Utilisation de l'apprentissage automatique pour prédire la qualité de service d'un centre d'appels-relais pour sourds et malentendants selon sa dotation en agents

ALSAMADI Samer[1], CELLIER Nicolas[1], MARTINEZ Cléa[1], PEHLIVAN Canan[1], FONTANILI Franck[1]
[1] IMT Mines Albi, Allée des Sciences 81000 Albi, +33767685563, prenom.nom@mines-albi.fr

Résumé. Pour évaluer si le nombre d'agents affectés à un centre d'appels permet d'atteindre la qualité de service visée, les travaux de recherche proposaient jusqu'à présent soit des modèles analytiques, soit des modèles de simulation. Grâce à l'abondance des données enregistrées dans les centres d'appels, il est aujourd'hui possible d'entrainer des modèles d'apprentissage automatique capables de prédire les performances du centre d'appel. Dans cet article, nous utilisons les données issues d'un centre d'appels-relais au service de la communauté des sourds et malentendants afin d'entrainer des modèles d'apprentissage automatique pour prédire la qualité de service selon le nombre d'agents alloués. Nous les comparons ensuite au modèle de file d'attente classique Erlang C, largement utilisé dans ce secteur d'activité. Les premiers résultats mettent en évidence la supériorité des modèles d'apprentissage automatique.

Resumen. Para evaluar si el número de agentes asignados a un centro de llamadas puede alcanzar la calidad de servicio deseada, la investigación ha propuesto hasta ahora modelos analíticos o modelos de simulación. Gracias a la abundancia de datos registrados en los centros de llamadas, ahora es posible entrenar modelos de aprendizaje automático capaces de predecir el rendimiento de los centros de llamadas. En este artículo, utilizamos los datos de un centro de llamadas que atiende a la comunidad de personas sordas y con dificultades auditivas para entrenar modelos de aprendizaje automático que permitan predecir la calidad del servicio en función del número de agentes asignados. A continuación, los comparamos con el modelo clásico de colas Erlang C, ampliamente utilizado en este sector de actividad. Los primeros resultados demuestran la superioridad de los modelos de aprendizaje automático.

Mots clés : Centre d'appels, apprentissage automatique, optimisation des hyperparamètres, évaluation des performances

Introduction

Les centres d'appel constituent un environnement riche en événements stochastiques pour les chercheurs qui souhaitent explorer la dynamique et la performance de ces systèmes. Parmi les domaines les plus étudiés dans les centres d'appel, on peut citer la prévision des appels entrants, ou encore la dotation et la planification des agents. Dans la littérature consacrée à la dotation et à la planification, les modèles analytiques de file d'attente et les modèles de simulation à événements discrets (SED) sont les méthodes les plus utilisées. Bien que ces deux types de modélisation aient leurs avantages, ils ne répondent pas à certains aspects nécessaires à la bonne planification des centres d'appels d'aujourd'hui. Par exemple, les modèles de file d'attente reposent sur des hypothèses sous-jacentes concernant le processus d'arrivée des appels, les temps de service et les temps de "patience" (temps d'attente avant abandon). Bien que ces hypothèses ne soient généralement pas tout à fait vérifiées dans les cas d'étude réels, elles fournissent des résultats qui s'avèrent satisfaisants auprès des

gestionnaires. Cependant, ces résultats sont discutables. Non seulement les performances de ces modèles sont très variables, mais ils ne permettent pas non plus de reproduire fidèlement le fonctionnement interne d'un centre d'appel moderne. Quant aux modèles à base de SED, ils ont démontré leur efficacité à évaluer les performances des centres d'appels. Malgré cela, en raison de la complexité croissante des structures des centres d'appels, certaines dynamiques peuvent être trop difficiles à modéliser correctement. Par ailleurs, l'utilisation de la simulation pour optimiser la dotation et la planification en personnel n'est pas très performante en raison de la durée d'exécution de chaque simulation qui peut s'avérer trop longue quand il s'agit de lancer plusieurs milliers de simulations.

L'enregistrement automatique de données étant une pratique adoptée par de nombreuses entreprises et entités de services, telles que les centres d'appels, il devient pertinent d'utiliser des modèles d'apprentissage automatique (Machine Learning = ML). Pour la dotation en agents, ces modèles basés sur des données historiques peuvent fournir des prédictions sur le futur, facilitant le processus de planification de l'offre nécessaire pour répondre à la demande à venir.
Ce travail de recherche consiste à exploiter les données disponibles de chaque appel téléphonique passé ainsi que les métriques d'évaluation de la performance, et en particulier la qualité de service, afin de construire des modèles de ML capables de déterminer la meilleure dotation en agents. Il est ensuite possible de planifier chaque agent sur l'horizon d'une journée de travail.

Ces travaux sont menés en collaboration avec un centre d'appels-relais pour les personnes handicapées qui concerne la communauté sourde et malentendante. Ce centre d'appels sera appelé ANGUS dans la suite de cet article. Grâce à la communication vidéo, les appelants sourds souhaitant contacter un correspondant entendant font appel à ANGUS qui leur fournit alors un agent dédié (interprète), capable de traduire presque instantanément la langue française parlée dans la langue des signes française (LSF) et vice versa. L'interprète sert de relais pour mettre en relation l'appelant sourd avec le correspondant entendant en fournissant une traduction quasi-simultanée dans les deux sens. ANGUS est un service ouvert en semaine, 9 heures par jour, entre 9h00 et 18h00. Le schéma de la figure 1 ci-dessous illustre ce service.

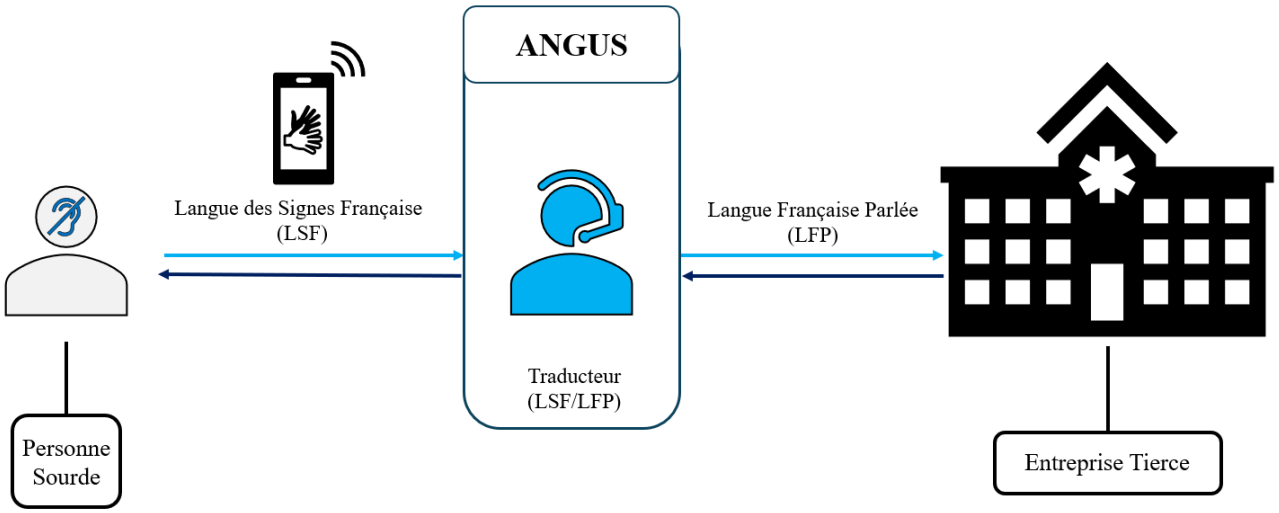

Figure 1 : Schéma des fonctionnalités d'ANGUS

1. Revue de la littérature

Ce domaine de recherche est très ancien, puisque les premiers modèles d'évaluation des performances utilisés dans le processus d'allocation des ressources d'un centre d'appels remontent à plus de cent ans (Erlang, A.K., 1917), lorsque le modèle Erlang C a été introduit pour la première fois. Ce modèle fournit une expression analytique permettant d'évaluer la qualité de service correspondant à une dotation en agents. Ce modèle a ensuite été modifié par (Palm, C., 1957) pour prendre en compte la possibilité d'abandon de l'appelant si son temps d'attente est trop long. Cette évolution est connue sous le nom de modèle Erlang A (A comme Abandon). Différents chercheurs ont ensuite travaillé sur des variantes de ces deux modèles, avec différentes hypothèses sous-jacentes sur le processus d'arrivée des appels, le temps de service et les distributions de temps de patience. (Tanir, O. et Booth, R.J., 1999) est l'un des premiers articles à utiliser la simulation à événements discrets (SED) pour l'évaluation des performances des centres d'appel. En faisant varier le niveau de dotation en personnel à chaque nouvelle simulation, les auteurs montrent qu'il est possible d'évaluer approximativement les performances d'un centre d'appels. Pour mener à bien un tel processus itératif, il convient d'adopter une approche de dotation en personnel. Cette approche consiste tout d'abord à choisir le critère d'évaluation (par exemple la qualité de service) puis ensuite de coupler un algorithme d'optimisation à la simulation qui va ainsi explorer et évaluer différentes solutions. (Cezik, M.T. et L'Ecuyer, P., 2008) utilise un modèle de SED avec un algorithme itératif de type "Branch and Bound" avec une programmation linéaire en nombres entiers (PLNE) pour résoudre le problème de dotation. (Gurvich, I., Luedtke, J. et Tezcan, T., 2010) utilise le modèle Erlang A dans le cadre d'un algorithme d'optimisation de la dotation par programmation stochastique. (Zan, J., Hasenbein, J.J., Morton, D.P. et Mehrotra, V., 2018) utilise le modèle Erlang C dans le cadre d'un programme d'optimisation stochastique comportant deux étapes.

Dans cet article, nous avons utilisé différents modèles, aussi bien les plus connus et répandus issus de la théorie des files d'attente (Erlang) que des modèles plus récents basés sur le ML, afin de comparer leurs performances lors d'une journée planifiée. À notre connaissance, le travail de (Li, S., Wang, Q., Koole, G., 2020) est la seule étude récente sur l'utilisation de modèles à base de ML pour la planification des agents des centres d'appels. Les auteurs proposent un modèle de ML qui est entraîné sur les résultats fournis par la simulation, ce qui permet d'accélérer le processus d'évaluation de la qualité de service (QoS).

2. Optimisation du modèle

Tout d'abord, comme pour tout projet basé sur le ML, un prétraitement des données doit être réalisé pour garantir l'intégrité de la base disponible. Une fois les algorithmes de détection des anomalies et les techniques de remplacement des données manquantes mis en œuvre, nous avons pu passer aux phases de construction et d'optimisation du modèle. Nous avons effectué une analyse comparative de différents modèles de régression numériques et basés sur le ML afin d'identifier le modèle le plus performant pour prédire la qualité de service sur chaque créneau horaire de 30 minutes. Les jours enregistrés dans la base de données sont divisés en tranches horaires de 30 minutes, de l'heure d'ouverture à l'heure de fermeture d'ANGUS, soit un total de 18 tranches horaires. Notre objectif est de prédire la qualité de service d'un créneau horaire donné, pour une allocation donnée d'interprètes. Les caractéristiques du modèle sont établies en utilisant les données existantes sur le nombre d'agents, les temps de service, et les temps de patience. Ces trois variables spécifiques ont été choisies en raison de leur effet connu dans la littérature sur le processus de planification d'un centre d'appels. Une analyse plus poussée de la corrélation entre ces colonnes de données et la variable cible, la qualité de service, indique le pouvoir prédictif potentiel de ces colonnes en matière de prédiction de la qualité de service.

Différents algorithmes ont été utilisés pour effectuer le processus de régression lors de la prédiction de la qualité de service d'une équipe d'agents donnée, et différents hyperparamètres régissent le fonctionnement de ces algorithmes. Il est donc crucial d'optimiser correctement ces hyperparamètres. Ce processus d'optimisation est réalisé à l'aide de la bibliothèque python HyperOpt introduite dans (Bergstra, J., Bardenet, R., Bengio, Y. et Kégl,

B., 2011). Au lieu d'effectuer une exploration de l'espace des solutions de type "*grid search*" dans laquelle toutes les combinaisons possibles d'hyperparamètres sont testées, nous avons choisis d'utiliser un algorithme d'optimisation de type bayésien qui ne teste que certains points de l'espace des solutions possibles en spécifiant un nombre fixe d'itérations. En ce qui concerne les mesures d'erreur entre la qualité de service réelle issue de la base de données du centre d'appels et la qualité de service issue de la prédiction par un modèle de ML, nous avons utilisé le pourcentage d'erreur absolue moyenne pondérée (Weighted Mean Absolute Percentage Error = WMAPE) afin d'éviter les problèmes qui peuvent survenir quand les valeurs de qualité de service sont proches de zéro. Afin de généraliser le modèle choisi et donc de valider qu'il fonctionne bien, non seulement pour la journée de planification en cours, mais aussi pour l'ensemble des données disponibles, nous avons utilisé une forme de validation croisée pour les données de séries temporelles, connue sous le nom de "backtesting". Dans le cadre de cette technique, les données disponibles sont divisées en plusieurs sections, après quoi nous entraînons nos modèles de prédiction de manière itérative sur une combinaison de ces sections afin de prédire la section de données suivante. Cela permet d'assurer une généralisation correcte du modèle et donc de garantir son pouvoir prédictif sur l'ensemble des données. Les mesures d'erreur obtenues pour chaque itération de formation sont enregistrées et leurs valeurs moyennes sont indiquées afin d'évaluer la performance d'un modèle. En ce qui concerne les modèles, nous avons comparé le modèle de file d'attente Erlang C, largement répandu et qui sert de référence pour ce travail, un modèle de régression linéaire multiple (MLR), un modèle de réseau neuronal artificiel (ANN), et un modèle d'arbre de décision Light Gradient Boosting (LGB).

3. Résultats

Les caractéristiques utilisées dans ces modèles sont générées à partir des valeurs des appels entrants, du nombre d'agents en poste, des temps de service moyens (AST), des temps de patience moyens (APT) et la qualité de service prédite pour le créneau horaire précédent (QoS_{lag}). Les hyperparamètres choisis pour être optimisés sont ceux qui sont notamment connus pour affecter les résultats des prédictions des modèles. Les résultats de l'optimisation des modèles sont présentés dans le tableau 1 ci-dessous. Le modèle Erlang C n'a pas d'hyperparamètres à optimiser puisque cela concerne uniquement les modèles à base de ML. On constate que le modèle Erlang C est celui qui présente les erreurs les plus importantes (WMAPE = 20,33) parmi l'ensemble des modèles testés. Le modèle LGB est le plus performant en termes d'erreur absolue moyenne, sachant que les modèles MLR et ANN fournissent aussi de très bons résultats. L'objectif final étant de générer des effectifs d'agents, nous sommes alors en mesure d'intégrer le modèle qui s'est avéré le plus performant en termes de prédiction de la qualité de service dans notre algorithme de dotation. Basé sur les exigences de QoS déterminées par l'administration du centre d'appel, cet algorithme sert d'heuristique qui explore l'espace des solutions possibles de dotation en agents et évalue la performance attendue de chacune de ces solutions en utilisant le modèle d'évaluation de la QoS que nous avons conçu tout au long de cet article.

Model	Caractéristiques	WMAPE
Erlang C	Calls, Agents, AST	20,33
MLR	Calls, Agents, AST, APT, QoS_{lag}	13,72
ANN	Calls, Agents, AST, APT, QoS_{lag}	13,94
LGB	Calls, Agents, AST, APT, QoS_{lag}	13,49

Tableau 1 : Résultats de l'optimisation des hyperparamètres du modèle

4. Conclusion

Dans cet article, nous avons évalué la prédiction de la qualité de service d'un centre d'appels pour la communauté des sourds et malentendants en comparant un modèle classique de file d'attente (Erlang C) avec différents autres modèles à base de ML. Les premiers résultats démontrent une supériorité des modèles de ML. La prochaine étape de nos travaux de recherche consistera donc à tester l'utilisation de ces différents modèles pour déterminer la dotation en agents, afin de vérifier si cette amélioration de la prédiction de la QoS peut se traduire par une amélioration de la dotation en agents.

5. Réferences

Bergstra, J., Bardenet, R., Bengio, Y. and Kégl, B., 2011. Algorithms for hyper-parameter optimization. Advances in neural information processing systems, 24.

Cezik, M.T. and L'Ecuyer, P., 2008. Staffing multiskill call centers via linear programming and simulation. Management Science, 54(2), pp.310-323.

Erlang, A.K., 1917. Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges. Post Office Electrical Engineer's Journal, 10, pp.189-197.

Gurvich, I., Luedtke, J. and Tezcan, T., 2010. Staffing call centers with uncertain demand forecasts: A chance-constrained optimization approach. Management science, 56(7), pp.1093-1115.

Li, S., Wang, Q., Koole, G., 2020. Optimal contact center staffing and scheduling with machine learning. Working paper.

Palm, C., 1957. Some observations on the Erlang formulae for busy-signal systems. TELE, 1, pp.1-168.

Tanir, O. and Booth, R.J., 1999, December. Call center simulation in Bell Canada. In Proceedings of the 31st conference on Winter simulation: Simulation---a bridge to the future-Volume 2 (pp. 1640-1647).

Zan, J., Hasenbein, J.J., Morton, D.P. and Mehrotra, V., 2018. Staffing call centers under arrival-rate uncertainty with Bayesian updates. Operations Research Letters, 46(4), pp.379-384.