Apprentissage Supervisé

Rapport de Projet

Auteurs :

- FHIYIL Soufiane
- MOUHDA Mohammed Reda

Année Universitaire : 2018/2019

Table des Matières

1- Introduction

L'apprentissage supervisé, dans le contexte de l'intelligence artificielle (IA) et de l'apprentissage automatique, est un système qui fournit à la fois les données en entrée et les données attendues en sortie. Les données en entrée et en sortie sont étiquetées en vue de leur classification, afin d'établir une base d'apprentissage pour le traitement ultérieur des données.

L'apprentissage est dit supervisé lorsque les données qui entrent dans le processus sont déjà catégorisées et que les algorithmes doivent s'en servir pour prédire un résultat en vue de pouvoir le faire plus tard lorsque les données ne seront plus catégorisées.

2- Données utilisés

Nous utilisons une dataset appelé « **Credit_card_Fraud** », ce jeu de données est constitué de transactions qui se sont produites en deux jours.

Nombre d'observations	Nombre de variables	Nombre de classes
284 807	31	2

3- Analyse exploratoire

On constate d'après la figure suivante, qu'il y'a un chevauchement entre les deux classes, on n'arrive pas à visualiser les partitions même si on a utilisé les deux premiers composantes principales qui contiennent la majorité de l'information.

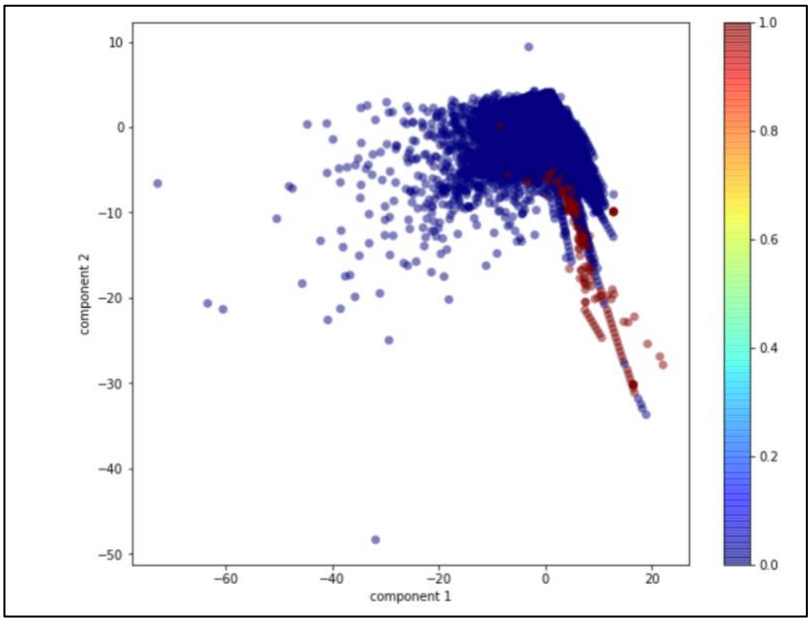

Figure 1 : Données visualisés sur les 2 composantes principales

D'après, le boxplot, on constate qu'il y'a beaucoup de valeurs extrêmes, les variables ont un peu près les mêmes médianes, et des valeurs (min, max) différents.

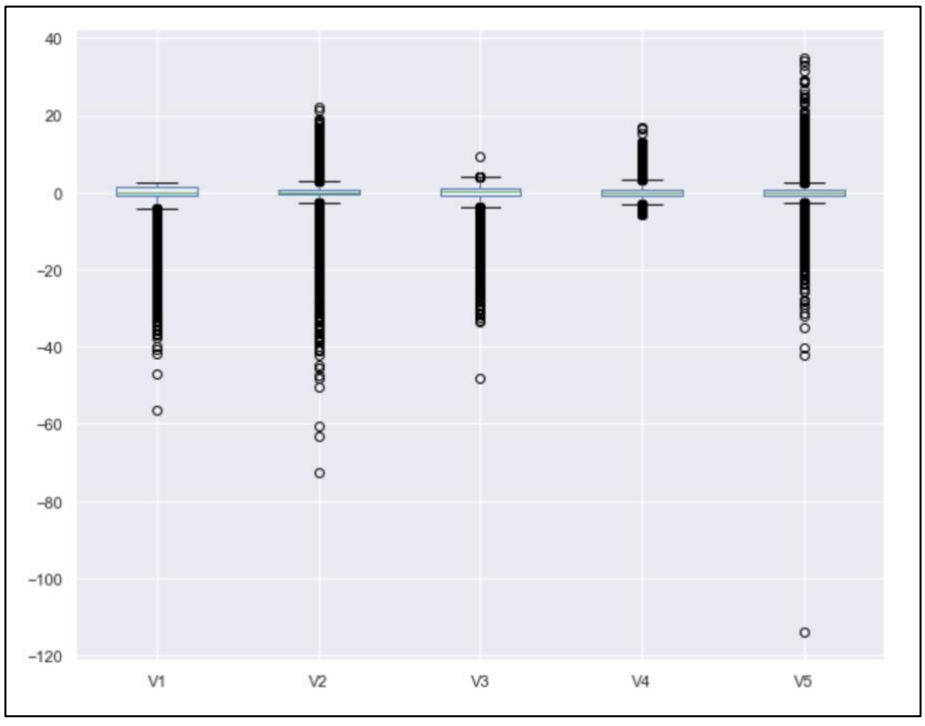

Figure 2 : Boxplot des premiers variables

4- Traitement des données non équilibrées

Le problème de déséquilibre appelé en anglais (Problem of Imbalanced Data) se produit lorsque l'une des classes ayant un échantillon plus que les autres classes, une grande différence de nombre ou de pourcentages de donné par rapport à tout l'ensemble .

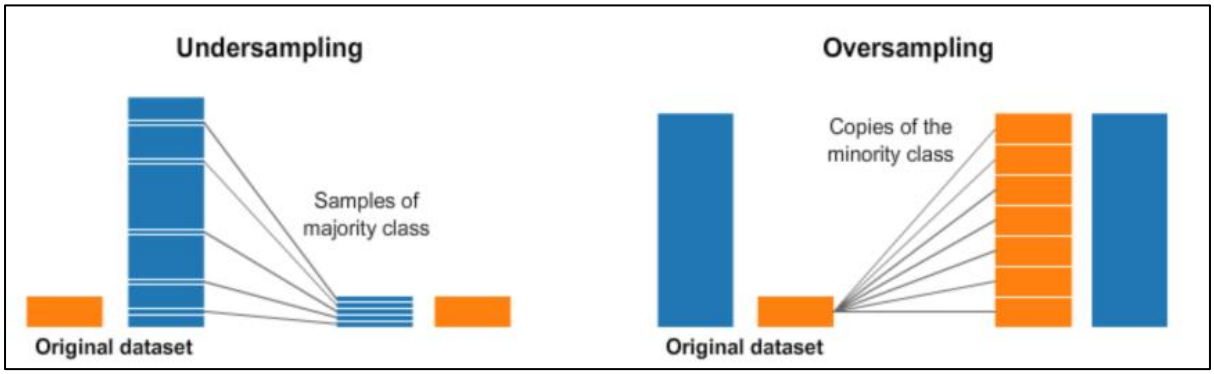

Figure 3 : les techniques de undersampling et oversampling

Puisque les données de la classe 0 sont assez larges que les données de la classe 1, on a essayé de faire un over-sampling et under-sampling à la fois.

- **Under-sampling :** Il consiste à retirer des échantillons de la classe majoritaire.
- **Over-sampling :** ajouter plus d'exemples de la classe minoritaire.

Pour éviter cette perte d'informations ou le sur apprentissage, nous allons utiliser les deux méthodes et les combiner afin d'équilibrer nos données.

5- Implémentation des algorithmes

Dans cette section, nous passons à l'implémentation des algorithmes tels que (Bayésien Naïf, KNN, LDA, QDA, Linear SVM, …), afin de comparer les performances de chaque algorithme en utilisant la métrique AUC.

Algorithmes/ métriques	**Bayesien Naïf**	**LDA**	**QDA**	**Linear SVM**	**KNN**	**Decision Trees**	**Random Forests**	**Logistic Regression**
AUC	0.91	0.9026	0.9187	0.9045	0.9045	0.8898	0.9011	0.9463
Précision classe 0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Précision classe 1	0.06	0.11	0.05	0.16	0.16	0.34	0.80	0.06
Validation croisée (moyenne)	0.9510	0.978	0.9825	0.99	0.99	0.998	0.999	0.989

D'après, le tableau suivant on constate que les algorithmes (Decision Trees, Random Forests) ont donné de meilleurs résultats par rapport aux autres algorithmes (Bayes, LDA, QDA, KNN, Linear SVM, Logistic regression), car la précision de la prédiction pour la classe 1 est élevé pour les arbres de décisions.

un arbre de décision grandit de manière itérative, donnant plus d'importance au nombre d'unités observées dans un nœud, tandis qu'une méthode de régression tente de faire correspondre toutes les observations à une ligne de distribution théorique. En conséquence, un arbre combine les probabilités de classification de chaque nœud obtenues par un ensemble de règles "hétérogènes", tandis qu'une régression se préoccupe davantage d'une formule "homogène" qui décrit plutôt l'ampleur de l'importance d'un facteur pour ajuster cette distribution.

Les arbres de décisions donnent de meilleurs résultats car il se basent sur la décision en fonction des nœuds.

6- Conclusion

Dans le cadre de ce projet, des différentes techniques ont été appliqués pour la classification des données. Nous avons pu constater que les résultats obtenus en appliquant une méthode à un autre diffère et cela dépend du fonctionnement de la méthode.

classifier des données non équilibrées n'est pas évident. Les taux de précisions obtenus sont souvent non significatifs. Cependant, dans ce cas, l'AUC peut nous aider à comparer mieux entre les modèles obtenus.

Pour conclure, nous proposons d'étudier différents méthodes pour équilibrer les données.

Et comme perspectives, les réseaux de neurones peuvent aussi être utilisés pour la détection de fraude.