Etapes d’un projet de Machine Learning

1. Prendre du recul pour avoir une vision d’ensemble.
2. Récupérer les données.
3. Découvrir et visualiser les données pour mieux les comprendre.
4. Préparer les données pour les algorithmes d’apprentissage.
5. Sélectionner un modèle et l’entraîner.
6. Régler avec précision le modèle.
7. Présenter la solution.
8. Lancer, surveiller et maintenir le système.

# Prendre du recul pour avoir une vision d’ensemble

Le but est dans un premier temps de construire un modèle de l’évolution du cours d’une action sur un mois.

L’objectif final est de prévoir l’évolution du cours d’une action sur 5 ans afin de savoir si cela vaut le coup d’investir dans le titre associé.

La prédiction a réaliser est le cours médian de l’action par jour, sur un mois.

Il est donc essentiel d’avoir une prédiction réaliste afin de ne pas perdre d’argent.

Si le modèle fonctionne très bien, ajusté à la journée avec un grand nombre de paramètres il pourrait permettre du day trading.

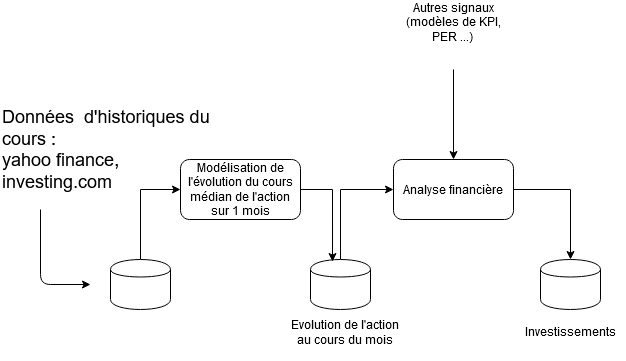


Figure : Pipeline de systèmes d'apprentissage automatique pour l'investissement boursier

Dans un premier temps, l’étude portera sur l’action PEUGEOT SA avec un jeu de données représentant l’évolution sur un an du titre. Le modèle retenu, bien que très peu fiable pour des raisons évidentes, est une régression linéaire. Puis le modèle évoluera vers une régression polynomiale, et d’autres représentation seront étudiées.

Ainsi, il s’agit d’un apprentissage **supervisé**, les exemples d’entraînement sont étiquetés : le cours réel de l’action pour chaque jour attendu est donné. Il s’agit d’une tâche de **régression** classique, **multiple** car le résultat attendu est la prédiction d’une valeur. Pour l’instant le modèle est univarié, c’est-à-dire qu’une **unique** prédiction pour chaque jour est réalisée, et pas plusieurs par jour.

L’apprentissage est d’abord réalisé comme si les données n’évoluaient pas rapidement, en première approximation car on considère que les variations du titre sont autour de 5% de la valeur médiane.

Les hypothèses sont donc :

* Evolution du titre étudiée sur un an
* Régression linéaire
* Multiple
* Univariée
* Apprentissage groupé (en batch) car données tenant en mémoire sans variations continues
* Modèle visant la prédiction sur un mois (au jour le jour, et à la fin du mois)

Puisqu’une méthode de régression est utilisée, la mesure de performance retenue est la racine carré de l’erreur quadratique (Root Mean Square Error, RMSE) :

Avec :

* m : le nombre d’observations du jeu de données (ex : si la RMSE est évaluée sur 2000 districts pour une prédiction de prix immobiliers, alors m = 2000),
* : le vecteur d’observation des valeurs des variables (non-compris l’étiquette) pour la ième observation du jeu de données et est son étiquette (la valeur de sortie souhaitée pour cette observation). Ainsi, si le premier district du jeu de données hypothétique précédent se situe à la longitude -118°, latitude 34°, possède 1500 habitants avec un revenu médian de 40.000 € et le prix médian des habitations est 160.000 € (en ignorant les autres variables), alors = 160.000 € et
* X est la matrice contenant toute les valeurs des variables des observations du jeu de données (pas les étiquettes), avec une ligne par observation et la ième ligne correspond à la transposée de  :
* h est la fonction de prédiction du système, lorsqu’il reçoit le vecteur des valeurs des variables pour une observations , il renvoie la valeur prédite .

Ainsi, si le système prédit que le prix médian des habitations dans le premier district est 158.000 €, alors . L’erreur de prédiction est donc . Il y a donc une erreur de prédiction de 2000 €.

est la fonction coût mesuré sur le jeu d’exemple en utilisant l’hypothèse h.

ATTENTION : Lorsque les valeurs sont très extrêmes, la mesure de la performance doit-être effectué autrement : Mean Absolute Error (MAE) … voir p41 *Introduction à Scikit-Learn*.

# Récupérer les données