DATA INSIGHTS

DATA LIFE & HEALTH

Risque absentéisme et Forecasting

Pourquoi s'intéresser au risque absentéisme ?

Comme présenté dans notre publication sur **le pilotage de l'absentéisme** (**Disponible ici**), l'absentéisme est devenu **un enjeu préoccupant et croissant dans le monde du travail.** Avec des objectifs de réduction des coûts liés aux absences pour les entreprises et des enjeux de réduction des risques pour les assureurs, l'absentéisme peut être **étudié, modélisé et anticipé** grâce à l'exploitation **des données DSN (Déclaration Sociale Nominative)**. L'étude de l'absentéisme passe par la création d'indicateurs permettant de quantifier et qualifier le risque au sein d'un portefeuille de salariés. **L'indicateur le plus connu reste le taux d'absentéisme**, un indicateur global qui permet d'avoir une vision générale et synthétique d'un portefeuille de salariés en termes d'absentéisme.

$$Taux\ d'absentéisme = \frac{Nombre\ de\ jours\ d'absence}{Nombre\ de\ jours\ d'exposition\ des\ salariés}$$

La modélisation du risque absentéisme

A partir de diverses informations (caractéristiques des entreprises et des salariés, consommation santé, données Open Data, etc), une modélisation du risque absentéisme est possible au travers de la modélisation d'une variable cible (le taux d'absentéisme par exemple). La Data Science et les différents **algorithmes prédictifs** à disposition (GLM, méthodes par arbres, réseaux de neurones...) permettent de **modéliser et d'expliquer le risque** au travers des variables explicatives ajoutées dans le modèle. La segmentation des individus en **profils de risque homogènes** via les arbres CART permet d'appréhender le risque en déterminant des profils plus ou moins sujets aux absences. Des modélisations sophistiquées telles que les Random Forest ou XGBoost permettent d'identifier des **variables d'importance marqueurs du risque absentéisme.**

Des modèles traditionnels au Forecasting...

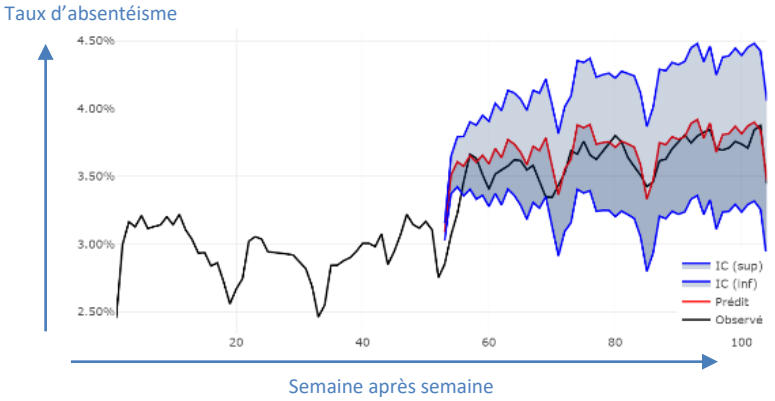
L'absentéisme comporte **une dimension temporelle difficilement prise en compte par les modèles classiques de Machine Learning**, réputés « statiques » dans l'apprentissage. L'ajout du caractère dynamique du risque absentéisme dans les modèles est donc nécessaire afin de pouvoir prédire le risque dans son intégralité. La méthode de **Forecasting** entre donc en jeu, privilégiant une méthode de prédiction moyenne des données futures **basée sur l'étude des données passées et de leur dynamique**.

Forecasting : utilisation des modèles de séries temporelles

Etudier la dynamique du risque au cours du temps revient donc à étudier sa **série temporelle**. L'étude peut porter sur l'évolution du taux d'absentéisme au cours du temps, un taux qui peut être calculé annuellement, trimestriellement, mensuellement ou de manière hebdomadaire. La modélisation de cet indicateur passe par l'utilisation des **modèles de séries temporelles** comme les modèles **ARMA ou ARIMA** prenant en compte l'évolution passée des données :

$$Y_t = \mu + \sum_k Y_{t-k}\varphi_k + \sum_j \varepsilon_{t-j}\theta_j$$

Des modèles de Machine Learning préalables ont permis de **connaître les variables les plus influentes** dans la prédiction du risque absentéisme. L'intérêt d'ajouter ces données est d'ajuster le modèle de série temporelle afin de donner **une prédiction moyenne dans le futur expliquée par des co-variables influentes** mises sous forme de séries temporelles. Ces nouvelles variables explicatives nommées **régresseurs,** permettent ainsi d'ajuster les prédictions au cours du temps, dans un nouveau modèle appelé **ARIMAX** (ou SARIMAX si l'étude de la saisonnalité de la série est prise en compte).



Contexte actuel

Le contexte actuel de pandémie vient perturber l'évolution habituelle du risque absentéisme au cours du temps. Le Forecasting par l'introduction de *régresseurs* apporte une nouvelle manière de l'appréhender. La prise en compte de **nouvelles variables temporelles** telles que les périodes de confinement, de couvre-feu, ainsi que des indicateurs de l'avancée de la pandémie, permet de suivre la nouvelle dynamique de ce risque et de **le projeter dans le temps sous ces différentes hypothèses.**

Cette publication a été réalisée sous la direction de
Nabil RACHDI, Head of Data Science
avec l'expertise de :
Alexandra BARRAL, Senior Manager
Jean-Pascal HERMET, Consultant
Fabien TRAVAILLOT, Consultant

addactis
THE RISKTECH FOR INSURANCE