Diapositive 1.

Bonjour,

Aujourd'hui je vais vous parler du travail que j'ai effectué lors de mon alternance chez AON France, qui consiste en l'Exploitation des données DSN pour des études d'absentéisme, voici le plan de notre présentation, on commence par…

Diapositive 2 Sommaire

Diapositive 3

C'est parti, pour vous mettre en contexte, parlons un peu de l'absentéisme de pourquoi il est devenu un problème compliqué à gérer dans les entreprises.

Diapositive 4

L'absentéisme est défini comme l'absence d'un salarié pendant une période donnée, sans en mentionner la cause ni la durée. Il s'agit d'un problème organisationnel qui ne connaît pas de frontières et qui touche tous les pays, affectant les sociétés au niveau de productivité et des coûts.

Pour mesurer l'absentéisme, il existe un indicateur appelé taux d'absentéisme, qui est calculé comme le rapport entre le nombre de jours d'absence d'un salarié et le nombre de jours de présence dans l'année.

Il est intéressant d'étudier l'évolution de cet indicateur au fil des années, pour cela on prend les études faites par Ayming France dans un échantillon des entreprises privées Françaises qui montrent l'augmentation qui a eu pendant les années et particulièrement comment s'est explosée à l'arrivée du COVID en 2020, à cause des détériorations en matière de santé des salariés et des modes du travail. On passe de 4 jours d'absence sur 100 a 7 d'absence sur 100 jours.

Cette augmentation a forts impacts économiques sur les sociétés, depuis le maintien du salaire des salariés absents et, si l'absence est trop longue, le remplacement du salarié

Mais y'a aussi des impacts qui ne sont pas possibles de quantifier, comme les affectations sur l'organisation de l'entreprise, la performance des équipes. Ces impacts coutent cher, c'est pour cela qu'on cherche à diminuer ces coûts.

C'est pour cela que l'objectif de mon alternance c'est d'identifier les populations a fort risque d'absentéisme et d'analyser les causes de ces risques là pour pouvoir proposer des pistes d'amélioration.

Diapositive 5

Pour attendre le but de l'étude on a divisé notre projet dans 2 grandes étapes.

La première c'est de traiter nos données source et construire notre base des données, On utilise le flux DSN, qui veut dire la déclaration sociale nominative, que je vais vous expliquer en détail qu'est-ce que c'est un peu plus tard dans la présentation.

Et la deuxième partie est la modélisation l'absentéisme et l'identification de groupes des salariés à risque.

Diapositive 6

Ensuite on continue avec la section description et traitement des données.

Diapositive 7

La DSN, déclaration sociale nominative, qu'est-ce que c'est ça. Au fait, avant 2017 si une entreprise devait faire et suivre des déclarations auprès de plusieurs organismes,

(par Example le pôle d'emploi, CNAV, CNAM et des assurances, entre autres).

Elle aurait dû envoyer des flux pour chaque organisme avec peut-être des formats qui ne sont pas les mêmes et en différentes périodes de l'année. Cette charge administrative est très compliquée à gérer, cela fait une charge lourde pour les entreprises.

C'est pour cela que à partir de 2017 l'état a mis en place un système appelé DSN.

Dans ce système l'entreprise mis en place un réceptacle qui centralise touts les flux et puis ce flux-là est envoyé charque mois a tous les organismes.

Un flux pour tout le monde, tous les organismes et acteurs sociaux ont accès aux mêmes données. Et toutes les informations sont trouvées là-dedans.

A partir de ce flux, en tant qu'assureur AON va chercher les informations pour pouvoir exploiter les études d'absentéisme.

Diapositive 8

Maintenant on va à voir à quoi ressemble un flux DSN, vous voyez, cette partie à gauche c'est un tableau compose par couples des clés et valeurs.

Ces valeurs et clés sont organisées dans un système de blocs, on peut voir que chaque bloc correspond à une information, par exemple, le bloc 0 correspond à identification de logiciel, le bloc 6 c'est les informations sur l'entreprise, le bloc 11 sont les informations sur l'établissement

On peut considérer La DSN comme paries des clés et valeurs, et derrière il faut qu'on trouve les informations correspondantes qui sont reparties entre blocs pour pouvoir faire notre transformation pour finalement construire nos bases des données

Une fois fait ça, on récupère les informations qui nous intéressent pour l'étude.

Diapositive 9

Nous avons fait un programme qui suit cette logique et transforme ces fichiers texte dans des tables,

tous ces tables sont liés uns avec les autres par des clés.

Le traitement est fait sur le logiciel Python. Là on traite une grande taille des données parce qu'on trouve des flux des tous les mois avec une grande masse des informations.

il faut alors, traiter en détail ces informations, construire de tables et fiabiliser les données.

A la fin de cette procédure on finit avec quinze fichiers ou encore plus, desquels on essaie de prendre des informations d'importance pour pouvoir à la fin obtenir nos deux tables finales, une qui a tout sur la démographie et une autre d'Arrêt de travail, (listar una o dos variables de la lista)

Diapositive 10

On a traité les données, maintenant on va choisir un client dans la liste faire une étude sur lui. On trouve ici les démographiques de ce client-là, c'est un client du secteur agroalimentaire qui a à peu près 6K salariés, la majorité- sont des hommes avec une Age moyenne de 42 ans, ancienneté moyenne de 9 ans dans la société, 1 salarié sur deux sont des ouvriers,

Diapositive 11

Ensuite, quelques chiffres de l'absentéisme de cette société, en 2020 y a 6,3% de taux d'absentéisme et puis cela augmente considérablement en 2021, ça passe a 7,2%. En regardant le détail on trouve que 40% des salariés sont absents, cela veut dire que 4 salariés sur 10 qui sont absents au moins une fois dans l'année, pour chaque salarie absent, la fréquence moyenne est à peu près de 2 fois par an, la durée moyenne de chaque absence est de 30 jours, et ça cout par jour à la société à peu près 200euros par salarié absent.

Diapositive 12

Ensuite on continue avec la section de modélisation de l'absentéisme.

Diapositive 13

En ce moment-là, on a vu et traité les données, on a sélectionné le client sur lequel on va tester on a déjà une base des données sur deux années, 2020 et 2021, là l'idée c'est que je vais travailler sur les données de 2020 et après je vais tester sur les données de 2021 pour voir si mon modèle est toujours adapté.

Alors, pour la modélisation on passe par 6 étapes, Voilà, j'ai construit mes modèles sur la base des données de 2020, j'ai appliqué les trois modèles GLM, CART, RANDOM FOREST sur ces données-là, puis je vais choisir un des trois modèles selon les résultats obtenus et après j'essaie d'améliorer le modèles et de retester, cela c'est toujours sur les données de 2020.

Ensuite je choisis les variables significatives parce que on a plein de variables comme le sexe, la fonction, entre autres,. Alors, je dois choisir les variables significatives qu'impactent l'absentéisme pour mon modèle

Après cette étape-là, j'applique mon modèle aux données de 2021, et comme conclusion je veux déterminer si mon modèle tient bien sur des nouveaux données, les données de 2021

Diapositive 14.

Avant de montrer les résultats des modèles, je vais faire un petit rappel des inconvénients et avantages des modèles utilisés.

- GLM est un modèle paramétrique, les deux derniers sont des modèles non paramétriques

- Le modèle paramétrique c'est facile à appliquer mais il faut que les données suivent des lois statistiques sinon le modèle ne prédit pas bien.

- Après les deux-là sont des modèles non paramétriques,
- Pour le premiere, on a l'avantage qu'il ne faut pas que les données suivent des lois mais c'est possible que ce modelle ne marche pas très bien sur des nouvelles données, donc c'est à voir si les résultats obtenus avec celui si sont satisfaisants ou pas.

- Et pour le dernier, on a aussi l'avantage qu'il ne faut pas que les données suivent des lois, mais cette approche la prend beaucoup de temps pour exécuter.

Diapositive 15

Ici, nous observons les résultats des modèles appliqués, nous pouvons voir qu'il n'y a pas une grande différence dans la performance des trois modèles, cependant, la précision de la forêt aléatoire est un peu meilleure par rapport aux autres. On choisit donc d'améliorer cette performance en optimisant les paramètres.

Diapositive 16

Ensuite, nous observons les résultats obtenus en général par l'approche du random forest, passant d'une performance de 69% dans une première tentative à une performance de 71% après avoir effectué un processus de test sur de nombreuses combinaisons de paramètres afin de trouver la plus performante.

Avec notre modèle entraîné et testé sur les données de 2020, on s'est posé la question de si ce modèle est capable de classifier correctement les nouvelles données, spécifiquement les tendances de l'année 2021, obtenant finalement des résultats tout à fait satisfaisants, montrant ainsi l'adaptabilité du modèle aux nouvelles informations.

Diapositive 17

À partir de notre modèle construit, nous pouvons identifier certaines données clés, telles que les variables les plus importantes lors de la formation du modèle. Cette approche est utile pour l'étape suivante, qui consiste à déterminer que les données les plus importantes pour déterminer l'absence d'un employé sont l'âge, l'ancienneté, la fonction, le salaire et la région où se trouve l'employé. Nous gardons donc ces variables pour les étudier dans l'étape suivante.

Diapositive 18

Ensuite on continue avec la section de modélisation de l'absentéisme.

Diapositive 19

voici les étapes suivies :

Pour identifier les populations à fort risque on utilise le model CAH, classification hiérarchique sur les données obtenues, après qu'on applique ce modèle on obtient un dendrogramme (qui est un arbre) sur lequel on va regarder et puis on peut choisir le nombre de groupes qu'on veut retenir et après on choisit le nombre de groupes on analyse les caractéristiques de ces groupes là pour identifier vraiment des populations à risque.

Diapositive 20

Comme nous l'avons dit précédemment, le dendrogramme me donne cinq groupes différents, chacun des groupes va être codé avec une couleur, d'après notre analyse il est possible déterminer quels sont les groupes avec le fort risque d'absentéisme.

Et le risque de chacun est classé en trois niveaux.

Très faible, faible et Risque élevé.

Et ils ont les caractéristiques suivantes :

- Le groupe rose est composé de dirigeants et de cadres, ils sont dans la catégorie **de risque la plus faible.**
- Le groupe bleu est composé de jeunes employés, âgés d'un an à trois ans, qui ne sont pas cadres. Ils font **partie de la deuxième catégorie de risque.**
- Dans le groupe jaune, qui est le plus important, la majorité des ouvriers et des employés administratifs ayant une ancienneté et un âge élevés, situés en région parisienne, se trouvent dans la deuxième catégorie de risque. Ces derniers présentent notamment le **risque le plus élevé.**
- Le groupe de gauche, le groupe orange. Là encore, on trouve des jeunes, non cadres, mais avec moins d'un an d'ancienneté et avec un **faible risque.**
- Enfin, on trouve le groupe vert, composé d'ouvriers, de professions intermédiaires, comme les électriciens, les techniciens et autres, d'un âge plus avancé, situés en dehors de la région parisienne. Ils **présentent le risque le plus élevé avec le groupe jaune.**

Diapo 21

Après d'avoir identifié les populations à fort risque d'absentéisme. Que fait-on par la suite ? L'idée c'est de comprendre. Et pour comprendre on essaie de calculer les indicateurs sur les % des salariés absents, la fréquence d'absence, et la durée, cela parce qu'on aimerait savoir si les problèmes d'absence sont plus liés à la fréquence, ou plutôt a la durée. Tout cela est pour identifier les causes et puis comprendre si l'absence est à cause de problèmes d'organisation, ou si c'est la charge familiale ou plutôt les absences sont trop fréquentes ou c'est un problème de charge de travail. Tous ces étapes sont faites pour pouvoir proposer des solutions adaptées.

Si par exemple on a des problèmes d'organisation, on peut proposer des accompagnateurs e carrière, ou la formation des managers ou même des bonus pour les managers fonction du taux d'absentéisme de leur équipe.

Si les absences ont un fort indicateur de fréquence d'absences on peur proposer la mise en place de récompenses aux salaries pas absents pour leur motiver à venir. Ou si le problème son à cause de que les salariés sont absents très long temps on peut proposer des aides au retour d'emploi. Si c'est un problème de charge familiale on peut aussi proposer quelques solutions, come mettre en place des crèches en entreprise ou accords de télétravail, entre autres.

En gros les pas à suivre sont de calculer des indicateurs pour identifier les causes de l'absence pour mieux guider en proposant des stratégies adaptées.

Diapo 22

On arrive à la conclusion de l'étude. Alors, pour conclure :

1. On peut dire que la DSN est une grande source qui offre des données de multiples dimensions, cependant, en termes de facilité, elle est très complexe à traiter.

- L'exploitation de cette source offre des perspectives d'analyses pour les clients et surtout ça va monter la valeur ajoutée de AON vis-à-vis des clients

2. Par un autre côté, dans cette étude, la forêt aléatoire est considérée comme une bonne alternative aux modèles GLM, traditionnels dans le monde de l'assurance. Pourquoi ?

- Parce qu'ils sont faciles à configurer, et ont des niveaux de précision satisfaisants qui peuvent être améliorés avec des techniques d'optimisation.

3. Enfin, l'objectif de l'étude n'est pas seulement de mesurer l'absentéisme mais aussi de proposer des solutions adaptées à chaque cas.