



ÉCOLE NATIONALE
DES SCIENCES
GÉOGRAPHIQUES



École Nationale des
Sciences Géographiques

Galigeo

Rapport de stage

Cycle des Ingénieurs diplômés de l'ENSG 3^{ème} année

Stage de fin d'étude ENSG
Estimation, analyse et prédition de flux piétons
BigData et Machine Learning



ÉCOLE NATIONALE
DES SCIENCES
GÉOGRAPHIQUES

Jules Pierrat

Septembre 2022

Non confidentiel Confidential IGN Confidential Industrie Jusqu'au ...

ÉCOLE NATIONALE DES SCIENCES GÉOGRAPHIQUES
6-8 Avenue Blaise Pascal - Cité Descartes - 77420 Champs-sur-Marne
Téléphone 01 64 15 31 00 Télécopie 01 64 15 31 07

*Ne pas imprimer svp.
Préférez la lecture sur écran*

Jury

Président de jury :

Victor Coindet Professeur de l'ENSG, Responsable du cycle TSI

Commanditaire :

M. Sébastien Connesson, COO de Galigeo

Encadrement de stage :

M. Jean-Michel Gaudin, Responsable du pôle Recherche et Développement à Galigéo

Enseignant référent :

M. Loïc Landrieu, Chercheur MATIS / IGN, Professeur de l'ENSG

Rapporteur expert :

qui est rapporteur du mémoire ?

Responsable pédagogique du cycle Ingénieur - TSI :

Victor Coindet Professeur de l'ENSG, Responsable du cycle TSI

Gestion du stage :

Delphine Genès, Relation entreprise de l'ENSG

Stage de fin d'étude du 2 mai 2022 au 28 Octobre 2022

Diffusion web : Internet Intranet Polytechnicum Intranet ENSG

Situation du document :

Rapport de stage de fin d'études présenté en fin de 3^{ème} année du cycle des Ingénieurs

Nombres de pages : 31 pages dont 4 d'annexes

Système hôte : L^AT_EX

Modifications :

EDITION	REVISION	DATE	PAGES MODIFIEES
1	0	09/2022	Création

Remerciements

Avant toute chose, je tiens à remercier le lecteur pour l'intérêt qu'il porte à mon rapport et j'espère qu'il trouvera ici tout ce pourquoi il est venu. Je veux remercier également les personnes m'ayant permis de réaliser dans les meilleures conditions ce stage ainsi que celles ayant contribué à l'élaboration de ce rapport.

Tout d'abord, j'adresse mes remerciements à mon professeur, **Mr Loïc Landrieux de l'Ecole Nationale des Sciences Géographiques**, mon maître de stage qui m'a suivi tout au long de ce stage, m'a guidé et éclairé dans mes décisions.

Je tiens à remercier mon maître de stage, **Jean-Michel Gaudin, Responsable du pôle recherche et développement à Galigeo** pour son suivi et l'intérêt qu'il a porté à mes travaux réalisés pendant le stage. Je remercie également **Raimana Teina, Data Scientist chez Galigeo** qui m'a guidé dans mes travaux et grâce à qui j'ai énormément progressé et appris durant toute ma période de stage. Mes remerciements vont également à **Sébastien Connesson, COO de Galigeo**, qui m'a permis de comprendre au mieux l'organisation de l'entreprise, les relations internes, les enjeux et les rapports aux clients.

Je remercie évidemment tout le reste de **l'équipe de Galigeo** pour son accueil, la confiance qu'ils m'ont accordée, leurs conseils et leur bienveillance. Je suis très heureux d'avoir pu travailler avec eux et me réjouis de continuer à le faire.

Enfin, après ces trois années fabuleuses je tiens à remercier toutes les personnes qui ont croisé mon chemin à **L'Ecole Nationale des Sciences Géographiques, mes professeurs, mes amis** et toutes les rencontres qui m'ont permis de grandir et de me préparer à cette nouvelle vie après les études.

Enfin, je tiens à remercier toutes les personnes qui m'ont conseillé et relu lors de la rédaction de ce rapport de stage : **ma famille, mon ami Antoine Rainaud** camarade de promotion.

Résumé

Ceci est mon résumé bla bla bla

Mots clés : clés, clés, clés

Résumé

This is my abstract blah blah blah...

Key words : key, key, key

Table des matières

Remerciements	3
Glossaire et sigles utiles	3
Introduction	4
1 Galigeo Ensg	5
1.1 Présentation de l'entreprise	5
1.2 Les objectifs de Galigeo	7
1.3 Organisation du stage	8
2 Le stage Ensg	10
2.1 Généralités	10
2.2 Prédition de flux piéton	10
2.3 Prédition de chiffre d'affaire	18
2.4 Autres missions	19
3 Bilan Ensg	21
3.1 Bilan de mes travaux	21
3.2 Bilan du stage	21
3.3 Bilan des trois ans à l'ENSG	21
Conclusion	22
A Planning du stage	29
B Statistiques DNN	31

Glossaire et sigles utiles

BI Business Intelligence

CNN Convolutional Neural Network

CEO Chief Executive Officer

COO Chief Operating Officer

CRM Customer Relationship Management

DNN Deep Neural Network

ENSG École Nationale des Sciences Géographiques

INSEE Institut national de la statistique et des études économiques

IRIS Îlots Regroupés pour l'Information Statistique

LSTM Long Short Term Memory

ML Machine Learning

POI Point of Interest

RetD Recherche et Développement

RNN Recurrent Neural Network

SaaS Software as Service

SVM Support Vector Machine

Introduction

Le géomarketing ou Location Business Intelligence en anglais est un pilier du marketing. Il étudie la variation des marchés dans l'espace. Les objectifs sont de modéliser offres et demandes en fonction de données économiques, sociales, culturelles, administratives et démographiques et leurs variations en fonction des géographies.

C'est un domaine essentiel pour les entreprises qui cherche à développer leurs espaces d'action. En effet, c'est une solution qui aide à la prise de décision pour le développement d'un business. Il permet de choisir les sites stratégiques les plus appropriés pour planter un nouveau commerce. La réalisation de modèle ou de simulation sont des outils essentiels en vue de comparer les atouts et risques d'une future implantation. En étudiant les espaces entourant toutes ses enseignes, une entreprise peut également anticiper la cannibalisation¹ ou la segmentation des portefeuilles² en prenant tout les paramètres réunis en un point de l'espace.

Il permet également d'établir des stratégies de marketing rentables et efficaces en établissant des profils de susceptibles consommateurs. En modélisant de manière précise et orienté dans le sens des besoins de l'entreprise ces profils, on obtient alors une idée complète des déplacements et des comportements réels des consommateurs. Les stratégies de prospection et communication sont donc amené à être plus efficace.

La concurrence est également bien étudiée et cela permet de projeter la pérennité de l'entreprise dans le temps en alertant sur l'évolution des réseaux de concurrents.

Le géomarketing est la solution efficace afin d'appréhender parfaitement les territoires impactés par une nouvelle implantation et ainsi suivre son évolution tout au long de sa croissance.

Ces dernières années, les solutions de géomarketing s'appuie de plus en plus sur des modèles de prédictions de plus en plus complexe et précis. La mise à disposition de modèle de flux piéton s'avère très utile pour permettre d'améliorer le géomarketing d'une compagnie. La demande concernant cette mesure est en augmentation et les entreprises vendeuses de solutions de géomarketing cherche à obtenir les meilleurs modèles prédictifs pour répondre au mieux aux besoins de leurs clients.

Les algorithmes de Machine Learning sont des outils puissants pour estimer spatialement les flux piétons utiles aux analyses de géomarketing. Ils nécessitent cependant des quantités de données très importantes pour obtenir les précisions nécessaires.

Mon stage a consisté en partie à la réalisation de ce modèle prédictif. Ce genre de mission est typique au métier de Géo Data Scientist³ et c'est donc autour de cette mission que je développe mon rapport de stage.

1. A compléter
2. A compléter
3. A completer

GALIGEO

1.1 Présentation de l'entreprise

1.1.1 Généralités

Galigeo est une société parisienne spécialisée dans le géomarketing. Créé en 2001, elle propose aux entreprises d'améliorer l'efficacité de tous leurs métiers grâce à ses logiciels combinant expertise cartographique et modélisation prédictive. Grâce à ses logiciels visualisant, analysant et agissant directement sur les bases de données opérationnelles (applications métier, BI, CRM, ...). Galigeo permet aux utilisateurs de se focaliser sur leur métier (retail¹, distribution, marketing, sécurité, ...).

Historiquement pionnier de la location Intelligence, Galigeo a poursuivi son développement ses dernières années en ajoutant la composante prédictive dans ses suites logiciels, composante basée sur les techniques innovantes de Machine Learning et d'Intelligence Artificielle.

En utilisant son expertise cartographique et de modélisation prédictive, Galigeo poursuit son développement en mettant à disposition de ses clients, des logiciels simples d'usage, à très forte valeur ajoutée métier.

De 2001 à 2006, Galigeo se consacre aux développements de solutions intelligentes en géodécisionnel pour faire de l'analytique avancée à partir de cartes géographiques. De 2006 à 2011, elle développe sa première solution logiciel, Galigeo Enterprise, qui permet par la suite, grâce à un nouveau pôle Conseil, de proposer des services spécifiques métier et adapté à chaque client autour du géomarketing. A partir de cette date, Galigeo va voir sa croissance augmenter sur le marché international en s'associant à différents partenaires. A partir de 2017 et jusqu'à aujourd'hui, Galigeo a choisi d'ajouter à ses solutions une composante prédictive très demandé sur le marché.

L'entreprise compte aujourd'hui une cinquantaine de clients très divers, grandes enseignes commerciales, services publiques, industries, etc. Galigeo peut fournir des solutions logiciels standards pour permettre à n'importe quelle entreprise de générer des rapports de géomarketing en utilisant des données interne et des données générales fournies par Galigeo. Elle réalise également des projet plus spécifique, propre à des besoins bien déterminés qui permette alors à ces client d'avoir une vraie valeur ajoutée et un géomarketing efficace.



FIGURE 1.1 – Exemple de clients Galigeo

Les bureaux de l'entreprise sont situé au 87 avenue d'Italie dans le treizième arrondissement de Paris.

1. A compléter



FIGURE 1.2 – Bureaux de Galigeo, 87 avenue d'Italie, 75013 Paris

1.1.2 Organisation interne

Galigéo est composé actuellement d'une vingtaine d'employés répartis dans 5 pôles : Administratif, Commercial, Marketing, Comptabilité, Consulting et Recherche et Développement.

Le pôle administratif est le pôle qui s'occupe de la gestion du budget, du personnel et des missions à Galigeo.

Le pôle commercial gère les relations clients, il propose des offres à de nouveaux ou ancien client, prospecte et cherche à faire grandir le cercle de clientèle de Galigeo.

Le pôle marketing imagine les produits, met en place les stratégies de pénétration du marché et réalise un catalogue de solutions que Galigeo peut fournir.

Le pôle comptabilité est responsable de la facturation des clients et la gestion interne des frais de personnels, des salaires, du matériel, etc.

Le pôle consulting est dédié à la réponse aux besoins techniques du client, il permet de rester proche du client. Il imagine les solutions retenues et les intègres dans les outils Galigeo mis à disposition pour l'entreprise.

Pour ma part, j'ai rejoint le pôle Recherche et Développement. L'équipe est composée de développeur, de testeurs, de designer, de data scientist, etc. Elle s'attache à améliorer les produits de Galigeo et à faire du support, de la maintenance et de l'innovation. Lorsqu'un consultant est en charge d'un projet pour un client, il s'appuie sur un ou plusieurs membres de l'équipe R&D pour conseiller ou réaliser les tâches techniques.

J'ai principalement travaillé avec Raimana Teina, Data Sientist chez Galigeo autour du grand projet actuel chez Galigeo « Prédiction de flux piéton » que je détaillerais dans la suite de ce rapport. Cependant j'ai également eu l'occasion de travailler sur des projets clients avec l'équipe consulting.

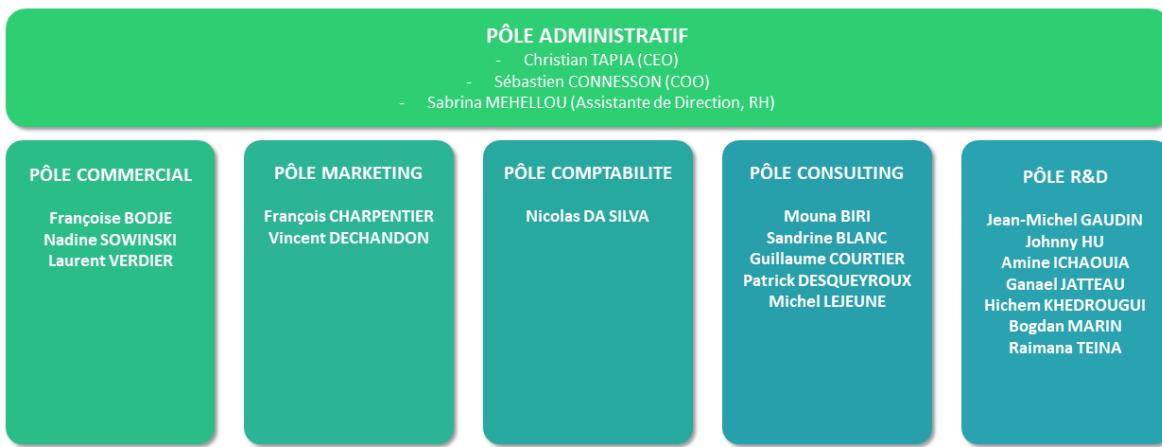


FIGURE 1.3 – Organigramme de Galigeo

1.2 Les objectifs de Galigeo

1.2.1 Les manques actuels

Aujourd'hui, Galigeo cherche à mettre en avant des modèles prédictifs au service du géomarketing pour ses clients. Une mesure importante en géomarketing est l'estimation de flux piéton à un endroit donné et sur une période donnée. Jusqu'ici, Galigeo utilisait un service tier afin d'obtenir une estimation de flux piéton. Cette solution de transition possède néanmoins de nombreux défauts. Tout d'abord Galigeo n'avait aucune visibilité sur les algorithmes utilisés pour faire cette estimation.

Il était alors difficile d'utiliser les résultats de cette estimation dans l'entraînement de nouveaux modèles de prédiction (estimation de chiffre d'affaires, estimation de parts de marchés, etc.). De plus, les coûts d'une telle solution restent élevés.

Il a donc été décidé de créer un modèle d'estimation de flux piéton propre à Galigeo sur lequel l'entreprise aurait accès à toutes les données d'entrées du modèle. Ainsi elle pourra modéliser d'autres variables essentielles au géomarketing plus facilement et avec une meilleure qualité. En effet, en connaissance des biais de notre modèle, ils seront plus faciles à corriger dans d'autres cas d'utilisation.

Galigeo souhaite également renforcer son équipe big data afin de mettre en valeurs de grosse quantité de données brutes inexploitable en l'état. En recrutant de nouveaux data scientist, elle libérera du temps à ses consultant et déchargera les équipes spécialisées dans la data actuellement en place.

Après un entretien au printemps 2022 chez Galigeo et cette explication des manques et objectifs de l'entreprise à court termes, j'ai fait part de mon envie à réaliser un stage au sein de l'équipe R&D. Ils ont retenu ma candidature et j'ai donc commencé mon stage de fin d'étude le 02 Mai 2022.

1.2.2 Les objectifs du stage

Durant ce stage j'avais donc comme première objectif de découvrir le géomarketing dans son ensemble. J'ai dû analyser et comprendre les cas d'usages métiers et me plonger dans le fonctionnement d'une équipe de développement logiciel.

J'avais également pour mission de concevoir et spécifier des solutions d'acquisition, de cleaning et de traitement des données. Comprendre le fonctionnement des produits Galigeo m'a permis de collecter et stocker la donnée de manière optimale afin de faciliter l'implémentation dans des bases de données.

J'avais également pour mission de développer des chaînes de collecte de la donnée en m'appuyant sur des méthodes de data engineering². Pour pouvoir par la suite utiliser au mieux la donnée dans des processus d'analyse et de traitement.

Pour la partie traitement de la donnée brute, mes objectifs étaient de développer des algorithmes et des traitements de la data grâce à des méthodes statistiques et de machine learning. Mon objectif était de transformer de la donnée brute inexploitable en une donnée pertinente pour le géomarketing des clients Galigeo.

Mon objectif était également de m'intégrer aux équipes de Galigeo afin de mieux comprendre les objectifs de l'entreprise à long termes et les directions à prendre.

1.2.3 Les objectifs à plus long termes

Galigeo cherche aujourd'hui à implémenter de plus en plus de modèles prédictifs dans ses solutions. Pour cela, elle veut s'appuyer sur la big data et la data science qui permettent d'obtenir des résultats de qualité et qui intéresse aujourd'hui le marché du géodécisionnel.

Galigeo cherche donc à développer son équipe R&D spécialisé dans la data afin de créer de nouveaux modèles compétitifs vis-à-vis des solutions concurrentes existantes et ainsi attirer de nouveaux clients et apporter des nouveautés aux plus anciens.

Aujourd'hui ces modèles prédictifs permettent d'évaluer une variable à un endroit spécifique (footfall³, chiffre d'affaires, cannibalisation, ...). On peut alors comparer plusieurs localisations et déterminer laquelle est la plus intéressante pour une entreprise.

Il sera alors intéressant de s'intéresser à des modèles de recherche qui permettent de trouver directement la meilleure localisation pour une variable donnée.

1.3 Organisation du stage

1.3.1 Planning

Le planning complet de mon stage est disponible en annexe A. Il est détaillé sur la période Mai - Septembre 2022. La rédaction de ce rapport a eu lieu avant la définition des tâches d'octobre.

1.3.2 Mes missions

Ma mission principale à Galigeo était de réaliser des modèles prédictifs de variables utiles au géomarketing. L'estimation du flux piéton en une adresse a été ma principale mission. Cependant j'ai également passé quelques temps sur la modélisation de chiffre d'affaires pour une marque de produit culturels et électroniques française. J'ai dû également modéliser la cannibalisation qu'entraînerait l'ouverture d'une enseigne proche d'une autre sur le territoire français.

Ces deux missions mettaient en pratique mes compétences en Machine Learning acquises à l'ENSG. Je reviendrais sur ma mission sur le flux piéton dans la partie 2.2 de ce rapport. Pour la partie modélisation de chiffres d'affaires et d'estimation de cannibalisation, j'y reviendrais dans la partie 2.3.

J'ai également réalisé d'autres plus petites missions pour Galigeo. J'ai eu la chance de travaillé quelques jours sur un projet d'analyse de données et sur un projet de data engineering pour une compagnie internationale. Je détaillerai rapidement ces deux derniers projet dans la partie 2.4 de mon rapport.

2. A compléter
3. A compléter

1.3.3 Relations internes et client

Dans le pôle R&D, le travail est organisé autour de la méthode Agile⁴. Un sprint⁵ dure environ 2 semaines et nous faisons un point tous les jours à 15h30 pour communiquer sur notre avancement et les difficultés rencontrés afin de pouvoir répondre rapidement aux besoins de chacun.

Je peux télétravailler 2 jours par semaine et cela fonctionne très bien car la plupart des réunions en présentiel sont également retransmises en visio-conférence. Galigeo utilise la suite Office 365 Professionnel ce qui permet d'organiser facilement le travail et les interactions.

J'ai également travaillé sur des projets où j'échangeais uniquement avec le pôle consulting. Dans ce cas-là, nous avions des réunions quotidiennes également avec le consultant en question et une réunion hebdomadaire avec le client.

4. A compléter
5. A compléter

LE STAGE

2.1 Généralités

Dans la suite de ce rapport je vais me consacrer sur les deux missions principales que j'ai réalisées à Galigeo. La première est l'estimation d'un flux piéton à une adresse donné. (2.2) Ce n'est pas une mission réalisée pour un client en particulier, c'est une fonctionnalité qui a été ajouté dans la dernière version des logiciels SaS de Galigeo et est donc accessible pour tous les utilisateurs du service.

La deuxième mission est une demande spécifique d'un client de Galigeo. J'ai eu accès à des données confidentiels à l'entreprise. Toutes les illustrations, schémas ou fragments de données, pour cette partie, seront donc fictifs

Enfin, je terminerai rapidement sur deux autres missions clients, plus secondaires pour lesquels je rentrerai moins dans le détail.

2.2 Prédiction de flux piéton

2.2.1 Mise en contexte

Le flux piéton est une très bonne variable pour estimer le flux de consommateur potentiel qui passe chaque jour devant une enseigne commerciale. Il est donc essentiel pour une entreprise qui fournit des services de géomarketing, de pouvoir estimer au mieux cette variable.

Jusqu'aujourd'hui, Galigeo s'appuyait sur des estimations calculées par une autre entreprise ce qui l'empêchait de corriger les biais et erreurs possibles. Elle n'avait pas la main sur les algorithmes d'estimations.

Ma mission a donc été de réaliser cet algorithme d'estimation du flux moyen de piéton sur une année autour d'une adresse donnée et ce pour l'ensemble du territoire national.

Il a fallu néanmoins prendre en compte les contraintes techniques de l'entreprise, les biais qui pouvait être présent dans la donnée utilisée et l'efficacité de l'algorithme. Si les temps de calculs sont trop longs, l'expérience utilisateur risque d'être impactée mais il faut garder un modèle puissant afin de minimiser les erreurs de prédiction.

2.2.2 La donnée

Pour estimer ce flux piéton, nous allons nous appuyer sur des mesures quotidiennes que Galigeo achète à un fournisseur. En effet, Galigeo reçoit quotidiennement des positions de cellulaires sur l'ensemble du territoire métropolitain. Elle en reçoit environ 70M par mois, chacune correspond à un évènement de visite. Cette donnée est collectée via des applications mobiles qui sont autorisées à transmettre au fournisseur de Galigéo la position du smartphone. Les positions sont anonymes mais

possède un identifiant de smartphone ce qui permet d'avoir aussi des informations de déplacements. La donnée brute est structurée comme ci-dessous :

Attribut	Description
idEvent	Id de l'évènement
uuid	Id du smartphone
latitude	Latitude de l'évènement
longitude	Longitude de l'évènement
accuracy	Précision de la localisation
arrival	Date d'arrivée à la localisation
departure	Date de départ de la localisation

TABLE 2.1 – Résumé de la structure d'un évènement de visite

Nous avons également utilisé de la donnée économique pour enrichir notre modèle. En effet Open Street Map propose une base open-source de POI structuré comme ci-dessous :

Attribut	Description
id_poi	Id du poi
type	Nature du POI (Magasins, Restaurants, Epiceries, ...)
latitude	Latitude du POI
longitude	Longitude du POI

TABLE 2.2 – Structure d'un POI

Il était également intéressant de rajouter de la donnée démographique à notre modèle pour cela nous avons utilisé les données de population de l'INSEE agrégé au niveau des IRIS géographiques.

Nous allons utiliser des algorithmes de Machine Learning pour estimer ce trafic piéton donc il nous faut des données vraies mesuré sur le terrain afin d'entraîner un modèle. Galigeo possède plus de 11000 mesures réparties plus ou moins équitablement sur le territoire même si la plupart d'entre elles sont en milieu urbain.

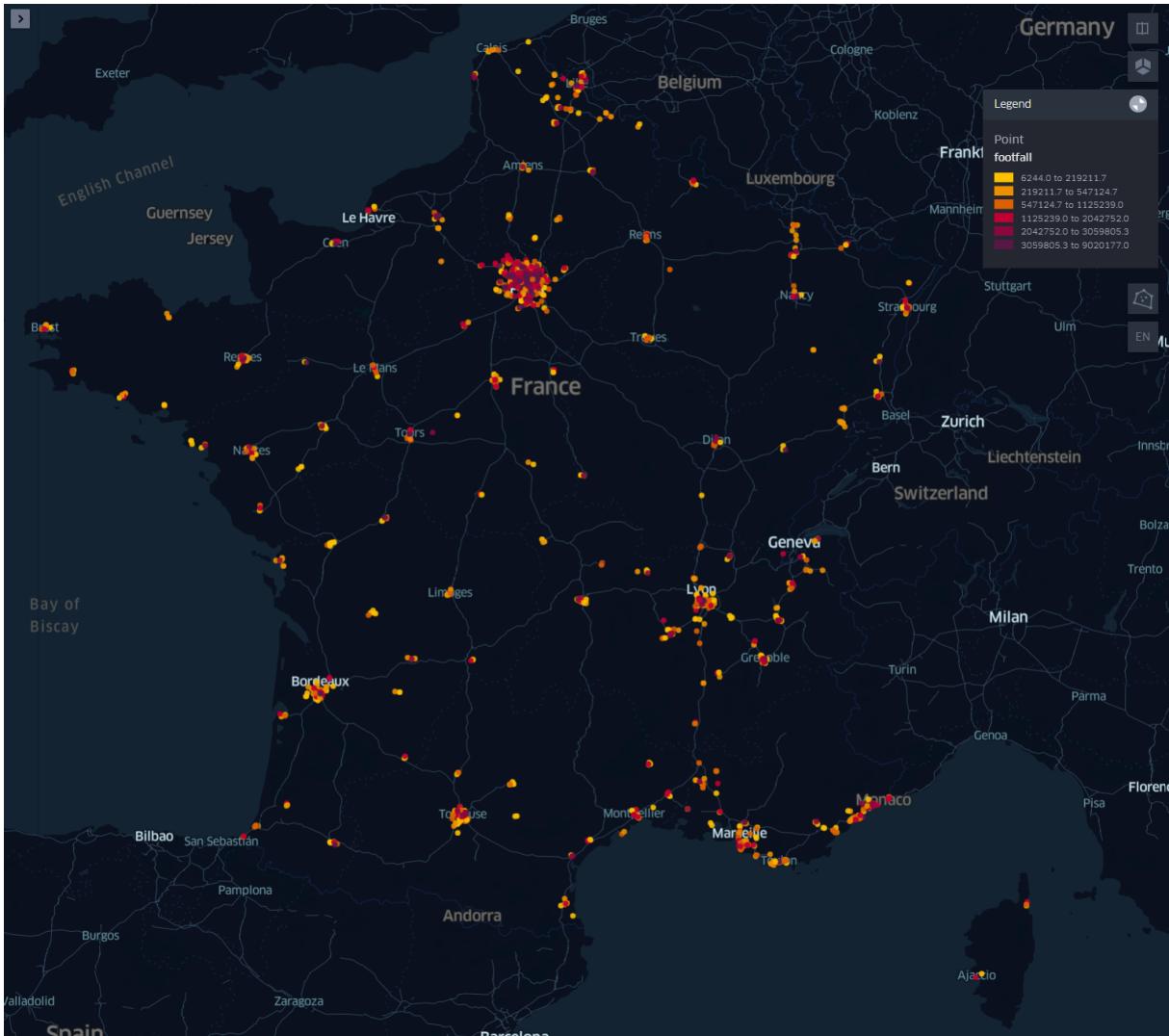


FIGURE 2.1 – Carte des données d'entraînement

2.2.3 Technologies utilisées

Agrégation spatiale

Une fois que nous avons l'ensemble de nos données on remarque qu'elles se distingue en deux groupes de par leur nature. Les données ponctuelles (Évènements de visites, POI, etc.) et les données surfaciques (Population par IRIS). Pour homogénéiser notre donnée et simplifier les futurs calculs il est important de segmenter l'espaces et ajouter un index spatial à nos données.

Nous nous sommes dirigé vers la solution des hexagones H3 [2] développé par Uber. C'est un système d'indexation géospatiale qui constitue un pavage hexagonal de la sphère multi-échelles et dont les index sont hiérarchiques.

Dans notre cas c'est la solution optimale car nous pouvons l'utiliser pour joindre nos données disparates, le format hexagonal facilite la modélisation des flux et est bien adapté pour appliquer le machine learning aux données géospatiales.

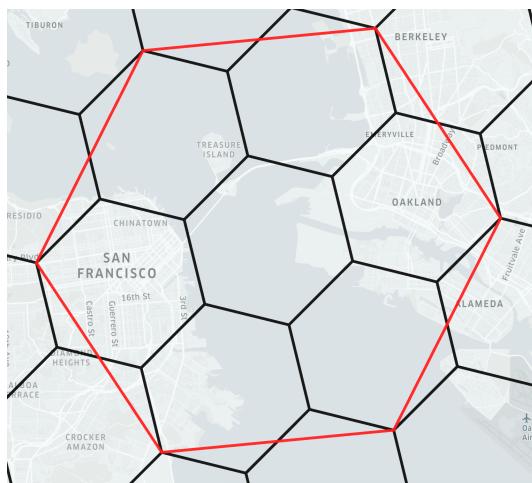


FIGURE 2.2 – Principe de multi-echelles des cellules H3

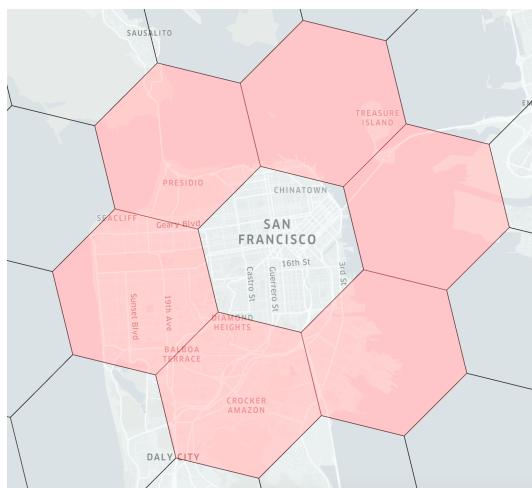


FIGURE 2.3 – Les 6 voisins d'une cellule H3 se rapproche d'un cercle, utile pour la modélisation de flux

Pour notre part on choisira le niveau d'échelle 11, ce qui correspond environ à un hexagone de 27 mètres de rayon.

Infrasture de base de donnée

Pour stocker toutes ces données, Galigeo utilise un service de Cloud Computing¹ nommé Google Cloud Plateform (GCP) qui permet de stocker une grande quantité de donnée et de faire des calculs plus ou moins complexe sur celle-ci.

Un des grands avantages de cette solution est l'intégration de Tensorflow, une bibliothèque open-source développé par Google pour faire du machine learning et de l'intelligence artificielle. Elle est très utile pour l'entraînement de réseaux de neurones profonds (DNN).

Nous pouvons également y combiner Keras, une interface open-source conçu pour simplifier l'utilisation de Tensorflow.

1. A compléter

Type de modèle utilisé

Étant donné les choix précédents, nous nous sommes donc orientés vers un modèle de Deep Neural Network (DNN).

Il est cependant important de noté que les modèles de Deep Neural Network, prouve leur meilleure efficacité par rapport à d'autre modèle de Machine Learning lorsque la donnée est non structurée et que la taille du dataset² est importantes (plusieurs millions de lignes)

Dans notre cas, ces conditions ne sont pas remplies mais nous verrons plus tard dans ce rapport les alternatives envisagées.

2.2.4 Préparation des données

Structuration des données

Cette étape est cruciale pour obtenir des bons résultats après la modélisation. Il s'agit de passer de la donnée brute présentée précédemment à une donnée structurée qui permettra d'entraîner le réseau de neurones.

Pour cela, nous avons projeté l'ensemble de notre donnée brute sur la grille H3 d'Uber. Pour chaque valeur de flux piéton mesuré, nous avons calculé le nombre de visites dans la même cellule ainsi que dans les cellules voisines. Nous avons agrégé ces données events par mois sur une année complète de septembre 2021 à août 2022.

Nous avons appliqué le même principe spatial pour les POIs. Les cellules voisines se calcul très facilement grâce aux fonctions open-source mise à disposition. Ainsi on peut calculer différents anneaux autour de notre cellule de départ.

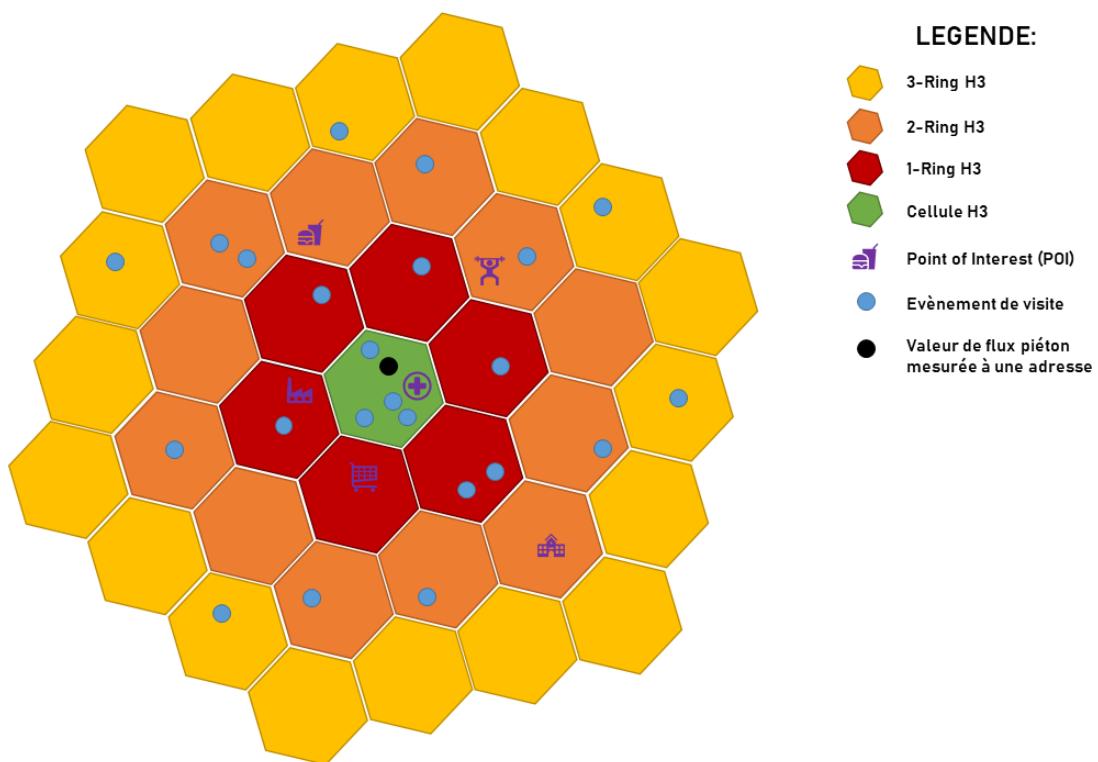


FIGURE 2.4 – Agrégation des données aux cellules H3 pour générer un dataset

Pour ce qui est de la donnée de population, nous avons simplement pris le centroïde de la cellule H3, et l'avons intersecté avec les géométries des IRIS françaises afin de déterminé dans quel IRIS la

2. A compléter

cellule se trouve.

Nous avons obtenu alors une structure de dataset comme ci-dessous. Une feature est un paramètre du modèle et le label est la variable à estimer.

Colonnes	Type	Description
h3_index		Index de la cellule H3
visits_MMAAAA	feature	Nb events dans la cellule H3 par mois (x12)
visits	feature	Nb events dans la cellule H3 sur l'année
visits_k1	feature	Nb events dans la cellule H3 et l'anneau 1 sur l'année
visits_k2	feature	Nb events dans la cellule H3 et les annaux 1, 2 sur l'année
visits_k3	feature	Nb events dans la cellule H3 et les annaux 1, 2, 3 sur l'année
visits_k4	feature	Nb events dans la cellule H3 et les annaux 1, 2, 3, 4 sur l'année
poi_N	feature	Nb de POI dans la cellule H3 par nature
poi_N_k1	feature	Nb de POI dans la cellule H3 et l'anneau 1 par nature
poi_N_k2	feature	Nb de POI dans la cellule H3 et les annaux 1, 2 par nature
poi_N_k3	feature	Nb de POI dans la cellule H3 et les annaux 1, 2, 3 par nature
poi_N_k4	feature	Nb de POI dans la cellule H3 et les annaux 1, 2 , 3 , 4 par nature
population	feature	Population in the IRIS of the H3 cell
footfall	label	Flux piéton mesuré en un point de la cellule H3

TABLE 2.3 – Structure du dataset

On obtient alors un dataset d'environ 11000 lignes et 14 colonnes.

Nettoyage des données

La deuxième étape de la préparation des données consiste à les nettoyer pour supprimer d'abords les lignes où il manquerait certaines variables puis celle dont les valeurs du label sont aberrantes.

On se retrouve alors avec un dataset réduit de 40%, il reste environ 6500 lignes.

Une fois ces deux étapes réalisées, nous pouvons sortir différents graphiques statistiques disponible en annexe : KDE, Matrice de corrélation, Répartitions des features.

2.2.5 Tuner et Entraînement du modèle

Notre dataset est maintenant prêt à être utilisé comme base d'apprentissage. Nous avons donc créé une structure de modèle. Cette structure possède des paramètres appelés Hyperparamètres. Chaque hyperparamètre peut prendre une valeur parmi une liste bien définie. Le principe du Tuner est de compiler ce modèle plusieurs fois avec des valeurs d'hyperparamètre bien différentes et choisies aléatoirement.

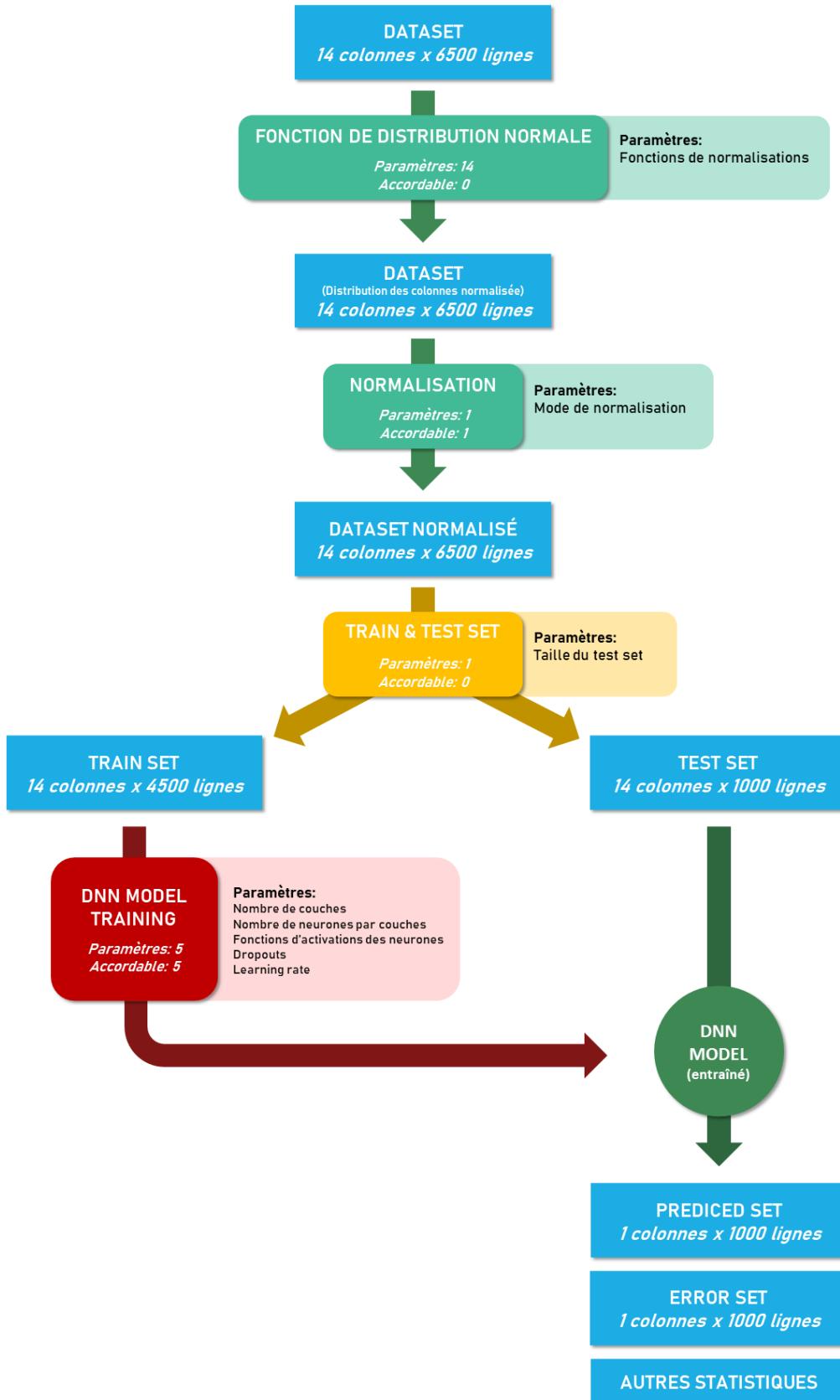


FIGURE 2.5 – Structure du modèle envisagé

La première étape de notre modèle est donc de normaliser la distribution de nos features. En effet, la distribution des valeurs pour chaque features n'est pas forcément "normale" et il est préférable de passer nos valeurs à travers une fonction pour la modifier.

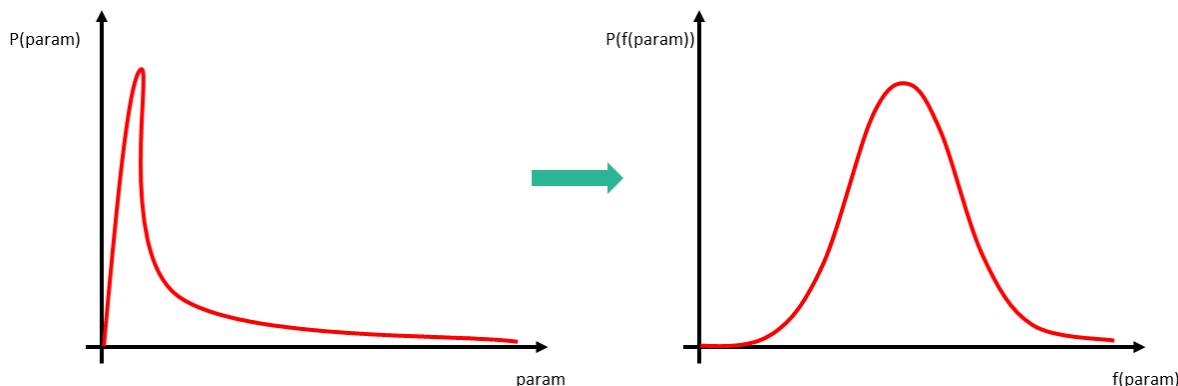


FIGURE 2.6 – Principe de normalisation de la distribution d'une feature

L'étape suivante est de normaliser les valeurs de nos features (cette fois ci, on ne modifie pas le label). Pour cela, il existe principalement deux types de normalisation : La normalisation min/max et la normalisation moyenne/écart-type.

Cela permet entre autres d'assurer une cohérence entre les données et permet à l'exécution d'être plus rapide et meilleure.

Une fois cette étape réalisée, nous divisons en deux partie aléatoire (la proportion est contrôlée) le dataset normalisé. La partie "trainset" permet d'entraîner le modèle et la partie "testset" permet de vérifier si le modèle est correctement entraîné.

L'objectif est maintenant d'obtenir un réseau de neurones entraîné. Pour cela, nous utilisons un modèle construit sur les hyperparamètres et lui passons l'ensemble du trainset en entré. En parcourant l'ensemble des données et en essayant d'estimer le label, le réseau de neurone apprend et règle ses paramètres internes (poids des axones, valeur interne des neurones, etc.) en appliquant une validation croisée.

De cette étape, nous récupérons un réseau de neurones profonds qui permet maintenant d'estimer notre label en fonction de nos features.

Nous passons donc notre testset dans notre réseau neuronal. Cela nous permet d'obtenir notre label estimé et de le comparer au vrai label. On obtient alors une liste de n erreurs (dans notre cas 1000), grâce auxquels nous pouvons mesurer la qualité du réseau de neurones. On peut calculer l'erreur quadratique moyenne (RMSE), un coefficient de détermination (R2) ou la précision du modèle ("accuracy").

2.2.6 Résultats

	Test set	Training set
RMSE	1 000 000	500000
R2 Score	0.45	0.86

TABLE 2.4 – Résumé des résultats

Pour estimer la qualité de notre modèle, il faut regarder les résultats de la colonne de gauche, "Test set". On remarque tout de suite que l'erreur moyenne quadratique est d'environ 1000000. Or, si l'on se réfère à l'annexe B, Statistiques du modèle, on remarque que le flux piétons médians de 1000000 également.

Le R2 score est également faible. Idéalement, il devrait se rapprocher de 1. L'estimation que fait le modèle n'est donc pas encore suffisante.

Cependant Galigeo renforce son dataset à chaque arrivée de donnée ce qui peut améliorer les résultats par la suite. Il se peut également que 6500 entrés soit trop peu et que l'enrichissement de mesure permette également de rendre plus précises les estimations du model.

Nous avons également discuté de nombreuses autres pistes pour obtenir des résultats plus intéressant mais je les détaillerai dans la partie 3.1 de ce rapport.

2.3 Prédiction de chiffre d'affaire

2.3.1 Contexte

Pour qu'une entreprise continue à croître et à se développer, elle peut étendre son aire d'attraction et ainsi cibler plus de clients. L'entreprise décide alors d'ouvrir une nouvelle enseigne. Le choix de l'emplacement est alors stratégique.

Pour ce projet, un client de Galigeo souhaite planter de nouvelles enseignes sur le territoire français. Pour ce faire il souhaite estimer le chiffre d'affaires et la cannibalisation³ qu'engendrera un magasin en fonction de sa localisation, de sa surface, du type de magasin et du flux piéton moyen à l'adresse.

Il a donc fallu créer un modèle prédictif du chiffre d'affaires et de la cannibalisation en s'entraînant sur les données des enseignes existantes. Cependant, cette entreprise possède 150 enseignes environ en France et cela semble insuffisant pour entraîner un modèle de machine learning. Nous avons donc utilisé une autre méthode pour démultiplier les données.

2.3.2 Données

La France est découpée en 15 500 IRIS (îlots Regroupés pour l'Information Statistique) et le client possède des données commerciales pour certains d'entre eux. Il est alors intéressant de travailler sur les IRIS et non sur les enseignes. En effet, nous avons à disposition :

- La liste des enseignes (surface, chiffre d'affaires, type, localisation, ...)
- La liste des concurrents (surface, type, localisation, ...)
- Des données socio-démographiques pour chaque IRIS (tranche d'âge, nombre de ménage, revenue moyen, ...)
- Un distancier des Iris (permet de connaître la distance et le temps de parcours entre deux centroïdes d'Iris)
- Une estimation de la valeur du marché dans chaque IRIS
- Une valeur de chiffre d'affaires pour chaque enseigne dans chaque IRIS (C'est ici le chiffre d'affaires tracé, une partie du chiffre d'affaires de chaque enseigne ne peut pas être localisé)

On peut alors créer un dataset où chaque élément correspond à un iris pour un magasin et agréger le reste de la donnée à cette structure. On cherchera alors à estimer une part de marché, qui correspond à la quantité de marché détenu par un magasin particulier. Chaque enseigne connaît l'origine géographique d'une partie de son chiffre d'affaire (20%), on peut donc connaître cette part

3. A compléter

de marché sous-évalué, en divisant ce chiffre d'affaire tracé par la valeur du marché dans un IRIS donnée.

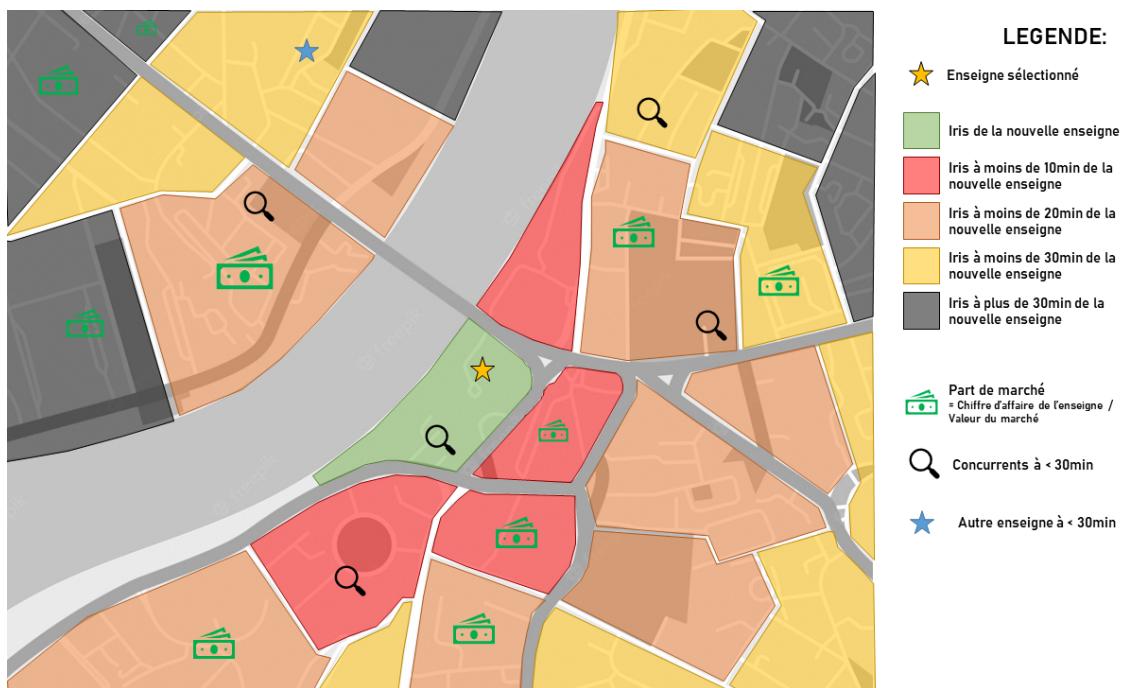


FIGURE 2.7 – Résumé de l'agrégation des données

2.3.3 Structure du modèle

Cette fois si, il était plus intéressant de ne pas choisir un algorithme de Deep Learning mais un algorithme de machine learning moins coûteux en ressource comme un régresseur « Random Forest » ou un « XGBoost ». On utilisera la même structure de modèle qu'à la partie 2.2 mais cette fois si la normalisation est moins importante car le modèle n'est pas composé de neurones mais d'arbres de décisions. Nous avons alors comparé les deux solutions et sélectionné l'algorithme XGBoost.

2.3.4 Résultats

2.4 Autres missions

2.4.1 Analyse de données - Base de donnée nationale des Bâtiments

Pendant mon stage, j'ai également pu travailler sur d'autre petite mission pour des clients de Galigeo pour lesquels il n'était pas forcément question de machine learning.

Un client de Galigeo réalise de la prospection chez des particuliers. Jusqu'ici, la prospection était organisée très simplement, un quartier était sélectionné et les prospecteurs passaient dans chaque habitation. Ce client cible les habitations se chauffant à l'électricité et il arrive souvent que certaines des habitations prospectées se chauffent au gaz ou autre. Dans un souci d'économie et de gain de temps, nous avons donc tenter de cibler les prospections.

En janvier 2022, une Base de Donnée Nationale des Bâtiments [1] est mise en ligne en open-source. C'est une base très complète avec plus de 90% des géométries renseignés. Environ 25% des bâtiments possède des informations énergétiques dont on ne connaît pas la qualité. Le client et Galigeo ont donc décidé de mener une campagne de prospection cet été en s'appuyant sur cette base de données afin de qualifier les données énergétiques disponibles. Nous avons donc fait de l'analyse de donnée sur quelques IRIS (à Rennes et à Paris) pour lancer la campagne de prospection.

Aujourd'hui les campagnes de prospection sont toujours en cours mais le résultat semble positif. Si la base de données est retenue pour diriger les campagnes de prospections du client, il faudra automatiser les processus d'analyse pour envoyer la donnée traitée dans leur application métier développé par Galigeo.

Sachant que seulement 25% des bâtiments ont des données énergétiques, il sera peut-être intéressant un jour d'utiliser un algorithme de classification par machine learning pour interpoler ces informations manquantes sur le reste de la base.

2.4.2 Data engineering - Scrapping et automatisation de chaîne d'acquisition de données

Une autre mission que j'ai pu réaliser était pour un client international de Galigeo. Ils utilisaient la solution de Galigeo pour leurs compagnies en France et souhaitent l'utiliser dans 13 nouveaux pays. La solution actuelle permet de sortir des rapports de géomarketing poussé qui s'appuie sur de très grandes quantités de données.

L'objectif était donc de récupérer de la donnée en quantité et d'automatiser son processus de publication dans une base de données. J'ai donc écrit des scripts permettant de scrapper des site web de concurrents pour récupérer de la donnée non structurée pour l'organiser. J'ai aussi dû automatiser des requêtes via des API afin de récupérer des données ponctuelles sur un pays entier.

Ce projet est encore en cours à la date où je rédige ce rapport. L'application et la base de donnée complète devrait être livré pour la mi-octobre.

BILAN

- 3.1 Bilan de mes travaux**
- 3.2 Bilan du stage**
- 3.3 Bilan des trois ans à l'ENSG**

Conclusion

Il est l'heure de conclure : bonne nuit !

Bibliographie

- [1] Centre Scientifique et TECHNIQUE DU BÂTIMENT. “Base de Donnée Nationale des Bâtiments”. In : *GitLab* (2022).
- [2] UBER. “H3 : A Hexagonal Hierarchical Geospatial Indexing System”. In : *GitHub* (2015).

Table des figures

1.1	Exemple de clients Galigeo	5
1.2	Bureaux de Galigeo, 87 avenue d'Italie, 75013 Paris	6
1.3	Organigramme de Galigeo	7
2.1	Carte des données d'entraînement	12
2.2	Principe de multi-échelles des cellules H3	13
2.3	Les 6 voisins d'une cellule H3 se rapproche d'un cercle, utile pour la modélisation de flux	13
2.4	Agrégation des données aux cellules H3 pour générer un dataset	14
2.5	Structure du modèle envisagé	16
2.6	Principe de normalisation de la distribution d'une feature	17
2.7	Résumé de l'agrégation des données	19
A.1	Planning du stage	30
B.1	Différentes statistiques du modèle de DNN	31

Liste des tableaux

2.1	Résumé de la structure d'un évènement de visite	11
2.2	Structure d'un POI	11
2.3	Structure du dataset	15
2.4	Résumé des résultats	17

Annexes

A	Planning du stage	30
B	Statistiques DNN	31

PLANNING DU STAGE

GANNT

ANNEXE

A

Planning du stage

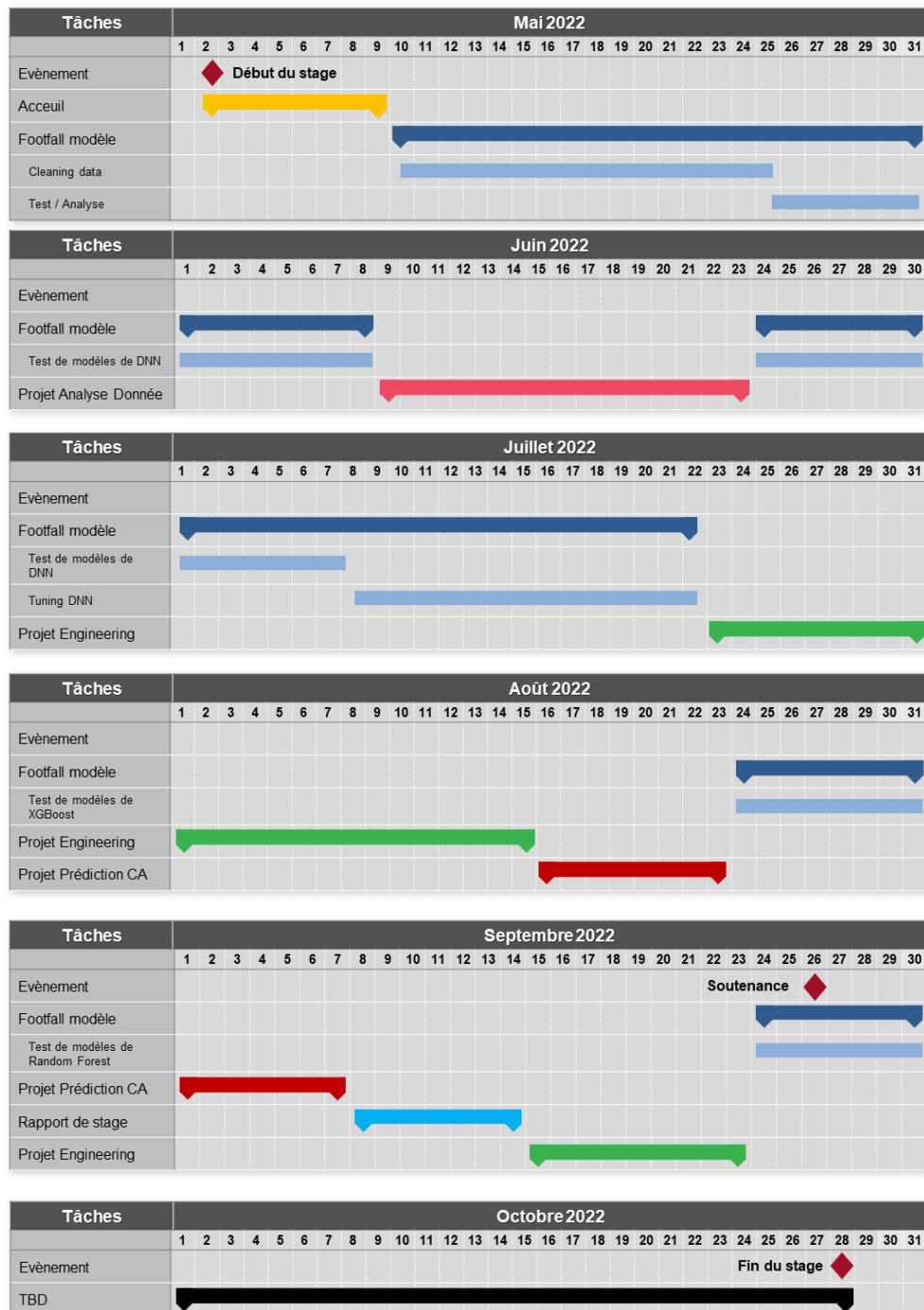


FIGURE A.1 – Planning du stage

STATISTIQUES ESTIMATION DE FLUX PIÉTONS

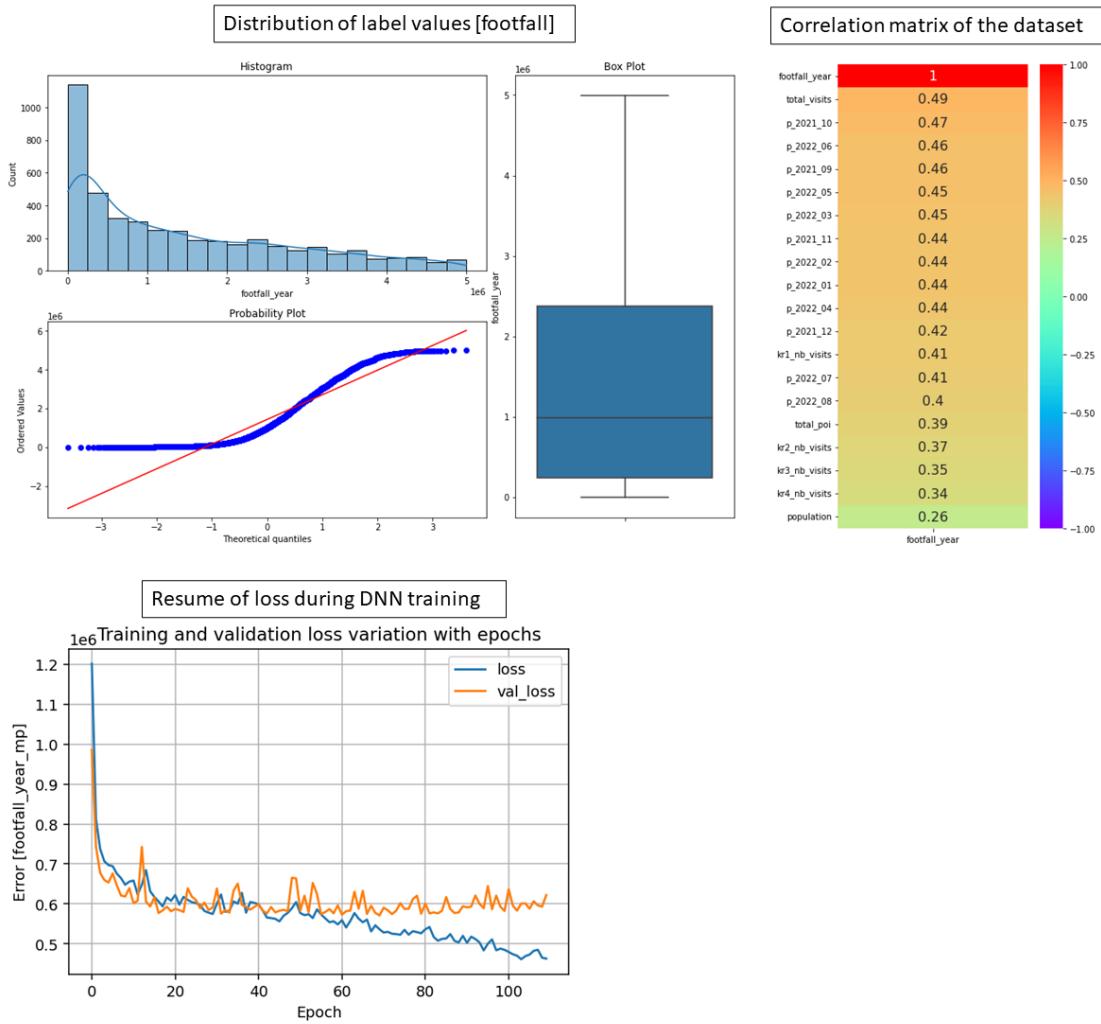


FIGURE B.1 – Différentes statistiques du modèle de DNN