Rapport de clôture du projet

Nom du projet : Machine learning techniques for credit risk

Numéro de l'équipe : 31

Partenaire du projet : ESILV and Politecnico di Milano

Propriétaire(s) du document	Rôle du projet (si défini)
Hugo Alquier	Référent
Antoine Capton	
Jean-Louis Delebecque	

Sommaire

1 Rapport de clôture de projet - Intro

Le rapport de clôture du projet contient des informations descriptives clés sur le projet. En tant que dernier document écrit sur le projet, il décrit la satisfaction du client à l'égard du projet. Il analyse le résultat du projet et le processus par lequel ce résultat a été produit. Son objectif est double : garantir que les activités de clôture sont menées correctement et faciliter le transfert d'expérience que le projet ne soit pas achevé ou qu'un suivi du projet soit envisagé à l'organisation du client.

2 Contexte

Le rapport de clôture du projet est complété par l'extraction des données déjà produites tout au long du projet : le dossier du projet, les documents de planification initiale, les rapports d'avancement.
La valeur ajoutée consiste en l'agrégation des données dans un seul document, l'analyse des différences globales, la synthèse de la satisfaction des clients.

3 Rappels descriptifs des projets

L'équipe projet est composée de 3 membres: Hugo ALQUIER, Antoine CAPTON et Jean-Louis DELEBECQUE, étudiant en 5ème année à l'ESILV dans les majeures IBO et Ingénierie Financière

Ce projet est en collaboration avec l'ESILV et Politenico di Milano

Pour ce projet nous devions prédire la potentialité de défaut d'un client (sa solvabilité) en utilisant des techniques de machine learning. Les attentes du partenaire étaient donc un travail d'analyse de données bancaires afin d'y appliquer différents modèles de machine learning.

4 Objectifs fixés par rapport aux résultats obtenus

4.1 Objectifs initiaux du projet

Le but initial de ce projet était de tester l'efficacité de nouvelles techniques pour vérifier la solvabilité des clients d'une banque. Pour ce faire nous devions appliquer plusieurs diverses techniques de machine learning sur un jeu de données imposé par notre client. Le rendu final devait être un modèle de machine learning fonctionnel ayant une précision intéressante et où les données traitaient bien du problème de risque de défaut de crédit. Les différentes étapes seront détaillées plus particulièrement dans la partie technique du rapport.

4.2 Résultats obtenus

D'un point de vue global, ce projet est une réussite tant au niveau technique qu'au niveau opérationnel.
En effet, il semble que les résultats obtenus grâce à l'implémentation de notre modèle soient cohérents et remplissent à la fois les attentes initiales de notre partenaire et répondent au problème explicité dans le sujet de base.
De plus, sur le plan temporel nous avons réalisé ce projet de A à Z en trois mois ce qui témoigne d'une efficacité de l'ensemble des membres. Ainsi, avec le recul nous pouvons retenir la synergie inhérente à l'organisation et la réalisation des tâches intermédiaires, comme point fort de notre projet, puisque nous avons toujours tenu les deadlines qui nous étaient imposées.
Cependant, comme décrit plus haut, dû à un délai relativement court, il n'est pas exclu que nous soyons passés à côté de certaines améliorations en omettant des tests différents sur notre base de données ou dans la sélection des données.

4.3 Liste des livrables

A la suite d'un imbroglio au démarrage de notre projet nous n'avons pu commencé tout de suite. Dans le but de combler ce retard, notre partenaire nous a aidé pour récupérer notre jeu de données initial puis nous a donné deux semaines pour l'étudier, le comprendre et l'analyser. Nous avons gardé ce rythme d'une entrevue toutes les deux semaines avec des objectifs intermédiaires à présenter et parfois juste une revue de l'avancement du projet.
Voici la chronologie de la mise en place de notre solution :

- Fin Octobre: Compréhension du sujet avant la rencontre avec notre PRM
- Début Novembre: Etat de l'art et recherche préliminaires
- Fin Novembre: Compréhension du jeu de données
- Mi-Décembre: Data engineering
- Janvier: implémentation de modèle

Nous avons été relativement libres dans notre gestion du temps vis à vis de notre PRM, aussi il a été agréablement surpris que nous lui présentions un premier modèle avant les vacances de Décembre. Cela nous a permis d'en discuter et de nous aiguiller pour le modelling final.

5 Revue technique

Pour ce projet nous devions donc prédire la potentialité de défaut d'un client (sa solvabilité) en utilisant des techniques de machine learning. Les attentes du partenaires étaient donc un travail d'analyse de données. Le rendu final devait être un modèle de machine learning fonctionnel ayant une bonne précision et où les données traitent du bien du problème de risque de crédits. Pour ce faire nous avons décomposé le problème en différentes étapes (méthodologie de Data Science) afin d'obtenir un modèle qui ne serait pas biaisé ou basé sur des informations erronées ou trop porteuses de sens .

I) Technical

Le travail a donc été décomposé en 8 grandes étapes en suivant la méthodologie Data Scientist que nous simplifions un peu pour le rapport.

1) La compréhension du problème :
Pour pouvoir sélectionner nos données nous avons d'abord posé clairement le problème: il nous fallait des données bancaires de clients ayant déjà été en retard de paiements et d'autres ou non, plus il y aurait de données supplémentaires (features) comme l'âge ou le revenu de ces clients plus le modèle pourrait être juste et pertinent.

2) La sélection des données :
Dû à des problèmes de lancement du projet, le partenaire nous a généreusement fourni un dataset (tableau de données) contenant toutes les informations qu'il nous fallait.
Ce dataset contenant 150 000 lignes (chaque ligne est un client) et 11 colonnes contenant des informations clés telles que: le revenu moyen, l'âge, l'historique de défaut (0 ou 1, ce sera notre colonne cible: variable que nous voudrions prédire), des informations sur le dépendance financière (famille, prêt), de combien de jours on-t-il déjà été en retard (30-59, 60-89 , plus de 90 jours) et combien de fois.
Il est à noter que ce dataset est présent sur Kaggle et à donc déjà fait l'objet de compétition il est est donc viable à 100%.

3) Data featuring ou nettoyage et préparation des données:
Nous avons donc effectué un travail classique:

- Repérer les valeurs manquantes et soit les supprimer soit les remplacer: nous les avons remplacer par la médiane des valeurs que nous avions.
- Trouver des valeurs aberrantes et les modifier
- Changer les types de variables si nécessaires: une date doit être une date et non un nombre.
- Utiliser la matrice de corrélation entre toutes les colonnes et notre colonne cible (défaut ou non) afin de voir les variables qui sont peu corrélées à notre cible et les retirer.

Pour ce faire nous avons utilisé des outils de tables des données (*pandas*), et de visualisation (*matplotlib* et *seaborn).*

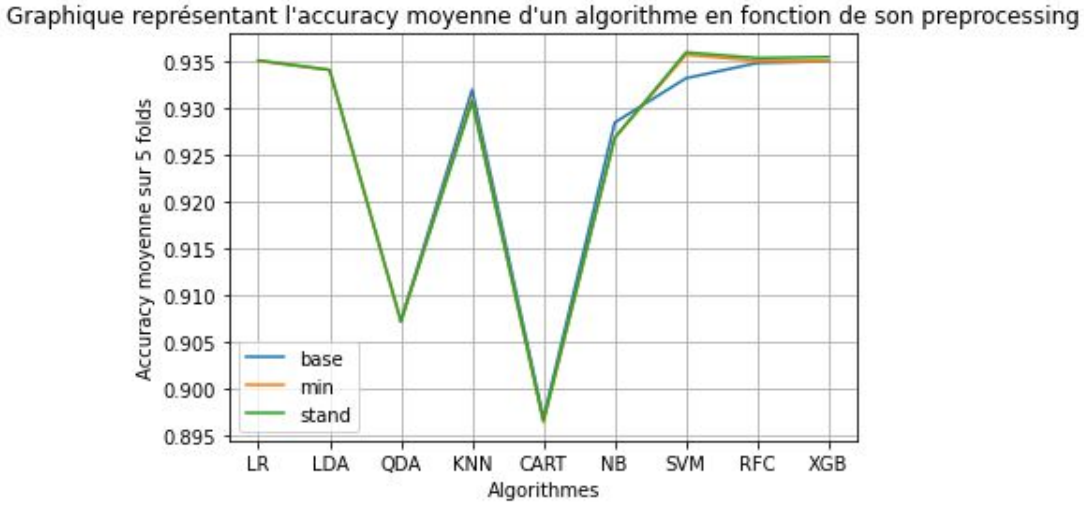
4) Modeling
Nous avions maintenant des données propres et prêtes à être utilisées.
C'est un problème de classification binaire, en effet la colonne cible a pour valeur soit 1 (défaut de paiements) ou 0.
Nous avons donc testé différents algorithmes de classification de la librairie *scikit-learn* très utilisés pour créer des modèles de machine learning.

Avant de modéliser nous avons séparé nos données en *train* et *test set*, pour que notre modèle s'entraîne avec 70 % des données (fit), puis on prédit des données avec le *test set* pour que les prédictions ne soient pas biaisées (le modèle ne doit jamais avoir vu les données de test).

Nous avons testé 9 modèles différents que que nous avons évalué à l'aide de la cross validation, cette technique nous permet de créer des modèles les tester 5 fois. En sortie nous avons donc 5 précisions nous prendrons leur moyenne comme précision.

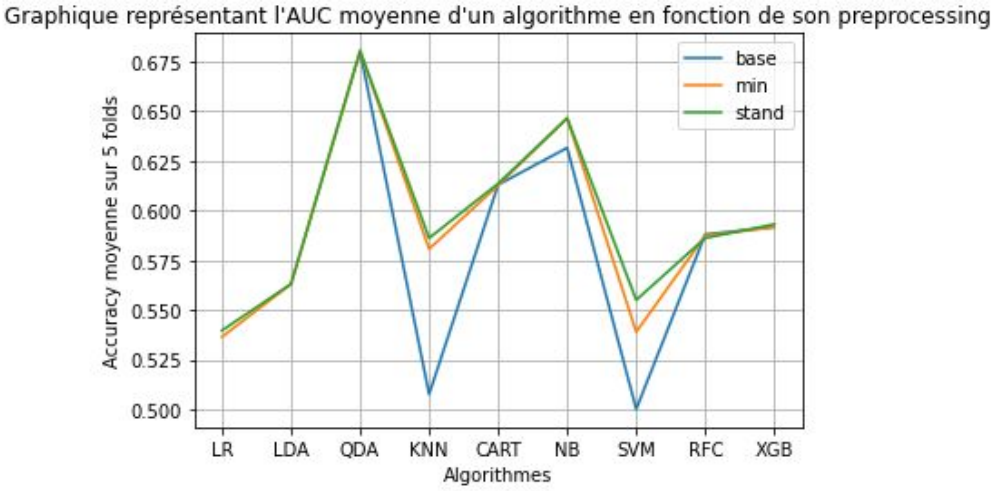
Il existe trois indicateurs de performance sur les modèles de classification : la précision, le recall et l'AUC : area under the curve. Nous avons utilisé la précision (pourcentage de bonnes prédictions) et l'AUC (aire sous la courbe du taux de bonnes prédictions sur le taux des mauvaises).
Comme toutes nos données étaient numériques nous avons essayé aussi deux scale (mise à l'échelle) différents : *standard* et *min*, cela peut parfois aider à avoir des meilleures performances.

Résulats:



On peut le voir ici toutes les modèles sont très performants : le SVM (support vector machines) a la meilleure précision

Regardons pour l'AUC:



Nous avons donc choisi le modèle QDA par rapport à son AUC plus élevé: 0.68 (meilleure AUC) et une bonne précision : 0.907 %.
Le SVM qui avait la meilleure précision a un des AUC les plus bas (0.56) c'est pour ça que nous ne continuons pas avec celui-ci.

Les modèles choisis (en bas des graphiques) sont les suivants:

- Régression Logistique
- LDA (Analyse de discriminant linéaire)
- QDA (Analyse de discriminant quadratique)
- KNN (K plus proches voisins)
- CART (Arbre de décision classifieur)
- NB (Naives Bayes)
- SVM (Supported Vector Machines)
- RFC (Random Forest Classifier)
- XGB (XGBoost variante du Random Forest)

Ce sont la plupart des modèles utilisés pour la classification binaire nous les avons donc tous testés

4) Optimisation du modèle

Pour optimiser les paramètres du modèle nous avons utilisé la méthode *GridSearch* qui permet de tester plusieurs fois un modèle avec différents paramètres à chaque fois et qui nous retourne les meilleurs paramètres.

Avant optimisation: 0.907 % de précision et 0.68 pour l'AUC

Après : 0.924 % de précision et 0.655 pour l'AUC, l'optimisation n'a donc pas été très concluante puisqu'on plus perdu en AUC que l'on a gagné en précision.

Nous avons terminé en essayant un KBest pour sélectionner le nombre de colonnes (features) optimales (voir dernière annexes).
Nous avons donc convenu de retirer 4 colonnes (les moins corrélées).
Avec comme résultat: 0.919 % de précision et 0.67 pour l'AUC ce sont nos meilleurs résultats.

5) Outil:

Pour faire du machine learning deux langages sont majoritairement utilisés python et R.
Nous avons choisi le python car il est plus répandu et permet donc l'accès à plus de bibliothèques et de fonctionnalités. Il permet aussi le travail en collaboration ce qui est beaucoup plus difficile sur R notamment grâce à l'outil google collabs qui nous permet de coder à plusieurs sur le même code.
Nous utilisons donc des notebooks pour coder ce qui est quasiment obligatoire pour du machine learning.

II) Organisation and système managérial

Répartition des tâches

L'équipe est composée de deux ingénieurs financiers et d'un Data Scientist la répartition des tâches c'est donc fait très logiquement par rapport aux différentes compétences de chacun.
Le Data scientist a fait les parties de modeling et optimisation du modèle tandis que les deux ingénieurs financiers se sont chargés de l'étude du problème et l'étude des données.
Les travaux annexes demandés par notre école ont été faits par les trois membres en collaboration.

Organisation et outils

L'équipe discutait via messenger et se partageait ses documents via Drive ou Microsoft Teams ou We Transfer pour le code. Nous échangieons plusieurs fois par semaine pour rendre compte des avancées et prochaines deadlines.
Nous avons établi avec notre partenaire que nous ferions une réunion toutes les deux semaines pour lui montrer nos avancements, ce que nous avons fait en visioconférence grâce à Microsoft Teams.
Nous avons utilisé google collabs pour être à plusieurs sur le même code (très pratique).

III) Conclusion

Le but de ce projet était de mener à bien un projet de machine learning et d'obtenir un modèle satisfaisant. C'est ce que nous avons fait en appliquant la méthodologie du Data Scientist, chaques étapes indispensables ont été respectées afin de bien poser le problème et de n'avoir aucun biais dans nos données. Puis nous avons pu tester la majorité des algorithmes de classification sur deux indicateurs de performances.
Le modèle a d'ailleurs été choisi sur les deux indicateurs et montre de bonnes performances. Nos choix techniques nous ont donc amenés à avoir de bons résultats finaux et un bon modèle, mais surtout à bien étudier les autres possibilités et s'assurer de la pertinence de notre démarche.

Les choix sur l'organisation se sont faits naturellement, nous nous connaissions mutuellement tous les trois et avons beaucoup l'habitude de travailler sur ce genre de projet.

6 Ressources prévues vs ressources utilisées

Comme évoqué précédemment, nous avons non seulement commencé le projet avec quelques semaines de retard mais nous étions aussi que trois membres contre quatre initialement prévu.
En effet, l'un des membres faisait partie de la liste des participants par erreur car étant en réalité en échange à l'étranger.
Finalement, avec le recul, ce sous-effectif s'est avéré être un avantage au regard du contexte. Il est vrai que la crise sanitaire d'une part et les obligations de chacun d'autre part, rendaient n'importe quel travail de groupe plus difficile à mener. N'être "que" trois personnes nous a certainement facilité l'organisation et nous a aussi évité d'éparpiller le travail entre de trop nombreuses personnes. En réduisant le nombre de personnes, on évite le risque d'un manque d'implication.
Enfin, il a été évidemment très compliqué voir impossible de se retrouver en physique, nous avons d'ailleurs uniquement travaillé en visioconférence avec notre PRM. Mais le sujet étant complètement basé sur de l'informatique nous ne pouvons pas considérer cela comme un frein.

7 Revue méthodologique

La méthodologie Agile se base sur ce principe simple : **planifier la totalité de notre projet dans les moindres détails avant de le développer est contre-productif**.

Dans notre cas cette phrase prend tout son sens, en effet nous étions en sous nombre comparés aux autres équipes projets, nous avons aussi commencé 1 mois en retard suite à des problèmes de communications avec les autorités compétentes de l'école. Après avoir réglé tous ces imprévus nous avons donc pris contact avec notre tuteur projet et avons commencé le travail demandé. Comme notre projet, divisé en plusieurs étapes, notre organisation s'est construite au fur et à mesure des deadlines. Nous avions un rendez-vous avec notre tuteur toutes les deux semaines, chaque rendez-vous correspondait à une deadline afin que notre tuteur voit concrètement les différents états d'avancement du projet.
Pour la répartition des rôles, cela s'est fait naturellement, nous nous connaissions avant le projet

et savions quelles compétences étaient les plus proéminentes chez chacun d'entre nous. Cela nous a permis d'agir avec efficacité tout au long du projet. Nous avons utilisé toutes les ressources numériques mises à disposition afin de gérer le projet tout en étant à distances les uns des autres à cause de la COVID 19. Les logiciels de partages de fichiers, de codes se sont révélés indispensables pour éviter la surcharge de travail d'un membre en particulier.

8 Risk Management

Nous avons connu des risques très tôt dans ce projet: nous n'étions que 3 et seulement une personne étudiait la Data Science, de plus suite à des problèmes de communication le projet a démarré un mois après le lancement des projets.

En première partie nous nous sommes occupés au plus vite de régler les problèmes de démarrage en se déplaçant à l'école et multipliant les emails.
Une fois en contact avec notre partenaire, nous avons directement reçu le dataset ce qui nous a permis de comprendre le problème en même temps que nos données, facilitant ainsi la préparation des données. Ainsi nous avons pu effectuer toutes les premières étapes d'un seul coup ce qui nous a permis de commencer à modéliser dans les temps (fin Décembre) et rattraper notre retard.
Pour ce qui est du fait que nous n'avions qu' un ingénieur Data Scientist cela n'a pas trop posé problème, il a eu des projets similaires au cours de l'année qui lui ont permis de savoir modéliser au cours de l'année. Les ingénieurs financiers ont aussi eu des cours de machine learning en python. Nous avons donc tous eu les outils nécessaires à ce projet dans nos cours de dernière année.
Nous n'avons pas connu de réel problème depuis le lancement du projet si ce n'est la puissance des ordinateurs qui a fait que certains calculs ont duré des heures et ont donc limité l'optimisation de certains modèles. Ce n'était pas un problème majeur puisque cela ne concernait pas notre meilleure modèle.

Au vue de nos résultats satisfaisants et des problèmes rencontrés, on peut dire que nous avons su gérer nos risques, en ce qui nous concernait il s'agissait de rattraper un retard et acquérir des compétences que nous n'avions pas en début d'année. Cela a pu se faire par l'optimisation de notre travail et l'apprentissage (grâce à des projets annexes).

9 Tâches post-projet

Selon le cahier des charges, le projet est fini puisque le modèle est prêt et performant.
Mais deux étapes pourraient être ajoutées à ce projet: le déploiement et la mise en place de retour (feedback).
Pour se faire nous pourrions mettre le modèle sur une API via Flask pour l'API et pickle pour exporter le modèle. Ce qui permettrait à des utilisateurs de transmettre leur données à l'API qui seront ensuite transmises à notre modèle afin qu'elles soient prédites. L'utilité ici est d'aider des banques à déterminer si des clients sont solvables (vont-t-ils rembourser leur prêt).
Le Feedback serait facile, il nous suffirait de collecter les données des banques après la mise en place du modèle et les comparer à nos prédictions. Puis s'en suivrait des multiples optimisations des features (rajouter des colonnes de données supplémentaires sur les clients) et des modèles afin de déterminer la solvabilité des clients.

D'autres travaux annexes comme travailler sur d'autres datasets ou données d'une banque sur lesquelles on pourrait améliorer notre modèle serait aussi pertinent, en effet notre modèle est conçu sur un dataset précis et attend donc des données du même format (dimension de la table).

10 Satisfaction du client et des utilisateurs

Notre partenaire est aussi notre PRM donc nous parlerons uniquement de sa satisfaction ici.
Suite à nos problèmes de démarrage notre partenaire à été compréhensif et coopératif en nous fournissant le dataset. Puis lors des premières réunions bi-mensuelles il était globalement satisfait de nos avancées sur les données (compréhension, nettoyage) et de notre compréhension du problème. A chaque réunion il nous donnait les étapes que nous devions faire pour la prochaine réunion. Nous avons toujours respecté ses deadlines.
Et c'est vers début décembre que nous avons testé nos premiers modèles et il était ravi de voir que nous en avions testé plusieurs et que nous étions proche de la fin du projet.
En conclusion nous dirions que notre partenaire est plutôt satisfait du projet, on a répondu aux attentes dans les temps et le livrable est prêt.

11 Recommandations pour l'achèvement du projet

Best practices:

- Être en communication constante. Le fil de conversation était très actif presque journalier, on a l'habitude de s'appeler lorsqu'on travaille aux mêmes horaires sur le projet.
- Tous partager en ligne (capacité d'écrire en même temps sur le document):
  les notes (collecter les informations et les centralisés) , le code, les différents rapports, les présentations.
- Répartir les tâches en fonction des majeurs de chacun
- Clarifier l'objectif final dès le début

Amélioration:
Nous avions une bonne organisation rythmée par les réunions avec le partenaire. Une amélioration possible serait d'établir un calendrier pour assurer un avancement plus linéaire mais nous n'étions que 3 qui devions travailler sur des parties indépendantes. Une réunion bi-mensuel pour tout mettre en commun était ce qui nous convenait le mieux individuellement.

Ce que l'on va changer

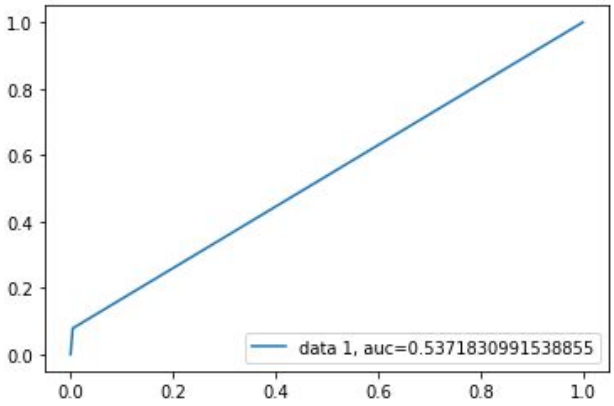
- Rencontrer physiquement le partenaire: un peu compliqué avec le covid mais c'est une bonne pratique dans un tel projet.

12 Annexes au rapport de clôture.

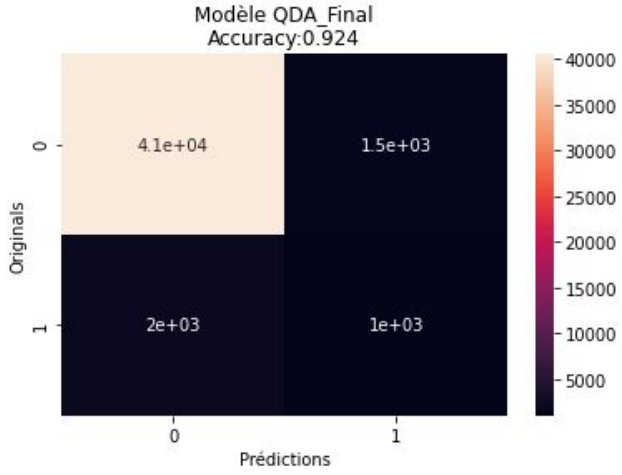
Lien github pour avoir tout le notebook: https://github.com/jdelebec/Pix-A5-2020

Data visualization:

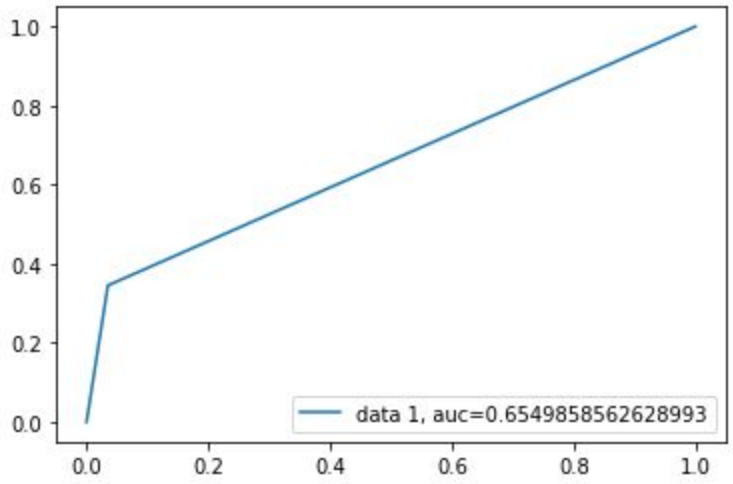
Courbe ROC de notre premier modèle: De Régression logistique, on peut y lire l'AUC



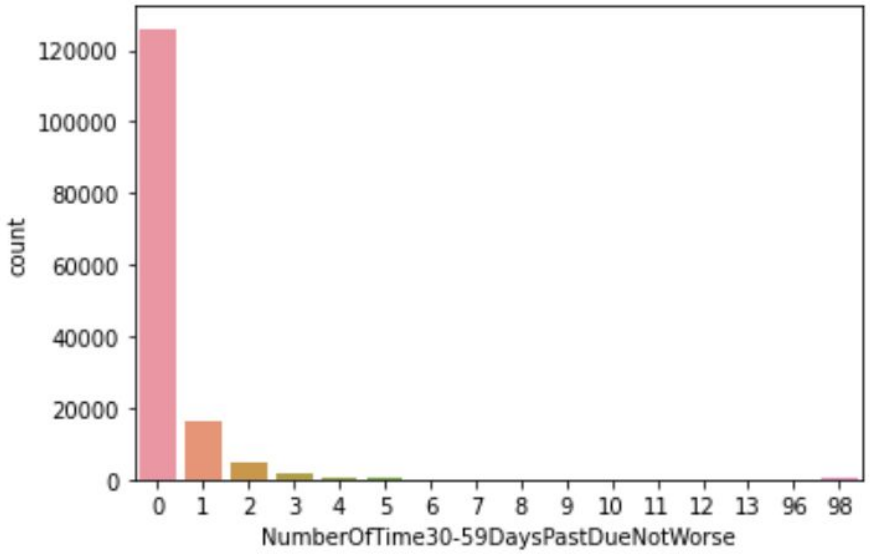
Confusion de matrice sur notre modèle final: QDA, après l'optimisation par GridSearch



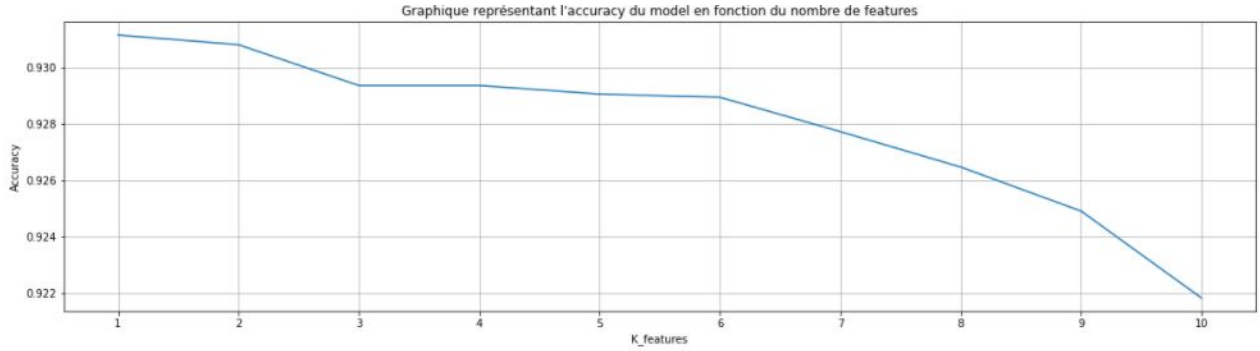
Courbe ROC de notre modèle final: QDA, après l'optimisation par GridSearch



Nettoyage des données: erreurs erronés des données avec les valeurs 96 et 98



KBest


Graphique représentant l'accuracy du model en fonction du nombre de features
Accuracy
0.930
0.928
0.926
0.924
0.922
1
2
3
4
5
6
7
8
9
10
K_features