DataScientest • com

Rapport Technique d'évaluation

Prédiction des nouvelles contaminations à la Covid-19 en France

Promotion décembre 2020

Participants :

Victor Lieutaud
Cédric Schlosser
Carine Pereira

Contexte

Depuis janvier 2020, la France, comme dans le reste du monde, fait face à l'épidémie de Covid-19 et ses conséquences économiques et humaines.
Pour gérer au mieux cette crise le gouvernement a mobilisé ses agents de santé publique afin de mettre en place un système de surveillance qui recueille quotidiennement des données sur l'évolution de l'épidémie dans le pays. L'objectif de cette démarche est d'analyser la menace afin de préparer la meilleure réponse possible.
Pour permettre à tous de participer à cet effort national, ces données sont accessibles au public depuis la plateforme https://www.data.gouv.fr/.
Dans ce contexte, les data sciences sont un outil précieux pour analyser l'évolution de l'épidémie mais aussi pour essayer de prédire sa trajectoire. En effet une prédiction précise des contaminations permettrait d'adapter au mieux la stratégie de réponses (vaccinations, confinements, aides aux entreprises, etc…) dans le but de minimiser l'impact humain du covid-19.

D'un point de vue technique :
Ce projet est un problème de régression sur des données temporelles que nous avons traité à l'aide de différents algorithmes de machine Learning supervisés.

D'un point de vue économique :
Si il est encore aujourd'hui difficile de chiffrer précisément l'ensemble des impacts directs (coûts hospitaliers, tests, vaccins…) et indirects (baisse de l'activité, aides aux entreprises, chômage partiel…) du covid-19, les experts s'accordent à un coût de plusieurs centaines de milliards d'euros. Obtenir de bonnes prédictions de l'évolution de l'épidémie serait donc un atout majeur pour aider à la prise de décision des mesures les plus efficaces pour limiter les dégâts économiques et psychologiques sur les français. Par exemple, une prédiction précise du nombre de contaminations aidera à limiter le nombre et la durée des confinements

Du point de vue Scientifique :
Ce projet concerne des données épidémiologiques. Il est nécessaire de comprendre la signification des différents indicateurs et leurs interactions afin de choisir ceux qui sont le plus pertinent pour la création du modèle.

Objectifs

Quels sont les principaux objectifs à atteindre ? Décrivez en quelques lignes.

Le principal objectif de ce projet est de créer un modèle prédictif de l'évolution de l'épidémie de Covid-19 en France. Pour cela d'autres objectifs sont à réaliser :

- Déterminer la variable cible qui sera la plus représentative de l'évolution de l'épidémie.
- Trouver, combiner, créer et nettoyer les données nécessaires pour notre dataset.
- Tester différents modèles de Machine Learning et choisir le modèle le plus efficace pour prédire notre cible.

Pour chacun des membres du groupe, préciser le niveau d'expertise autour de la problématique adressée ?

Nous sommes 3 à avoir réalisé ce projet :

- Victor Lieutaud : spécialisation risques/finance de marché, à l'aise avec les time series. Expérience de data scientist depuis quelques mois
- Cédric Schlosser : Bio-informaticien de formation, expertise en biologie.
- Carine Pereira : Biostatisticienne dans une entreprise pharmaceutique, ingénieur en biotechnologie, expertise dans la biologie et les facteurs de transmission des virus.

Êtes-vous entré en contact avec des experts métiers pour affiner la problématique et les modèles sous-jacents ? Si oui, détaillez l'apport de ces interactions.

Prise de contact avec un expert une experte de la question des times series parmi l'équipe de DataScientest. Apport de l'échange : sujet très complexe pour nous dans une optique d'utilisation d'un modèle "Multivariate Time Series". Cela a orienté le choix de nos deux modèles finaux.

Avez vous connaissance d'un projet similaire au sein de votre entreprise, ou bien dans votre entourage ? Quel est son état d'avancement ? En quoi vous a-t-il aidé dans la réalisation de votre projet ? En quoi votre projet contribue-t-il à l'améliorer ?

Victor : Je travaille actuellement dans mon entreprise sur un projet de prévision de flux de trésorerie, et sur un projet d'aide à la décision sur les marchés financiers. J'ai donc pu réutiliser une partie de la pipeline de ma recherche sur le modèle de flux de trésorerie. Cela permet aussi une meilleure compréhension de l'approche, puisque dans le modèle de trésorerie on utilise la variable temps mais pas juste dans une approche time series (d'où le randomforest regressor et le XGBoostRegressor)

Data

Cadre

Quel(s) jeu(x) de donnée(s) avez-vous utilisé pour atteindre les objectifs de votre projet ?

Plusieurs jeux de données ont été combinés pour ce projet.

Jeu de données principales :
Synthèse des indicateurs de suivi de l'épidémie COVID-19
URL :
https://www.data.gouv.fr/fr/datasets/synthese-des-indicateurs-de-suivi-de-lepidemie-covid-19/

Fichier choisi :
table-indicateurs-open-data-france-2021-07-27-19h05.csv

Raison du choix du fichier :
Ce jeu de données comprend l'essentiel des indicateurs de synthèse permettant le suivi de l'épidémie de COVID-19 sur l'ensemble du territoire Français.

Jeu de données secondaire :

Données relatives aux personnes vaccinées contre la Covid-19
URL :
https://www.data.gouv.fr/fr/datasets/donnees-relatives-aux-personnes-vaccinees-contre-la-covid-19-1/

Fichier choisi :
vacsi12-fra.csv

Raison du choix du fichier :
Ce jeu de données comprend l'essentiel des indicateurs concernant les vaccinations et permettant le suivi de l'épidémie de COVID-19 sur l'ensemble du territoire Français.

Données ajoutées :

Des colonnes ont été ajoutées manuellement pour indiquer les périodes de confinement et la saison (simplifiée en été et hiver). Nous avons décidé d'ajouter ces deux facteurs car il semble logique qu'ils aient une influence sur la propagation du virus.

***Ces données sont-elles disponibles librement ? Dans le cas contraire, qui est le propriétaire de la donnée** ?*

Les données utilisées ainsi qu'un grand nombre de données relatives au Covid-19 sont disponibles en open source sur le site du gouvernement :
https://www.data.gouv.fr/fr/pages/donnees-coronavirus/

Les données sont mises à jour quasiment quotidiennement depuis le début de l'épidémie en février 2020.

Décrivez la volumétrie de votre jeu de données ?

Le fichier table-indicateurs-open-data-france-2021-07-27-19h05.csv fait 21 colonnes et 553 lignes (au 27/07/2021) correspondant à un an et demi de données sur l'épidémie.

Le fichier vacsi12-fra.csv fait 8 colonnes et 212 lignes (au 27/07/2021)

Après nettoyage des données nous avons un dataframe de 28 colonnes et 436 lignes (au 27/07/2021)

Pertinence

Avez-vous eu à nettoyer et à traiter les données ? Si oui, décrivez votre processus de traitement.

Le jeu de données provient de la fusion de deux fichiers sur la colonne date.

Beaucoup de données n'ont pas été recueillies au début de l'épidémie, incluant des variables importantes, la conséquence est un grand nombre de NaNs. Nous avons donc supprimé les lignes précédent la date du 19/05/2020.

Les colonnes cv_dose1 et esms_cas sont pratiquement vides avec respectivement 1 et 11 valeurs non null et ont donc été supprimées. Tout comme la colonne fra n'apportant aucune information utile.

Le fichier des vaccinations (nombres de vaccinations, taux de vaccination…) ne contenait aucune valeur manquante avant la fusion des fichiers. La fusion étant faite sur la date de nombreux NaN sont générés avant la date des premières vaccinations. Ces valeurs manquantes sont donc remplacées par des zéros.

Pour rappel la colonne 'conf_j1' correspond au nombre de nouveaux cas confirmés (J-1 date de résultats)

La colonne conf_j1 avait 40% de valeurs manquantes, toutes au début de l'épidémie et ne présentent pas une forme caractéristique d'une loi de probabilité spécifique. Cette colonne a donc été supprimée.

Le reste des NaN ont été remplacées à l'aide d'un algorithme de type MICE (fonction IterativeImputer de sklearn). Cet algorithme permet de donner des valeurs aux NaN en se basant sur les valeurs des autres colonnes.

Quelles variables vous semblent les plus pertinentes au regard de vos objectifs ?

Les variables qui semblent le plus pertinentes au premier abord sont :

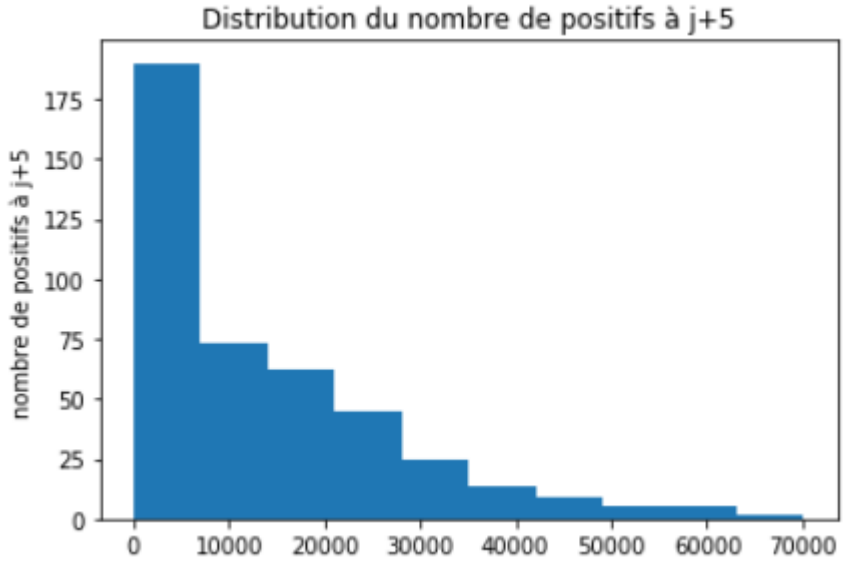
- tx_incid : Taux d'incidence (nombre de personnes testées positives pour la première fois depuis plus de 60 jours rapporté à la taille de la population).

- tx_pos : Taux de positivité des tests virologiques (Le taux de positivité correspond au nombre de personnes testées positives (RT-PCR et test antigénique) pour la première fois depuis plus de 60 jours rapporté au nombre total de personnes testées positives ou négatives sur une période donnée ; et qui n'ont jamais été testées positive dans les 60 jours précédents.)
- R : Facteur de reproduction du virus (le nombre moyen de personnes qu'une personne infectée peut contaminer. Si le R effectif est supérieur à 1, l'épidémie se développe ; s'il est inférieur à 1, l'épidémie régresse).
- pos : Nombre de personnes déclarées positives
- pos_7j : Nombre de personnes déclarées positives sur une semaine
- Le nombre de patients hospitalisés / en réanimation / décédé : incid_hosp, incid_rea, incid_dchosp.
- date : la date du jour
- Confinement_oui : variable qui indique si un confinement était en cours ou non
- couv_dose1 = le pourcentage de la population ayant reçu la première dose.
- couv_complet = le pourcentage de la population ayant reçu les deux doses et étant couvert complètement par le vaccin.

Quelle est la variable cible ?

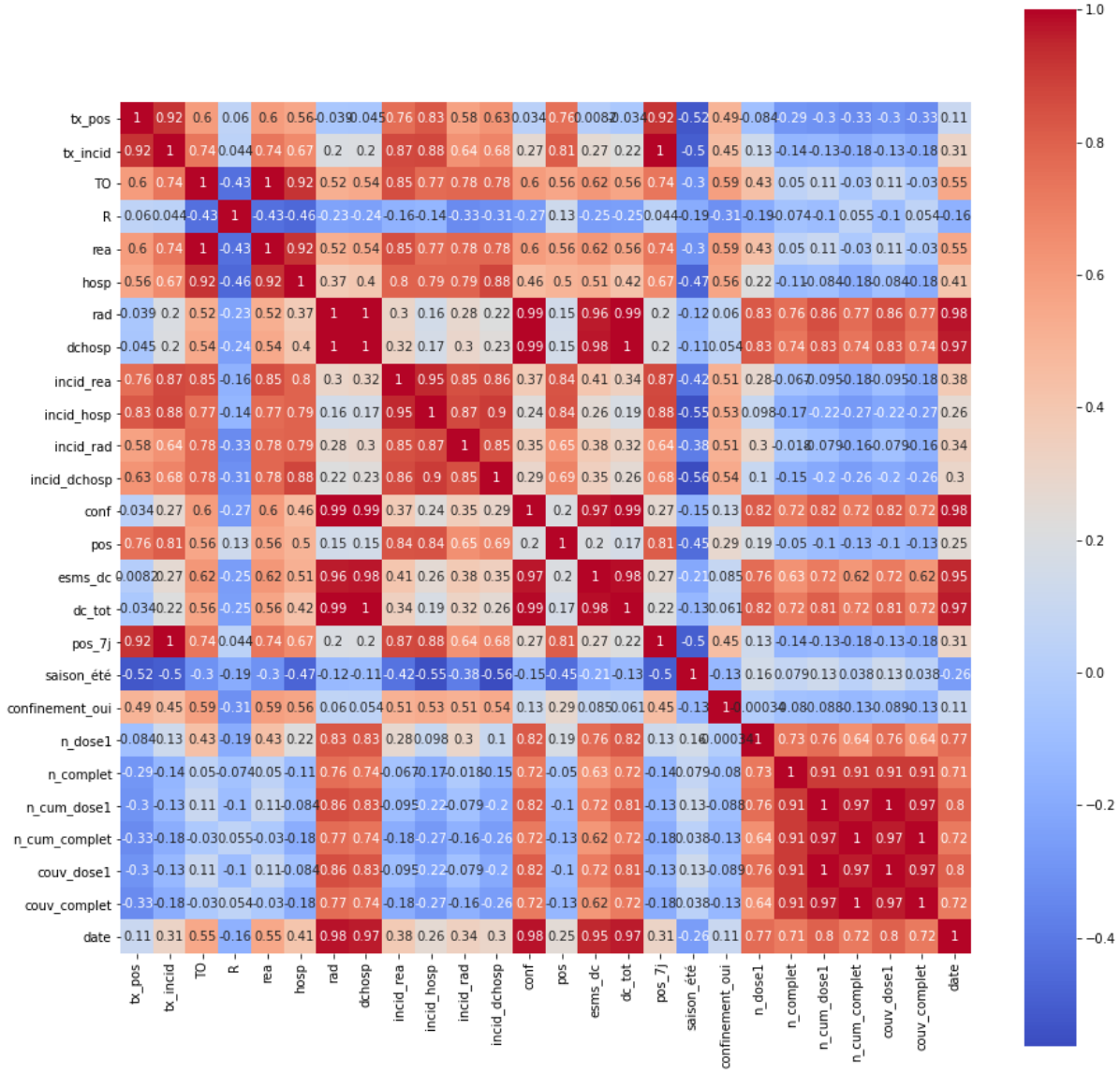
Nous créons la variable cible pos+5 qui correspond au nombre de de personnes déclarées positives à J+5.
L'idée derrière cette variable cible est que les personnes contaminées à T0 seront déclarées positive à T+5, car le temps moyen d'incubation est de 5 jours. Il ne faut pas oublier en regardant les résultats que ceci est une approximation, le délai de cinq jours étant une moyenne et tous les patients contaminés ne sont pas testés.

Décrivez la distribution de ses valeurs ?

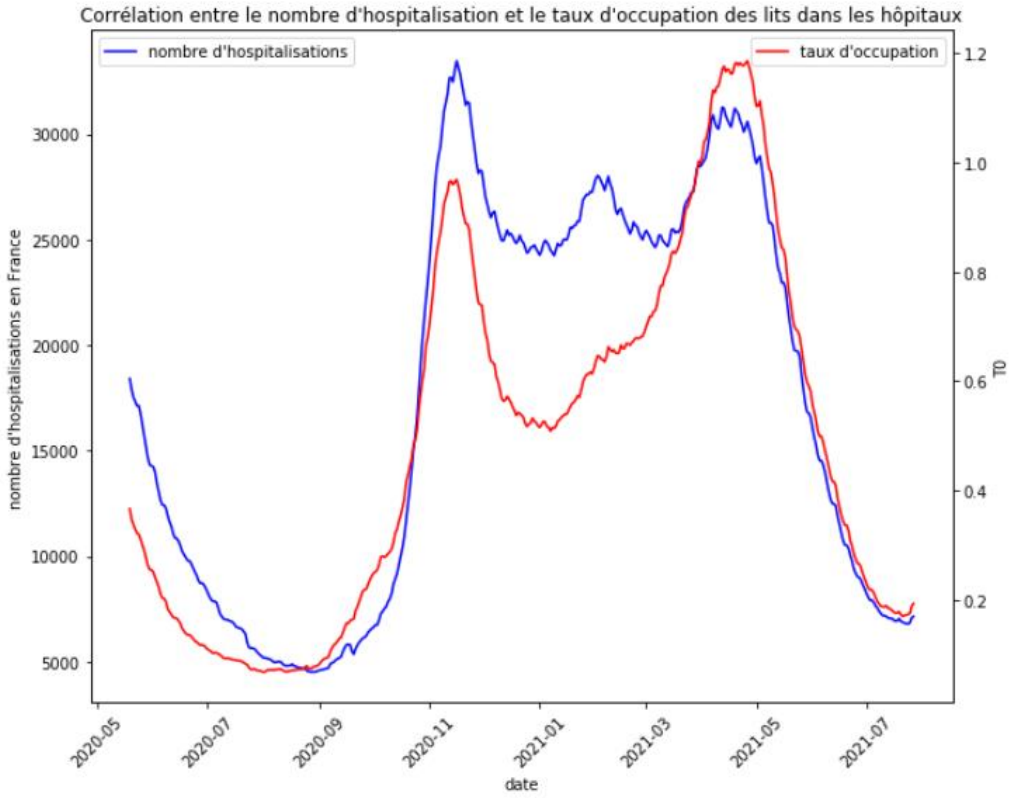


Comme attendu, il y a plus de valeurs faibles (correspondant aux moments où l'épidémie a été maîtrisée) que de valeurs élevées du nombre de contaminations à j+5.

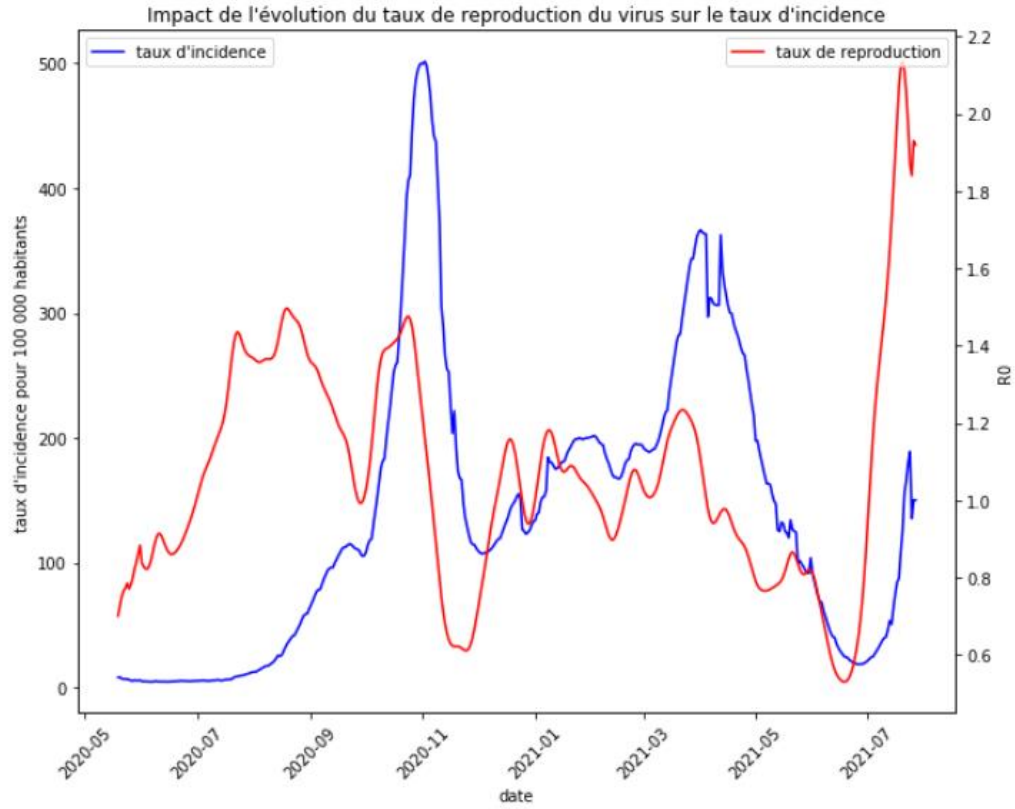
Avez-vous identifié des relations entre différentes variables ? Entre variables explicatives ? Et entre vos variables explicatives et la/les cible(s) ?



On constate une forte corrélation entre les variables tx_pos, tx_incid, T0, rea, hosp, incid rea, incid_hosp, incid_rad, incid_dchosp, pos, pos_7j. Ces variables représentant des taux (incidence, positivité au virus, occupation des hôpitaux) et des nombre de patients (nouveaux patients hospitalisés, en réanimation ou décédés, nombre de patients cumulés, nombre de nouveaux patients positifs à une date donnée) une forte corrélation positive entre elles semble tout à fait logique. La figure ci-dessous illustre la corrélation entre le nombre de patients hospitalisés et le taux d'occupation des hôpitaux.



Le graphique ci-dessous illustre l'impact du R0 sur le taux d'incidence.



Ce graphique montre une augmentation du taux d'incidence corrélée au taux de reproduction du virus. En effet, comme dit précédemment, le taux de reproduction du virus donne une indication sur le niveau de circulation du virus impactant directement le taux d'incidence. On remarque également sur ce graphique l'impact des différents confinements.

On remarque facilement que toutes les variables liées à la vaccination (n_dose1, n_complet, n_cum_dose1, n_cum_complet, couv_dose1, couv_complet) ont une corrélation positive entre elles extrêmement forte. Cela n'est pas étonnant au vu de l'étroite liaison de ses variables. Ces variables sont également corrélées négativement avec les variables du paragraphe précédent à l'exception des variables cumulées (ex : ldchosp, le total des décès en hopital). Le vaccin ayant pour but de ralentir la propagation, cette corrélation négative est le résultat qu'on espérait voir.

On remarque également une corrélation positive très forte entre la date et les variables cumulées ce qui encore une fois semble logique.

Quelles particularités de votre jeu de données pouvez-vous mettre en avant ?

Nos données sont des données temporelles.
Nous avons un grand nombre de variables très fortement corrélées (valeurs proche de 1).
Nos données provenant de l'épidémie de covid-19 l'échelle de temps est relativement faible (un an et demi environ) ce qui se traduit par un nombre de lignes restreint (moins de 450).

Projet

Classification du problème

À quel <u>type</u> de problème de machine learning votre projet s'apparente-t-il ? (Classification, régression, clustering ….)

Il s'agit d'un problème de régression.

À quelle <u>tâche</u> de machine learning votre projet s'apparente-t-il ? (détection de fraude, reconnaissance faciale, analyse de sentiment …) ?

Une tâche de prédiction de prix de stocks. Si le prix de notre stock est de x aujourd'hui, à combien sera-t-elle évaluée dans 15 jours.
Notre objectif est ici de prédire la future valeur du nombre de patients positifs à j+5.

Quelle est la métrique de performance principale utilisée pour comparer vos modèles ?

Nous avons utilisé l'erreur absolue moyenne (MAE).

Avez-vous utilisé d'autres métriques de performances qualitative ou quantitative) ? Si oui, détaillez.

Non.

Choix du modèle & Optimisation

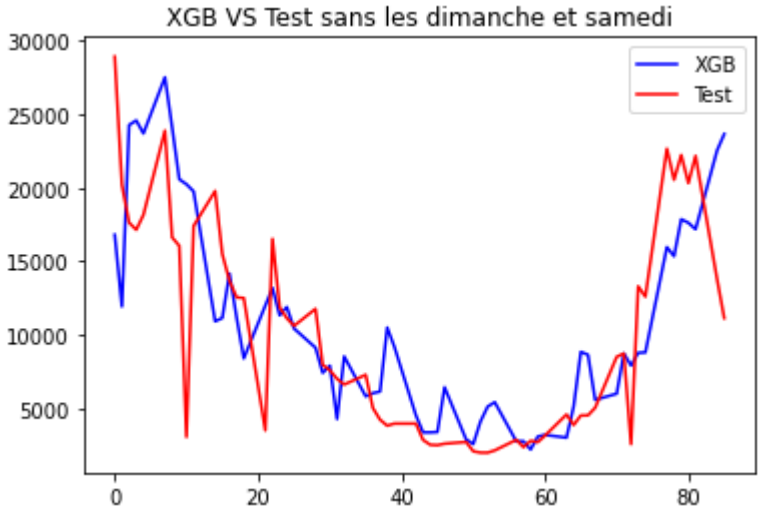
Quels algorithmes avez-vous essayé ?

Nous avons essayé deux algorithmes :

- RandomForest (sklearn RandomForestRegressor)
- XGBoost (sklearn XGBRegressor)

Décrivez celui / ceux que vous avez retenu et pourquoi ?

Le modèle retenu est XGBRegressor qui a une MAE significativement plus faible que RandomForestRegressor et qui capte bien la tendance des données.

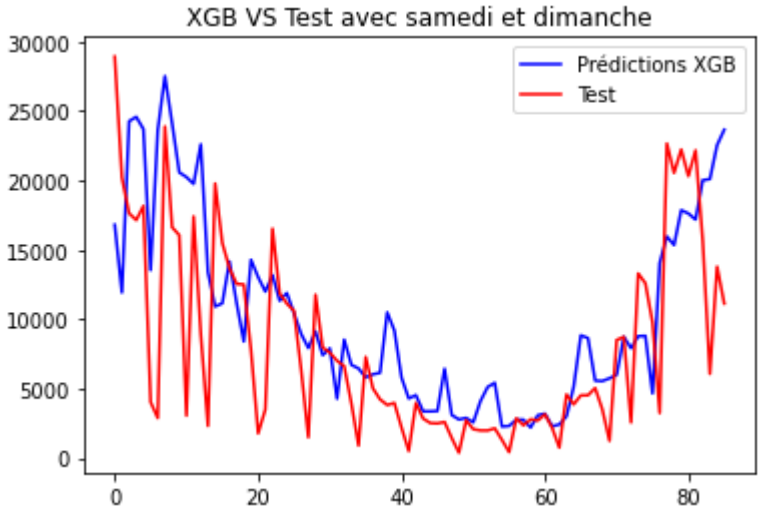


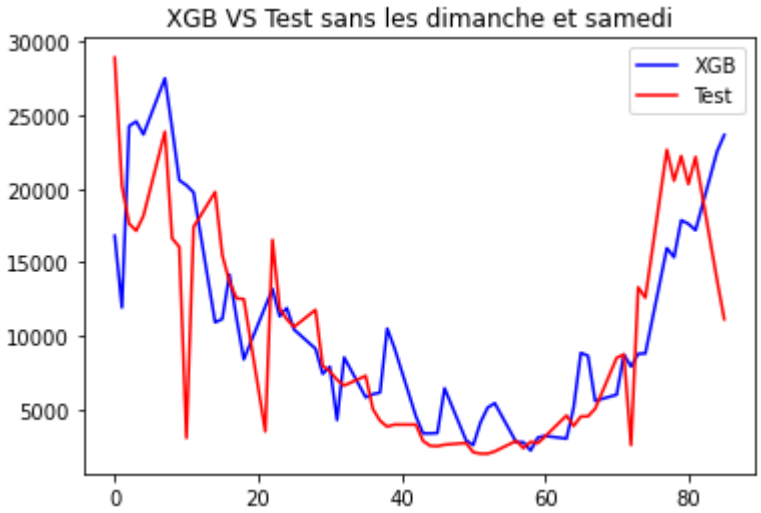
Qu'est ce qui a engendré une amélioration significative de vos performances ?

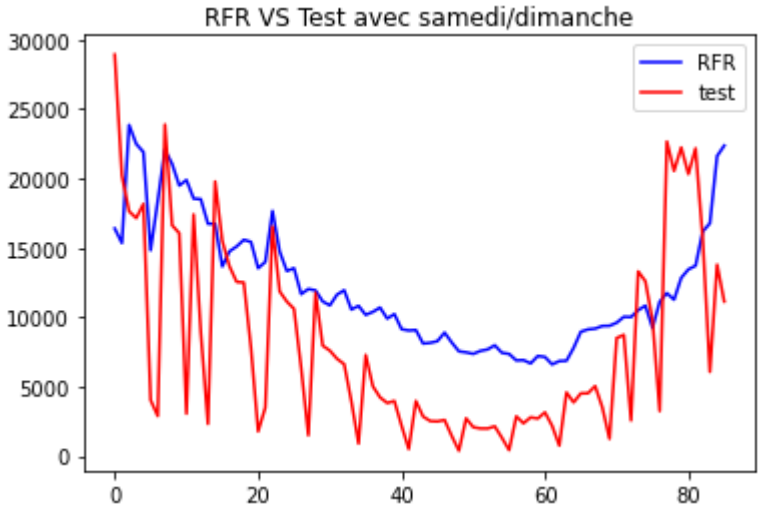
L'utilisation de XGBRegressor et la recherche des meilleurs paramètres a permis une nette amélioration des performances.

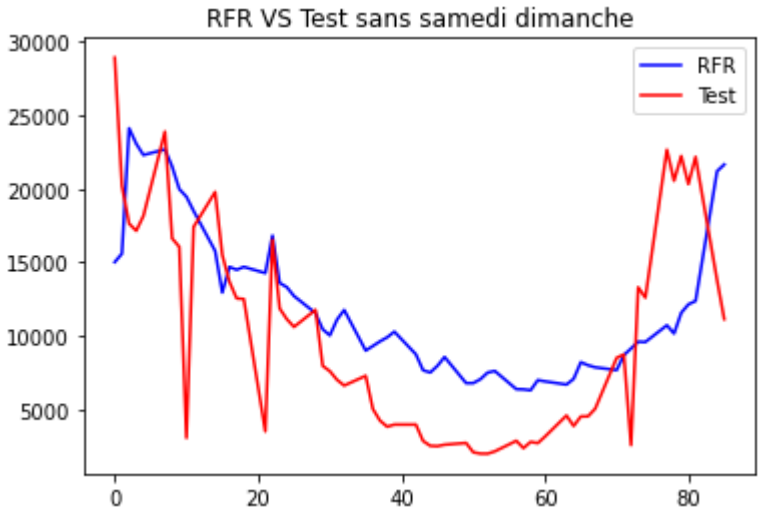
Avez-vous analysé les erreurs de votre modèle ?

On a pu constater sur les données de tests que les samedis et les dimanches causaient de gros pics négatifs très probablement à cause d'un nombre plus faible de tests ces jours-là.


XGB VS Test avec samedi et dimanche
Prédictions XGB
Test
30000
25000
20000
15000
10000
5000
0
0
20
40
60
80


XGB VS Test sans les dimanche et samedi
XGB
Test
30000
25000
20000
15000
10000
5000
0
20
40
60
80





Cela a-t-il contribué à son amélioration ? Si oui, décrivez.

La suppression des samedis et dimanches et le réentraînement du modèle sur les nouvelles données a permis d'obtenir une MAE encore significativement plus faible (3593.244 au lieu de 4836.806).

Détaillez quelle a été votre contribution principale dans l'atteinte des objectifs du projet.

Cedric : recherche de data, data viz (JDD 1 et JDD2) & data preprocessing
Carine : data viz (JDD final) & recherche de data
Victor : recherche & code sur les modèles, l'algorithme MICE

Description des travaux réalisés

Répartition de l'effort sur la durée et dans l'équipe

cf Diagramme de Gantt en annexe

Bibliographie

Sur quels éléments bibliographiques (articles de recherches, blog, livres, etc...) vous êtes-vous appuyé pour réaliser votre projet ?

informations sur les time series multivariées :
https://www.analyticsvidhya.com/blog/2018/09/multivariate-time-series-guide-forecasting-modeling-python-codes/

Difficultés rencontrées lors du projet

Quel a été le principal verrou scientifique rencontré lors de ce projet ?

Le virus responsable de la Covid-19 a été identifié il y a presque 2 ans maintenant. Il est clair que de nombreuses inconnues existent à son propos, qui pourraient améliorer l'exactitude de la prédiction des modèles de machine learning.
De plus, des facteurs environnementaux et biologiques accélérant la circulation ont été mis en évidence au fur et à mesure de l'année 2020 écoulée. Par exemple, des variants ont été mis en évidence mais leur pourcentage au sein de la population n'a été enregistré qu'ultérieurement. Nous disposons donc de très peu de données concernant certains facteurs pouvant jouer sur les contaminations, que nous n'avons donc pas pu prendre en compte dans notre modèle.

Pour chacun des points suivants, si vous avez rencontré des difficultés, détaillez en quoi elles vous ont ralenti dans la mise en place de votre projet :

Prévisionnel : (tâche qui ont pris plus de temps que prévu etc ….)

Le choix des données et la définition d'un objectif clair. Un très grand nombre de jeux de données différents étant accessibles, un choix des données les plus pertinentes a dû être fait. De la même façon choisir une variable cible appropriée à nécessité quelques essais.

Jeux de données : (Acquisition, volumétrie, traitement, agrégation etc….)

Non.

Compétences technique / théoriques : (Timing d'acquisition des compétences, compétence non proposée en formation etc…)

Les modèles basés sur les times séries se sont avérés trop difficiles à implémenter ce qui nous à pousser à utiliser les modèles RandomForest et XGBoost.

Pertinence : (de l'approche, du modèle, des données etc ...)

Non.

IT : (puissance de stockage, puissance computationnelle, etc….)

Non.

Autres

Non.

Bilan & Suite du projet

En quoi votre projet a-t-il contribué à un accroissement de connaissance scientifique ?

L'épidémie de Covid-19 est encore un phénomène très récent et si de plus en plus de modèles de machines learning apparaissent sur internet, leur nombre est encore limité et il est encore difficile de savoir quelles approches de prédictions sont les plus efficaces (modèle de ML ou deep learning utilisé, nature des données choisie).
Dans ce contexte, notre projet propose une approche de plus.

Pour chacun des objectifs du projet, détaillez en quoi ils ont été atteints ou non.

Notre objectif était de créer un modèle prédictif de l'évolution de l'épidémie de Covid-19 en France.
Pour cela nous avons choisi de nous concentrer sur la prédiction du nombre de cas positifs, spécifiquement en créant la variable pos+5 qui représente le nombre de cas positifs à J+5 (durée moyenne de l'incubation).
Nous avons parmi toutes les données disponibles sur le site du gouvernement choisi de combiner les deux jeux de données qui nous paraissaient les plus appropriés pour notre prédiction.
Nous avons testé deux modèles de ML, RandomForest et XGBoost. Le deuxième modèle (XGBoost) obtient de meilleurs résultats que le premier avec une MAE plus faible et il permet de capter la tendance des données.

S'ils ont été atteints, dans quel(s) process(es) métier(s) votre modèle peut-il s'inscrire ? Détaillez.

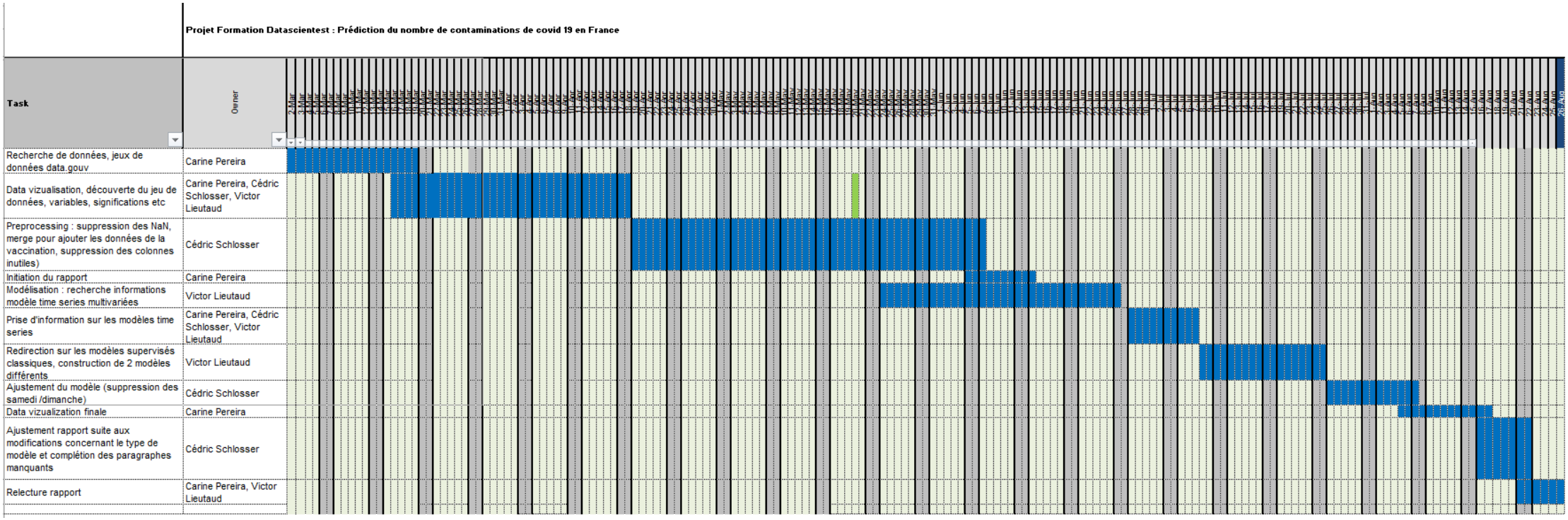
Une bonne prédiction du nombre de contaminations à 5 jours permet de mieux connaître la dynamique de l'épidémie, si de nouvelles contraintes doivent être mises en place ou tester l'efficacité de ces contraintes dans le temps.
On peut donc imaginer une inscription dans un process décisionnel pour les gouvernements, les laboratoires pharma qui doivent produire des vaccins, des hôpitaux… à condition d'arriver à augmenter la frontière temporelle à plus de 5 jours.

Dans le cas contraire, quelles pistes d'amélioration suggérez-vous pour améliorer les performances de votre modèle ?

Comme évoqué précédemment, les samedi et dimanches ont causé quelques problèmes, supprimer les jours fériés serait une bonne idée. Une autre alternative à tester serait de « flaguer » les jours non travaillés en ajoutant une nouvelle colonne binaire plutôt que les supprimer.
Les réseaux de neurones récurrents auraient également pu être testés (cette notion a été vue à la fin de la formation, nous n'avons donc pas eu le temps de tester cela).

Annexes

Diagramme de gantt



Description des fichiers de code

Traitement_des_Donnees.ipynb :
Fichier permettant la création et le nettoyage du jeu de données à partir des deux fichiers téléchargés depuis les pages suivantes :
https://www.data.gouv.fr/fr/datasets/synthese-des-indicateurs-de-suivi-de-lepidemie-covid-19/
https://www.data.gouv.fr/fr/datasets/donnees-relatives-aux-personnes-vaccinees-contre-la-covid-19-1/
Le processus de scaling est effectué dans le fichier pipeline_modele.py.

Data_viz.ipynb :
Fichier permettant l'analyse des données à l'aide de graphiques.

pipeline_modele.py :
Fichier permettant l'entraînement et le paramétrage des deux types de modèles (RandomForestRegressor et XGBRegressor)