BankChurner story

1.1 Problem Context

A bank manager is uncomfortable with more and more customers leaving their credit card services. They would really appreciate it if someone could predict who will be affected so that they can proactively go to the customer to provide them with better services and turn customer decisions in the opposite direction.

1.2 The goal

This project is carried out in a sequence of steps, the first of which consists of an exploratory analysis, where the objective is to know the behavior of the variables and to analyze attributes that indicate a strong relationship with the cancellation of credit card service customers. After the second part, which consists of applying resource engineering techniques, the third act consists of applying a machine learning algorithm to find the best resources for building the model. At the end of the project, after the completion of all steps, a machine learning model will be developed, capable of predicting, based on the data of a system, whether a customer will leave the credit card service or not.

1.3 Data set

This data set consists of 10,000 customers mentioning their age, salary, status marital, credit card limit, credit card category, etc.

We have only **16,07%** of customers who have canceled. Therefore, it is a little difficult to train our model to predict customer turnover.

# **2 Exploratory Data Analysis (EDA)**

As stated in the ‘Project Description’, this step aims to discover the main elements responsible for the cancellation or non-cancellation of credit card service customers. For this first session to be carried out successfully, I will apply some statistical techniques (Descriptive Analysis) and visualization, which will provide important and satisfactory insights to proceed with the rest of the analysis.

## 2.1 Correlation Analysis - Spearman

So that this analysis is not too extensive, I will apply the non-parametric statistical test of spearman, thus obtaining the correlation coefficient, which measures statistical dependence between two variables. In this way, we can verify from the beginning which variables should receive the most attention, saving time in analyzing variables that do not have a strong influence on the rate of customers who leave the credit card service.

* For the purpose of this exercise, Attrited Customer and Existing Customer classes are releveled to 1 and 0 respectively.

Table

Description automatically generated

The variables that demonstrate a considerable negative association in relation to the dependent attribute, and that are the target of investigation, are: Total\_Trans\_Ct, Total\_Ct\_Chng\_Q4\_Q1, Total\_Revolving\_Bal and Avg\_Utilization\_Ratio. A negative association (<0) indicates that the attribute has a relevant level of significance in the customer’s permanence.

With a positive association we have the variables: Contacts\_Count\_12\_mon and Months\_inactive\_112\_mon.

## 2.2 Number of Cancellations For Each Qualitative Variable Category

Before starting to analyze the quantitative variables, I will go through the treemap graphs (visualization technique to represent hierarchical data using nested rectangles), to know the behavior of the categorical variables of the data set. More precisely, I will investigate which categories of qualitative variables have the highest number of credit card cancellations.

**With the graphics below we can filter the following information:**

* **46,28%** of credit card cancellations are from married people and **7,38%** are from customers who have marital status as divorced.
* **35.16%** of credit card cancellations are from people with an annual income less than 40K and **7.17%** are from customers who have an annual income greater than or equal to 120K.
* **30.88%** of credit card cancellations are from people who have graduated and the lowest number of cancellations (**4.45%**) are from customers who have post Doctorate.
* **93.17%** of credit card cancellations are from customers who have the Blue card type and the lowest number of cancellations (**0.19%**) are from customers who have the Platinum card type.

## 2.3 Behavior of Total Transactions Made by Customers

Proceeding, but now analyzing quantitative variables. As seen in Pearson’s statistical test, where the variables ‘Total\_Trans\_Ct’ have a correlation coefficient of -0.376, ‘Total\_Revolving\_Bal’ with -0.241, ‘Avg\_Utilization\_Ratio’ with 0.24 and ‘Total\_Ct\_Chng\_Q4\_Q1’ with -0.312, indicating that they all positively influence customers stay. This session whose name is ‘Behavior of the total customer transaction’ aims to understand the behavior of the quantitative variables listed above in relation to the target variable ‘Attrition\_Flag’, which informs whether the customer has left the card service or not.

Chart, histogram

Description automatically generated

Chart, box and whisker chart

Description automatically generated

One way to interpret the separating measures from the boxplot chart above is:

* **50%** of customers who left credit card services had a number of transactions in the last 12 months less than or equal to 43, remembering that the maximum number of transactions in the last 12 months of customers who left the service is 72. While the median of the people who remained with the card services is 71 transactions.
* **75%** dos clientes que deixaram os serviços de cartão de crédito tiveram um número de transações nos últimos 12 meses igual ou inferior a 51. O terceiro quartil de pessoas que permaneceram com os serviços de cartão é de 82 transações.

Chart, scatter chart

Description automatically generated

Graphical user interface, chart, application, table, Excel

Description automatically generated with medium confidence

Chart, histogram

Description automatically generated

Chart

Description automatically generated with medium confidence

A picture containing chart

Description automatically generated

Chart, bar chart

Description automatically generated

Histogram

Description automatically generatedHistogram

Description automatically generated with medium confidence

Chart, bar chart

Description automatically generated

Graphical user interface

Description automatically generated with medium confidence

-> not sure about this graph since correlation is low.

Relationship between categoric variable vs churn => Information Value

신용평가 모형에 좋음

Information Value assese the overall **predictive power** of the variable being considered, and therefore can be used for comparing the predictive power among competing variables

Text

Description automatically generated with medium confidence

Graphical user interface, application

Description automatically generated

Graphical user interface, table

Description automatically generated

Strong predictor : Education level, Dependent count

Medium : income category

Chart, bar chart

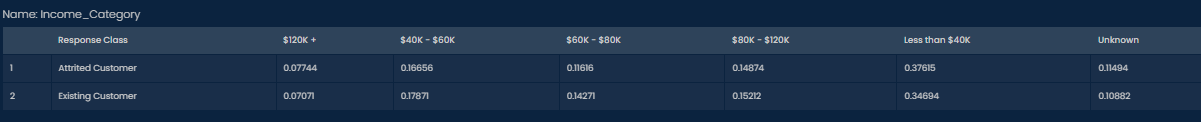
Description automatically generated

Chart, bar chart, histogram

Description automatically generated

Graphical user interface

Description automatically generated with medium confidence



이제 어떤 어떤 컨디션에 가장 attrition이 높을 것이다라는 결과가 나옴.

그러면 이제 테스트를 한번 더 해서 증명을 해내자.

To deepen the analysis of the characteristics of the qualitative attributes of customers and try to find some pattern, I will group all the particularities of customers who abandoned credit card services and search among all possible combinations within the qualitative attributes space of the data set, if there is a profile with certain qualities that stands out among all the others, that is, I will try to find a customer profile with certain particularities, that has an evasion rate that is very far from the others.

Graphical user interface

Description automatically generated

Or Chi-Squared test or Naïve Bayes Classifier or Logistic Regression(come up with equation building a predictive model)?

다 해보고 제일 make sense 하는 걸로 하면 될 듯!

이제 진짜 일요일에 끝내자!