Noemie Turlotte
Angelina Benyahia
Marie-Lou Leroy

Credit Scoring Report

Account Churn Prediction
Using Machine Learning Classification

November 2024

I. Introduction - Business case and problem

A. Business case

Société Générale is currently facing a significant business challenge: an increasing number of clients, including long-term customers with valuable financial histories, are leaving the bank for competitors such as Crédit Agricole and BNP Paribas. This migration threatens the banks' revenue and market position, as these clients are drawn to more personalized offers and superior digital experiences that Societe General has yet to match. The risk managers' current approach, which was effective in the past, no longer meets the tailored engagement expectations of today's customers, leaving them feeling disconnected and prompting their departure.

B. Problem

Our role as risk managers is to predict customer churn for the company Societe Generale through a machine learning classification model, which will identify customers at high risk of closing their accounts. Our model will allow the company to take preventive measures and improve customer retention.

C. Tools used

This project was conducted through google collabs, using Python and a number of libraries such as numpy and scipy for numerical operations, pandas for data manipulation, matplotlib and seaborn for visualization, and scikit learn which contains the machine learning methods.

II. Data preprocessing and feature engineering strategy

The dataset is an excel file named “account_churned_project.xlsx” that contains 10,127 entries and 21 columns, detailing various attributes related to customer demographics, accounts and behaviors.

Key columns include:

- identification: Unique identifier for each customer.
- churn_flag: Status indicating whether the customer is still active or has churned.
- age, gender, number_dependants: Demographic information.
- education, civil_status, income: Socioeconomic details.
- account_category, account_age: Account-specific information.
- balance, card_Limit, open_to_use: Financial metrics.
- total_transaction_amount, total_transaction_count: Transactional behavior metrics.

In data preprocessing, we aim to prepare the data for modeling by addressing data quality issues, transforming data types, and scaling values. Here are the four steps that will ensure our model does not encounter any difficulties or bias while exploiting the data:

1. Handling Missing and Inconsistent Data
Missing or inconsistent values can lead to model bias or errors. Fortunately, there were no missing values in our file.

2. Encoding Categorical Variables
Machine learning models can only process numerical data, so categorical values need to be converted.

What we changed:

- We use binary encoding to convert gender and churn_flag into binary (e.g., 0 for Male, 1 for Female in gender; 0 for Existing Customer, 1 for Churned in churn_flag).
- For categorical variables with more than two categories, such as education, civil_status, and income, we used One-hot encoding to create binary columns for each category

3. Scaling Numerical Data
Scaling ensures that numerical features have comparable ranges, improving model performance.

What we changed:
We applied standard scaling to columns with large ranges, like balance, card_Limit, total_transaction_amount, etc., transforming them to a mean of 0 and standard deviation of 1.

We also normalized variables like account_age or open_to_use to a 0-1 range, which is particularly useful for distance-based models (e.g., KNN).

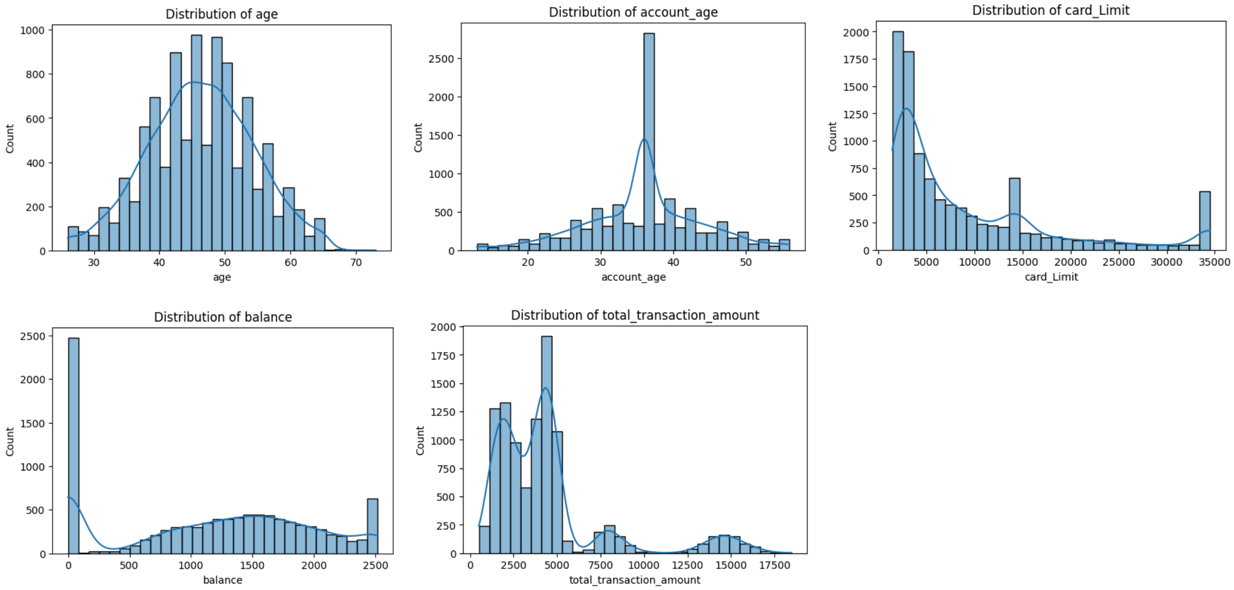
4. Feature Engineering Transformations
Derived features may better capture important aspects of customer behavior.

What we changed:
Card Utilization: We created a new feature, card_utilization, by calculating balance/card_Limit to measure how much of their credit limit a customer uses.

III. Exploratory data analysis

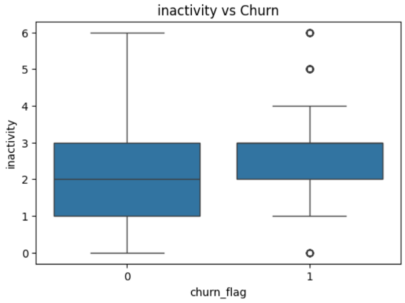
Exploratory Data Analysis (EDA) allows us to gain insights into the factors influencing churn by performing statistical analysis and visualizing relationships within the dataset.

We studied some of the features to get a better understanding of the dataset

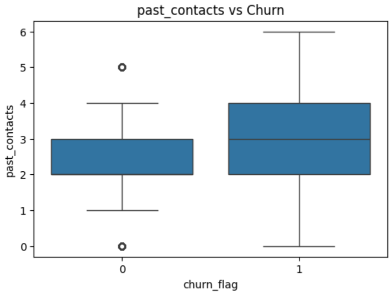


By creating box plots for each key feature against churn status, we can visually compare the distributions of the feature values for churned customers versus existing customers. This helps identify potential patterns or differences in the features that might be indicative of churn. For example, if a box plot shows a significant difference in the median or interquartile range of a feature between the two churn groups, it suggests that the feature could be an important factor in predicting customer churn. Box plots can also reveal outliers and the overall spread of the data for each group.

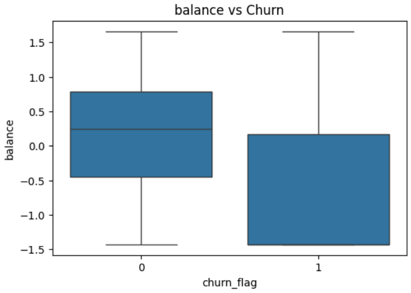
The boxplots with the greatest differences between existing and attrited customers are the following :



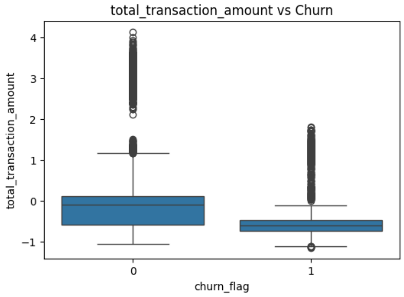
inactivity: The box plot likely shows a higher median and a wider IQR for churned customers compared to existing customers. This suggests that churned customers tend to have longer periods of inactivity.



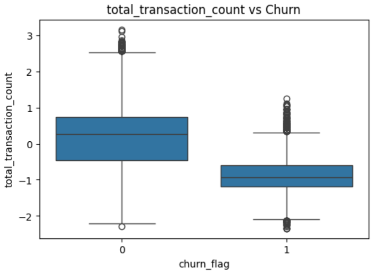
past_contacts: we observe a higher median and possibly a wider IQR for churned customers. This indicates that churned customers might have had more contact with customer service or support in the past.



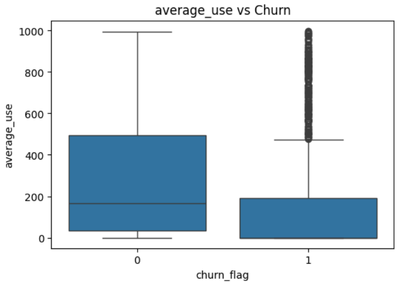
balance: this shows a lower median and potentially a narrower IQR for churned customers. This suggests that customers with lower balances are more prone to churn.



total_transaction_amount: The box plot shows a lower median and a narrower IQR for churned customers compared to existing customers. This implies that churned customers tend to have lower overall transaction amounts.



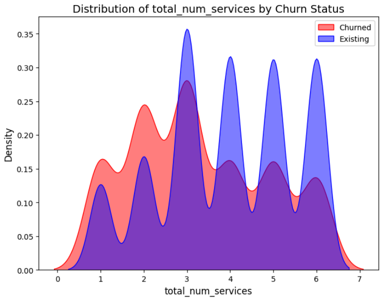
total_transaction_count: Similar to 'total_transaction_amount, the box plot for 'total_transaction_count' might display a lower median and a narrower IQR for churned customers, indicating they make fewer transactions.

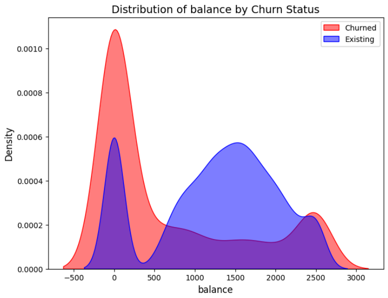


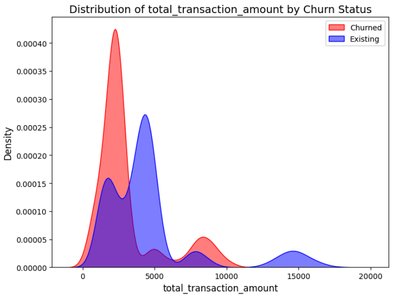
average_use: The 'average_use' box plot might demonstrate a lower median for churned customers, suggesting they have lower average usage of the service or product

We created overlayed histograms or KDE plots for churned vs. existing customers. The primary goal is to visually compare the distributions of key features for customers who churned versus those who remained. By visually comparing the shapes, peaks, and spread of the distributions, we can gain insights into how these features differ between churned and existing customers. ultimately leading to a better understanding of churn drivers and potential mitigation strategies.

The most remarkables histograms are :



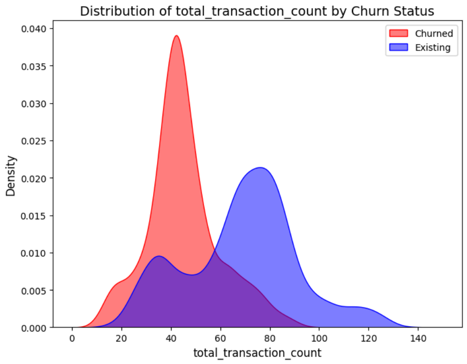


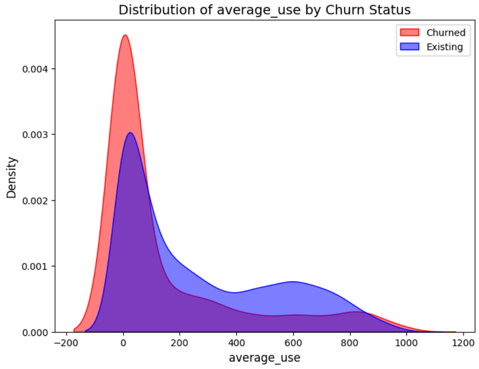


total_num_services: The KDE plot likely shows a higher peak for existing customers at a lower number of services, while churned customers might have a slightly higher peak at a higher number of services, or a more spread-out distribution. This could indicate that customers with a specific number of services are more or less likely to churn.

balance: The 'balance' KDE plot might reveal a shift in the distribution for churned customers towards lower balances. This supports the idea that customers with lower balances are more prone to churn.

total_transaction_amount: The KDE plot for 'total_transaction_amount' could show a similar pattern to 'balance', with churned customers having a distribution shifted towards lower transaction amounts. The peak for churned customers might be lower and potentially to the left of the peak for existing customers. This suggests lower spending by churned customers.

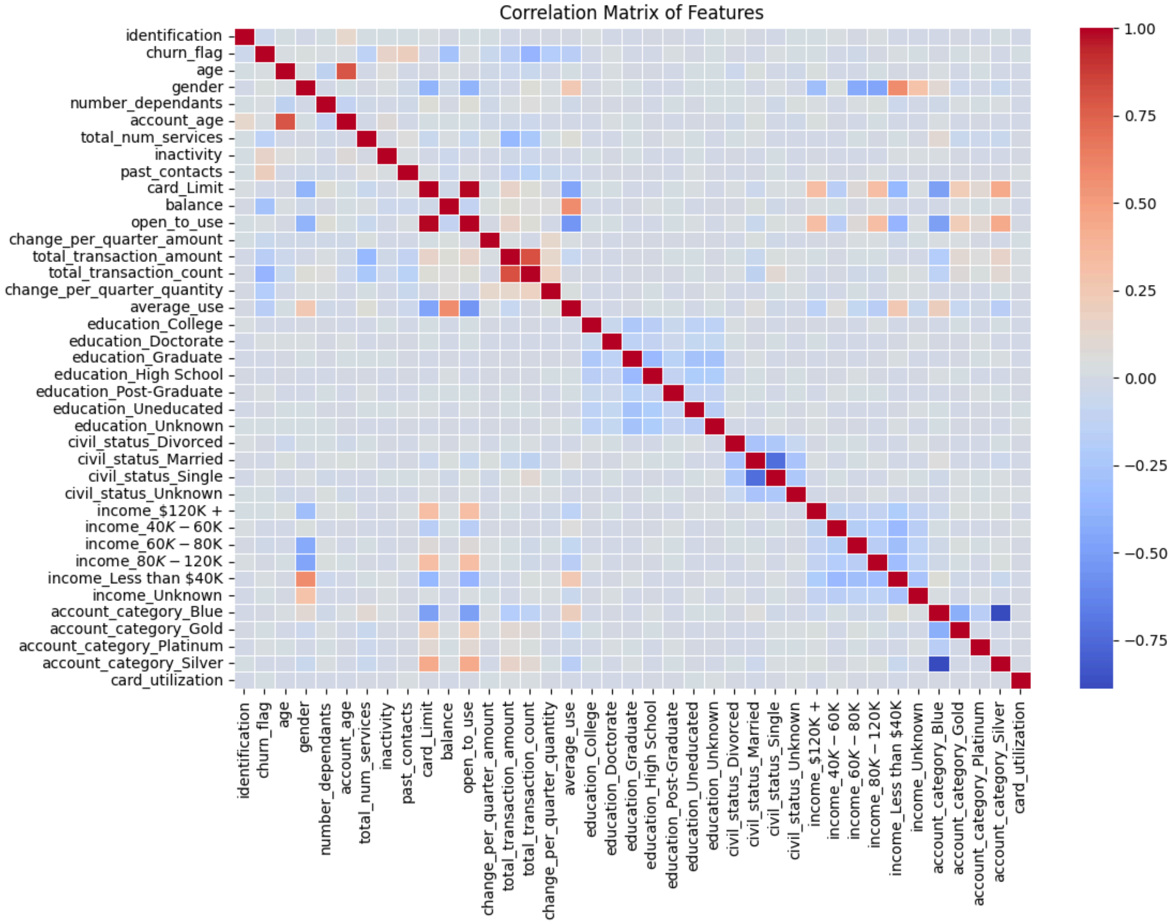




total_transaction_count: This plot also shows a shift towards lower transaction counts for churned customers. The peak for churned customers might be lower and to the left of the peak for existing customers, indicating they make fewer transactions overall.

average_use: The 'average_use' KDE plot displays a lower peak for churned customers, shifted to the left compared to existing customers. This suggests that churned customers have a lower average usage of the service or product.

We created the correlation matrix



We then tried to identify the 10 features that have the highest absolute correlation with customer churn ('churn_flag') based on the linear relationship between them and churn. The list given is: total_transaction_count,balance, past_contacts, change_per_quarter_quantity, total_transaction_amount,average_use, inactivity,total_num_services, change_per_quarter_amount.

But while this approach can be useful for an initial understanding of feature importance, it has some limitations:

- Correlation primarily captures linear relationships between variables. It might miss important features that have non-linear relationships with the target.
- Highly correlated features might carry similar information, leading to redundancy. Selecting all of them could introduce noise and reduce model performance.
- Correlation analysis considers features individually and doesn't capture interactions between features that might be predictive of the target.
- Sometimes, correlations can be observed by chance, especially in high-dimensional datasets. These spurious correlations might not represent true relationships.

Machine learning models, particularly those designed for classification tasks, can address these limitations and potentially provide better results for feature selection and prediction.

IV. Model strategy

To predict customer churn effectively, a carefully designed model strategy is essential to balance accuracy, interpretability, and scalability. Given the nature of the problem and the dataset, the following considerations shaped our approach:

1. Nature of the Problem

Customer churn prediction is a binary classification problem, where the target variable (churn_flag) indicates whether a customer has left the bank. The primary goal is to identify high-risk customers accurately and take timely actions to retain them. Specific characteristics of the dataset inform the choice of models:

- Mixed Data Types: The dataset contains both numerical (e.g., balance, transaction_amount) and categorical variables (e.g., gender, education).
- Imbalanced Classes: Customers who have churned likely represent a smaller proportion of the data, which necessitates methods that handle class imbalance.
- Complex Relationships: Customer behavior often exhibits non-linear patterns, requiring models that can capture intricate interactions between features.

2. Model Selection Rationale

A mix of simple and complex models is tested to ensure a thorough exploration of performance trade-offs, especially between interpretability and predictive power:

Model	Justification
Logistic Regression	Serves as the baseline model due to its simplicity and interpretability. It offers insights into feature importance through coefficients and works well with imbalanced data when combined with techniques like class weighting or oversampling.
Random Forest	A powerful ensemble method that handles non-linear relationships and ranks feature importance effectively. It is less sensitive to overfitting compared to single decision trees.
K-Nearest Neighbors (KNN)	Useful for capturing local patterns in data. However, its sensitivity to high dimensionality and larger datasets may limit its applicability in production settings.

We are going to test three different classification models to compare results and evaluate which one is the most accurate. A **logistic regression model** could provide a baseline prediction by identifying linear relationships between features and churn risk. **The K-Nearest Neighbor (KNN) model** will classify customers based on similarity to other clients in the dataset, allowing us to assess churn risk using proximity in feature space. Finally, the **Random Forest model** will leverage ensemble learning to capture non-linear patterns and interactions between variables, potentially offering more robust predictions. By comparing these models, we aim to determine which approach is most effective for accurately predicting customer churn and supporting Société Générale in improving retention strategies.

V. Presentation of the model results and performance

Reminder :

True Positives (TP): Cases where the model correctly predicted churn (positive class).

False Positives (FP): Cases where the model predicted churn, but the customer did not churn.

False Negatives (FN): Cases where the model missed predicting churn for customers who actually churned.

Precision measures the proportion of correctly predicted positive cases out of all cases predicted as positive by the model.
Precision=True Positives (TP)/(True Positives (TP)+False Positives(FP))

Recall measures the proportion of correctly predicted positive cases out of all actual positive cases in the dataset.
Recall=True Positives (TP)/(True Positives (TP)+False Negatives (FN))

F1-Score: This is a balanced measure between precision and recall.

Support: This shows the actual number of customers in each class in the test set (2551 existing customers and 488 churned customers).

Macro Average: This averages the precision, recall, and F1-score across both classes, giving equal weight to each.

Weighted Average: This also averages the metrics but considers the number of samples in each class (giving more weight to the majority class).

A. Logistic regression

a. Interpreting the Classification Report

```
Logistic Regression Metrics:
              precision    recall  f1-score   support

           0       0.96      0.84      0.90      2551
           1       0.50      0.81      0.62       488

    accuracy                           0.84      3039
   macro avg       0.73      0.83      0.76      3039
weighted avg       0.88      0.84      0.85      3039
```

For Class 0 (Existing Customers):
Precision: 0.96 Out of all the customers predicted as "Existing," 96% were actually existing customers.
Recall: 0.84 Out of all the actual existing customers, the model correctly identified 84% of them.
F1-score: 0.90 This is a balanced measure of precision and recall, indicating an overall good performance for this class.
Support: 2551 This is the number of actual existing customers in the test dataset.

For Class 1 (Attrited Customers):
Precision: 0.50 Out of all the customers predicted as "Attrited," only 50% were actually attrited customers.
Recall: 0.81 Out of all the actual attrited customers, the model correctly identified 81% of them.
F1-score: 0.62 This is lower than the F1-score for class 0, indicating a relatively weaker performance in identifying attrited customers.
Support: 488 This is the number of actual attrited customers in the test dataset.

Overall Metrics:
Accuracy: 0.84 The model correctly classified 84% of all customers in the test dataset.
Macro Average: This averages the unweighted mean per label.
Weighted Average: This averages the support-weighted mean per label.
Interpretation:

The model is quite good at identifying existing customers (high precision and recall for class 0).
It struggles more with correctly identifying attrited customers, as indicated by the lower precision for class 1. This suggests that the model might generate some false positives for churn.
The overall accuracy is decent, but it's important to consider the performance on both classes, especially if identifying attrited customers is a primary goal.

b. Interpreting the Warning

The ConvergenceWarning shows that the Logistic Regression algorithm might not have found the optimal solution within the default number of iterations (max_iter).

c. Reasoning and steps to consider:

Feature Scaling: The data likely has different scales, so use StandardScaler or MinMaxScaler to normalize/standardize before training.

The model shows reasonable overall accuracy but struggles with precision for the churned class, meaning it might flag too many non-churned customers as churned. To achieve better results, we did some hyper parameter tuning by changing the number of iterations, in a new updated Logistic Regression model, giving the model more steps to converge.

d.Updated Logistic Regression

Precision : Out of all the customers predicted as "Existing," 96% were actually existing customers. This is excellent, indicating a very low rate of falsely classifying existing customers as churned. For the predicted “Attrited” customers, 51% were actually attrited customers. This is an improvement from the previous 0.42, but still indicates a significant number of false positives.

Recall: For both existing and attrited customers, the model can capture a significant portion of the churned customers, with a recall of 0.85 and 0.82. This means that the model correctly identified 85% of the actual existing customers, and 82% of the actual attrited customers.

F1-Score: For class 0 (Existing Customers), the F1-score is 0.90. This is a balanced measure of precision and recall and shows a very good overall performance for this class. Class 1 (Attrited Customers) scores 0.62. This is an improvement from the previous 0.54 and suggests that the model has a better balance between precision and recall for the churned customers class now.

Overall:

Compared to the previous results, the updated model shows improvement, particularly in precision for the churned customers (increased from 0.42 to 0.51) and the F1-score for the churned customers (increased from 0.54 to 0.62). The overall accuracy has also improved slightly to 84%. However, the convergence warning still persists, indicating there might be further room for improvement by trying different solvers or scaling techniques or more feature engineering. Despite this, the model is performing better, correctly identifying a good portion of churned customers with fewer false positives than before.

B. K-nearest neighbors

a. Understanding the Metrics

```
KNN Metrics:
              precision    recall  f1-score   support

           0       0.89      0.80      0.84      2551
           1       0.32      0.50      0.39       488

    accuracy                           0.75      3039
   macro avg       0.61      0.65      0.62      3039
weighted avg       0.80      0.75      0.77      3039
```

Precision: In this case, it's 0.32, meaning only 32% of those predicted as churned were actually churned.

Recall: Here, it's 0.50, meaning the model caught 50% of the actual churned customers.

F1-score: The F1-score is 0.39 for churned customers.

Accuracy: The KNN model has an accuracy of 0.75 or 75%.

Weighted avg: The average precision, recall, and F1-score calculated by weighting each class's contribution based on its support (number of instances).

Interpretation

The model has a high precision (0.89) and recall (0.80) for "Existing Customers" (0), indicating it does a good job at correctly classifying customers who are not churning. However, the model has a lower precision (0.32) and recall (0.50) for "Attrited Customers" (1). This means it's making more mistakes when trying to identify customers who are likely to churn. The low precision suggests that many customers predicted as "Attrited" are actually not churning, while the recall indicates it's missing a significant portion of the actual churned customers.

Overall Accuracy can be misleading: Even though the overall accuracy is 75%, it's crucial to look at the performance on the "Attrited Customer" class separately. Since the dataset likely has more "Existing Customers" than "Attrited Customers," the accuracy can be high even if the model does poorly on the minority class (churned customers).

In Summary, while the KNN model has decent overall accuracy, it is not effectively identifying customers at risk of churning. We used techniques like SMOTE (Synthetic Minority Over-sampling Technique) to balance the classes or explore other models like Random Forest or Gradient Boosting, which often perform well in imbalanced datasets.

C. Random forest

a. Understanding the Metrics

b.

```
Random Forest Metrics:
              precision    recall  f1-score   support

           0       0.97      0.96      0.97      2551
           1       0.81      0.83      0.82       488

    accuracy                           0.94      3039
   macro avg       0.89      0.90      0.89      3039
weighted avg       0.94      0.94      0.94      3039
```

Precision:

For class 0 (Existing Customer): 0.97. This means that 97% of the customers predicted as "Existing" were actually existing customers.

For class 1 (Attrited Customer): 0.81. This means that 81% of the customers predicted as "Attrited" were actually attrited customers.

Recall:

For class 0 (Existing Customer): 0.96. This means that the model correctly identified 96% of the actual existing customers.
For class 1 (Attrited Customer): 0.83. This means that the model correctly identified 83% of the actual attrited customers.

F1-score:
For class 0 (Existing Customer): 0.97
For class 1 (Attrited Customer): 0.82

Support:
Class 0 (Existing Customer): 2551
Class 1 (Attrited Customer): 488

Accuracy: The overall correctness of the model's is 0.94, meaning the model correctly predicted the churn status for 94% of the customers in the test dataset.

Macro Average: The average precision, recall, or F1-score across all classes, without considering class imbalance.

Weighted Average: The average precision, recall, or F1-score across all classes, weighted by the number of instances in each class. This takes class imbalance into account.

In summary, the Random Forest model demonstrates strong performance, achieving high precision, recall, and F1-score for both classes. The model is particularly effective at identifying existing customers (class 0). While its performance is slightly lower for attrited customers (class 1), it still achieves a respectable F1-score of 0.82. The overall accuracy of 0.94 indicates a good overall prediction capability.

c. Interpretation

The Random Forest model is performing very well overall, with an accuracy of 94%. It is particularly good at identifying existing customers (high precision and recall for class 0). It is reasonably good at identifying churned customers as well (decent precision and recall for class1), although the precision could potentially be improved.
In the context of churn prediction, the model has a good balance between correctly identifying churned customers (recall) and avoiding false alarms (precision). The relatively high F1-score for churned customers indicates that the model is doing a decent job in this important aspect.

d. Next Steps

Overall, the Random Forest model seems to be providing a strong baseline for our churn prediction task.

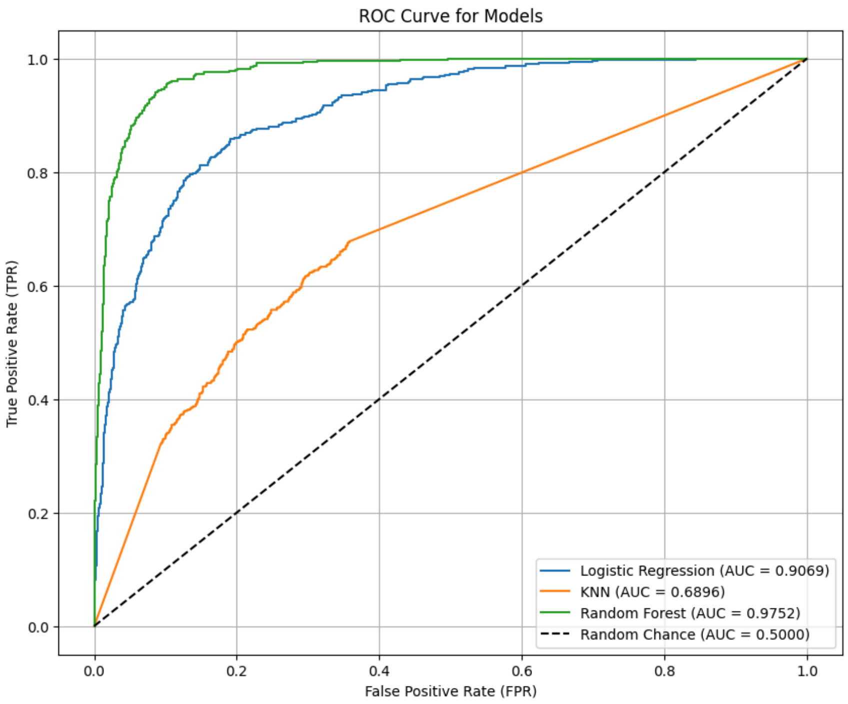
VI. Model selection and why ?

A. Model Selection Based on Comparison of Results

Previously, many results were presented for each model. When comparing the results of all three models, the Random Forest model shows the best scores; it has the best precision (0.97) and recall (0.96) for the existing customers class (class 0). It also demonstrates a good recall (0.83) for the attrited customers class (class 1). This means it correctly identifies the majority of customers in both classes, with relatively few false positives and false negatives. The F1-score of Random Forest is also the highest, with 0.97 for class 0 and 0.82 for class 1. This metric combines precision and recall, indicating a good balance in the classification of both customer types. This makes Random Forest a reliable model for both classes. Random Forest has the highest accuracy at 94%, meaning it correctly classifies a large number of data points. In comparison, Logistic Regression (84%) and KNN (75%) have lower accuracies, showing that Random Forest is more reliable for correctly predicting both classes. The macro and weighted averages of Random Forest are also superior, reaching 0.89 and 0.94, respectively, for precision, recall, and F1-score. This reflects the balanced performance of the model, considering all classes.

B. Area Under the Curve

In this part, we take a deeper look into the AUC : The Area Under the Curve is a metric used to evaluate the performance of a classification model, and further proves that the Random forest model is the right choice for this project.



a. Logistic Regression
 (AUC = 0.9063)

The Logistic Regression model demonstrates excellent discriminatory power. An AUC of 0.9063 suggests that the model has a high probability of correctly distinguishing between churned and existing customers. It's significantly better than random chance (AUC of 0.5), indicating a strong ability to predict customer churn.

b. K-Nearest Neighbors (AUC: 0.6891)

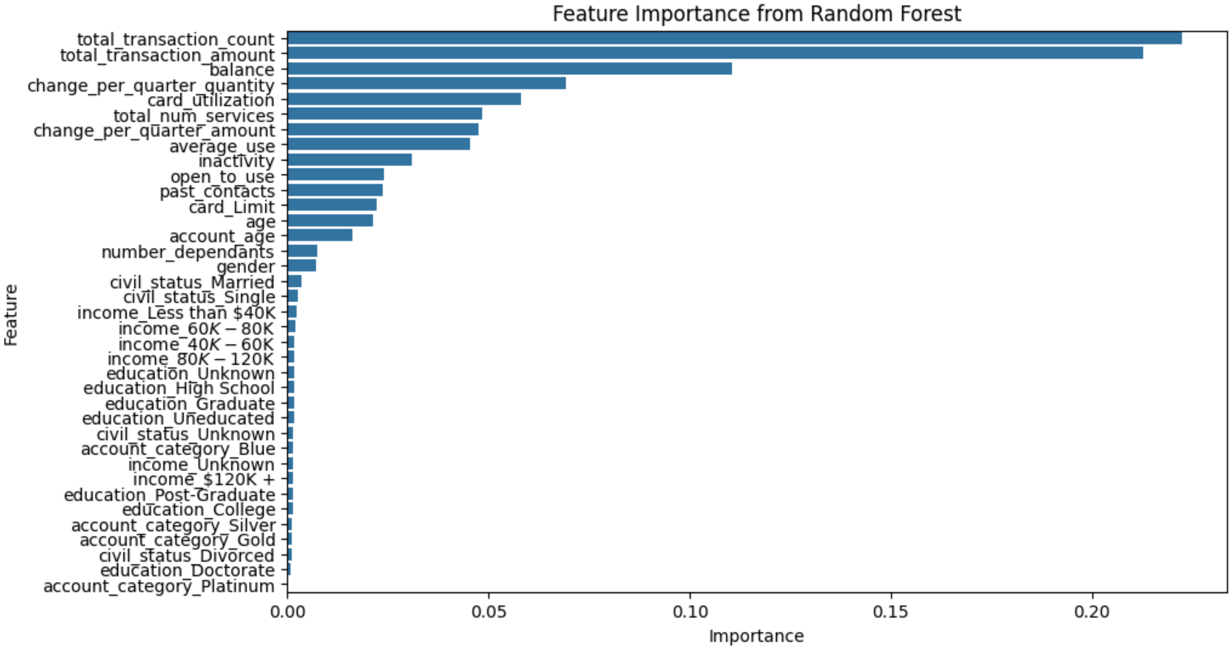
The K-Nearest Neighbors model shows moderate performance. An AUC of 0.6891 is above random chance but not as strong as the Logistic Regression or Random Forest models. It indicates a reasonable ability to differentiate between churned and existing customers, but there's room for improvement.

c. Random Forest (AUC: 0.9755)

The Random Forest model exhibits outstanding performance. An AUC of 0.9755 is very close to 1, signifying exceptional discriminatory ability. This model has a very high likelihood of correctly classifying customers as churned or existing, making it the best performer among the three based on AUC.

In summary, **Random Forest** is the top-performing model with better precision, recall, F1-score, higher accuracy, and better overall averages. It is particularly effective at correctly identifying attrited customers, which is crucial for this prediction task. It is followed by **Logistic Regression**. **K-Nearest Neighbors** demonstrates moderate performance but may not be as reliable as the other two for predicting customer churn in this scenario.

As we selected the Random Forest model, we can now express the feature importance in the model and define with a maximum accuracy the features with the higher impact on customer churn.



VII. Conclusion

Through our analysis and modeling efforts, we successfully identified key factors driving customer churn and developed a predictive framework using machine learning. Among the tested models, the Random Forest classifier emerged as the most effective, achieving superior precision, recall, F1-score, and AUC metrics. Its robust performance makes it the ideal tool for accurately identifying at-risk customers, enabling Société Générale to take timely, targeted actions. By leveraging this model, the bank can transition from a reactive to a proactive approach in customer retention, addressing churn risks before they materialize.

To combat the issue of churn, it is imperative that Société Générale evolves beyond traditional methods and adopts innovative, data-driven strategies. This not only ensures the retention of valuable customers but also enhances the overall client experience. Such efforts are crucial to maintaining the bank's competitive position in an increasingly dynamic financial landscape.

Appendix

VIII. Customer Retention Strategy

To mitigate customer churn, Société Générale must adopt a comprehensive and tailored retention strategy that addresses the needs of all customers while focusing on those most at risk.

General Measures for All Clients:
To improve satisfaction across the board, the bank should enhance its digital offerings by investing in advanced, user-friendly platforms. Features like real-time financial advice, personalized dashboards, and seamless transaction processes will cater to the digital preferences of modern clients. Furthermore, customer support services should be upgraded by providing 24/7 multi-channel assistance, including chatbots and live representatives, to resolve issues promptly and effectively. These enhancements should be complemented by loyalty programs that reward long-term customers with exclusive benefits, such as fee reductions or cashback offers, creating a sense of value and appreciation.

Targeted Measures for At-Risk Clients:
For customers identified as high-risk by the model, Société Générale should focus on personalized outreach. Tailored communication, such as individualized messages or offers addressing specific financial needs, can significantly improve engagement. High-value clients showing signs of churn should be assigned dedicated relationship managers to provide focused attention and support. Additionally, offering financial reviews to reassess product fit and propose better-aligned solutions can address dissatisfaction. Flexible account options, such as those tailored for retirement or major life transitions, may also prevent churn by ensuring products evolve with customer needs.

To further encourage retention, the bank should offer short-term incentives like fee waivers, bonus interest rates, or exclusive promotions to dissuade customers from leaving. Win-back campaigns targeting recently churned clients could be implemented to re-engage and recover lost customers through compelling offers and personalized outreach.

By combining these general and targeted measures, Société Générale can create a customer-centric ecosystem that not only reduces churn but also strengthens its competitive edge, fosters loyalty, and secures long-term growth.