StayWithMe | Churn Prediction

Lorenzo Meloncelli - Valerio Schettini - Jany Khan Gony - Carlo Ardito

1. Introduction

The goal of a corporation is constant growth. To achieve that, Customer Relationship Management (CRM) should collect and analyze data across different channels to improve the Customer Life Cycle and reduce Churn Rates. In particular, the latter represents the percentage of customers who decided to cancel a subscription or service. Churn Rate is an essential metric throughout multiple industries, and corporations must reduce it.

This project aims to predict who will leave the StayWithMe (SWM) Bank, going from an Active Customer to an Attrited one.

2. Exploratory Data Analysis

To fully achieve our goal, it is crucial to start with a comprehensive Exploratory Data Analysis of the available data in order to uncover patterns, spot anomalies, test hypotheses and verify assumptions.

The data provided consists of 10,127 observations for each of which there are 17 different features:

- Basic info:
    - CLIENTNUM: Unique identifier for the customer holding the account.

- Target:
    - Attrition_Flag: Specifies whether the account was closed (Attrited Customer).

- Demographic Variables:
    - Customer_Age: Demographic variable - Customer's Age in Years.
    - Gender: Demographic variable - M=Male, F=Female.
    - Dependent_count: Demographic variable - Number of dependents.
    - Education_Level: Demographic variable - Educational Qualification of the account holder (example: high school, college graduate, etc.).
    - Marital_Status: Demographic variable - Married, Single, Divorced, Unknown.
    - Income_Category: Demographic variable - Annual Income Category of the account holder (< 40K, 40K - 60K, 60K−80K, …).

- Variables (Product):
    - Card_Category: Product Variable - Type of Card (Blue, Silver, Gold, Platinum).
    - Months_on_book: Period of relationship with the bank.
    - Total_Relationship_Count: Total no. of products held by the customer.
    - Months_Inactive_12_mon: No. of Months in the last 12 months.

○ Contacts_Count_12_mon: No. of Contacts in the last 12 months.
○ Credit_Limit: Credit Limit on the Credit Card.
○ Total_Trans_Amt: Total Transaction Amount (Last 12 months).
○ Total_Trans_Ct: Total Transaction Count (Last 12 months).
○ Avg_Utilization_Ratio: Average Card Utilization Ratio.

2.1 Data Visualisation

The following figures represent the relation between the variable 'Attrition Flag' and most of the other variables that are present in the dataset.

Fig1. Percentages of 'Existing Customers' and 'Attrited Customers'

Fig2. The total number of products held by 'Existing Customers' and 'Attrited Customers'

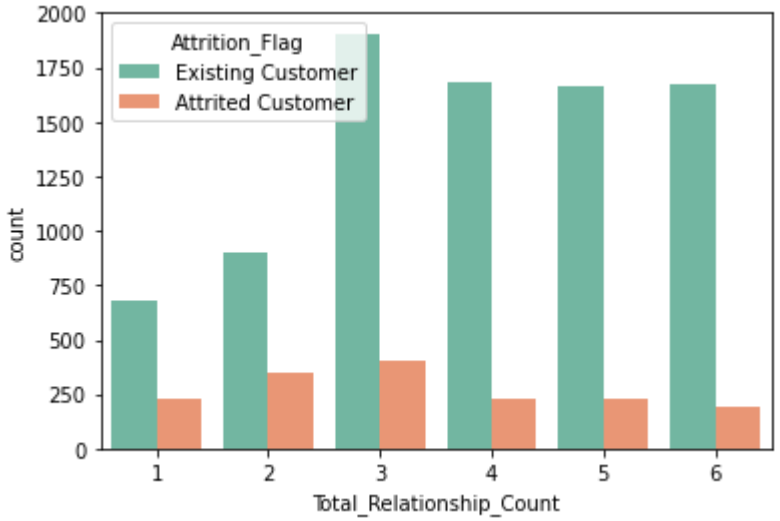


Fig3. Total transaction amount done by ‘Existing Customers’ and ‘Attrited Customers’ in the last 12 months.

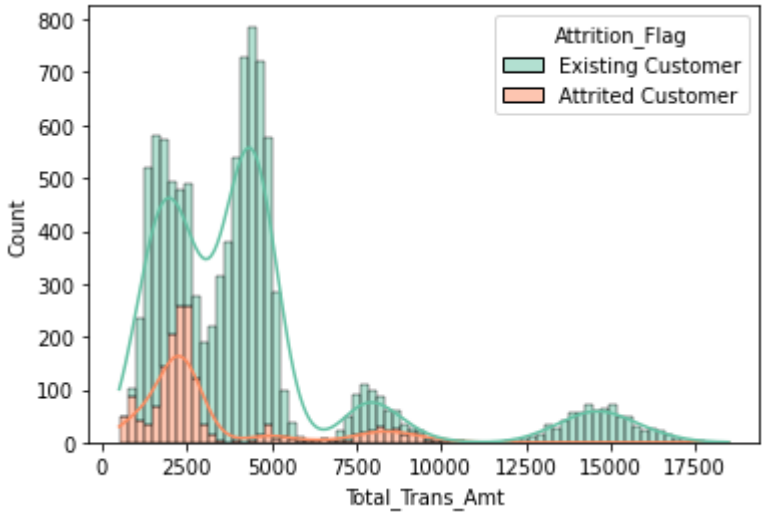
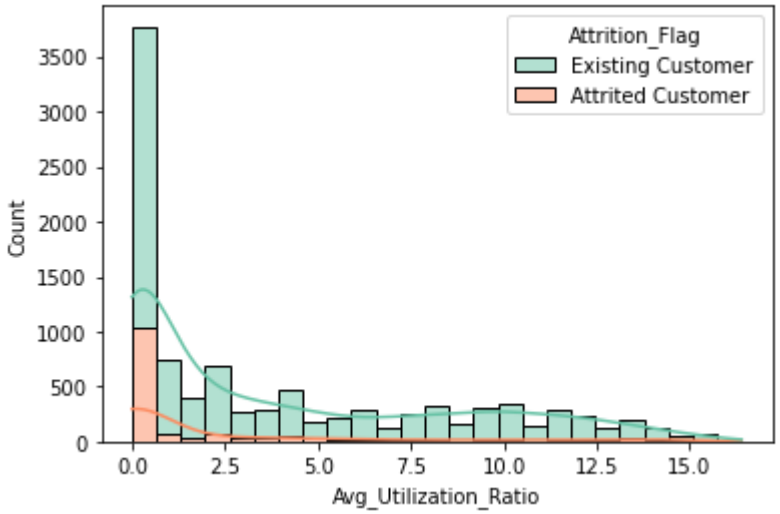


Fig4. Average card utilization Ratio by ‘Existing Customers’ and ‘Attrited Customers’ in the last 12 months.

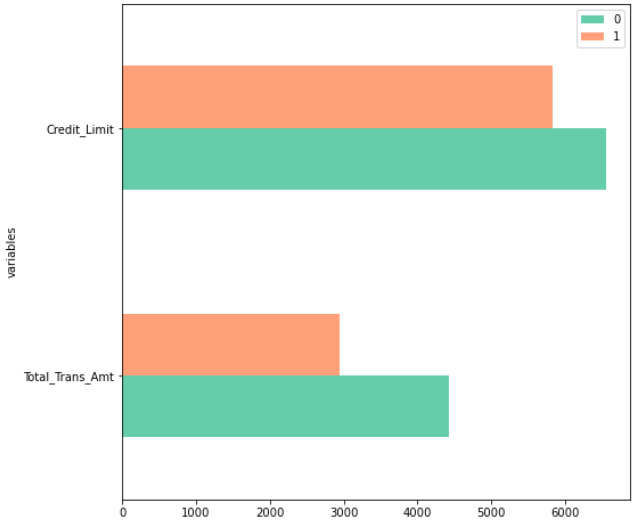


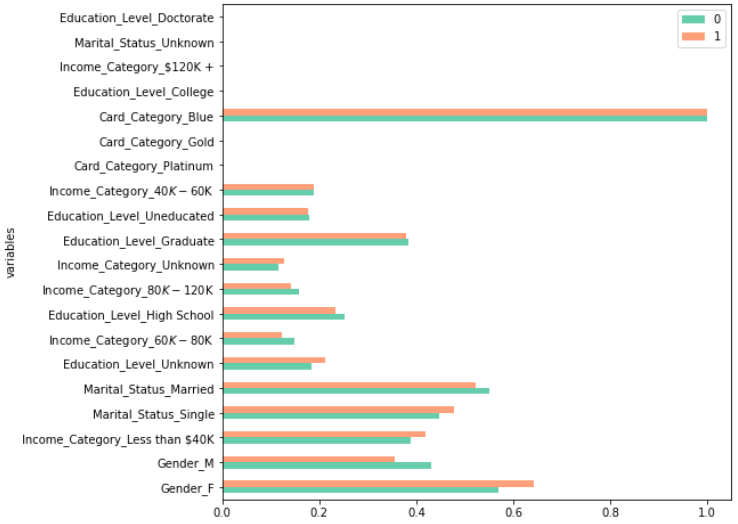
The main discovery we made concerns the dependent variable. For instance, the dataset seems to consist mostly of attrited customers. On the one hand this is good, as not many customers want to leave the company in question. On the other hand, this bias can influence many machine learning algorithms, leading some to ignore the minority class entirely.

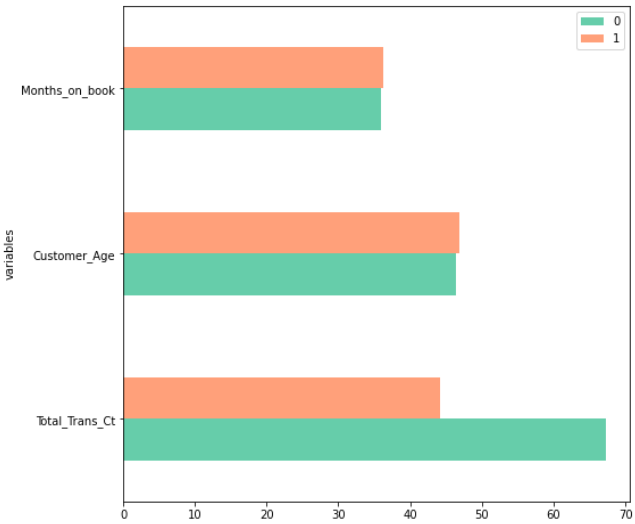
2.2 Mean Analysis

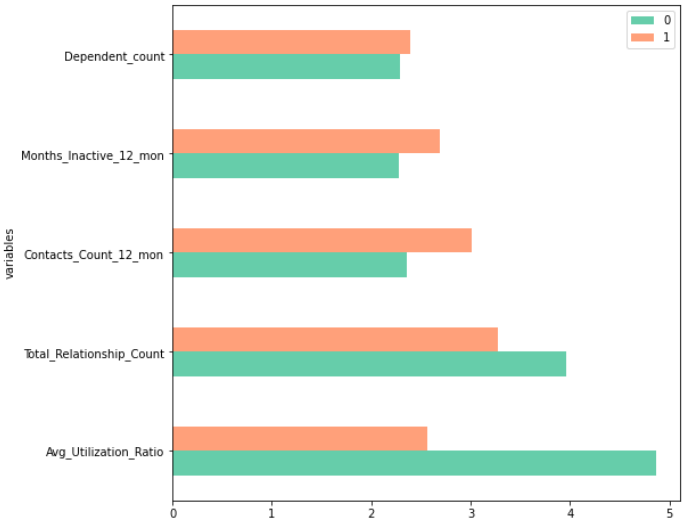
We have also implemented what we call a “Mean Analysis”, so as to reveal some more details about the “Churn Persona”.
For a better visualization, we first created subsets made of variables whose values fall within a similar range. Then, for each subset, we grouped the dataset into active customers and attried ones.

Then, for both groups, and for all subsets, we computed the average value of each feature present in the dataset. The results were finally sorted in descending order based on the difference between the means. In this way we obtained a ranking of the variables whose values, on average, differed the most between the active customers and the attrited ones.







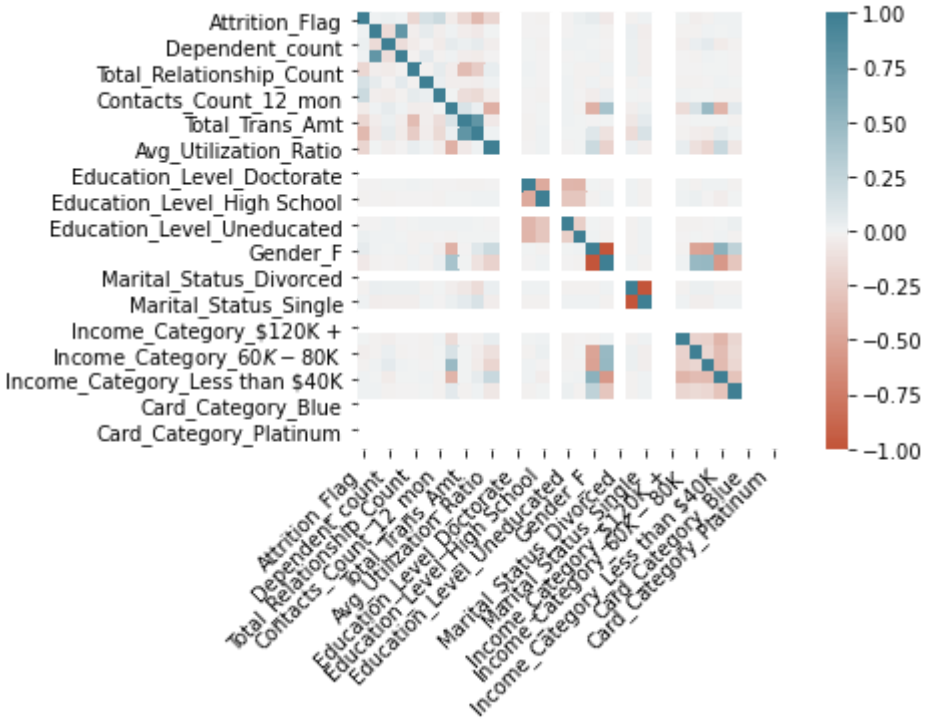


So far, highlighting that the average card utilization ratio, together with the total transaction count and the total transaction amount of the attrited customers are, on average, lower than those of the existing customers, our analysis has already provided useful insights to the SWM Bank for monitoring its clients.

2.3 Analysis Of The Correlations

Finally, we have analyzed the correlations between all of the features present in the dataset. Even if 'Correlation doesn't imply causation', this heatmap provides even more insights, and helped us decide whether to drop some variables for our prediction models.

Fig5. Blue shades represent a positive correlation, while red shades represent a negative correlation.



We have tried to fit the models that will be described in the following sections removing the most correlated variables from the dataset. However, after noticing a performance deterioration, we decided to leave them all, so as to retain all the information.

3. Data Preparation

For the sake of tuning our prediction models, it is first necessary to prepare our data.
We first replaced outliers with the respective average. Then, to overcome the problem of the unbalanced dataset, we randomly selected examples from the majority class and deleted them from the training dataset (random undersampling).

4. Models

Attached to this report is a Jupyter notebook (StayWithMe _ Churn Prediction) with the running code to demonstrate how, among other things, we set up the models.

We decided to use Logistic Regression and four different types of machine learning algorithms:

- K-Nearest-Neighbor
- Support Vector Machine
- Random Forest
- Artificial Neural Network.

In order to maximize the performance of models, we first had to tune them with Grid Search, a method that consists of manually defining a subset of the hyperparametric space and exhausting all combinations of the specified hyperparameter subsets. Each combination's performance is then evaluated using cross-validation and the best performing hyperparametric combination is chosen.

We have observed that not identifying an attried customer is far more expensive than identifying as attried someone that is not. This derives from the fact that keeping a customer is far less expensive than acquiring one. For this reason, in our model evaluation cycle, we used the Recall metrics to decide which prediction method is the best. Precisely, out of all positive predictions that could have been made, recall quantifies the number of correct positive predictions.

4.1 Calibrated Probability

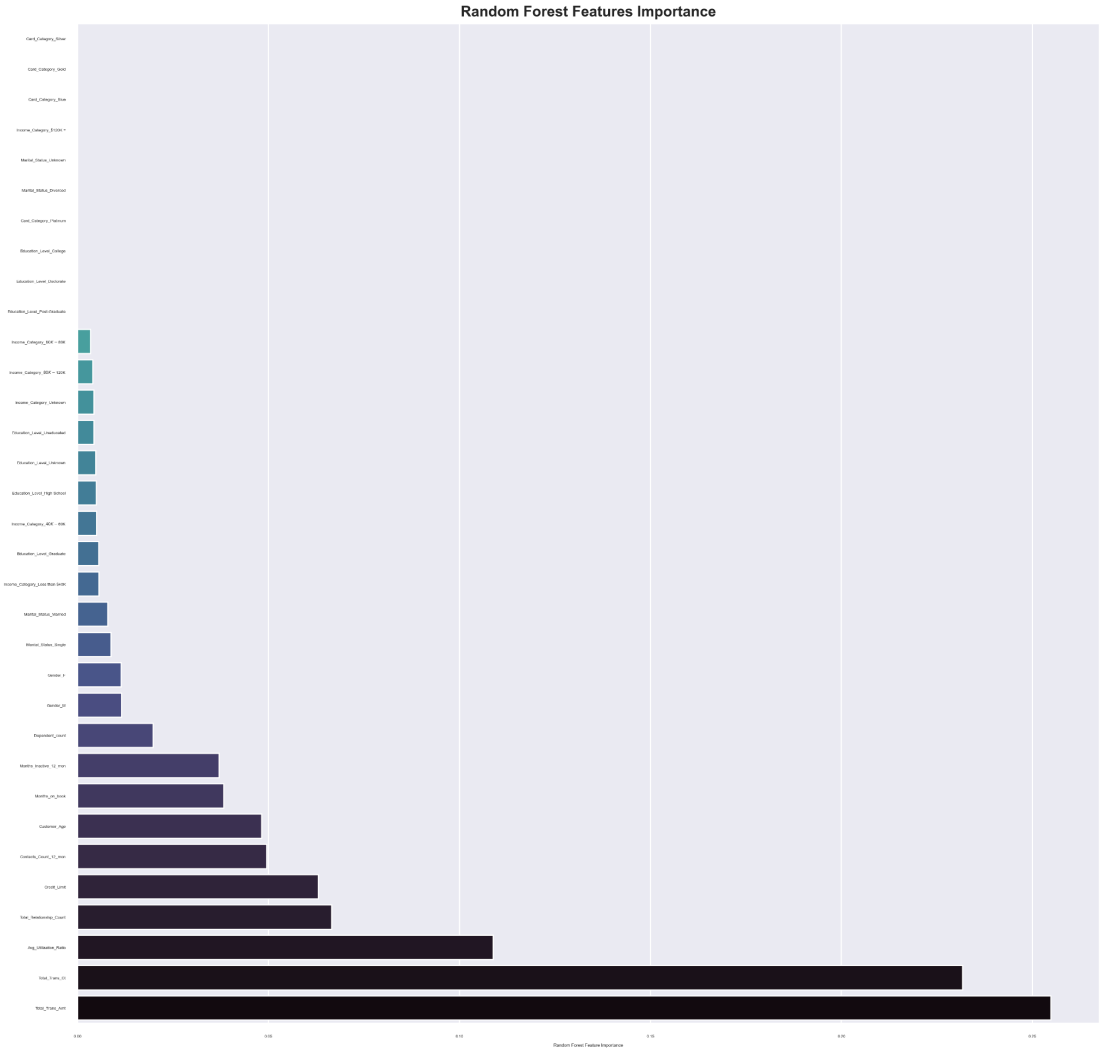
A highlight of our code is the possibility of modifying the threshold that our models use to make their predictions. As standard, the threshold we used for the following results is 0.5,which means that if the probability of being attried exceeds 0.5, the customer is classified as such. However, within the decision-making process of the Bank, managers can tune it at their discretion. For example, if the costs of "managing" a customer identified as attried were not so high, it could be useful to decrease the threshold to 0.4 and therefore increase the number of people identified as such.

This was possible thanks to Probability Calibration, the process of calibrating an ML model to return the true likelihood of an event.

4.2 Random Forest

The model that performed better in terms of recall is the Random Forest. This learning method is used to accomplish many tasks, thanks to its construction of Decision Trees that operates as an ensemble. The random forest is a straightforward method for understanding how variables contribute to the decision to leave.

We have plotted the Random Forest visualization of the variables that have influenced the Churn decision:



5. Results

Other than Recall, we have also computed other performance metrics, such as precision and accuracy. Below are our results:

	Recall	Accuracy	Precision
Random Forest	**0.92**	**0.90**	**0.87**
KNN	0.83	0.77	0.72
SVC	0.84	0.85	0.85
Artificial Neural Network	0.82	0.85	0.86
Logistic Regression	0.84	0.84	0.83

The Artificial Neural Network did not perform as expected because it was trained on the dataset created using the random undersampling method explained above. We think that the consistent reduction in the number of observations was the major cause for the unsatisfactory results. Another attempt we made was that of leaving the dataset unbalanced, adjusting the threshold for classification in order to increase our recall at the expense of both precision and accuracy. Again, even though we observed a slight increase in performance, the random forest remained the model that offered the best results.

6. Future Developments

The algorithms we have implemented can be further improved with the feedback deriving from the SWM Bank Management. The possibility to modify the threshold provides a great personalisation of the analysis improving the decision-making process of the CRM department of SWM. Further improvements could allow, with the right learning system, to automatically receive new data and produce more and more accurate forecasts.