**Summary**

Fourteen classification machine learning models were trained, and the Gradient Booster Classifier was ultimately selected as the best. The model demonstrated good performance, achieving an 86% accuracy with both the training and testing datasets. The main predictive factors identified were Age and the Number of products purchased.

**Introduction**

The performance of the marketing strategies impacts directly on market share, fidelity and revenue, keys for keeping a health companies. One of the most important ways to evaluate it is through the quantification of customers leaving your services or ceasing to buy your products, also known as the customer churn rate. The use of Machine Learning technologies to model the churn rate would allow to explore main factors that explain this phenomenon, predict future churn, and prevent it.

The purpose of this analysis was to train machine learning classification models with explanatory variables associated with the churn/no-churn target variable from a banking company to make predictions.

**Methods and Analysis**

The Bank Customer Churn database was used, which includes 12 factors attempting to explain the churn of 10,000 customers:

Customer Identification number (CustomerId)

Surname of each customer (Surname)

Index of the customer credit behavior (CreditScore)

Customer Location (Geography\_1.0)

Gender

Age

Years that customer has been a client of the Bank (Tenure)

Balance of customer account at the time data was obtained (Balance)

Number of products the customer has purchased in this bank (NumOfProducts)

Indicator referring whether the customer has a credit card (HasCrCard)

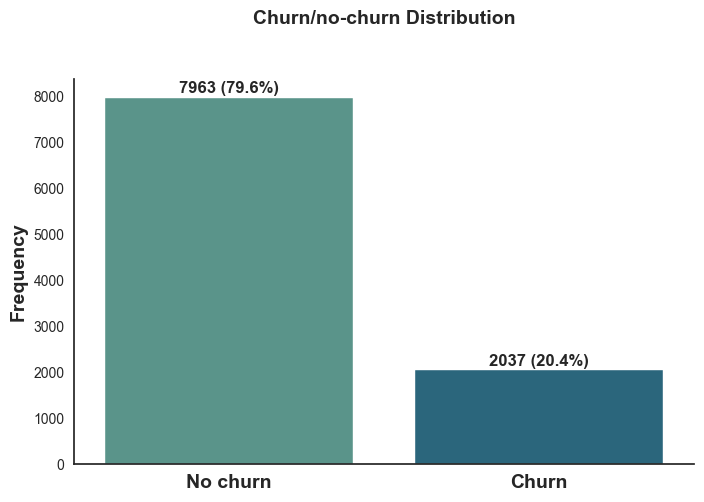
Indicator of customer activity at the time data was obtained (IsActiveMember)

Estimated Salary (EstimatedSalary)

Indicator referring whether the customer left the Bank (Exited). This is the churn variable.

To perform the analysis this dataset was divided in a training data (80%) and testing data (20%).

Due to the imbalanced distribution of churn/no-churn, the SMOTE method of correction was applied to train the model. This algorithm creates new data set based on the K-nearest neighbors of the minority class (No Churn) at the moment of fit the models.



Machine Learning models were trained and ranked to choose the best one based in the performance to predict the training data. Based on the chosen model test were made to evaluate the generalization power through the ability to predict unknown data.

Due to the binary nature of data Churn (Exited variable) 14 classification models were evaluated for this study:

Ada Boost Classifier

Decision Tree Classifier

Dummy Classifier

Extra Trees Classifier

Gradient Boosting Classifier

K Neighbors Classifier

Light Gradient Boosting Machine

Linear Discriminant Analysis

Logistic Regression

Naive Bayes

Quadratic Discriminant Analysis

Random Forest Classifier

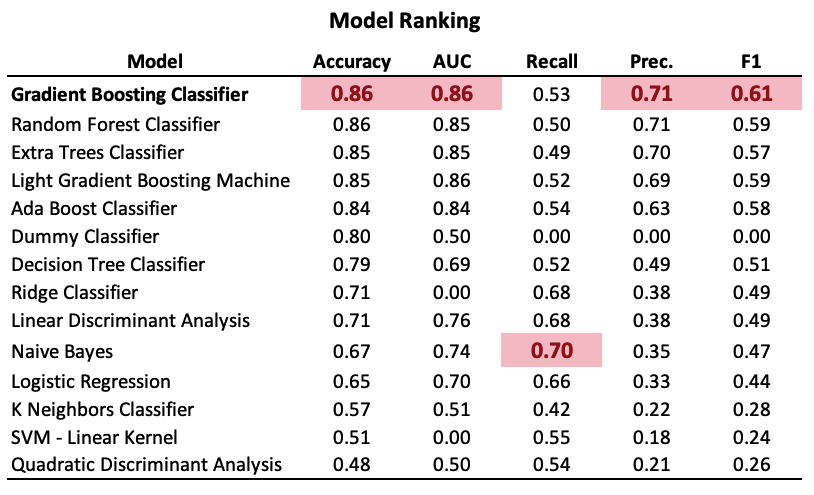
Ridge Classifier

SVM - Linear Kernel

The Python libraries Pandas and NumPy were used for data manipulation, Matplotlib and Seaborn for visualization, and PyCaret to perform the machine learning models.

**Results**

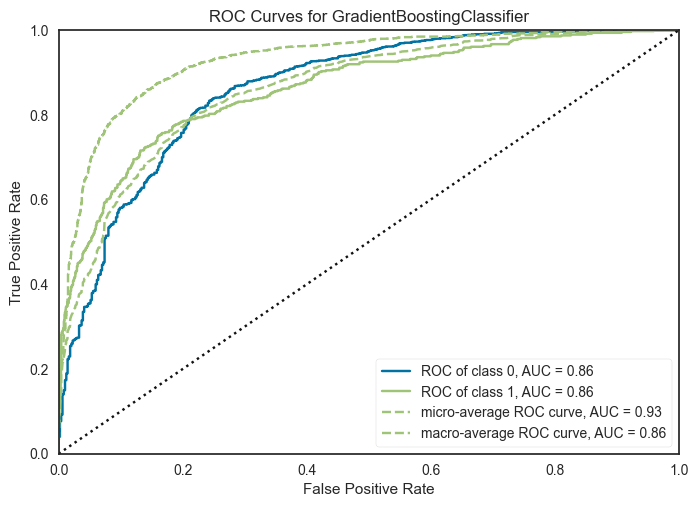
The performance index of each model is ranked and showed below. The red cells are the maximum values for each index among the models.



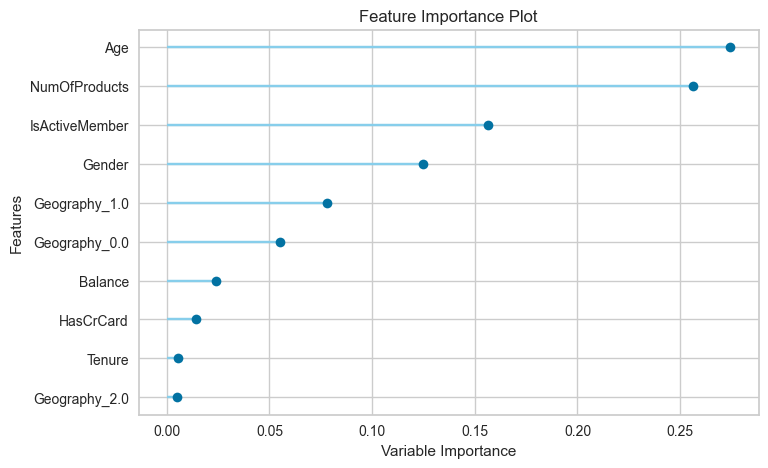
The accuracy, AUC, and F1 indexes—important measures to compare the overall performance of models—show that the Gradient Boosting Classifier is the best model among all. Particularly, the Accuracy is 0.86, which means around **86% of the training samples were predicted correctly**.

Based in these results the model chosen to continue with the analysis is the Gradient Boosting Classifier. This is a model which combine the prediction of multiple weak learners (decision trees) sequentially. This improves the fitting more than each tree could reach separately.

The ROC curve of this model shows a pronounced curve, with an AUC of 0.85, which is equivalent to saying that there is an 85% probability of ranking a randomly chosen positive instance higher than a randomly chosen negative instance.

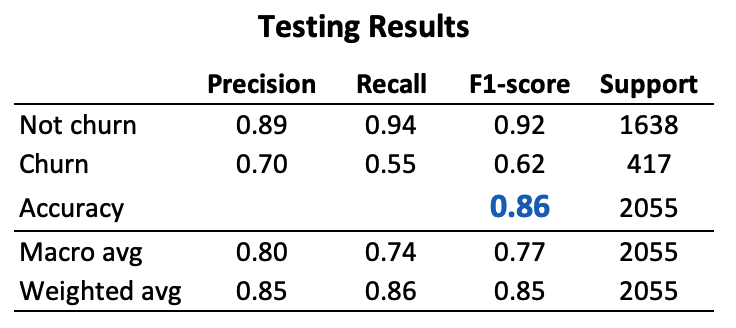


The ranking of the factors those explain better the variability of churning showed below. The three main factors are Age, and Number of Products boughed.

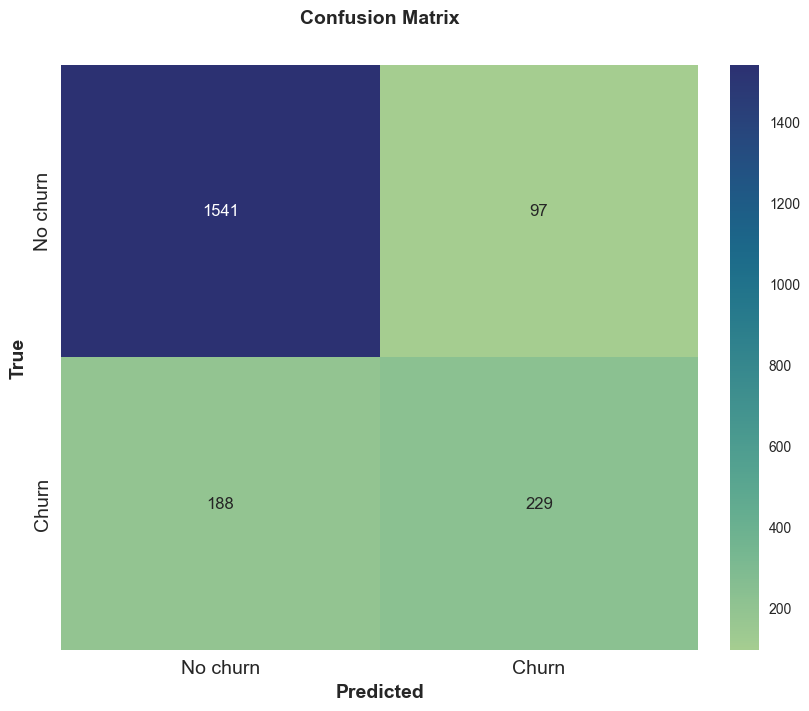


Finally, the test was made using testing dataset. It shows how good is the selected model to predict results with data it “doesn’t know”, or saying in other words, how good is this model when we try to test its generalization.

The results of this test show an Accuracy of 0.86, which means that **86% of the testing data were predicted correctly.** The testing accuracy is the same of from the training data performance, so we can surely say that there is no significative overfitting.



The confusion matrix shows the difference between of good prediction of No Churn and Churn.



**Conclusion**

Using a machine learning model, this study yielded good general performance, especially in identifying customers who don’t churn.

The main factors (Age and the number of products purchased) suggest a subsequent segmentation analysis to define types of customers based on the type of products they buy and generation preferences.