PROJET DE CREDIT SCORING

Recherche du meilleur score & résultats

Présentation : Boubacar Fakoly DOUMBIA | UIE, Master Data Science

12/10/2021

Introduction

Résumé

Dans cette analyse, nous allons présenter les résultats obtenus sur les modèles après une analyse descriptive et exploratoire faites sur les données associées à des crédits à la consommation. Les résultats chiffrés avec l'exposition des critères de validations utilisés seront aussi présentés.

Nous allons passer en revue à la construction des modèles, de la validation du score, une étude des performances et tirer une conclusion à la fin.

.

Objectif du projet

C'est de rechercher le meilleur score, de justifier pourquoi il s'agit effectivement du meilleur et d'expliquer son intérêt et son utilisation pratique pour une banque, tout en sachant que le critère à modéliser est la variable 'incident' de paiement.

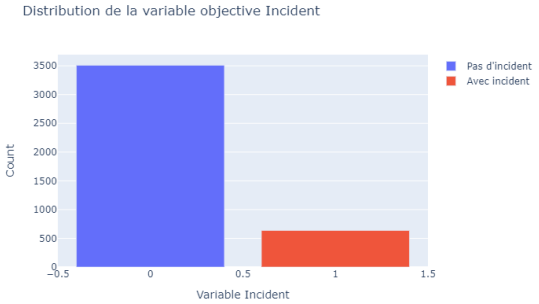
Connaissance et nettoyage des données

La base de données contient 4146 observations, avec 10 variables (revenu, depnaiss, datenaiss,duree,montcred, situfam, cb,numero, incident). Nous avons rencontré une ligne vide et cette dernière a été enlevée de la base, les types de données sont tous de type float à part la variable revenu, et par la suite cette variable a été convertie en float comme les autres. Il existe quelques variables à occurrence binaire telles que : situfam, cb et la variable cible incident. On a aussi enlevé la variable numero pour le reste de l'étude après vérification du nombre d'occurrence répétée qui est 1 pour chaque valeur de la variable.

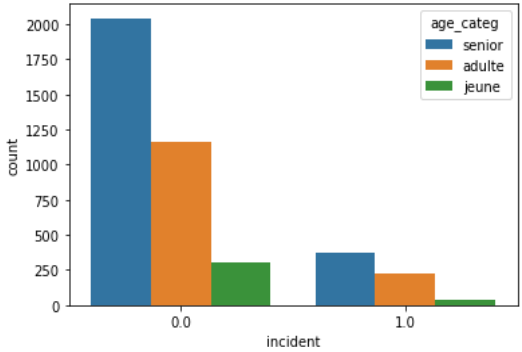
Les dates de naissance vont de 1912 à 1986, avec une moyenne aux alentours de 1949. La durée du crédit va de 6 à 18 mois et 12 mois en moyenne, les montants vont de 400 à 19438 et 2715 en moyenne, la variable ancienn(ancienneté) va de 1 an à 73 ans en moyenne 21.

Analyse exploratoire et descriptive

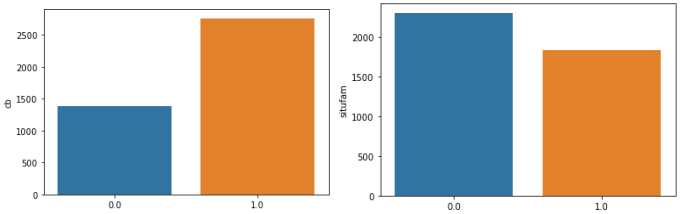
Les observations sont reparties entre les consommateurs avec incidents qui représentent 636 cas et sans incidents avec 3509 cas.



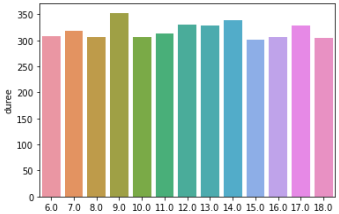
Quand on fait la réparation des consommateurs avec incident et sans incident, on constate que les consommateurs de la tranche d'âge de plus de 65 ans (senior) sont majoritaires, suivis de ceux de la tranche des consommateurs entre 41 et 65 ans (adulte), ensuite les jeunes de 0 à 40 ans au plus.



En prenant les variables avec de plus petit nombre d'occurrence, nous savons que les consommateurs en possession d'une carte bancaire sont les plus nombreux que ceux-là qui n'en ont pas. Et les célibataires sont aussi nombreux par rapport à ceux en couple. La durée du crédit reste peu variante et moins signifiante par sa faible variable, cependant il y'a plus de consommateurs qui ont une durée de crédit de 9 mois.

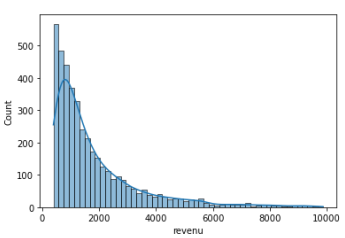


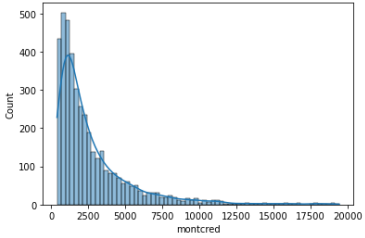
A gauche les cartes bancaires et à droite les statuts matrimoniales

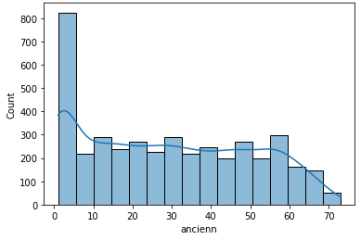


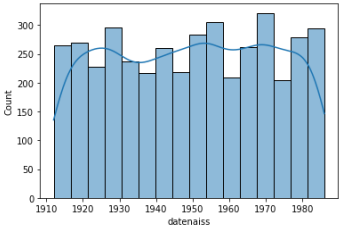
La durée du crédit

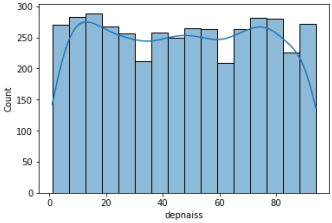
L’analyse graphique faites aux autres variables sur leur population nous donnent d'informations intéressantes sur les variables revenu et montcred avec de variations significatives indiquées par une forte concentration de population de consommateurs à faible montant et moins de population quand le montant augmente. En revanche les autres variables témoignent de peu d'information signifiant à part la variable ancienn qui montre une forte concentration de consommateurs entre 0 à 5 ans, et varie peu pour le reste.





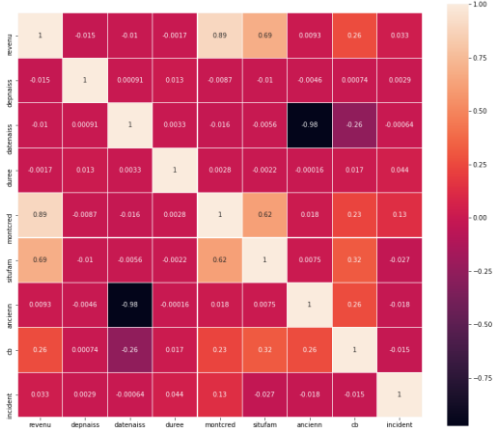






La distribution faite à la variable montcred en général, nous montre une forte population avec moins de montant du crédit sans incident et une faible population pour les montants élevés, nous observons la même chose pour les montants de crédit avec incident, pareil pour la distribution générale. Le même phénomène présent avec montcred peut s'expliquer avec la variable revenu.

Nous pouvons vérifier certaines hypothèses faites dans les parties précédentes en s'appuyant sur le taux de corrélation entre les variables.



Le heatmap présentant les résultats de la corrélation, nous montre bien une plus grande corrélation des variables, revenu (0.033), duree(0.044), montcred(0.13) par rapport à la variable cible incident. Une grande corrélation est aussi observée entre le revenu et le montant du crédit (0.89), entre le revenu et la possession de la carte du crédit (0.26), et entre le revenu et la situation matrimoniale (0.69). Autrement dit les autres valeurs en couleur rouge clair, orange et jaunâtre sont des valeurs intéressantes.

Par suite on enlève les variables depnaiss (departement de naissance) et la date de naissance datenaiss car elles ont très peu d'impact avec les autres variables explicatives qu'avec la variable cible.

Construction du modèle

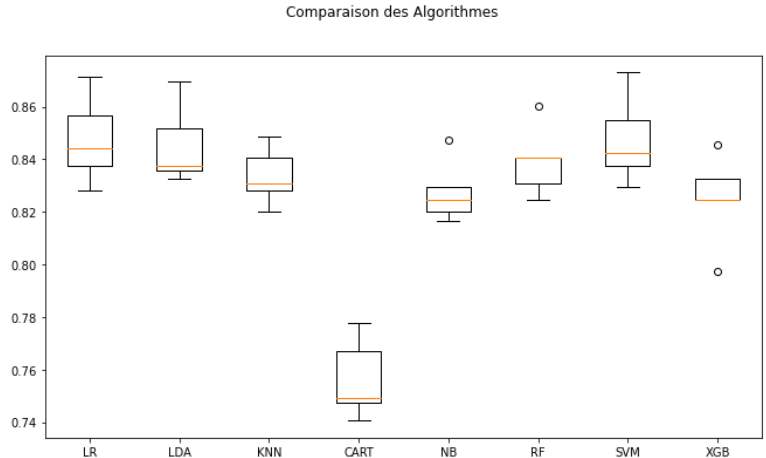
Présélection des modèles

Les données ont été divisées en données d'entrainement 75% et en données de test 25%. Et le seed pour la génération est fixé à 11.

Nous avons établi un algorithme de sélection automatique de bonne performance avec une liste de modèle présélectionnés. C'est la première phase, la mesure des scores choisie est la métrique $Acurracy = True\ Positive\ /\ (True\ Positive + True\ Negative) * 100$. Cette mesure sera utilisée pour le reste de la modélisation.

Ci-dessous le tableau de résultats des performances obtenus.

Nom du modèle	Score
Régression Logistique (LR)	**0.847490**
Plus proche voisin (KNN)	0.833656
Forêt Aléatoire (RF)	0.836548
SVM (Support Vector Machines)	**0.847490**
Classificateur d'arbre de décision (CART)	0.762229
Classificateur XGB	0.824970
Analyse discriminante linéaire (LDA)	0.845559
Naive Baye Gaussian (NB)	0.827539


Comparaison des Algorithmes
0.86
0.84
0.82
0.80
0.78
0.76
0.74
LR
LDA
KNN
CART
NB
RF
SVM
XGB

Les résultats ont révélé que la régression logistique et le SVM ont les meilleurs scores. Nous effectuerons une cross validation sur ces deux modèles en recherchant les meilleurs paramètres pour améliorer la performance du modèle et à la fin, mesurer de nouveau les performances et les nouveaux scores obtenus.

Validation du score et étude des performances

Les résultats obtenus à la suite de la modélisation faite sur les données avec les deux modèles choisis, nous révèlent une égalité de performance avec les deux modèles, la régression logistique et le SVM avec la métrique Accuracy qui nous donne une valeur de **0.8428158148505304** pour les deux.

La **matrice de confusion** pour la régression logistique donne 873 bonnes prédictions sans incident (TP) et 1 bonne avec incident (TN), contre 161 mauvaises prédictions avec incident (FP) et 2 mauvaises prédictions sans incident (FN), au total 874 bonnes prédictions et 163 mauvaises prédictions.

Le SVM donne 871 bonnes prédictions sans incident, 3 bonnes avec incident, contre 159 mauvaises prédictions avec incident et 4 mauvaises sans incident et au total 874 bonnes prédictions et 163 mauvaises prédictions.

La $\boldsymbol{Precision} = \boldsymbol{TP}/(\boldsymbol{TP} + \boldsymbol{FP})$, nous obtenons avec SVM un score de 0.85 pour la classe des sans incident et SVM, et 0.43 avec la classe avec incident. Avec la régression logistique, nous avons 0.84 pour les classes sans incident, 0.33 avec la classe avec incident.

Avec $\boldsymbol{Recall} = \boldsymbol{TP}/(\boldsymbol{TP} + \boldsymbol{FN})$, nous avons 1 et 0.02 successivement les valeurs des classes sans incident et avec incident pour SVM, et 1 et 0.01 successivement les valeurs des classes sans incident et avec incident pour la régression logistique.

Le score $\boldsymbol{F1} = \boldsymbol{2} * (\boldsymbol{Recall} * \boldsymbol{Precision})/(\boldsymbol{Recall} + \boldsymbol{Precision})$, nous donne 0.91 et 0.04 successivement les valeurs des classes sans incident et avec incident pour SVM, et 0.91 et 0.01 successivement les valeurs des classes sans incident et avec incident pour la régression logistique.

Conclusion

Avec les scores obtenus, nous réalisons de performances supérieures avec le SVM, qui avec toute la métrique appliquée dépasse la régression logistique. Même s'il faut aussi constater que ces deux modèles n'ont pas un grand écart de score.

A noter que lors des entrainements avec les variables *datenaiss* et *depnaiss*, la régression logistique a obtenu un meilleur *accuracy* que le SVM, c'est après suppression de ces colonnes dont les causes ont été expliquées tout en haut du document, que SVM a atteint cette performance et dépasse la régression logistique.

Nous obtenons enfin notre meilleur score avec SVM.

Il est connu que le SVM est facile d'emploi, il permet de traiter des données avec beaucoup de variables (donc une très grande dimension). Il présente aussi de bon comportement en prédiction. Nous avons donc remarqué avec les résultats obtenus lors des entrainements. La mise en œuvre d'un tel algorithme peu couteux en temps est très efficace pour une banque.


UIE