MACHINE LEARNING POUR TIME SERIES : Application au Covid-19

Mai 2020

coperneec

par Pascal Pierrot, Data Scientist

L'application de l'apprentissage automatique à la prévision et au traitement de séries temporelles a connu un essor ces dernières années grâce notamment à la puissance de calcul des ordinateurs et au nombre croissant de données disponibles. Les secteurs d'applications sont nombreux : finance, économie, énergie, retail...etc.
Voici quelques-unes de ces méthodes au travers d'un exemple d'actualité : la prévision de nouveaux cas de Covid-19 à court et moyen terme.

1 - Contexte

Le nouveau coronavirus suscite encore beaucoup d'interrogations concernant son évolution épidémique et les facteurs influençant cette dernière. L'analyse des pandémies historiques ne suffit pas toujours à prévoir l'évolution car le contexte et les virus sont différents. Le Machine Learning n'a pas vocation à remplacer un modèle épidémique paramétré sur des projections au long terme mais peut être utilisé comme une approche complémentaire intéressante. Nous pouvons, en effet, tirer parti des données de pays se situant dans une phase avancée de l'épidémie et formuler ainsi un problème d'apprentissage supervisé pour l'anticipation de l'évolution épidémique à plus court terme (de un jour à une semaine) et aussi analyser quelles variables exogènes impactent cette évolution.

2 - Analyse des données disponibles et choix d'une variable à prédire

Données utilisées : Les données sont mises à disposition sur le site Kaggle par Johns Hopkins University Center for Systems Science and Engineering. Elles comprennent l'évolution du nombre de cas et de victimes par zone géographique ainsi que des données météo et des indicateurs par pays : santé, hôpitaux, population, densité, taux d'immigration et dates de la prise de mesures restrictives comme la fermeture des écoles ou le confinement.

Variable cible : Le choix d'une variable à prédire dépend de la qualité, de la disponibilité mais également de l'intérêt de la donnée. Nous avons choisi de prédire ici les nouveaux cas à court et moyen terme.

Horizon de prédiction : Le choix de l'horizon de prédiction dépend également des données disponibles, en l'occurence un horizon entre un jour et une semaine semble raisonnable au vu de l'avancement de l'épidémie.

Stationnarité : Afin de travailler avec des séries stationnaires, nous allons prédire la variation relative des nouveaux cas plutôt que la valeur absolue cumulée.

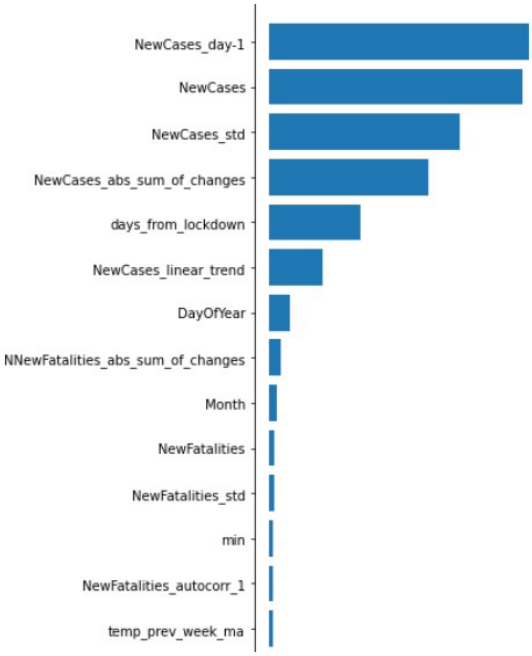
3 - Importance des variables

En plus de l'évolution des courbes au cours des derniers jours, nous distinguons deux types de variables explicatives : les variables dites statiques d'une part, dont les valeurs ne varient pas au cours du temps mais ayant une influence globale sur les courbes (indicateurs sur la santé, le climat ou la population d'un pays par exemple) et les variables dynamiques d'autre part (nombre de jours écoulés depuis le confinement, données météo, mortalité). Avant de calibrer les modèles, il nous faut étudier et sélectionner ces variables.

Méthodologie : Des analyses de corrélations de type Kendall adaptées aux données non normalement distribuées permettent de trouver des relations (linéaires) entre variables statiques et variables cibles. Pour les variables dynamiques nous utilisons la méthode Mean decrease accuracy : le principe est d'analyser l'impact d'une permutation des valeurs d'une variable sur la précision d'un modèle. Plus la précision du modèle décroît lors de cette permutation, plus elle sera considérée comme importante. Le regroupement des variables fortement corrélées entre elles (clusters) lors de cette opération permet d'améliorer cette technique.

Résultats : L'analyse des corrélations a notamment fait ressortir une corrélation positive entre les taux de cancers, de maladies respiratoires, d'obésité et de malnutrition et le taux moyen de croissance des nouveaux cas de Covid. Une corrélation positive également entre des mesures restrictives précoces et la rapidité de propagation (proportion de population infectée après une certaine période). Par ailleurs, un climat plus chaud semble être corrélé avec une propagation de cas plus lente. L'analyse des données dynamiques montre quant à elle l'importance de l'évolution des nouveaux cas précédents, du nombre de jours écoulés depuis le début des restrictions. Enfin, les données météo ne semblent pas jouer un rôle déterminant.

Visualisation de l'importance des variables par la méthode Mean decrease accuracy appliquée sur les données « out-of-bag » d'un algorithme RandomForest :



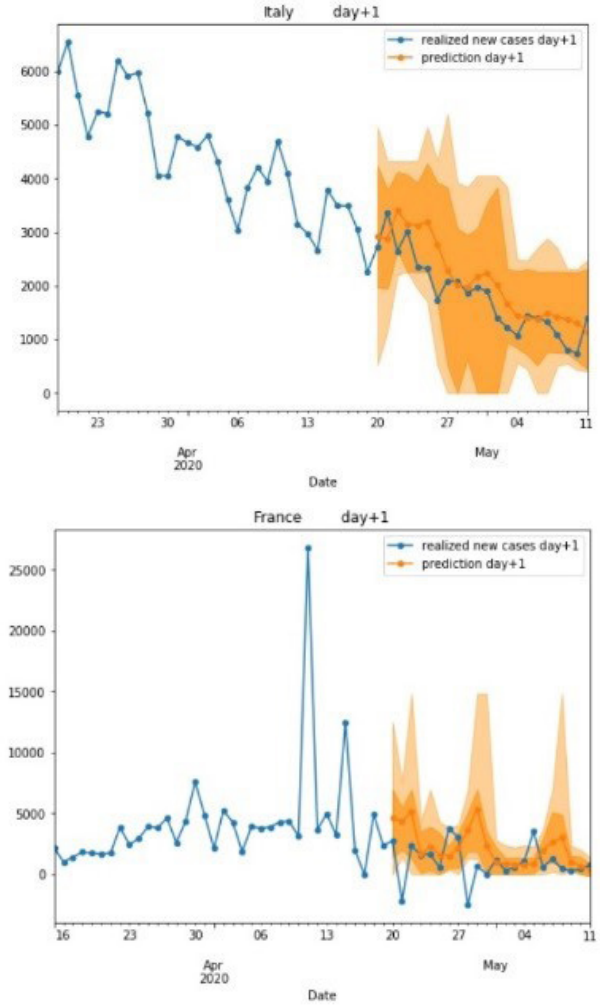
4 - Construction des modèles et résultats

Choix de l'approche : Nous formulons le problème d'apprentissage supervisé en utilisant la variation du nombre de nouveaux cas entre le présent et un instant futur comme variable cible et ces mêmes valeurs sur les derniers jours ainsi que les autres variables sélectionnées précédemment comme variables explicatives. Un même modèle est entraîné sur l'ensemble des courbes de tous les pays, de manière à minimiser l'erreur des moindres carrés entre le nombre de nouveaux cas réels et estimés.

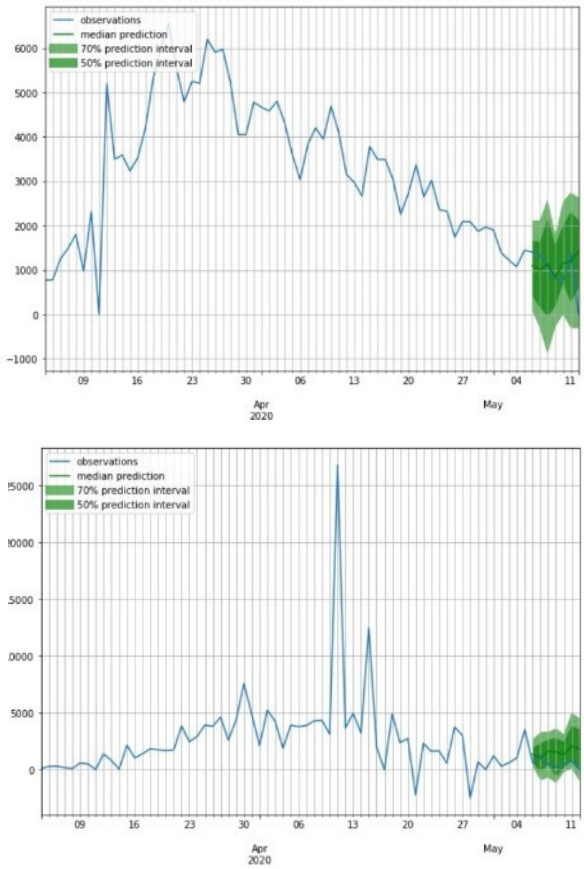
Choix des modèles : Nous avons d'abord utilisé un modèle de type Bagging (RandomForest) ainsi que des modèles gradient boosting: LightGBM, XGBoost tout en limitant la croissance des arbres de décision pour éviter le surapprentissage. L'intérêt est de tester des modèles ayant des objectifs différents : l'un tente de réduire la variance lorsque l'autre vise à réduire le biais. Nous avons aussi utilisé la régression par quantile sur ces modèles. Un troisième type de modèle est également testé : Il s'agit de prédictions probabilistes basées sur des réseaux de neurone. Le modèle prédit une distribution et non plus une seule valeur et vise à maximiser la vraisemblance associée à une distribution (loi conditionnelle) choisie sur les valeurs futures de la série et non plus à minimiser l'erreur des moindres carrés (cas d'une simple régression). Lors de la prédiction, un tirage est réalisé dans cette distribution ce qui donne une plage d'incertitude avec moyenne et quantiles.

Backtest et résultats : Une fois entraînés, les modèles sont testés sur une période plus récente. Le modèle Random forest est celui ayant obtenu le meilleur score sur cette période avec une erreur absolue moyenne de 164 pour la variation des nouveaux cas et de 2.1 % pour la variation relative.

Visualisation des résultats de prédiction pour le jour suivant sur une période de test :



Résultats du modèle RandomForest (1000 arbres, max_depth=20) avec les quantiles 30-70% et 10-90% pour l'Italie et la France.



Le résultat du modèle MLP Feedforward (100 neurones, 10 epochs) avec une loi t-student comme distribution de sortie utilisant la librairie GluonTS.

Même si les résultats obtenus grâce à cette approche Machine Learning sont intéressants, l'heure n'est pas encore à la remise en cause des modèles épidémiologiques utilisés plus classiquement.
Notre approche est d'abord complémentaire et vise à enrichir la prévision à court terme par l'utilisation d'une masse de données plus importante avec l'intégration de données exogènes (par exemple : l'adaptation des comportements et les mesures prises par les pays se trouvant à un stade plus avancé de l'épidémie) afin que le modèle s'ajuste dynamiquement au cours du temps.

A propos de Coperneec

"From revolution to performance"

Coperneec est un cabinet de conseil cross-sectoriel spécialiste de la valorisation de la Data. Nous intervenons sur l'ensemble de la chaîne des savoir-faire autour de la Data Science, la Data Analyse et du Data Management.
Nos méthodes et techniques scientifiques éprouvées permettent de résoudre des problématiques dans tous les secteurs de l'industrie.

Notre vocation : extraire la connaissance à partir des données et pérenniser les avancées technologiques qui en découlent. La R&D est au cœur de notre ADN et les expertises de nos consultants (data scientists, data analysts, data engineers) sont en permanence challengées afin d'accompagner au plus près les révolutions technologiques et scientifiques.

Contactez-nous

Aymeric Lisbonne
Partner
alisbonne@coperneec.com
06 88 69 67 75

coperneec

est une marque de

canopee group