Rapport du projet final — Prédictions des résultats électoraux

Travail présenté à Monsieur Christian Gagné
GIF-4101 – Introduction à l'apprentissage machine
Automne 2019

Équipe 8

Alexandre Richard	*111156356*
Félix Bouchard	*111160393*
Bryan Elliott Tam	*111133067*
Alexandre Lortie-Rochette	*111 146 499*

Faculté des sciences et de génie

1. Présentation du problème

Les prédictions électorales ont été historiquement peu fiables et ont eu de la difficulté à capter certains phénomènes importants. De plus, la conduite de sondages téléphoniques auprès des individus est longue, dispendieuse et biaisée. En effet comme avance Seth Stephens-Davidowits un ancien analyste chez google les individus ont tendances à mentir à leurs amis, leur famille, aux sondages et même soi-même. Ceci combiné a une augmentation du phénomène de bien-pensance, il est important de développer des méthodes alternatives au sondage pour déterminer les intentions et préoccupation réelle des individus. Dans cette optique le travail cherche à déterminer si les recherches effectuées sur Google permettent de prédire les intentions de votes des citoyens canadiens.

La base du jeu de données utilisées se base sur les résultats des élections canadiennes entre 2004 et 2019, soit la période qui concorde avec les données disponibles sur Google Trends. Les données qui nous intéressent sont disponibles sur le site de Wikipédia.org pour présenter les informations sur les chefs de parti, le nombre de votes reçus, le nombre de députés élus, l'augmentation du nombre de voix selon l'année précédente.

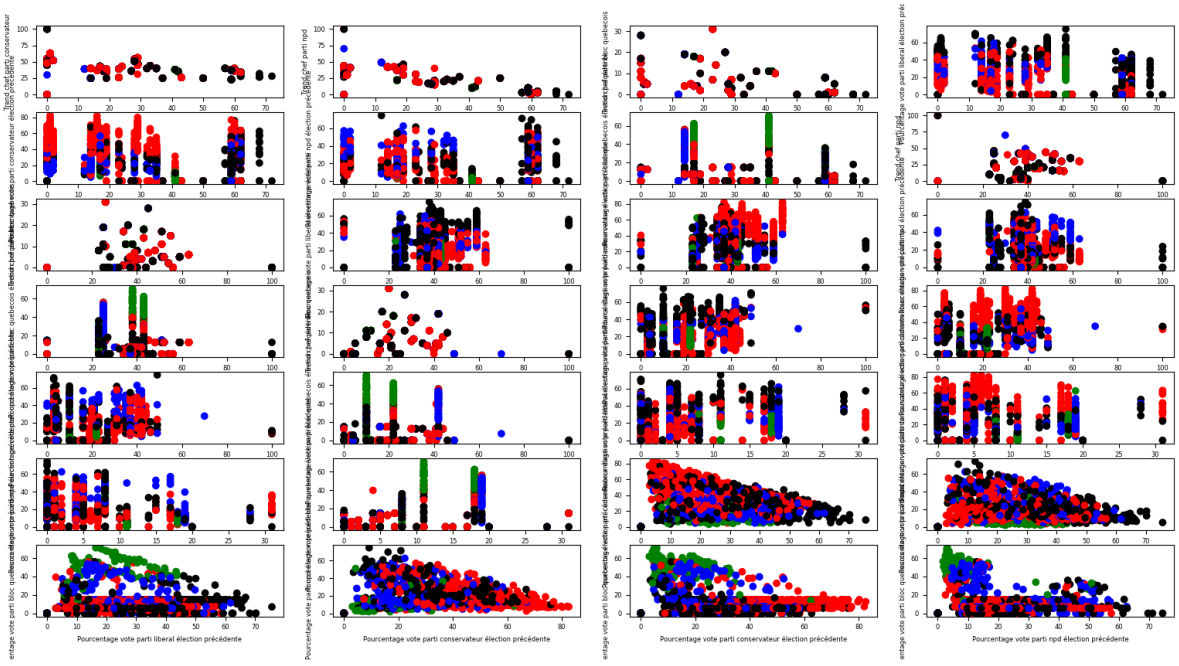
2. Approche proposée

En raison du faible nombre de donnés et un nombre de variables presque qu'illimité il est nécessaire de faire une présélection des variables. Cette présélection est en fonction des sujets les plus souvent abordés durant une élection. Cette information combinée au nombre de recherche Google sur le sujet nous a permis de sélectionner 6 sujets que les citoyens canadiens recherchent sur Google. Les sujets sont la politique, la santé, l'éducation, l'environnement, l'économie et les chefs de parti. Ces sources d'information s'ajoutent aux résultats de l'élection précédente pour être utilisées pour prédire le résultat de l'élection dans chaque circonscription. En raison du faible nombre d'élections disponible qui par conséquent limite la taille de l'échantillon. Nous avons consacré nos énergies sur des méthodes de classification pouvant bien fonctionner avec un minimum de points comme la méthode knn et la méthode svm.

3. Méthodologie expérimentale

La première étape à réaliser dans ce travail a été de choisir les sources d'information, les collecter et les normaliser. Dans le paragraphe précédent, la politique, la santé, l'éducation,

l'environnement, l'économie et les chefs de parti ont été sélectionnés comme indicateur pour faire notre classification. Pour les évaluer, il est nécessaire de déterminer les mots clés qu'un individu intéressé par un de ces sujets utilise dans ces recherches internet. Pour réaliser cette tâche, nous avons analysé les termes les plus recherchés sur Google et les avons inscrits dans une base de données. Nous avons par la suite conduit des essais sur ces termes et avons conservé les termes les plus prometteurs pour chaque sujet. Initialement, l'objectif était de couvrir plus de six sujets sur Google Trend et utiliser plus de cinq mots clés par sujet cependant, les limites de l'api utilisé, pour faire ces recherches, nous a empêché de comparer plus de cinq termes à la fois et le nombre de requêtes totales était aussi limité. Pour donner suite à la collecte complète de données, l'équipe a tracé les résultats des élections selon toutes les combinaisons possibles des paramètres dans la figure suivante.

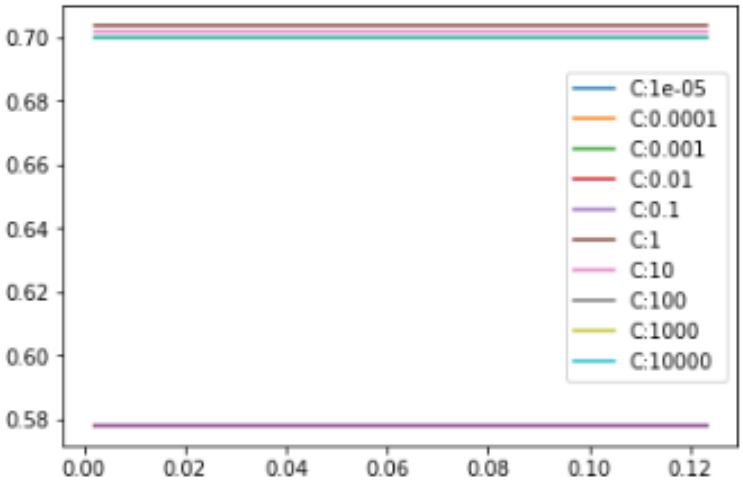


Légende : Noir : Libéral, Rouge : Conservateur, Bleu : NPD : vert : Bloc Québécois

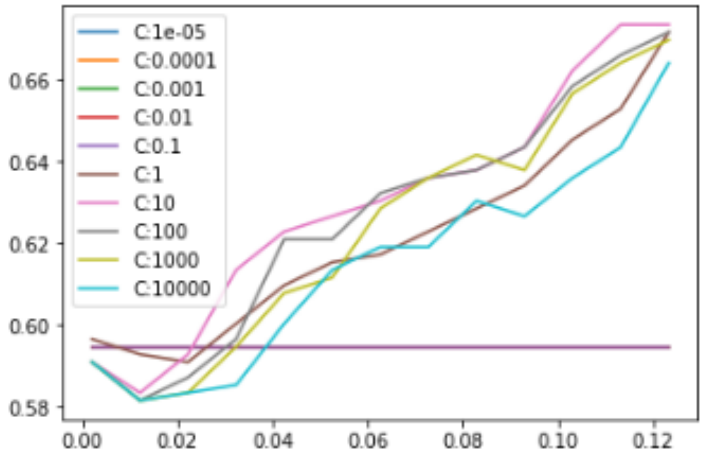
La figure illustre l'importance de normaliser les données certaines ont une variance élevée et sans la normalisation cela peut conduire à des biais dans l'entrainement particulièrement considérant que la méthode KNN et SVM sont basés sur la distance. Pour effectuer la classification les deux méthodes sont entrainées sur les données séparées en données d'entrainement et en données de test en raison du peu de données disponibles on utilise KFold pour utiliser la stratégie Leave one out dans notre modèle.

4. **Résultats expérimentaux**

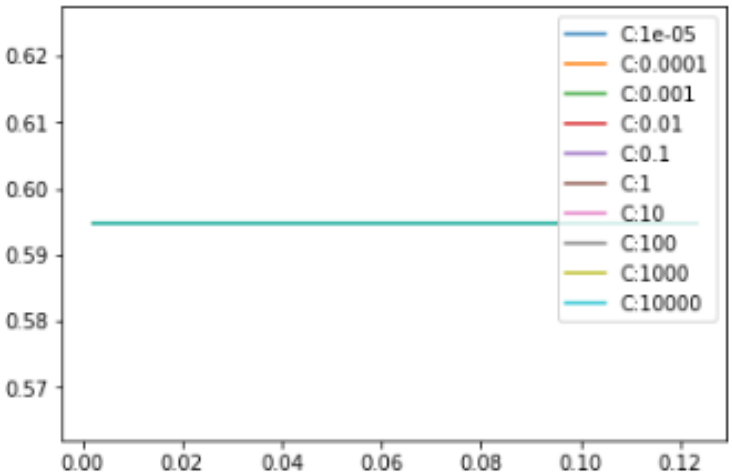
SVM linéaire :



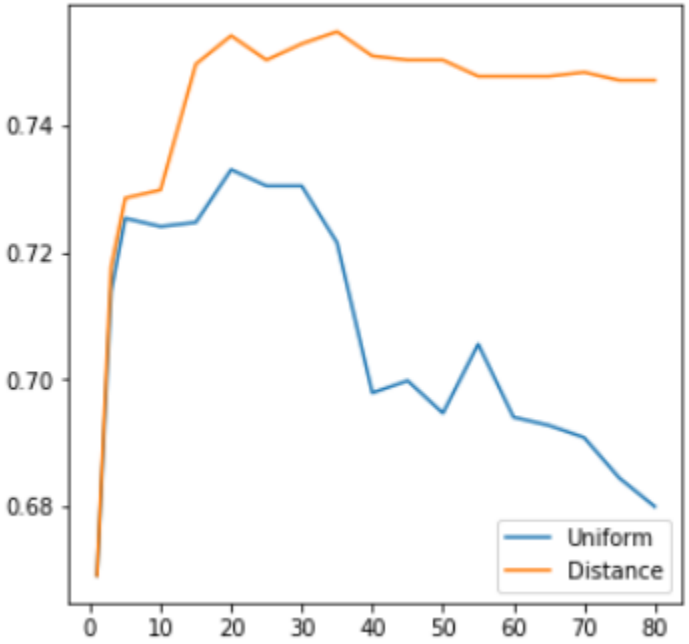
SVM RBF :



SVM SIGMOID :



KNN : Score en fonction du nombre de voisins considérés.



5. Analyse des résultats

Le modèle de classification KNN avec K=35 prédit correctement quel parti sera élu dans une circonscription 75.5% du temps. Considérant que les méthodes traditionnelles prédisent généralement à plus ou moins 6% du résultat réel nos méthodes de prédictions font piètre figure. Cependant, un taux de classement de 75.5% démontre qu'il existe des relations entre les intentions de vote des individus et les recherches internet qu'ils effectuent. Pour le SVM la méthode la plus performante est linéaire avec un score de 70%. Comme pour KNN cela indique qu'il existe une relation entre les recherches internet et l'intention de vote. Plusieurs raisons peuvent justifier le résultat décevant de nos algorithmes testés. Premièrement, les tendances de recherche disponible sur Google Trend sont disponibles par province ou par ville. Faire la relation entre la ville et une circonscription s'est avéré extrêmement complexe et les mesures ont donc été prises pour les provinces créant une généralisation d'une mentalité qui peut différer d'une circonscription à l'autre. Deuxièmement, la limite imposée par l'api de Google nous empêchait de collecter une grande quantité de variables et de les comparer. Lors du test avec 25 mots clés en français et 25 mots en anglais, l'api a refusé de traiter nos demandes, car elles étaient trop nombreuses. Dans un monde idéal, la méthode à utiliser pour résoudre ce type de problème aurait été de sortir l'ensemble des thèmes de juridiction fédérale et utiliser la sélection arrière séquentielle pour diminuer la dimensionnalité et conserver

uniquement l’information pertinente et utiliser cette dernière pour entrainer nos modèles de classification.