Activité durant l'année de stage long

Arthur Imbert

Octobre 2017

1 Kynapse

J'ai effectué mon premier stage de césure chez Kynapse, du 13 juin au 7 décembre 2016. C'est une petite société de conseil spécialisée dans le Big Data et la transformation numérique. A mon arrivée, nous n'étions qu'une demi-douzaine. En décembre, l'équipe avait doublé. Kynapse avait ses bureaux dans les locaux de l'entreprise Open qui détenait financièrement des parts dans la startup et l'intégrait à son offre commerciale. Cela nous permettait d'exploiter certaines ressources administratives et humaines d'Open, ainsi que leur réseau commercial.

L'équipe de Kynapse est composée pour moitié de data scientists, issus d'écoles d'ingénieurs, et de managers, issus d'écoles de commerce. Je travaillais sous la supervision d'Irène Balmès la responsable de l'équipe de data scientists. A la fois parce que l'équipe était restreinte, mais également parce que l'activité de Kynapse en offrait l'opportunité, j'ai passé l'essentiel de mon stage à naviguer entre des missions purement techniques et d'autres plus commerciales, en contact direct avec le client. C'est l'une des raisons qui a rendu ce stage extrêmement enrichissant sur le plan personnel et professionel.

Au niveau technique, j'ai essentiellement travaillé sur deux grosses missions : la détection de chutes et la désagrégation des courbes de charge électriques.

La mission sur la détection de chutes a été menée en relation avec une équipe spécialisée en IoT (Internet of Things) d'Open, basée à Lyon. Une entreprise développe des harnais de sécurité destinés à être portés par des travailleurs en hauteur (sur des échelles, des toits, des échafaudages, etc.). Il est prévu d'équiper ce harnais avec un double capteur accéléromètre-gyroscope (comme ceux présents dans les smartphones). L'objectif est de déterminer automatiquement si le travailleur chute en fonction des 6 signaux mesurés par les capteurs (les 3 axes de l'accéléromètre et les 3 axes du gyroscope). La difficulté réside dans le fait que les travailleurs sont amenés à être actifs et notamment à sauter et à se déplacer brusquement durant leur journée. Cela pouvait conduire nos modèles à émettre de nombreuses fausses alertes. Pour être opérationnelle, une solution ne devait prédire aucun faux négatif (aucune chute ne devait être manquée) et très peu de faux positifs (de fausses alertes). Enfin, comme le système était destiné à être embarqué, le modèle utilisé devait être peu couteux en terme de mémoire. Nous avons réalisé un gros travail de *feature engineering* afin de créer de nouvelles statistiques à partir des 6 axes collectés quasiment en temps réel. Outre des variances et des moyennes, nous avons calculé pour chacun de ces axes des coefficients de dissymétrie (skewness, kurosis) et même des corrélations entre ces axes. Toutes ces *features* ont nourri un simple modèle de forêts aléatoires dont nous avions limité la profondeur afin de l'embarquer facilement. A mon départ de Kynapse, les résultats étaient prometteurs, même si la mise en production n'avait pas encore commencé. Plusieurs tests ont été réalisés par les travailleurs, ainsi que des simulations de chutes sur mannequins. Un résultat surprenant a été la détection, anormalement élevée, de faux positifs lorsque les travailleurs effectuaient une journée sans qu'aucune chute n'ait été enregistrée. En réalité, ils effectuaient des actions beaucoup trop brutales pour les capteurs (comme se laisser glisser d'une échelle). Nous n'avions pas prévu de discriminer de telles actions dans nos modèles car elles étaient supposées être intedites sur les chantiers.

La mission sur la désagrégation des courbes de charge électriques est plus complexe. A partir des données d'un compteur électrique (nécessairement agrégées à la demi heure), nous cherchions à estimer la consommation du ménage, par appareil ou catégorie d'appareils (chauffe-eau, chauffage, etc.). Nous avons récupéré des données externes sur internet (notamment sur la consommation individuelle de différents appareils) et travaillé à clusteriser les différents moments de la semaine en fonction de la consommation. Pour la désagrégation en tant que telle, après une revue de littérature sur le sujet, nous avons utilisé un algorithme de *Metropolis Hastings*. Le résultat était un graphique dit "camembert" où nous estimions la part de chaque appareil dans la consommation agrégée observée pour la journée.

Outre ces missions, j'ai également réalisée, avec le reste de l'équipe, des visualisations sur les Jeux Olympiques. Nous avions utilisé des données scrapées ou téléchargées sur internet. Enfin, j'ai suivi les cours données par Irène Balmès à des clients sur des sujets autour de l'apprentissage statisitque et des réseaux de neurones.

Concernant mes missions plus commerciales, elles consistaient essentiellement à répondre à des appels d'offre (parfois avec une étude de cas technique) ou à rencontrer des clients. Trois missions sortent du lot : la réponse à un appel d'offre pour une entreprise de transport de passagers, l'organisation d'un séminaire pour des promotteurs immobiliers et la rédaction d'un article publié dans *Les Echos* à l'occasion des élections américaines.

Pour la réponse à l'appel d'offre, tout un chapitre technique était attendu, avec notamment une étude de cas concernant de la maintenance prédictive, ainsi que la proposition d'une architecture Big Data capable de gérer de grand volumes de données mesurées en temps réel et originellement biaisées. En effet, les bornes qui permettaient à cette entreprise de valider les tickets de ses usagers n'avait pas de système horaire homogène. Par exemple, certaines passaient à l'heure d'hiver automatiquement, d'autres étaient décallées d'une heure. En collaboration avec un autre membre de l'équipe, bien plus expérimenté sur la question, j'ai pu commencer à m'initier à quelques concepts clés d'une architecture Big Data.

Historiquement, Open développe un logiciel largement utilisé par les promotteurs immobiliers. Ils souhaitaient également organiser un séminaire sur les possibles transformations numériques à apporter à l'activité du promotteur immobilier (mêlant solutions IoT, cartographie et data science). A cette occasion, une personne issue de Kynapse était souhaitée afin de présenter de potentiels études de cas autour de la *data science* qui pourraient intéresser les promoteurs. Je me suis chargé de collecter des informations sur la promotion immobilière (à travers des rendez-vous clients notamment), ainsi que sur les potentielles données auxquelles les promoteurs avaient accès et qui leurs apportaient une véritable plus-value. J'ai ensuite effectué une visualisaton cartographique, ainsi qu'une étude économétrique des prix de l'immobilier (notamment à base de régression quantile, en exploitant des données de l'INSEE) pour proposer un complément intéressant à la traditionnelle grille de prix utilisée par les promoteurs. Enfin j'ai présenté ces travaux devant une quinzaine de promoteurs à l'occasion du séminaire, en compagnie d'Irène Balmès.

Fortement inspiré par l'actualité de l'élection américaine et par les publications du site américain FiveThirtyEight, j'ai rédigé un article sur l'utilisation croissante des données dans la sphère politique. A l'origine destinée au blog de Kynapse, il a été mis en avant par la responsable de communication du groupe Open afin de paraître dans la rubrique Idées des *Echos*.

2 Autres

Pour le reste de mon année de césure, j'ai effectué une formation professionnelle durant l'hiver. J'ai passé le mois de janvier à Chamonix en vu de compléter et de valider des épreuves pour le monitorat de ski alpin (diplôme que je mène en parallèle de mes études depuis la fin du lycée). Cette période n'a pas été comptée comme stage.

A partir du 1er mars 2017 j'ai rejoint l'INRIA et l'équipe PARIETAL afin de travailler sur ledit rapport : l'intégration de données hétérogènes. En parallèle de cette étude, j'ai participé à un sprint d'une semaine dans les locaux de Criteo pour répondre à des *issues* émises sur la librairie python de machine learning *scikit-learn*. Enfin, j'ai travaillé, lors d'un séminaire avec le reste de l'équipe, sur des données issues du *Cambridge Centre for Ageing and Neuroscience (Cam-CAN)*. Cela consistait essentiellement à prétraiter des scores de comportement destinés à être associés ensuite à des données de neuroimagerie par le reste de l'équipe.