2023

Joachim/Pierre/Walid

Simplon

03/11/2023



Détermination de la Satisfaction des passagers des compagnies aériennes

Table des matières

[A. Analyse exploratoire des données 2](#_Toc149825231)

[1. Valeurs manquantes 2](#_Toc149825232)

[2. Valeurs aberrantes 3](#_Toc149825233)

[3. Corrélation 3](#_Toc149825234)

[B. Création du modèle 4](#_Toc149825235)

[1. SGD Classifier 4](#_Toc149825236)

[2. Kernel Approximation 5](#_Toc149825237)

## Analyse exploratoire des données

L’analyse des données se trouve dans le rapport html situé à la racine du projet : report.html.

On a choisi d’établir un rapport automatisé à cause du délai très court pour réaliser le projet.

### Valeurs manquantes

On remarque qu’il n’y a qu’une seule colonne où des valeurs sont manquantes, c’est la colonne « Arrival Delay in Minutes »

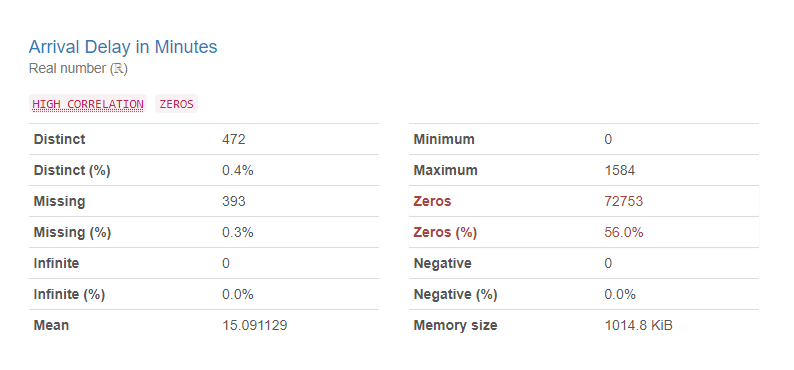


Figure Retard à l'arrivée en minutes.

On voit qu’il y a 0.3% de valeurs manquantes et que la moyenne est d’environ 15 minutes.

D’après la figure 2, plus des trois quarts des valeurs sont comprises entre 0 et 15.

On va donc garder cette colonne et remplacer les valeurs manquantes par la moyenne car cela n’influencera que très faiblement les résultats.

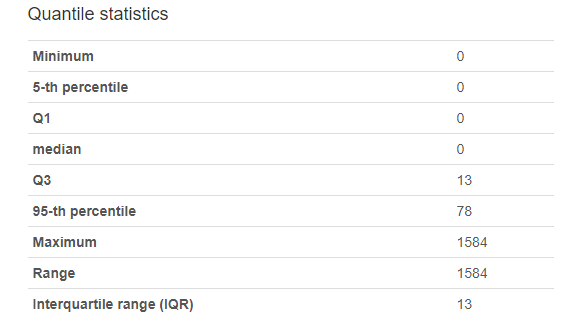


Figure statistiques quantile pour le retard à l’arrivée.

### Valeurs aberrantes

D’après notre analyse, il n’y a aucune valeur aberrante. Nous avons levé nos doutes quant aux valeurs des distances de vols (31 km) et temps de retards ( > 24h).

### Corrélation

Sur le tableau des corrélations du rapport html, on observe trois fortes corrélation (> 0.5) avec la satisfaction client pour les colonnes : « Inflight wifi service (0.526) », « Online boarding (0.618) » et « class (0.503) ». Elle semble avoir un fort impact sur la satisfaction client, à vérifier.

A l’inverse, on observe de très faibles corrélation (< 0.1) avec la satisfaction client pour les colonnes : « Departure/Arrival time convenient (0.068) », « Departure Delay in Minutes (0.018) », « Arrival Delay in Minutes (0.018) » et « Gender (0.011) ». Ce qui laisse à penser qu’elles n’ont que très peu d’impact voir pas d’impact du tout sur la satisfaction client, à vérifier.

Au vu des exigences du cahier des charges où apparaissent même les colonnes les moins corrélées avec la satisfaction client nous décidons de toutes les garder.

## Création du modèle

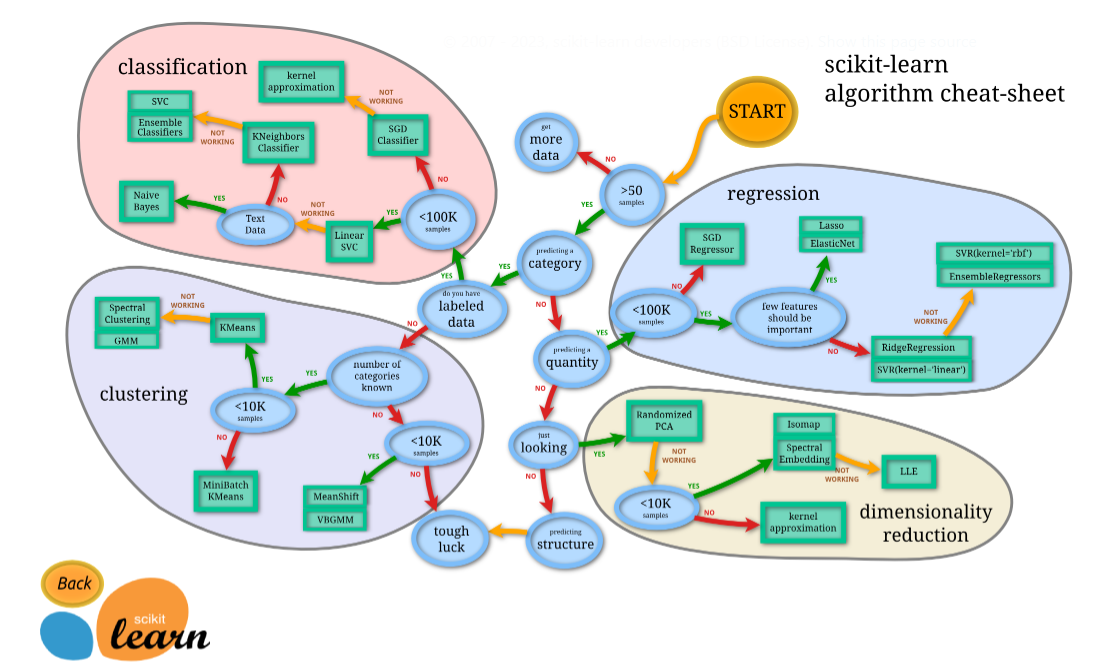


Figure Choix du model de machine Learning

En vertu de la figure 3, nous choisissons d’effectuer les tests avec deux modèles qui sont SGD Classifier et Kernel approximation, car nous cherchons à faire de la classification avec plus de de 100K de données.

### SGD Classifier

La Descente de Gradient Stochastique (SGD, pour Stochastic Gradient Descent) est une approche simple mais très efficace pour ajuster des classificateurs et des régresseurs linéaires sous des fonctions de perte convexes telles que les Machines à Vecteurs de Support (SVM) linéaires et la Régression Logistique.

Cette méthode est particulièrement utile lorsque l’on veut séparer deux groupes de points sur un graphique (fig.4).

La SGD a été avec succès appliquée à des problèmes d'apprentissage automatique à grande échelle et épars, souvent rencontrés dans la classification de texte et le traitement du langage naturel. Étant donné que les données sont éparses, les classificateurs de ce module s'adaptent facilement à des problèmes comportant plus de 10^5 exemples d'entraînement et plus de 10^5 caractéristiques.

Strictement parlant, la SGD est simplement une technique d'optimisation et ne correspond pas à une famille spécifique de modèles d'apprentissage automatique. C'est seulement une manière de former un modèle. Souvent, une instance de SGDClassifier ou SGDRegressor aura un estimateur équivalent dans l'API scikit-learn, potentiellement en utilisant une technique d'optimisation différente.

Les avantages de la Descente de Gradient Stochastique sont les suivants :

1. Efficacité.
2. Facilité de mise en œuvre (de nombreuses possibilités d'optimisation du code).

Les inconvénients de la Descente de Gradient Stochastique comprennent :

1. La SGD nécessite plusieurs hyperparamètres tels que le paramètre de régularisation et le nombre d'itérations.
2. La SGD est sensible à la mise à l'échelle des caractéristiques.

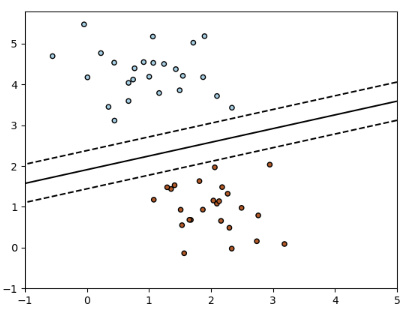


Figure Séparation des groupes avec SGDClassifier

### Kernel Approximation

Ce sous-module contient des fonctions qui approximent les mappages de caractéristiques correspondant à certains noyaux, tels qu'ils sont utilisés, par exemple, dans les machines à vecteurs de support. Les fonctions de caractéristiques suivantes effectuent des transformations non linéaires de l'entrée, qui peuvent servir de base pour la classification linéaire ou d'autres algorithmes.

Ces fonctions aident à transformer les données d'une manière qui les rend adaptées à des algorithmes de classification ou d'autres types d'analyses. Ces transformations sont non linéaires, ce qui signifie qu'elles ne suivent pas une relation simple comme "plus c'est grand, plus c'est important". Au lieu de cela, elles capturent des relations plus complexes entre les données.

L'idée principale est que ces fonctions peuvent permettre de traiter des données de manière plus efficace et précise, en les préparant pour l'analyse. Cela peut être particulièrement utile lorsque l’on travaille avec de grandes quantités de données, car cela peut réduire les coûts de calcul et accélérer le processus d'apprentissage.

Les SVM (Support Vector Machines) à noyau standard ne sont pas adaptées aux ensembles de données volumineux, mais en utilisant une approximation de mappage de noyau, il est possible d'utiliser des SVM linéaires beaucoup plus efficaces.

En particulier, la combinaison d'approximations de mappage de noyau avec SGDClassifier peut permettre l'apprentissage non linéaire sur de grands ensembles de données.

### Random Forest (Forêt d’arbres décisionnels)

Les forêts d’arbres décisionnels ou forêts aléatoires (Random Forest) sont une technique d’apprentissage ensembliste qui s’appuie sur des arbres de décision. Le modèle random forest implique la création d’arbres décisionnels (decision tree) multiples en utilisant ensembles de données fractionnés à partir des données d’origine. Et en sélectionnant aléatoirement un sous-ensemble de variables à chaque étape de l’arbre décisionnel. Le modèle sélectionne ensuite le mode de toutes les prédictions de chaque arbre décisionnel.

Quel est l’intérêt de cette méthode ?

En s’appuyant sur un modèle de prévalence de la majorité (c’est-à-dire sur lequel la majorité l’emporte), il réduit le risque d’erreur d’un arbre individuel.

### Model choisit

Le modèle choisit est Random Forest car il offre un f1-score de 96% contre moins de 90% pour les deux autres et l’objectif était un f1-score d’au moins 95%.