École Nationale Supérieure d'Informatique et d'Analyse des Systèmes

Rapport du projet de fin de 2ème année

Filière Génie Logiciel

Sujet

Prédiction de churn avec le Machine Learning

Réalisé par :
BENTAZAR Rihab
EL BIACHE Houda

Jury :
Mme Hdioud Boutaina (encadrante)
Mme Berrada Bouchra

Année universitaire 2020 - 2021

Remerciements

Ce projet est le fruit des conseils et critiques bienveillantes d'un grand nombre de personnes. Il nous est agréable de nous acquitter d'une dette de reconnaissance auprès de toutes ces personnes, dont l'intervention a favorisé son aboutissement.

Tout d'abord, nos sincères remerciements s'adressent à notre chère professeure encadrante, Madame HDIOUD Boutaina. Ses compétences indéfectibles, de même que ses directives clairvoyantes et ses conseils instructifs nous ont toujours été singulièrement précieux au cours de ce projet.

Il nous est aussi agréable de remercier notre jury Madame BERRADA Bouchra d'avoir accepté de juger notre travail, aini que notre chef de filière, Madame EL ASRI Bouchra et les enseignants de notre filière qui ont veillé à notre formation en mettant à notre service leur savoir et leur érudition.

Toute notre considération et tout notre respect vont également à toute personne ayant contribué à l'élaboration de ce projet et à sa réussite.

Résumé

Le présent document est l'aboutissement de notre travail dans le cadre du projet de fin de deuxième année, dont l'enjeu principal est de mettre en pratique nos connaissances et compétences acquises en analyse des données et en Machine Learning. Le projet consistait à la prédiction du churn par le biais des algorithmes du Machine Learning.

Le churn (attrition en Français), étant un mot Anglais qui réfère à la perte de clients ou d'abonnés et qu'on retrouve principalement dans l'univers des entreprises de télécommunications ou dans celui des banques.

Nous avons entamé notre travail, tout d'abord, par la définition de la problématique et la compréhension du problème métier. Ensuite, nous avons procédé à l'import, le nettoyage, la préparation, ansi que l'analyse des données. Puis, nous avons construit notre modèle et évalué ses performances, suite à la comparaison de quelques algorithmes, pour finalement arriver à le déployer au travers d'une application web.

Ce rapport qui a pour but de décrire le déroulement de notre projet contient l'ensemble des éléments mis en jeu et permet de synthétiser le résultat du travail accompli.

Abstract

The current document is a report which summarizes the work done as part of our end of the second year project which main aim was to put into practice the skills and the knowledge we acquired in terms of Data Analysis and Machine Learning. This project consisted on predicting churn with Machine Learning algorithms.

Churn (attrition in French), being an English word which refers to the loss of customers or subscribers and which is mainly found in the world of telecommunications or banking companies.

First, we started our work, by defining and understanding the problematic. Then we proceeded to data importing, cleaning, preparation, and analysis. Following the comparison of some algorithms, we then built our model and evaluated its performance, to end up being able to deploy it through a web application.

This report, which purpose is to describe the progress of our project, contains all the elements involved and makes it possible to summarize the work accomplished, as well as its results.

Table des figures

Table des matières

Chapitre 1

Présentation du projet

1.1 Introduction

La prédiction du churn (ou attrition) de la clientèle est une condition essentielle pour une entreprise prospère.

La plupart des entreprises ayant une activité par abonnement surveillent régulièrement le taux de désabonnement de leur clientèle.

En effet, maintenir les taux de désabonnement aussi bas que possible est ce que chaque entreprise poursuit, et la compréhension de ces mesures peut aider les entreprises à identifier les désabonnements potentiels à temps pour les empêcher de quitter la base de la clientèle.

Dans ce projet, on va se focaliser sur le churn externe ou switch qui constitue un vrai problème pour les entreprises dans le cas de clients quittant pour partir chez le concurrent.

1.2 Problématique

Construire et conserver une clientèle fidèle peut être un défi pour toute entreprise, en particulier lorsque les clients sont libres de choisir parmi une variété de fournisseurs au sein d'une catégorie de produits ou de services.

De plus, fidéliser les clients existants est généralement plus évident et rentable que le fait de s'aventurer en quête d'en acquérir de nouveaux.

Pour cette raison, l'évaluation de la fidélisation des clients est cruciale pour les entreprises.

Il est essentiel non seulement de mesurer le niveau de satisfaction client mais aussi de mesurer le nombre de clients qui cessent de faire affaire avec une entreprise ou un service.

1.3 Objectif

Le churn désigne les clients ou les utilisateurs qui ont quitté les services ou qui migrent vers un concurrent du même secteur.

Il est très important pour toute organisation de conserver ses clients existants et d'en attirer de nouveaux en cas de perte d'un segment de la clientèle, ce qui peut ce révèler être mauvais en terme d'affaires.

L'objectif est d'explorer la possibilité de l'apprentissage automatique pour la prédiction de l'attrition afin de conserver un avantage concurrentiel dans l'industrie.

1.4 Contexte et périmètre

L'une des études de cas les plus célèbres et les plus utiles de la prédiction du churn se trouve dans l'industrie bancaire.

Il est tout autant important pour les banques, que pour toute autre entreprise, tous domaines confondus, d'analyser les données pertinentes des clients et de développer un modèle de prédiction robuste et précis pour fidéliser la clientèle et élaborer des stratégies pour réduire les taux d'attrition des clients.

1.5 Identification du type d'apprentissage automatique

Le type de sortie ou l'output que l'on attend de notre programme est bien une valeur discrète ; une catégorie (churn ou not churn).

Il s'agit donc d'un problème de classification binaire.

La classification binaire fait référence aux tâches de classification qui ont deux étiquettes de classe.

En règle générale, les tâches de classification binaire impliquent une classe qui est l'état normal (not churn) et une autre classe qui est l'état anormal (churn). La classe pour l'état normal reçoit l'étiquette de classe 0 et la classe avec l'état anormal reçoit l'étiquette de classe 1.

1.6 Méthodologie

Un projet de Machine Learning ne s'aborde pas comme un projet logiciel classique.

En effet la modélisation par apprentissage diffère totalement de la programmation stricte basée sur des règles et exceptions.

Il en va donc naturellement de même sur la façon d'aborder ce type de projet.

Réussir donc un tel projet de Machine Learning revient à principalement respecter les étapes ci-dessous :

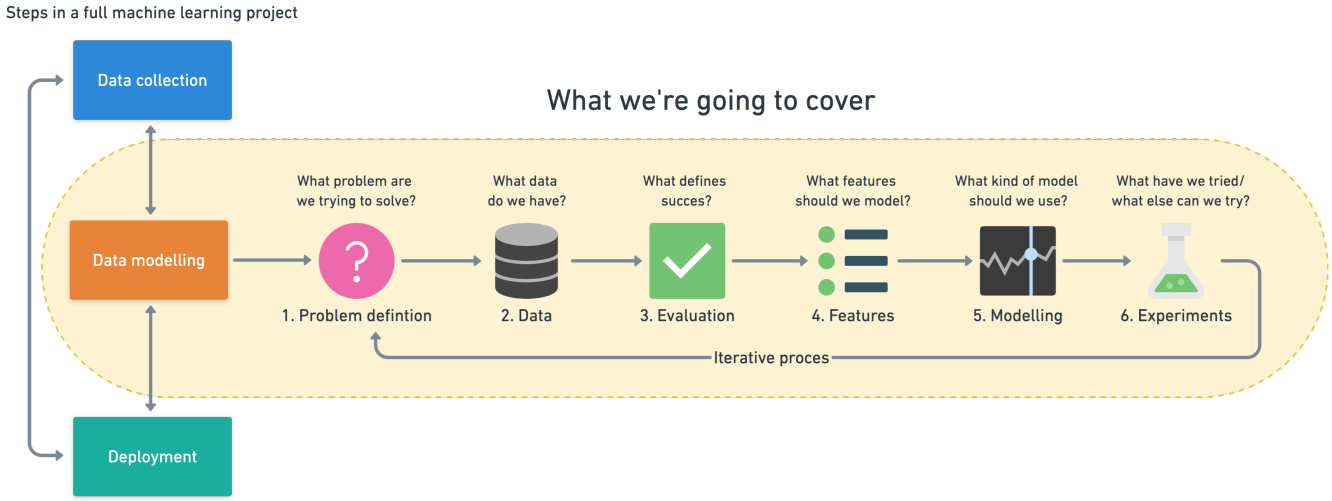


FIGURE 1.1 – Méthodologie de projet Machine Learning

Chapitre 2

Import et nettoyage des données

2.1 Dataset

Les données sont au cœur de l'apprentissage statistique (ML). Elles sont cruciales pour les algorithmes d'apprentissage qui en dépendent fortement.

L'ensemble de données a été fourni par Kaggle et est disponible ici : `https://www.kaggle.com/santoshd3/bank-customers`.

Certaines informations, telles que le nom de l'entreprise et les données privées des clients, ont été gardées anonymes pour des raisons de confidentialité et n'affecteront pas les performances du modèle.

2.1.1 Aperçu

	RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited	Reason for exiting company
0	1	15634602	Hargrave	619	France	Female	42	2	0.00	1	1	1	101348.88	1	High Service Charges/Rate of Interest
1	2	15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0	Nil
2	3	15619304	Onio	502	France	Female	42	8	159660.80	3	1	0	113931.57	1	Long Response Times
3	4	15701354	Boni	699	France	Female	39	1	0.00	2	0	0	93826.63	0	Nil
4	5	15737888	Mitchell	850	Spain	Female	43	2	125510.82	1	1	1	79084.10	0	Nil

FIGURE 2.1 – Aperçu du dataset

Notre collection de données contient 10000 lignes (échantillons de données) and 15 colonnes (attributs).

2.1.2 Attributs

1/ RowNumber : numéro de la ligne (type : int)
2/ CustomerId : identifiant du client (type : int)
3/ Surname : nom de famille (type : object)
4/ CreditScore : cote de crédit bancaire ou cote de solvabilité du client (type : int)
5/ Geography : pays ou région du client (type : object)
6/ Gender : sexe du client (type : object)
7/ Age : age du client (type : int)
8/ Tenure : nombre de mois que le client est resté avec la banque (type : int)
9/ Balance : solde bancaire du client (type : float)
10/ NumOfProducts : nombre d produits du client (type : int)
11/ HasCrCard : si le client a une carte de crédit ou pas (type : int)
12/ IsActiveMember : si le client est actif ou pas (type : int)
13/ EstimatedSalary : le salaire estimé du client (type : float)
14/ Exited : si le client a quitté la banque ou pas (type : int)
15/ Reason for exiting company : la raison pour laquelle le client a quitté la banque (type : object)

2.2 Nettoyage des données

Le Data Cleaning (nettoyage de données) est l'étape la plus importante avant d'analyser ou modéliser des données mais elle peut être très fastidieuse.

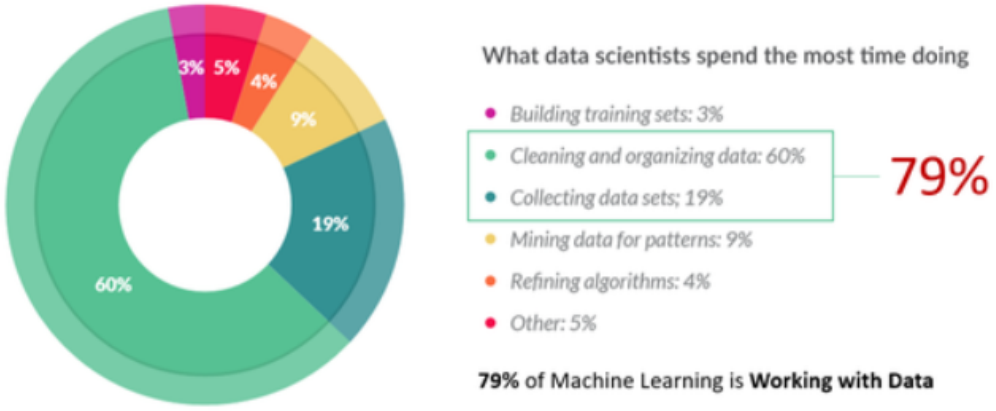


FIGURE 2.2 – Data Cleaning

L'utilisation de la bibliothèque Pandas de Python facilite cette tâche vu les outils prédifinis pour ce genre de traitement.

Les principales étapes de cette phase de nettoyage consistent principalement à :
-> Supprimer des colonnes non utilisées ou impertinentes.
-> Renommer les colonnes selon notre convenance.
-> Remplacer la valeur des lignes et la rendre plus significative.

Nous avons ainsi supprimé la colonne RowNumber puisque cette dernière n'est nullement importante ni pertinente pour notre modèle prédictif.

Puis, nous avons renommé quelques colonnes pour les rendre plus claires et plus significatives.

Ensuite, on a remplacé quelques valeurs : le champ HasCreditCard, à titre d'exemple, se compose des valeurs 1 et 0 étant 1 comme Yes et 0 comme No, mais cela semble souvent ambigu, donc nous en avons changé les valeurs pour rendre la compréhension plus simple.

Nous avons également vérifié si notre dataset ne contenait pas de valeurs nulles ou dupliquées. Ce qui n'est pas le cas.

Voici un aperçu de notre dataset à la fin de cette phase :

	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCreditCard	IsActiveMember	EstimatedSalary	Churn	Reason for exiting company
0	15634602	Hargrave	619	France	Female	42	2	0.00	1	Yes	Yes	101348.88	Yes	High Service Charges/Rate of Interest
1	15647311	Hill	608	Spain	Female	41	1	83807.86	1	No	Yes	112542.58	No	Nil
2	15619304	Onio	502	France	Female	42	8	159660.80	3	Yes	No	113931.57	Yes	Long Response Times
3	15701354	Boni	699	France	Female	39	1	0.00	2	No	No	93826.63	No	Nil
4	15737888	Mitchell	850	Spain	Female	43	2	125510.82	1	Yes	Yes	79084.10	No	Nil

FIGURE 2.3 – Aperçu du dataset suite au nettoyage des données

Chapitre 3

Visualisation, Analyse et Evaluation des données

Cette phase consiste à comprendre les ensembles de données en résumant leurs caractéristiques principales à travers des graphes pour la visualisation et l'exploration des données.

Pour celà, nous allons mettre en pratique les bibliothèques Python : Matplotlib et Seaborn.

3.1 Variable target : Churn

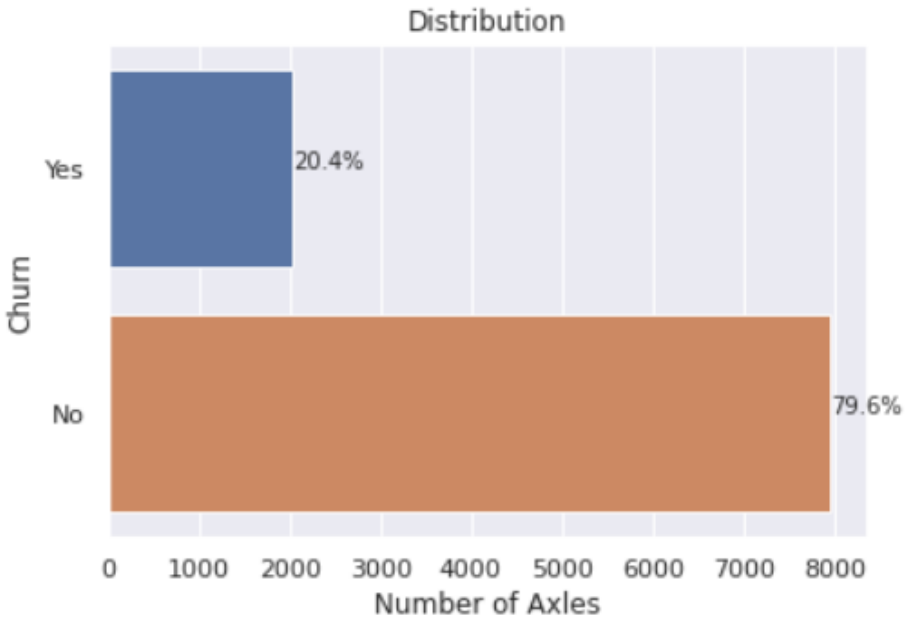
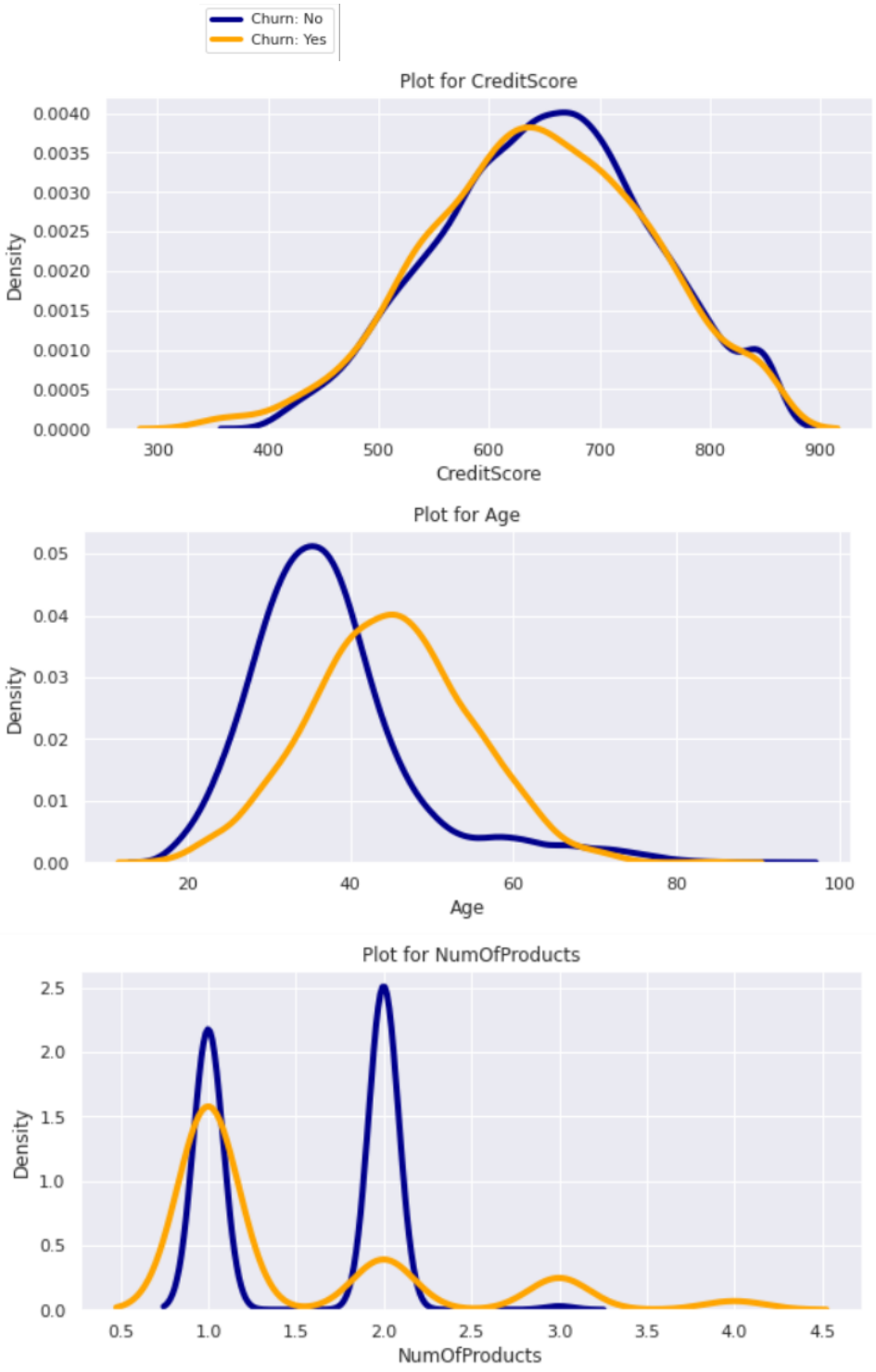


FIGURE 3.1 – Plot of Churn Distribution

Nous avous donc clairement un problème de classification binaire avec une cible légèrement déséquilibrée :

Churn : No — 79.6 pour cent
Churn : Yes — 20.4 pour cent

3.2 Données numériques



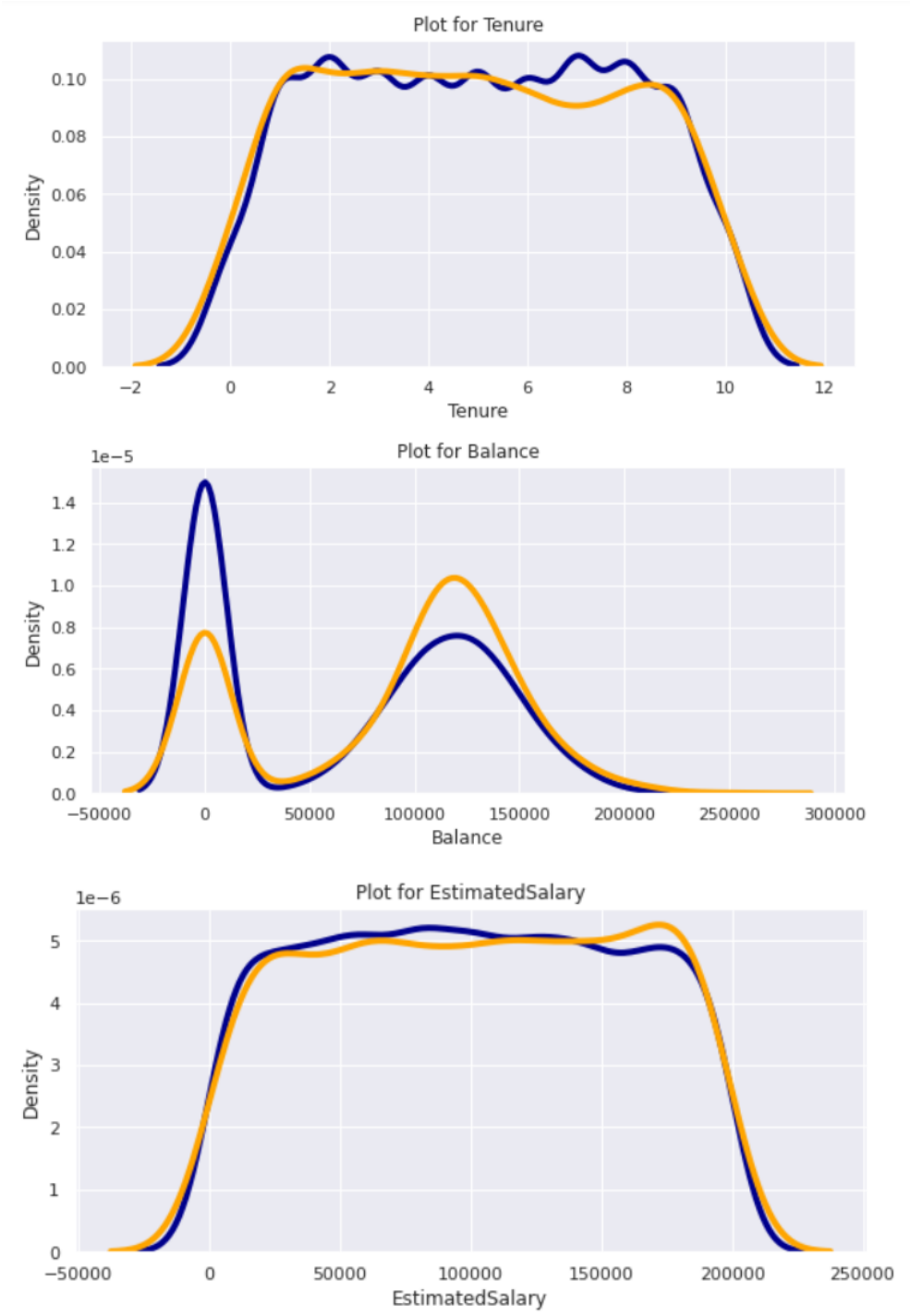


FIGURE 3.2 – Plot for numerical values

D'après les graphes précédents, nous pouvons conclure que :

— Les clients ayant un credit score entre 400 et 900 ont plus de chance de churn
— Ainsi que les clients adultes (de 25 à 65 ans)
— De même pour les clients ayant uniquement 1 produit
— Les clients ayant un solde bancaire nul ou moyen ont plus tendance à quitter la banque en question

3.3 Données binaires

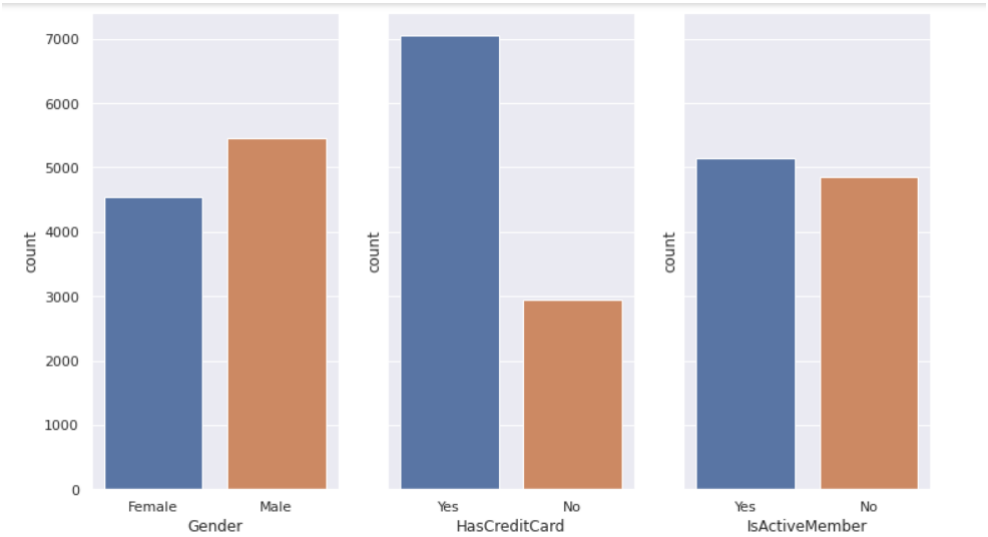


FIGURE 3.3 – Plot for binary values

— Notre dataset est légèrement plus populé par le genre masculin que féminin
— Nous avons plus de clients ayant une carte de crédit
— Nous avons presque tout autant de membres actifs qu'inactifs

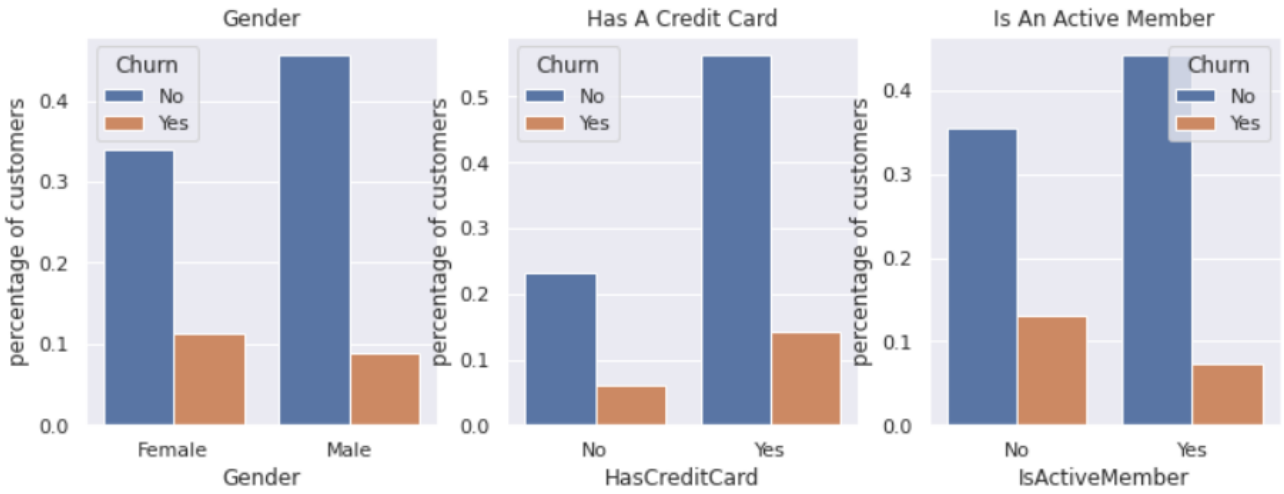


FIGURE 3.4 – Plot of churn distribution for binary values

— Les femmes ont tout autant de chances de churn que les hommes
— Les clients ayant une carte de crédit sont plus sujets au churn
— Idem pour les membres inactifs

3.4 'Geography'

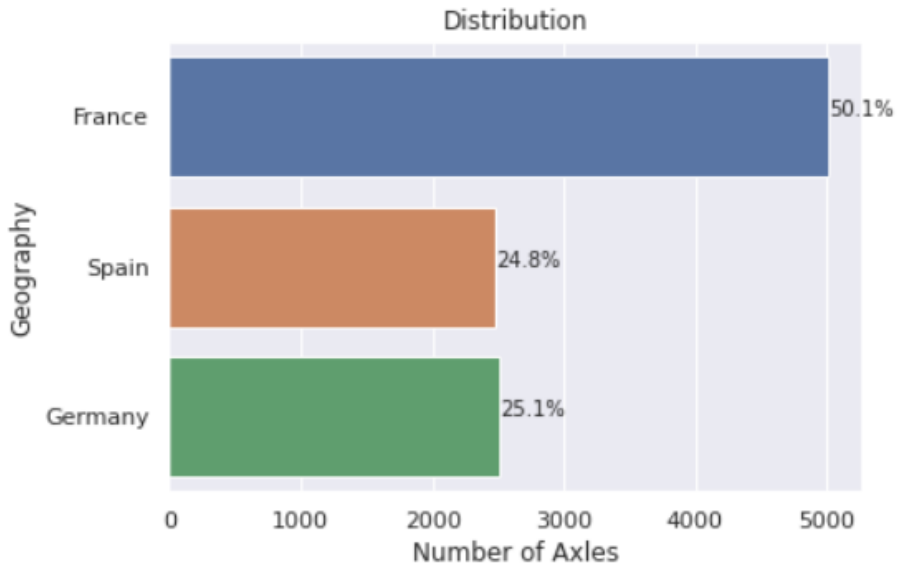


FIGURE 3.5 – Plot for 'Geography' feature

La moitié de notre population de clientèle étudiée est Française, la moitié restante est soit Espagnole soit Allemande

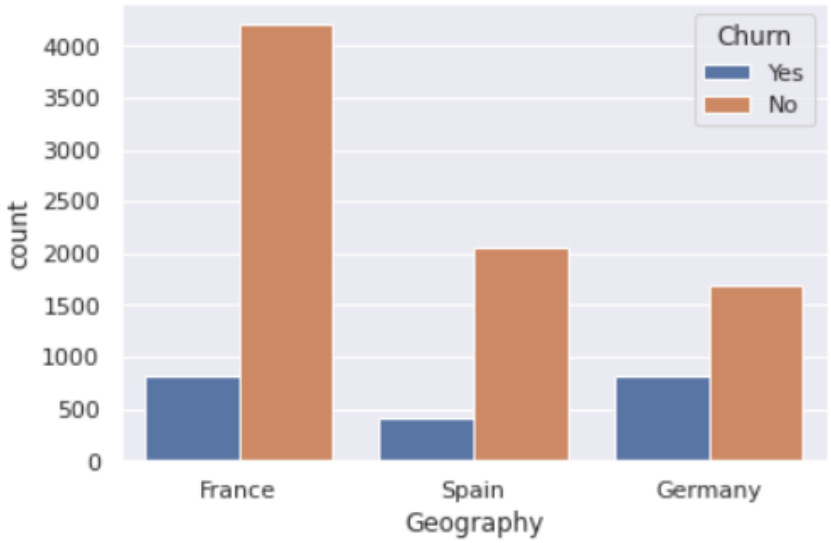


FIGURE 3.6 – Plot of Churn Distribution for 'Geography' feature

Les Allemands au plus tendance au churn, suivis de près par les Français

3.5 Corrélation

En probabilités et en statistique, la corrélation entre plusieurs variables aléatoires ou statistiques est une notion de liaison qui contredit leur indépendance.

3.5.1 Correlation Heat map

La Heat map est une technique de visualisation de données qui montre l'ampleur d'un phénomène sous forme de couleur en deux dimensions. La variation de couleur peut être due à la teinte ou à l'intensité, donnant au lecteur des indices visuels évidents sur la façon dont le phénomène est groupé ou varie dans l'espace.

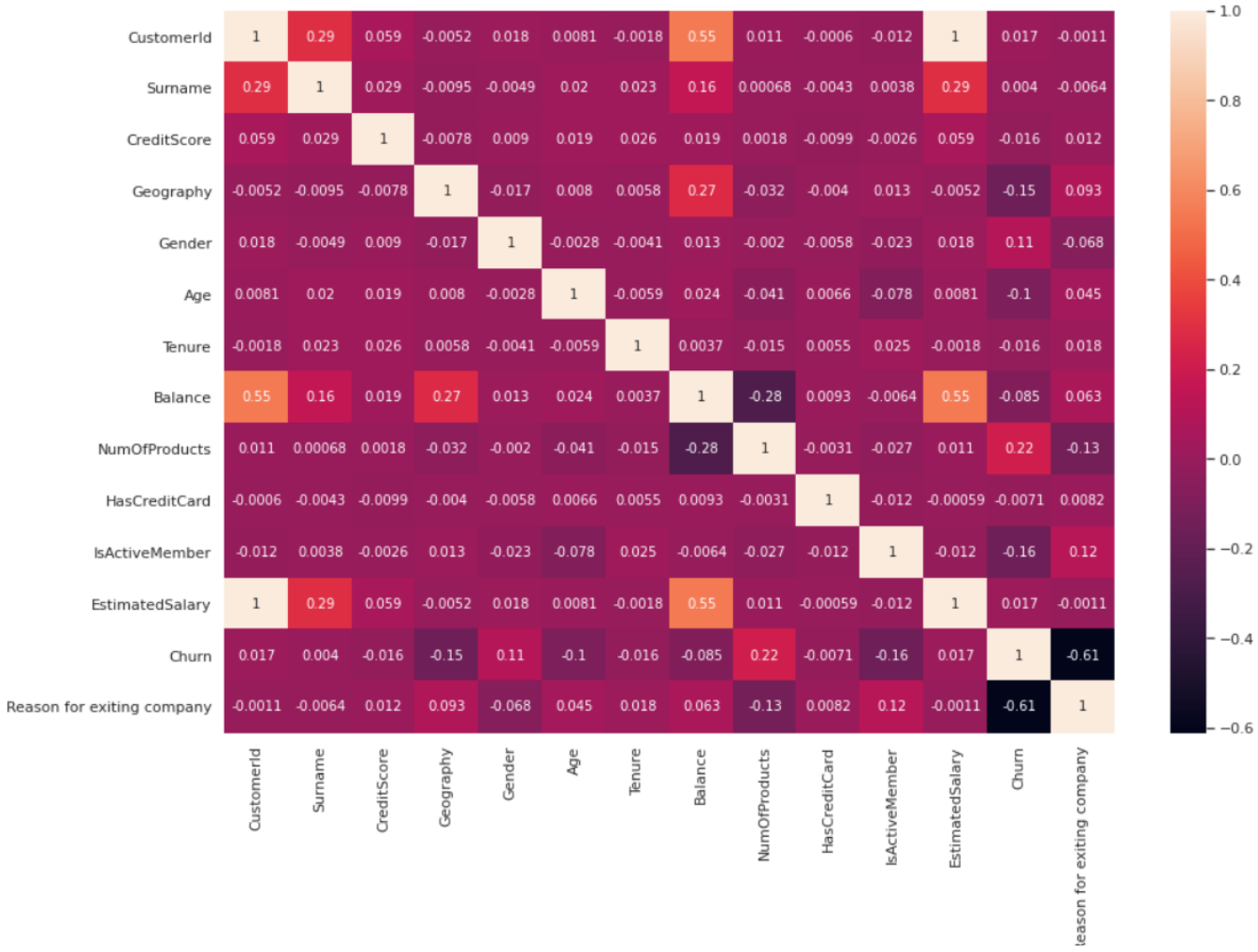


FIGURE 3.7 – Correlation Heat map

Interprétation :

Plus le coefficient est proche de 1, plus la relation linéaire positive entre les variables est forte.
Plus le coefficient est proche de -1, , plus la relation linéaire négative entre les variables est forte.
Plus le coefficient est proche de 0, plus la relation linéaire entre les variables est faible.

3.5.2 Corrélation avec le churn rate

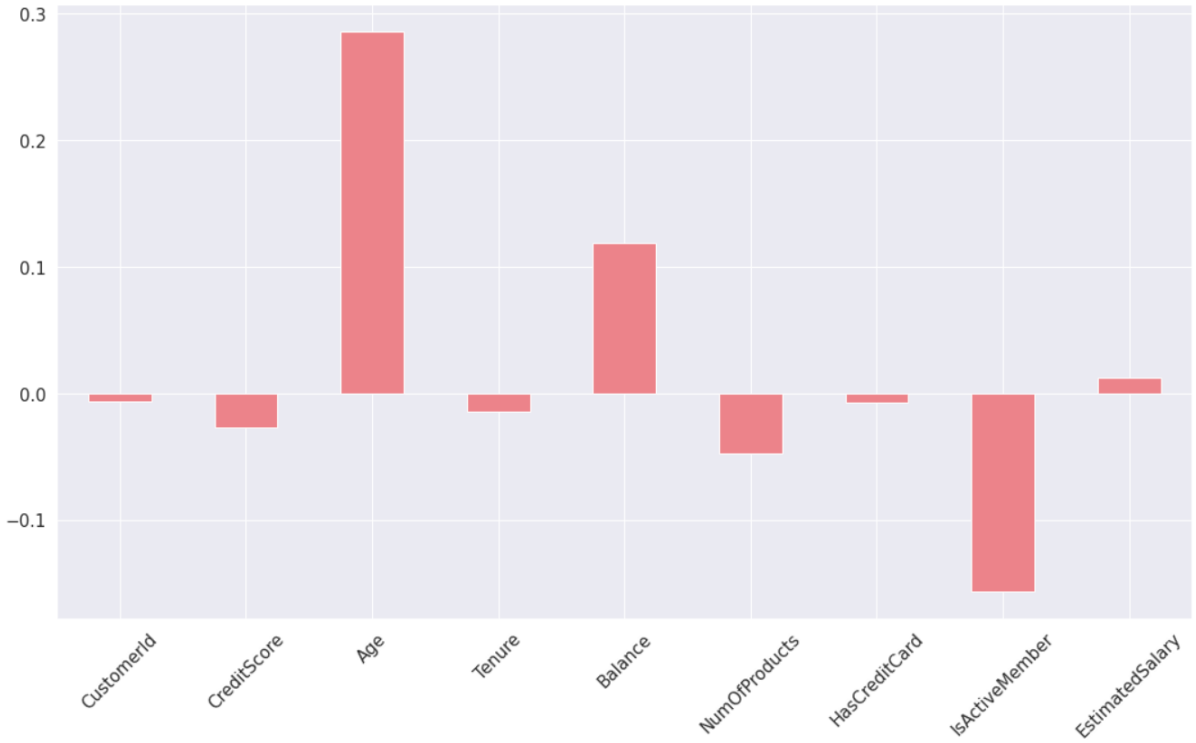


FIGURE 3.8 – Correlation with churn rate

— L'age et le solde bancaire du client sont fortement corrélés positivement avec le churn rate
— Son activité est fortement corrélée négativement avec le churn rate

Chapitre 4

Data preprocessing

4.1 Categorical encoding

4.1.1 Label encoding

Cette approche est très simple et consiste à convertir chaque valeur d'une colonne en un nombre.

On l'applique donc aux colonnes 'Geography' et 'Gender' :

	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCreditCard	IsActiveMember	EstimatedSalary	Churn
0	619	0	0	42	2	0.00	1	Yes	Yes	101348.88	Yes
1	608	2	0	41	1	83807.86	1	No	Yes	112542.58	No
2	502	0	0	42	8	159660.80	3	Yes	No	113931.57	Yes
3	699	0	0	39	1	0.00	2	No	No	93826.63	No
4	850	2	0	43	2	125510.82	1	Yes	Yes	79084.10	No

FIGURE 4.1 – Label encoding

Bien que le Label Encoding soit direct, il présente l'inconvénient que les valeurs numériques peuvent être mal interprétées par les algorithmes comme ayant une sorte de hiérarchie/ordre.

Ce problème d'ordre est traité dans une autre approche alternative courante appelée «One-Hot Encoding».

4.1.2 One-Hot Encoding

Dans cette stratégie, chaque valeur de catégorie est convertie en une nouvelle colonne et affectée d'une valeur 1 ou 0 (notation pour vrai/faux) à la colonne.

Appliquons celà à notre dataframe :

	Geography_France	Geography_Germany	Geography_Spain
0	1	0	0
1	0	0	1
2	1	0	0
3	1	0	0
4	0	0	1

	Gender_Female	Gender_Male
0	1	0
1	1	0
2	1	0
3	1	0
4	1	0

FIGURE 4.2 – One-Hot encoding

4.2 Oversampling Technique

SMOTE (Synthetic Minority Over-Sampling TEchnique) est une technique utilisée pour traiter des ensembles de données déséquilibrées.

Le principe de SMOTE est de générer de nouveaux échantillons en combinant les données de la classe minoritaire avec celles de leurs voisins proches.

4.3 Train Test Split

Pour la suite de notre travail, nous enlèvons les colonnes 'CustomerId', 'Surname' et 'Reason for exiting company' pour qu'elles n'impactent pas sur la construction du modèle prédictif étant donnée qu'elles ne sont pas du tout significatives pour ce dernier.

Nous divisons également les données en subsets d'entraînement (80 pour cent du dataset) et de test (20 pour cent du dataset).

Chapitre 5

Model

5.1 Model Selection

Nous procédons à la comparaison des algorithmes de classification de base : Logistic Regression - Gaussian NB - Random Forest - Kernel SVM - Decision Tree Classifier.

5.1.1 1ère itération

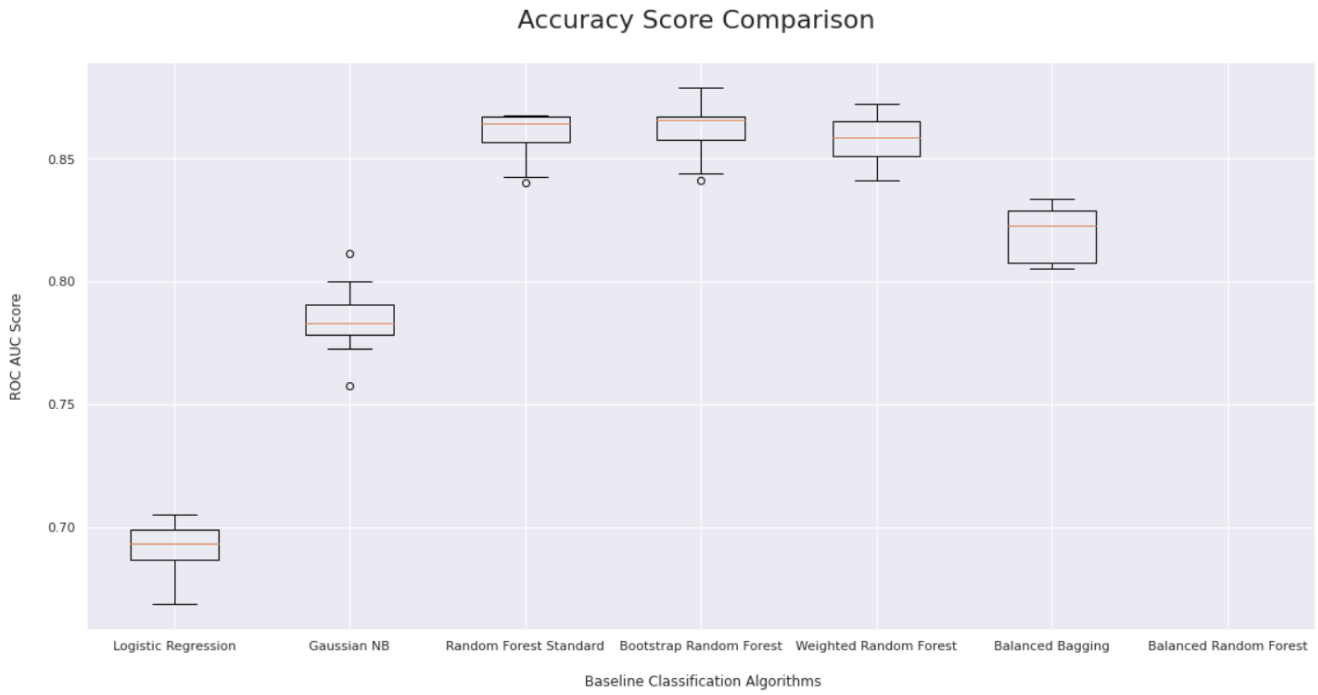


FIGURE 5.1 – Accuracy score comparison

	Algorithm	ROC AUC Mean	ROC AUC STD	Accuracy Mean	Accuracy STD
2	Random Forest Standard	85.43	2.07	85.95	0.98
4	Weighted Random Forest	85.43	2.05	85.80	0.95
3	Bootstrap Random Forest	85.34	1.68	86.15	1.11
5	Balanced Bagging	83.49	1.45	81.94	1.09
1	Gaussian NB	74.72	1.92	78.46	1.40
0	Logistic Regression	74.66	2.23	69.20	1.03
6	Balanced Random Forest	NaN	NaN	NaN	NaN

FIGURE 5.2 – Scores comparison (first iteration)

Dès la première itération des algorithmes de classification de base, nous pouvons voir que Random Forest a surpassé les autres modèles pour le dataset en question avec les scores AUC moyens les plus élevés.

5.1.2 2ème itération

	Model	Accuracy	Precision	Recall	F1 Score	F2 Score
2	Random Forest	0.8625	0.777311	0.454545	0.573643	0.495713
1	Gaussian NB	0.7835	0.361702	0.083538	0.135729	0.098722
0	Logistic Regression	0.7855	0.347222	0.061425	0.104384	0.073529

FIGURE 5.3 – Scores comparison (second iteration)

À partir de la 2ème itération, nous pouvons définitivement conclure que Random Forest est un modèle de choix optimal pour le dataset donné, car il présente relativement la combinaison la plus élevée de scores de precision, de recall et de F2 scores ; donnant le plus grand nombre de prédictions positives correctes tout en minimisant les "false negatives".

Par conséquent, essayons d'utiliser Random Forest et d'évaluer ses performances.

5.2 Identification du nombre optimal d'arbres pour le modèle Random Forest

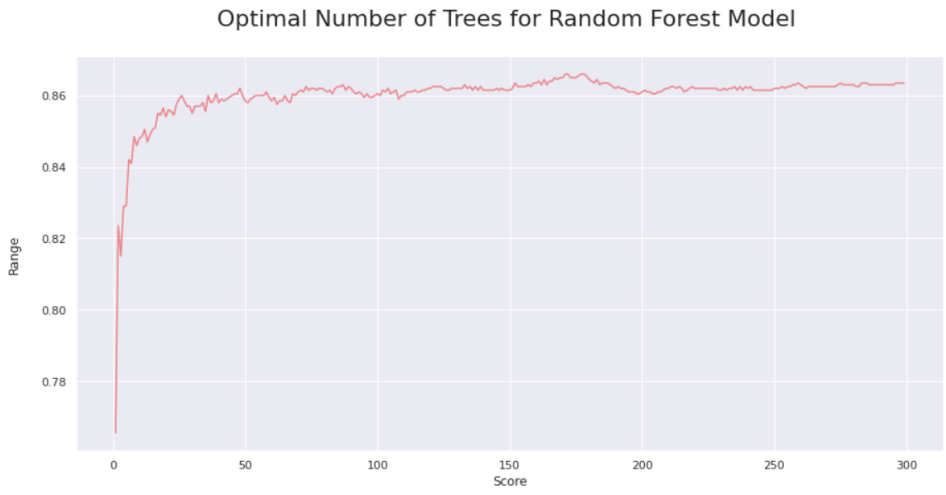


FIGURE 5.4 – Optimal number of trees for Radom Forest model

Notre étude nous a permis de déduire que le nombre optimal d'arbres est : 172

5.3 L'algorithme Random Forest

L'algorithme des « forêts aléatoires » (ou Random Forest parfois aussi traduit par forêt d'arbres décisionnels) est un algorithme de classification qui réduit la variance des prévisions d'un arbre de décision seul, améliorant ainsi leurs performances.

Dans sa formule la plus classique, il effectue un apprentissage en parallèle sur de multiples arbres de décision construits aléatoirement et entraînés sur des sous-ensembles de données différents.

Le nombre idéal d'arbres, qui peut aller jusqu'à plusieurs centaines voire plus, est un paramètre important : il est très variable et dépend du problème.

Concrètement, chaque arbre de la forêt aléatoire est entrainé sur un sous ensemble aléatoire de données selon le principe du bagging, avec un sous ensemble aléatoire de features (caractéristiques variables des données) selon le principe des « projections aléatoires ». Les prédictions sont ensuite moyennées lorsque les données sont quantitatives ou utilisés pour un vote pour des données qualitatives, dans le cas des arbres de classification.

L'algorithme des forêts aléatoires est connu pour être un des classifieurs les plus efficaces « out-of-the-box » (c'est-à-dire nécessitant peu de prétraitement des données).

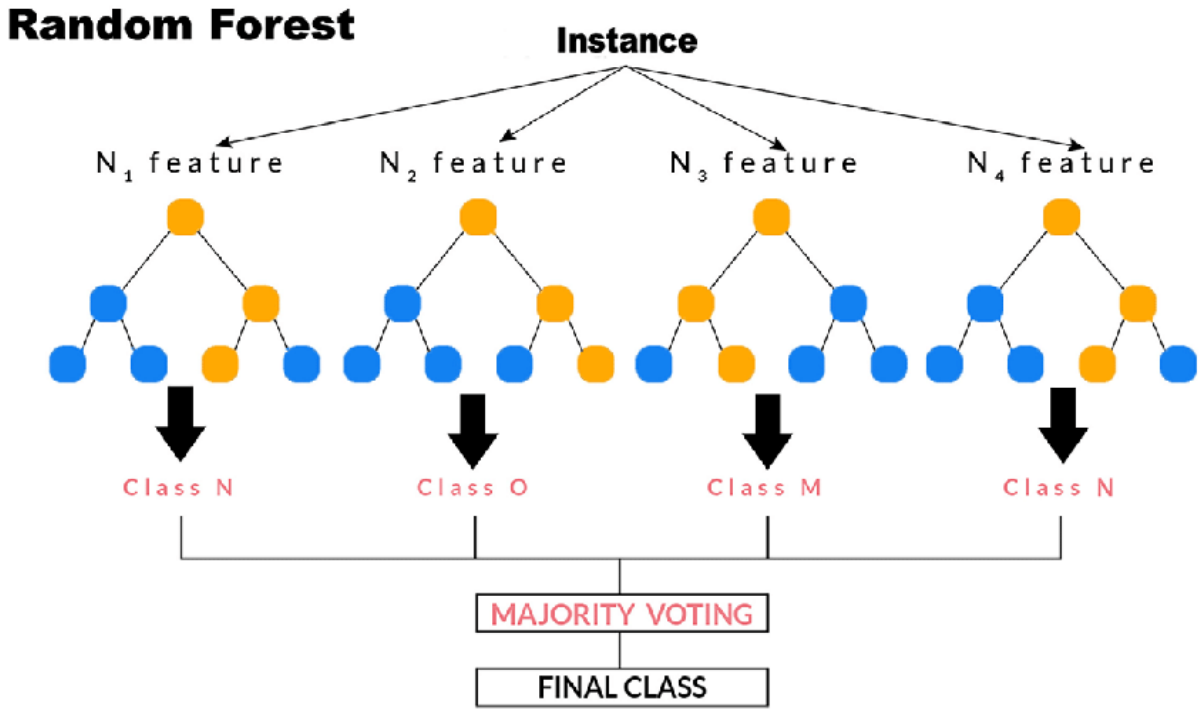


FIGURE 5.5 – Random Forest algorithm

5.4 Evaluation du modèle et résultat final

5.4.1 Evaluation

	precision	recall	f1-score	support
0	0.88	0.97	0.92	1593
1	0.78	0.46	0.58	407
accuracy			0.86	2000
macro avg	0.83	0.71	0.75	2000
weighted avg	0.86	0.86	0.85	2000

0.864
<Figure size 2016x1440 with 0 Axes>

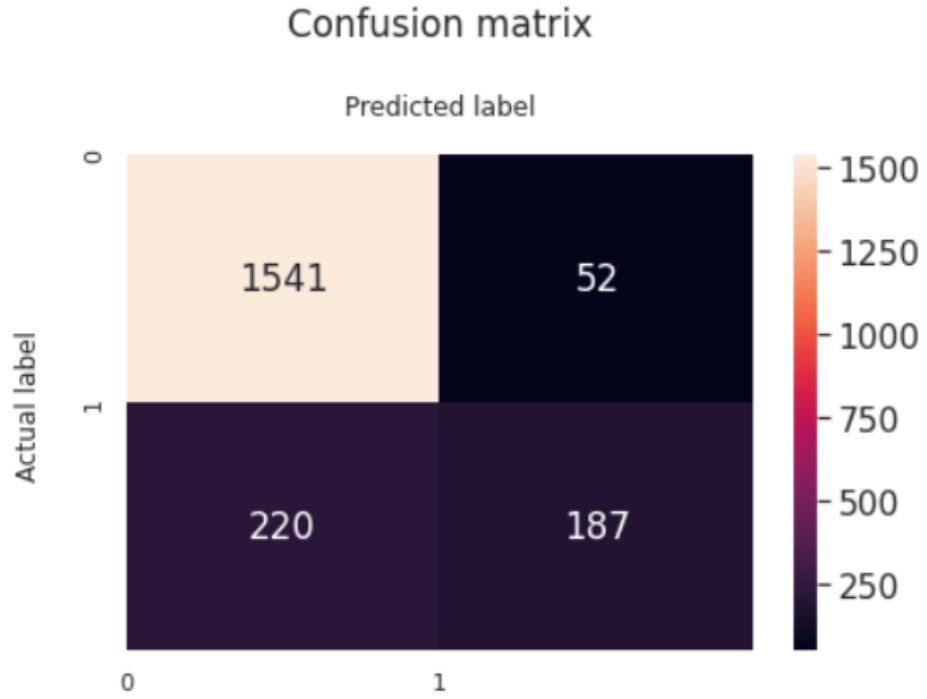


FIGURE 5.6 – Model Evaluation

Nous avons atteint une accuracy globale de près de **86 pour cent** avec une simple implémentation directe du modèle sans effectuer de feature engineering, feature selection, ou hyperparameter tuning.

Si nous appliquons toutes ces techniques, nous pourrions facilement obtenir une accuracy supérieure à 90 pour cent et améliorer davantage notre modèle.

5.4.2 Résultat final

	CustomerId	Churn	predictions	propensity_to_churn(%)	Ranking
1	15647311	0.0	0	7.55	7
11	15737173	0.0	0	13.24	5
27	15700772	0.0	0	2.44	9
31	15706552	0.0	1	84.67	1
32	15750181	0.0	0	40.49	2
...	...	...	...	...	...
9977	15579969	0.0	0	6.70	7
9981	15672754	1.0	0	0.70	10
9983	15656710	0.0	0	10.80	6
9995	15606229	0.0	1	91.99	1
9998	15682355	1.0	0	4.36	8

[2000 rows x 5 columns]

FIGURE 5.7 – Final Results

5.5 Balanced Random Tree

Etant donné que notre dataset n'est pas équilibré, comme nous avous pu le voir dans le chapitre 3, il serait plus optimal d'opter pour le modèle **Balanced Random Tree**.

Chapitre 6

Outils et déploiment

6.1 Outils et technologies

Pour la mise en exécution du travail exposé dans les chapitres précédents, plusieurs outils ont étés utilisés, dont notamment :

6.1.1 Google Colab

FIGURE 6.1 – Logo de Google Colab

Colaboratory, souvent raccourci en "Colab", vous permet d'écrire et d'exécuter du code Python dans votre navigateur. Il offre les avantages suivants :

— Aucune configuration requise
— Accès gratuit aux GPU
— Partage facile

Que vous soyez étudiant, data scientist ou chercheur en IA, Colab peut vous simplifier la tâche.

6.1.2 Python

FIGURE 6.2 – Logo de Python

Python est un langage de programmation interprété, multi-paradigme et multiplateformes. Il favorise la programmation impérative structurée, fonctionnelle et orientée objet.

6.1.3 Librairies Python

FIGURE 6.3 – Logo des librairies Python

6.1.4 Flask

FIGURE 6.4 – Logo de Flask

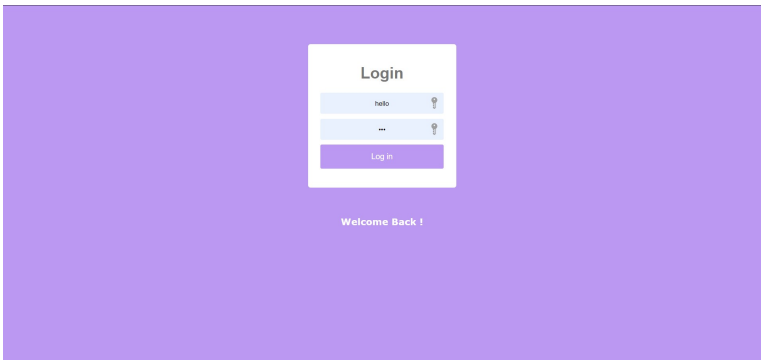
Flask est un micro framework open-source de développement web en Python. Il est classé comme micro-framework car il est très léger. Flask a pour objectif de garder un noyau simple mais extensible. Il n'intègre pas de système d'authentification, pas de couche d'abstraction de base de données, ni d'outil de validation de formulaires. Cependant, de nombreuses extensions permettent d'ajouter facilement des fonctionnalités.

6.1.5 LaTeX

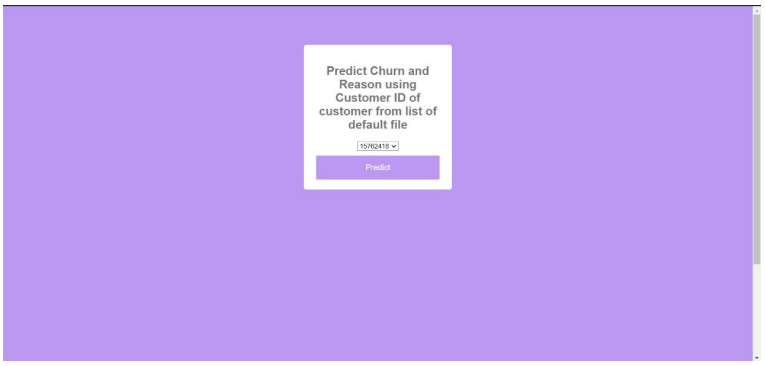
FIGURE 6.5 – Logo de LaTeX

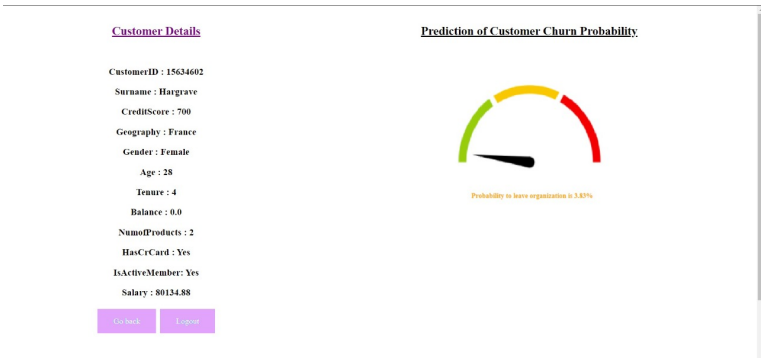
LaTeX est un langage et un système de composition de documents. Il permet de rédiger des documents dont la mise en page est réalisée automatiquement en se conformant du mieux possible à des normes typographiques.

6.2


Login
hello
...
Log in
Welcome Back !


Use Default Test Data File To Predict
Predict
Logout


Predict Churn and Reason using Customer ID of customer from list of default file
15762418
Predict


Customer Details
CustomerID : 15634602
Surname : Hargrave
CreditScore : 700
Geography : France
Gender : Female
Age : 28
Tenure : 4
Balance : 0.0
NumofProducts : 2
HasCrCard : Yes
IsActiveMember: Yes
Salary : 80134.88
Go back
Logout
Prediction of Customer Churn Probability
Probability to leave organization is 3.83%

Conclusion

En guise de conclusion, notre projet de fin de deuxième année consistait à la réalisation d'un modèle prédictif du chun en mettant en oeuvre le Machine Learning.

Pour réaliser ce projet, nous avons commencé par la compréhension de la problématique métier. La phase suivante était celle de l'import, le nettoyage, la préparation, ainsi que l'analyse des données , suivi de la phase de construction et d'évaluation du modèle, pour finalement arriver à le déployer au travers d'une application web.

Or, notre réalisation est loin d'être parfaite, voire terminée. En effet, des améliorations pourraient éventuellement être envisageables par la suite. On pourrait notamment améliorer davantage les résultats de notre prédiction en tirant bénéfice des techniques de feature engineering, feature selection, et hyperparameter tuning, ainsi qu'en essayant de résoudre le problème que pose le dataset déséquilibré en optant pour un modèle tel que Balanced Random Forest.

En somme, ce projet nous aura permis d'affiner nos capacités, de renforcer nos compétences en matière d'apprentissage automatique et de nous familiariser avec ses différentes techniques, comme il nous aura permis d'améliorer notre aptitude à collaborer et à travailler en binôme, malgré les nombreuses difficultés auxquelles nous avons fait face.