UNIVERSITÉ DE NANTES

IAE NANTES
ÉCONOMIE & MANAGEMENT

Note de synthèse

Projet économétrie des variables qualitatives

Election & Big data

Mots clés : Cambridge Analytica, modèle multinomial non ordonné, partis politiques

DEL'CHATEAU Jean-Baptiste

2020/2021

Présentation du sujet

Le BIG DATA et les élections. Nous sommes partis de l'affaire Cambridge Analytica, malheureusement célèbre sur la collecte illégale de données, notamment pour l'élection de D. TRUMP et lors du Brexit. Ce n'est qu'en 2018 que cette affaire a fait grand bruit, alors que celle-ci a officiellement débuté en 2008, avec deux publications dans deux journaux. Cambridge Analytica est une société de conseil stratégique basée à Londres. Pour la collecte des données, un quizz a été inventé par A. KOGAN (psychologie à l'université de Cambridge), grâce auquel ce psychologue a eu accès à plus de 200 millions de données, mais 200 millions de personnes n'ont pas répondu au quizz. Ce nombre de données récoltées vient du fait que les comptes Facebook des répondants mais aussi de leurs amis ont été piratés. Les personnes répondant à ce quizz, recevaient une récompense pécuniaire, ce questionnaire était basé sur des questions directement liées à l'individu, comme par exemple des traits de caractère tel que l'impulsivité ou l'enthousiasme … Ces données récoltées ont été récupérées par les équipes de TRUMP, ce qui a permis aux personnes indécises de se faire influencer. Pour résumer "grossièrement", par l'analyse des réponses aux questionnaires, une personne va être déclarée indécise, cependant, sachant que sa plus grande préoccupation était une plus grande sécurité dans les villes des Etats-Unis, alors des publications ou publicités vont défiler sur son Facebook, prônant la mise en place par TRUMP de mesures visant à diminuer l'insécurité, cette dernière va donc être influencée par les réseaux sociaux et possiblement voter TRUMP. C'est ce même procédé qui a été réalisé pour des milliers de personnes, chacun a vu des milliers de publicités sur leur réseau social, ce qui a permis de faire pencher la balance pour TRUMP. Il s'est passé le même phénomène pour le Brexit.

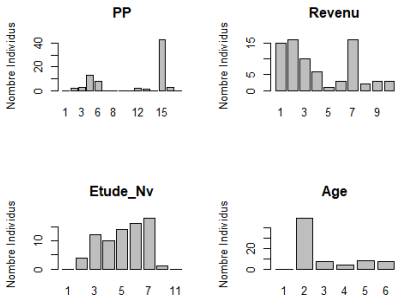
Nous sommes donc partis de cette affaire pour savoir si en France cela était possible. Pour ce projet, nous avons créé un sondage, suite auquel, nous avons récolté 189 réponses. Dans notre sondage, nous avons demandé à nos sondés quelle était leur tendance politique, ce qui nous a servi de variable à expliquer, mais nous leur avons également demandé quel était le parti politique de chacun de leurs parents et quel était l'axe politique qui les motivait. Pour faire référence à l'affaire Cambridge Analytica, nous avons également posé des question concernant l'individu, les mêmes questions qui était posées dans le questionnaire de KOGNAN. Nous avons donc une variable à expliquer avec 4 valeurs et 23 variables explicatives.

Statistiques descriptives

Avant de commencer les statistiques descriptives ou l'économétrie, nous avons dû changer notre base de données. Nous avons effectué un chiffre ou nombre à chaque modalité et pour chaque variable. Sur le graphique, on ne verra que des chiffres et pas de tranches d'âge ou de revenus par exemple.

Sur notre échantillon, nous avons réalisé plusieurs statistiques descriptives, mais dans cette synthèse nous allons mettre en avant celle qui nous paraît la plus pertinente. Nous allons voir l'analyse statistique du parti politique (PP) du sondé, de son revenu, de son niveau d'étude et de son âge en fonction des axes politiques. Nous avons 9 axes politiques, nous n'allons pas mettre les 9 graphiques mais seulement un à titre d'exemple. Le graphique présenté ci-dessous, concerne un de nos axes politiques, l'écologie. Si nous faisons l'analyse de ce graphique, nous voyons que d'une part la population la plus jeune (18-25 ans) et d'autre part les plus haut et plus bas revenus sont les plus sensibles à l'écologie. Les personnes se sentant le plus concernées par ce sujet ont un niveau d'étude assez homogène, aucun niveau d'étude ne se détache.

Graphique 1 : histogramme de l'analyse statistique par rapport à l'écologie



Sur le dossier, c'est en faisant cette partie que nous nous sommes rendu compte que nous avions différents biais avec notre base de données. En effet, une population entre 18 et 25 ans trop nombreuse, un nombre de sondés pas assez élevé pour que notre étude soit représentative, un trop grand nombre de personnes ne se classant dans aucun parti politique et se sentent perdues en politique.

Méthodologie

Nous avons utilisé des modèles non ordonnés, notamment le modèle Logit multinomial général. Pour sélectionner le meilleur d'entre eux, nous avons décidé d'utiliser le ratio de vraisemblance et le pseudo R^2 de McFadden. Lorsque que le pseudo R^2 de McFadden est compris entre 0.2 et 0.4, nous allons alors dire que le modèle a un bon ajustement.

Résultats

Ainsi nous avons donc décidé d'étudier le phénomène d'influence des personnes via les réseaux sociaux. Les résultats de notre sondage nous ont permis de constater que nous avions des biais sur nos données, dont un important provenant de l'âge et de la catégorie socio-professionnelle du sondé. Parmi nos 189 sondés nous retrouvons une majorité d'étudiants pouvant ainsi fausser les modèles. Nous ne pouvions pas faire d'inférence statistique avec les résultats que nous avons trouvés.

Pour faire nos modèles, nous avons pris notre variables à expliquer soit le parti politique (pp) plus les variables dépendantes qui sont, le parti politique du parent 1-2 et les différents axes politiques. Tous nos modèles sont partis de cette base, puis par la suite nous avons ajouté des variables explicatives et ensuite nous avons choisi le meilleur modèle, celui dont le pseudo R^2 de McFadden était le plus proche de 0.4. Nous avons commencé par faire des modèles avec toute la base de données et nous n'avons trouvé aucun bon modèle. Nous nous sommes rendu compte que certains partis politiques n'avaient aucun vote. Nous avons donc décidé de faire des modèles en retirant tous les partis sans vote. Nous nous sommes retrouvés avec un meilleur modèle. Pour chaque modèle, nous avons pris une référence (un parti politique), qui nous servira pour comparer et pour interpréter par rapport à cette référence. Nous avons essayé de faire

un modèle avec comme référence un parti politique n'ayant qu'un seul votant (le MoDem) et nous avons également fait un modèle en prenant comme référence la réponse ayant le plus de voix (le votant ne se classe dans aucun parti).

Pour revenir à l'affaire Cambridge Analytica, avec les réponses de notre sondage, presque la moitié des répondants sont indécis et ne se classent dans aucun parti politique, ce sont des personnes dans le même état d'esprit qui ont été influencées par les équipes de TRUMP et pour les voix pour le BREXIT. Cela peut être très dangereux, et peut prendre des proportions énormes. A l'heure actuelle avec la pandémie, les gens sont de plus en plus sur leur téléphone ou ordinateur, c'est donc d'autant plus facile de les influencer, du fait qu'on est pratiquement sûr que ces personnes vont regarder leur téléphone plusieurs fois par jour. Les gens sont de plus en plus connectés, ce qui peut être un avantage mais ce qui peut être également très dangereux.

Limites et perspectives

Il existe de nombreuses perspectives sur ce sujet, faire le sujet avec plus de données, et des données plus représentatives de la France, pour pouvoir faire de l'inférence statistique et généraliser le modèle. La base de données connaît aussi des limites. Etant donné que notre base n'était pas représentative et hétérogène, nous ne pouvions pas sortir un bon modèle. Il y a une limite psychologique aussi, ce n'est pas parce que des pubs vont être publiées sur Facebook que les gens vont suivre les pubs, il y a une grande partie aléatoire, en fonction de l'influençabilité des personnes. Certaines personnes même indécises, ne sont pas influençable et donc ne suivront pas les publications Facebook et préférerons voter blanc au lieu de voter pour un parti.