UNIVERSITÉ DE NANTES

IAE NANTES
ÉCONOMIE & MANAGEMENT

Note de Synthèse

Est-il possible de prédire le vote d'un citoyen à partir de données collectées ?

Mot Clés : Variable Qualitative, Modèle Multinomial Non Ordonnée Général, Election, Big Data, Cambridge Analytica

Année 2020-2021

Kyllien ROMAND

Présentation du Sujet

L'étude suivante a pour but de démontrer les risques que peuvent présenter le big data et l'analyse de celui-ci dans un vote démocratique (élection, référendum). Basé sur le scandale de Cambridge Analytica ; scandale qui a mis au jour la possibilité de manipulation de la population à partir des données des réseaux sociaux sur l'élection américaine de 2017 mais aussi sur le référendum du Brexit, cette manipulation étant déjà prouvé avec l'utilisation des données à des fins marketing, cependant la nouveauté avec ce scandale est l'utilisation d'un modèle psychologique afin de déterminer la matrice de personnalité de chaque individus (modèle du BIG FIVE).

Méthode Économétrique : Modèle Multinomial Non Ordonnée Général

Ce type de modèle permet de prendre en compte comme variable expliquée une variable de type qualitative, discrète et non ordonnée (le parti politique ne peut être quantifier), ces modèles appelés parfois modèles logit polytomiques non ordonnées sont une famille du modèle de base : le logit multinomial. Ce modèle permet donc de prendre en compte chaque variable liés à l'individu en fonction de chaque parti politique et donc de vérifier, par exemple, si un type de catégories sociales ou un type de tranche de revenu est plus explicatif pour un parti politique.
Afin de différencier tous les modèles mis en place lors de l'étude économétrique, il est étudié le pseudo-R^2 de McFadden qui correspond au degré d'ajustement et donc de l'explication des divergences du modèles, et le log de vraisemblance.

Les Données

Pour étudier ce phénomène et la possibilité de prédiction de vote d'un individu, un sondage a été mis en place avec pour question principale quel est le parti pour lequel le sondé irait voter pour l'élection présidentielle française de 2022, la population sondées est donc l'ensemble de la population française ayant la possibilitée de voter en 2022.

Afin d'utiliser la méthode économétrique, décrite précedemment, il était nécessaire d'avoir des variables qui étaient dépendantes aux votes des individus, les variables choisis ont été le vote des deux parents puisqu'il est considéré que dans la plupart du temps les idées politiques sont d'abord établis par les idées de ses parents, puis se forgent au fur et à mesure du temps sauf si un conflit avec ces parents apparaissent, ce qui amènent à une totale contradiction dans les idées politiques. La dernière variable dépendante à la variable expliquée est les différents politiques qu'ils considèrent comme étant important à leur yeux, ces axes sont en général très important dans le choix d'un vote, puisqu'un citoyen va voter pour la candidat qui représente le mieux ces axes ou idéaux politiques, par exemple : un individu qui a pour axes politiques principaux la "sécurité" et "l'immigration" est généralement une personne ayant des affiliations de droites ou d'extrêmes droites, cependant la sécurité n'est plus aussi vrai qu'avant puisque l'insécurité est martelée sur les médias et par de nombreux partis de divers bords politiques.

Des variables concernant le modèle psychologique BIG FIVE ont été utilisé, afin que les sondés puissent se quantifier sur les 5 axes de personnalités en fonction de la force, de 1 à 10, que représente leur attrait par rapport à un axe : leur extraversion, leur ouverture d'esprit, leur agréabilité, leur névrosisme et leur côté consciencieux.

Puis des questions concernant les individus ont été posées tel que leur âge, leur tranche de revenu, leur catégorie sociale, leur niveau d'étude … Ces questions ont pour but de vérifier si à l'aide des données publiques (données présentes sur les réseaux sociaux), il est possible de catégoriser les individus et d'utiliser ces dernières afin de déterminer pour quel parti politique ils ont le plus d'affinité. Par exemple, dans les mœurs, il est dit que plus on vieillit ou plus la tranche de revenu est élevée plus on a de chance de voter à droite, ou que les partis extrêmes sont des partis dits populaires.

Résultats Statistiques du Sondage

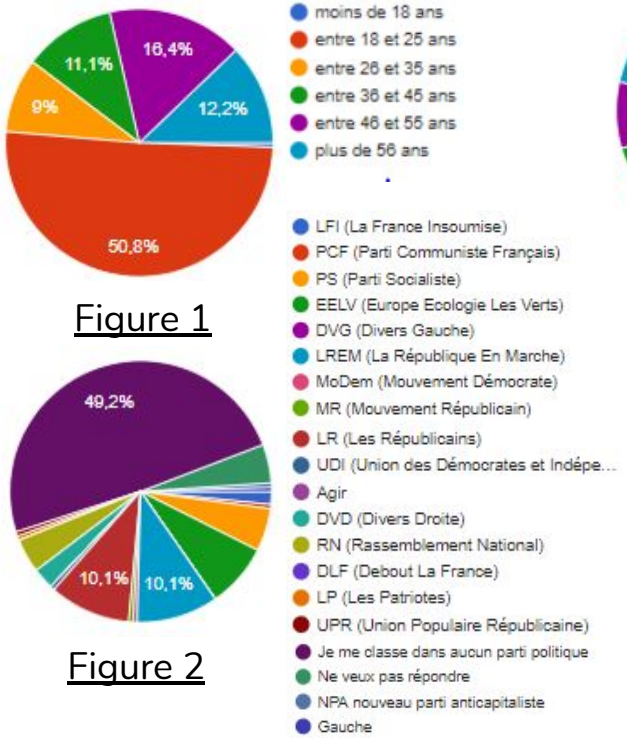

Figure 1

Figure 2

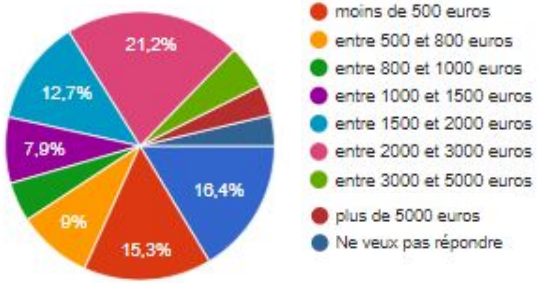

Figure 3

Figure 1 : Age des Sondés
Figure 2 : Parti Politique
Figure 3 :Tranche de Revenu

Ces 3 figures permettent de représenter la population sondés (189 réponses), et de mettre en avant le fait que quasiment 50% des sondés ne se considèrent pas dans un parti politique, et que toujours 50% des sondés ont entre 18 et 25 ans.
L'échantillon est donc pas du tout représentatif de la population française.

Discussion Autour des Résultats Économétrique

À l'aide de modèle économétrique il est difficile d'affirmer s'il est possible ou pas de prédire le vote d'un citoyen, d'une part le nombre de sondés étant trop faible il est impossible de faire de l'inférence (189 réponses), puis l'échantillon n'est pas représentatif de la population française puisqu'il est grande partie constitué d'étudiant provenant pour la plupart de l'IAE de nantes. De plus, d'autres biais techniques sont apparus lors de la conception des modèles.

Cependant il a pu être remarqué que le vote des parents et aussi les axes politiques sont hautement significatives qu'importe le modèle étudié, le vote des parents n'étant pas une donnée accessible elle n'est pas un risque pour le citoyen, à contrario des axes politiques qui peuvent parfois être détecté à l'aide des pages ou groupes suivis sur les réseaux sociaux ou des partages qui sont effectués, et pour lesquels l'individu adhère.

De plus, différents résultats permettent de mettre à jour différents clichés existant dans le paysage politique français, car il a pu être remarqué que l'âge expliqué au seuil de risque 10% une grande partie des personnes votant pour le parti Rassemblement National, ou que le fait de s'informer via les Journaux Télévisé était hautement significatif dans le cadre des votant pour le parti LREM, mais aussi que la catégorie socioprofessionnelle des parents pouvait expliquer le vote pour les partis de Divers Droite et du Rassemblement National.

Ces résultats sont à prendre avec précaution car de nombreuses limites sont apparus tout au long de l'étude, il est donc impossible d'affirmer ou de refuser l'idée que la prédiction de vote d'un citoyen peut être effectué à partir de donnée collectée, par contre il est possible d'affirmer que le big data et l'analyse de celui-ci permet d'influencer le choix de vote d'une personne comme la montré l'affaire Cambridge Analytica, c'est pourquoi il faut faire attention aux données que l'on partage sur Internet car elles peuvent être utilisé contre nous et à notre insu.