Introduction :

L’objectif de ce document est de décrire l’approche adoptée et les résultats obtenus dans la recherche d’une méthode de prévision du résultat de matchs de boxe en fonction de certains critères des boxeurs.

Section 1 : Ressources utilisées :

Le jeu de données utilisé est disponible à l’URL suivant: <https://www.kaggle.com/slonsky/boxing-bouts>

Extrait :



Section 2 : Préparation des données

Le jeu de donnée fait l’objet de plusieurs modifications avant d’être interprété.

1. Certaines variables sont directement éliminées :

* Les détails des notes des juges 6 dernières colonnes) sont éliminées car redondantes avec l’information de boxeur gagnant
* Le détail du type de victoire (KO, décision,…) est éliminé car il fait partie du résultat et ne pourra être utilisé comme paramètre d’entrée lors de nos prédictions
* La notion de droitier/gaucher (Stance Orthodox / Southpaw) est éliminée car elle semble erronnée dans le dataset, en effet tous les enregistrements contiennent une valuer identique pour les boxeurs A et B

1. Le format du fichier est ajusté. Plutôt que de contenir séparément les données des boxeurs A et B pour chaque variable, nous allons fusionner l’information en 1 colonne par variable, qui contiendra de delta entre les 2 valeurs du point de vue du boxeur A. Par ex. pour un enregistrement ou Height\_A=180 et Height\_B=190, l’information est remplacée par : gap\_height = -10.
2. La variable expliquée (Y) devient alors « A\_win » = 1 ou 0 (1 lorsque le boxeur A gagne, 0 lorsqu’il perd ou qu’il y a match nul).
3. Le fichier contient environ 80% d’enregistrement où le boxeur A est vainqueur. Il est probable que l’attribution du code A ou B ait été attribué de manière non aléatoire lors de la création du dataset. On peut supposer par exemple que le code A ait systématiquement été attribué au boxeur ayant le meilleur classement dans la division internationale, ou autre. En conséquence, notre variable expliquée (Y) prend la valeur 1 dans 80% des cas. Afin d’éviter une distortion de notre processus d’analyse, le dataset est donc dédoublé en ramenant toutes les statistiques de boxeur B au niveau du boxeur A. Par exemple, la ligne *A\_won = 1, gap\_height = +12,…* est doublée, et son doublon est ajusté comme suit : A\_won = 0, gap\_height = -12,… etc… Notre dataset contient ainsi une part équivalente de Y = 0 et Y = 1.
4. Certaines valeurs aberrantes sont identifiées (age > 100, allonge supérieur a 500cm, etc..). Ces valeurs sont éliminées avec les règles suivantes :
   1. l’écart d’âge est supérieur à 50 ans ( => 1 enregistrements supprimé)
   2. l’écart d’allonge est supérieur à 100cm (=> 19 enregistrements supprimés)
5. afin de pouvoir être analysé avec une méthode de classification (régression logistiques) , les valeurs sont normalisées à l’aide d’une fonction automatiques AutoScaler fournie par Sklearn (source : <http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html>)
6. Enfin, le fichier contient beaucoup de valeurs non remplies. Le graphique ci-dessous illustre les parts de données remplies et manquantes sur le dataset sur les 387 430 enregistrements dont il dispose :

On constate une part de Null importants dans les variables :

* Weight (poids),
* Reach (allonge)
* Height (taille).

Afin de ne pas fausser l’analyse, nous éliminons tous les enregistrements du dataset ou l’une des variables n’est pas renseignée. Ceci a pour effet de limiter le jeu de données **7 300** enregistrements environ, mais cela nous apparait suffisant pour l’analyse.

Section 3 : Application du modèle

Le jeu de données obtenu après toutes les manipulations préparatoires est du format suivant:

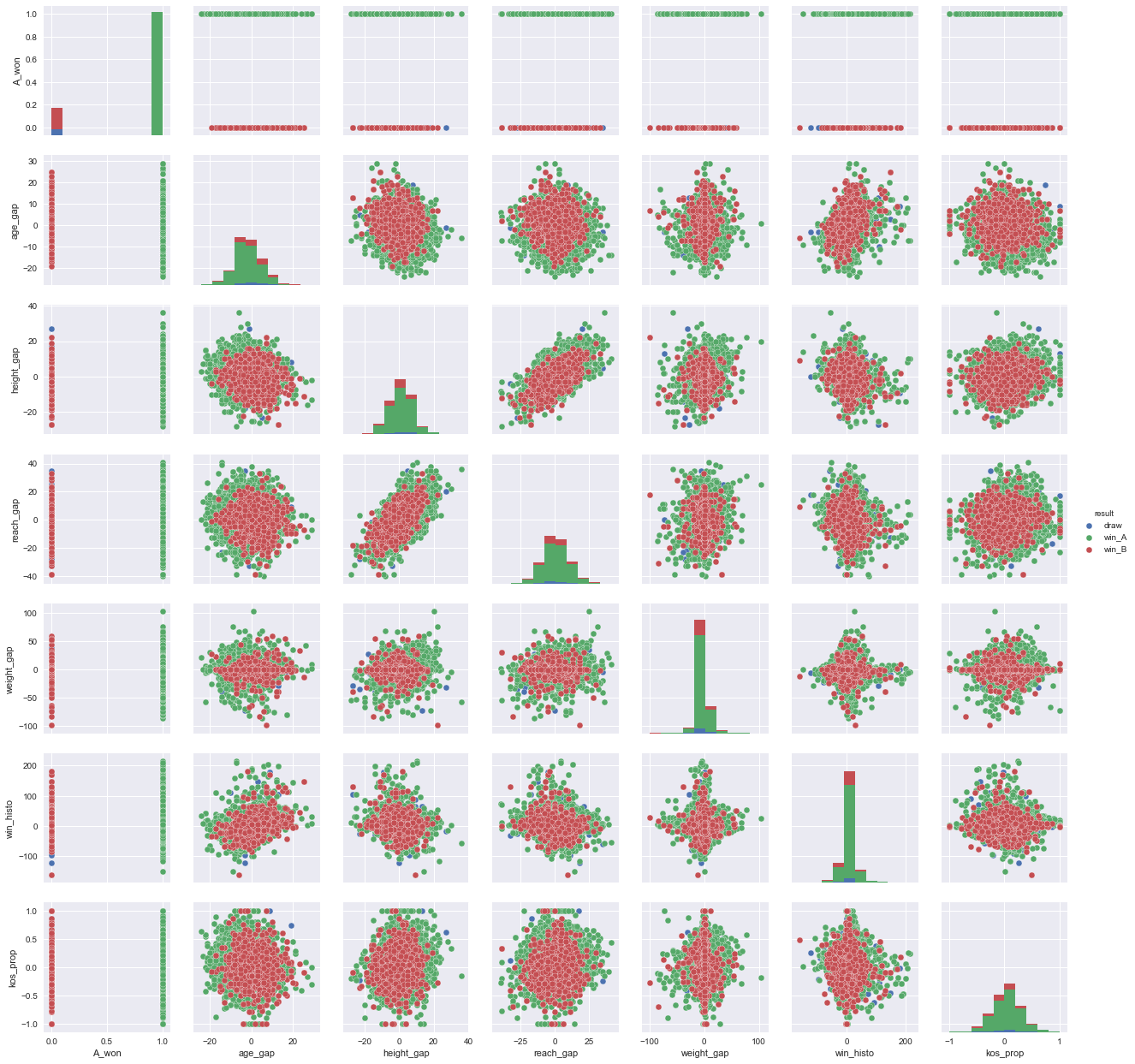
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Index | A\_won | age\_gap | height\_gap | reach\_gap | weight\_gap | win\_histo | kos\_prop |
| 0 | 1 | 8.0 | 4.0 | -1.0 | 0.0 | -11 | 0.198 |
| 1 | 1 | -5.0 | -10.0 | -6.0 | 0.0 | -1 | 0.068 |
| 2 | 0 | -4.0 | 1.0 | -1.0 | 0.0 | 17 | 0.082 |

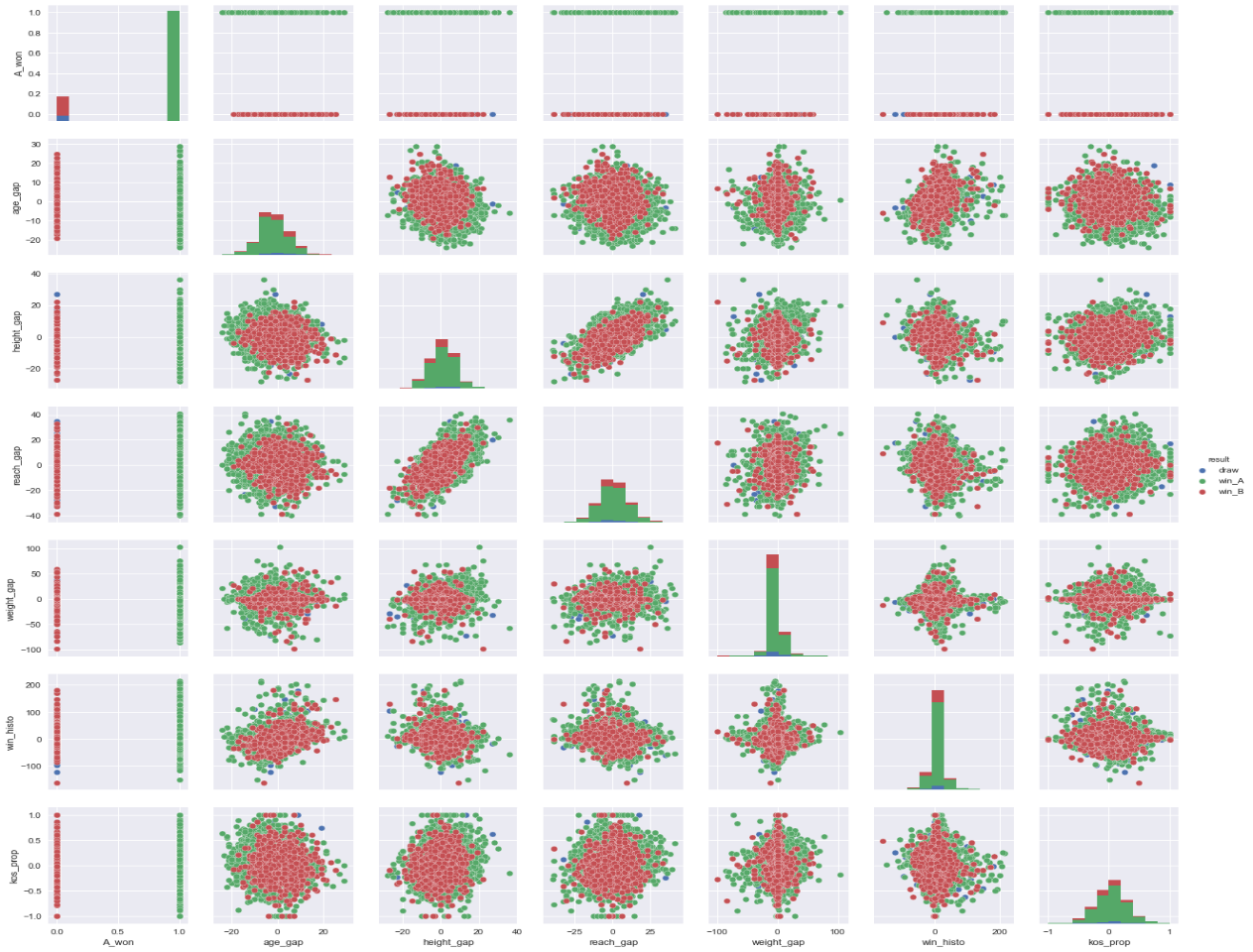
Une régression logistique appliquée à chaque dimension individuellement fait apparaitre les corrélations suivantes:

|  |  |  |
| --- | --- | --- |
| **3** variables explicatives a une forte corrélation au résultat : **l’âge , le palmarès, les victoires par KO** | **2** variables explicatives ont une légère corrélation au résultat : **la taille et l’allonge** | **1** variable explicative avec un corrélation au résultat dans certains cas |
|  |  |  |
| Les boxeurs + jeunes que leurs adversaires gagnent plus souvent, la jeunesse semble être un avantage sur l’expérience. Ensuite, les boxeurs au meilleur palmarès (et notamment avec le + de victoire par KO) remportent + de combats. | Les boxeurs ayant un avantage de taille et/ou d’allonge gagnent plus souvent. | L’écart de poids positif semble être un avantage mais pour une partie seulement du dataset (on note l’amplitude de la zone de variance bleutée). |

Note : l’avantage de du poids semble annulé face à des boxeurs plus jeunes, comme le montre le graphique croisé ci-dessous :

Lost / won

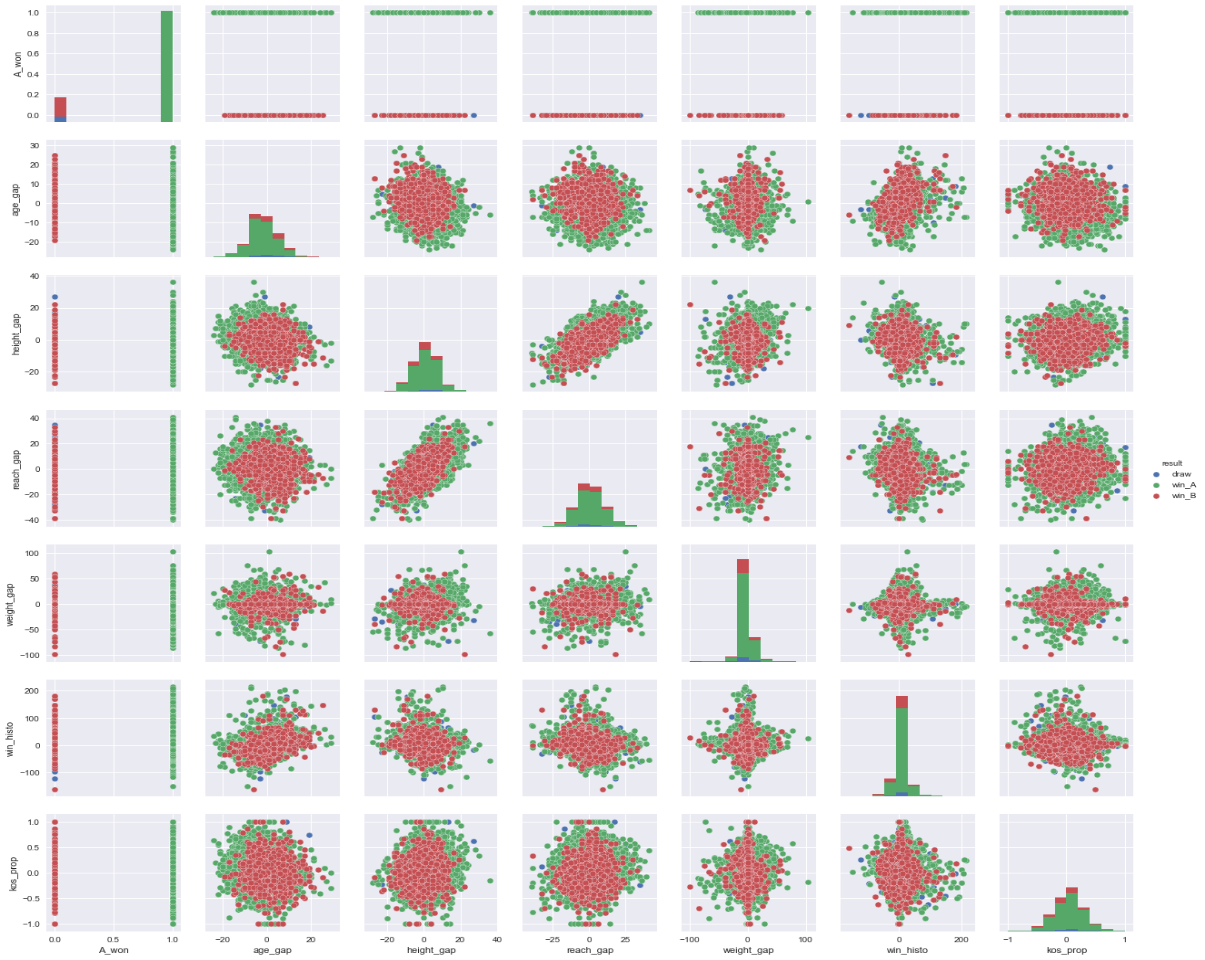




Globalement + de victoires chez les boxeurs + lourds…

Weight gap

Age gap



.... mais la victoire est indépendante du poids pour les boxeurs + jeunes

En conclusion, le dataset est analysé avec une méthode de classification de type régression logisitique fournie par sklearn (source : <http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html>)

L’analyse porte sur 80% du dataset extrait de manière aléatoire, 20% du dataset étant reservé a un jeu de test de validation.

## Section 4 : évaluation du modèle

Le modèle de prédiction est appliqué sur le jeu de test (20% du data set) et prédit correctement la victoire des boxeurs dans **67%** des cas.

La matrice de confusion est la suivante :

Total prévisions correctes: **67,5%**

Soit 1988 valides sur 2942.

Vs “si 1 prédit tout le temps”: 51% (1489 valides sur 2942)

Vs “si 0 prédit tout le temps”: 49% (1453 valides sur 2942)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | reel 0 | reel 1 |  |  |
| test 0 | 988 | 489 |  |  |
| test 1 | 465 | 1 000 |  |  |
|  | *Total 1 453* | *Total 1 489* |  |  |

En conclusion le modèle opère de manière plus performante qu’une prédiction statique.

La pertinence de chaque variable est analysée avec la fonction Logit de statsmodel (source <https://www.statsmodels.org>). Cette analyse révèle un écart important de pertinence entre les variables utilisées :

Les 2 variables fournissant les plus mauvais scores (poids et allonge) sont éliminées successivement et le modèle relancé.

On notera que le résultat de Adjusted Square mean descend, indiquant que le retrait des variables n’apporte pas de bénéfice à la prévision. L’analyse de la matrice de confusion confirme que la suppression de ces variables explicatives n’apporte pas de bénéfice au modèle en l’état.

## Section 5 : optimisation du modèle.

L’analyse des variables par statsmodel a révélé un score faible de pertinence des variables « poids » et « allonge ». Le retrait de ces variables du dataset n’a pas apporté de bénéfice aux prévisions en tant que tel, néanmoins il est important de noter que ces 2 variables ont largement contribué à la réduction du dataset lors de la préparation des données, car elles contenaient une part importante de valeurs vides.

Une nouvelle version du dataset est donc préparée dans laquelle les variables Poids et Allonges sont éliminées en tout début de préparation de données. En conséquence, la suppression des enregistrements contenant des données vides réduit moins le dataset d’étude. Dans cette version, le dataset préserve **90 000** enregistrements.

Le bénéfice apporté par ce nombre + grand d’enregistrements (90 000 au lieu de 7 300) est visible sur la nouvelle matrice de confusion obtenue :

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | reel 0 | reel 1 |  | Total prévisions correctes: **69,3%**  Soit **+1,7 point** par rapport au premier score obtenu |
| test 0 | 12 739 | 5 665 |  |  |
| test 1 | 5 595 | 12 759 |  |  |
|  | *Total 18 334* | *Total 18 424* |  |  |

Conclusion

En conclusion, notre capacité à prédire l’issue d’un match de boxe est d’environ **69,3%** si l’on connait

* L’âge des 2 participants
* Leur écart de taille
* Leur écart de palmarès passé, dont la proportion de victoire par KO