of “existing clients” → their behavior should be similar too (on average).
- We do not group “Unknown” because it is the one with the lowest rate of “existing customers”.

d. Income_Category Variable:

```

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

- The “\$120K+” category corresponds to customers earning \$120k or more per year and has the lowest “existing customer” rate (0.826685)→ is indicating that this type of "Income_Category" is the one that falls into more clients who do not want to continue with the bank's services.
- The “Less than \$40K” category corresponds to clients who earn less than 40 thousand dollars per year and has a low rate of “existing clients” (0.828138), very similar to the previous one.
- All other categories have a higher “existing customer” rate.




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

The groups made were the following:

- “\$120K+”, “Less than \$40K” and “Unknown” can be grouped together because they have a very similar “existing customer” rate → their behavior should be similar too (on average).
- “\$80K - \$120K” and “\$40K - \$60K” can be grouped together because they have a very similar “existing customer” rate → their behavior should be similar too (on average).
- We do not group “\$60K - \$80K” because it is the one with the highest rate of “existing customers”.

e. Card_Category Variable:

```
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

- The “Platinum” category corresponds to customers who have a card of that color and is the one with the lowest rate of “existing customers” (0.75) → is indicating that this type of “Income_Category” is the one that falls into more clients who do not want to continue with the bank's services. In addition, we observe that very few people access it (20), so we should group it with the next category. This same situation occurs with the “Gold” category since it only has 116 observations. → we group it together with “Platinum” category.
- The other categories have a higher rate of “existing customers” and many more observations.

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

The groups made were the following:

- “Platinum” and “Gold” can be grouped together because of what we defined above.
- We do not group “Blue” and “Silver” we leave them alone because, although they have the highest rate of “existing customers”, they differ a little between them.

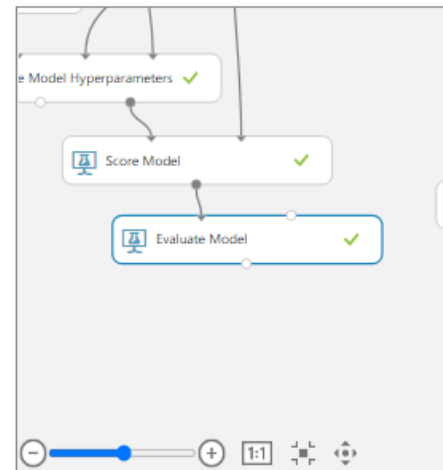
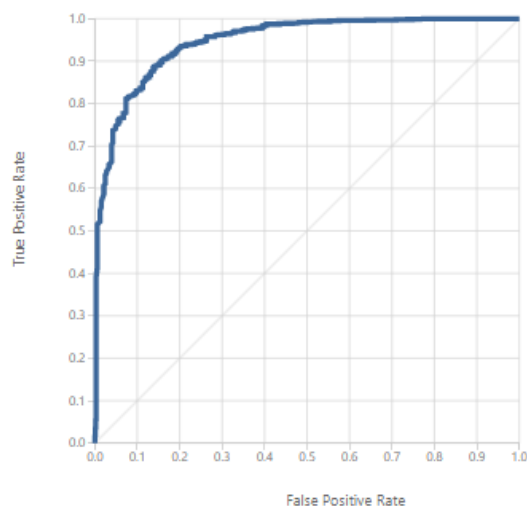
```

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;

```

We create a dummy variables for each categorical variable using the groups mentioned above.

ROC PRECISION/RECALL LIFT



True Positive	False Negative	Accuracy	Precision	Threshold	AUC
1653	45	0.920	0.934	0.5	0.947
False Positive	True Negative	Recall	F1 Score		
116	211	0.973	0.954		

Positive Label Negative Label
Existing Customer Attrited Customer

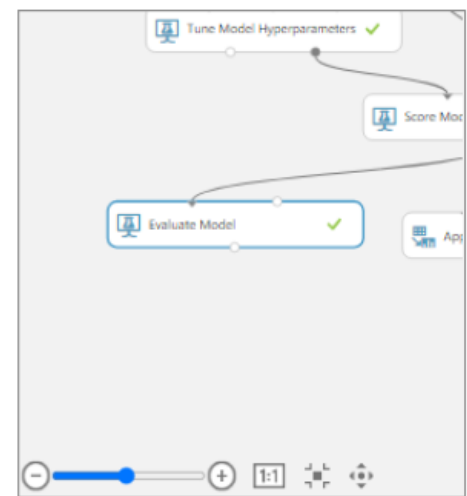
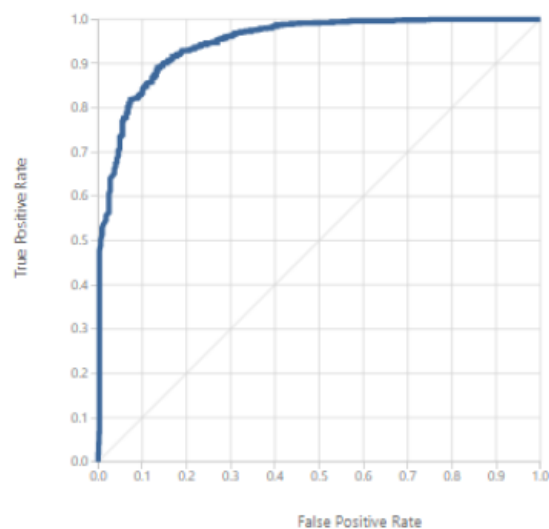
Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulative AUC
(0.900,1.000]	1349	24	0.678	0.816	0.879	0.983	0.794	0.465	0.927	0.047
(0.800,0.900]	161	23	0.769	0.884	0.928	0.970	0.889	0.598	0.856	0.106
(0.700,0.800]	80	23	0.820	0.912	0.947	0.958	0.936	0.704	0.786	0.170
(0.600,0.700]	36	20	0.847	0.920	0.953	0.948	0.958	0.767	0.725	0.228
(0.500,0.600]	27	26	0.874	0.920	0.954	0.934	0.973	0.824	0.645	0.305
(0.400,0.500]	20	16	0.891	0.922	0.955	0.927	0.985	0.886	0.596	0.353
(0.300,0.400]	11	31	0.912	0.913	0.950	0.912	0.992	0.921	0.502	0.447
(0.200,0.300]	7	34	0.932	0.899	0.943	0.896	0.996	0.949	0.398	0.550
(0.100,0.200]	6	55	0.962	0.875	0.931	0.871	0.999	0.987	0.229	0.718
(0.000,0.100]	1	75	1.000	0.839	0.912	0.839	1.000	1.000	0.000	0.947

We observe that model 4 improves regarding with the previous one since the AUC increases.

Model 5: Interactive Variables

```
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;
```

Here we see the interactions of customers' available card balance with the dummy variables we created earlier.



True Positive	False Negative	Accuracy	Precision	Threshold	AUC
1653	45	0.922	0.936	0.5	0.948
False Positive	True Negative	Recall	F1 Score		
113	214	0.973	0.954		
Positive Label		Negative Label			
Existing Customer		Attrited Customer			

Score Bin	Positive Examples	Negative Examples	Fraction Above Threshold	Accuracy	F1 Score	Precision	Recall	Negative Precision	Negative Recall	Cumulative AUC
(0.900,1.000]	1361	22	0.683	0.823	0.883	0.984	0.802	0.475	0.933	0.042
(0.800,0.900]	153	24	0.770	0.886	0.929	0.971	0.892	0.604	0.859	0.104
(0.700,0.800]	74	26	0.820	0.910	0.946	0.957	0.935	0.699	0.780	0.177
(0.600,0.700]	35	19	0.846	0.918	0.951	0.947	0.956	0.759	0.722	0.232
(0.500,0.600]	30	22	0.872	0.922	0.954	0.936	0.973	0.826	0.654	0.297
(0.400,0.500]	18	19	0.890	0.921	0.955	0.927	0.984	0.878	0.596	0.354
(0.300,0.400]	13	28	0.911	0.914	0.951	0.913	0.992	0.923	0.511	0.438
(0.200,0.300]	8	33	0.931	0.902	0.944	0.898	0.996	0.957	0.410	0.539
(0.100,0.200]	5	57	0.961	0.876	0.931	0.872	0.999	0.987	0.235	0.713
(0.000,0.100]	1	77	1.000	0.839	0.912	0.839	1.000	1.000	0.000	0.948

The AUC is slightly increased by adding interactive variables.

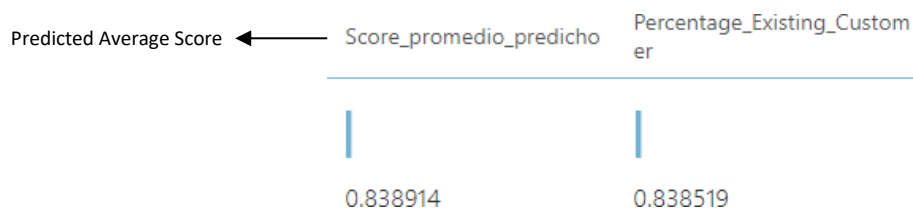
Conclusions of the analysis:

	AUC	True Positive	False Negative	False Positive	True Negative	Accuracy	Precision	Recall	F1 Score
Base model	0,935	1646	52	126	201	0,912	0,929	0,969	0,949
Model 1: Only numerical variables	0,928	1641	57	132	195	0,907	0,926	0,966	0,946
Model 2: Numerical variables + Feature Engineering	0,943	1650	48	120	207	0,917	0,932	0,972	0,952
Model 3: Numerical variables + Feature Engineering + Filter based	0,944	1647	51	121	206	0,915	0,932	0,97	0,95
Model 4: : Numerical variables + Feature Engineering + Filter based and dummies	0,947	1653	45	116	211	0,92	0,934	0,973	0,954
Model 5: Numerical variables + Feature Engineering + Filter based and dummies and interactives	0,948	1653	45	113	214	0,922	0,936	0,973	0,954

The model that we would choose is number 5, since it is the one with the highest AUC value.

We compare the Predicted Average Score versus the Percentage of "existing customers" in the sample (actual value):

```
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;
```



In global terms, the predicted value (0.838914) and the real value (0.838519) give very similar values, so we can say that the model is acceptable. → 0,04%.





A difference between the predicted value and the actual value of up to $\pm 5\%$ is what we will accept.

Now we do the analysis by deciles and find out if the prediction by sections is also good:

```
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;
```

Orden = Order
 Grupo = Group
 Frecuencia_real = Real_frequency
 Frecuencia_estimada = Estimated_frequency

	Grupo	Frecuencia_real	Frecuencia_estimada	count(*)	
These would be the bad deciles because they have the fewest "existing customers" if you look at actual_frequency. We set the limit at 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	
	5	0.955	0.975036	200	
These would be the good deciles because they have the most "existing customers" if you look at actual_frequency.	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	

We observe that the estimated frequency is consistent with the real frequency, then section by section we see that the estimated_frequency also orders from lowest to highest as the real_frequency and that, in addition, the estimated_frequency is quite similar to the real_frequency.