Validation de la robustesse du modèle de prédiction électorale

CLESSN

Invalid Date

Table des matières

1 Résumé exécutif

Ce rapport présente les résultats de l'évaluation de la robustesse du modèle de prédiction électorale. Le modèle a été testé sur un ensemble de validation comprenant 17 591 observations, représentant 30% des données disponibles. Les principales métriques de performance sont:

- **Accuracy globale**: 54,19%
- **Accuracy du premier ou deuxième choix**: 82,72%
- **Recall moyen (macro)**: 45,09%
- **Intervalle de confiance à 95% pour l'accuracy**: 48,45% - 54,04%

Le modèle présente une performance acceptable pour la prédiction du vote, particulièrement lorsqu'on considère le premier ou le deuxième choix prédit. Les performances varient selon les partis et les régions, avec des résultats particulièrement bons dans l'Atlantique et les Prairies.

2 Introduction

Ce document présente les résultats détaillés de l'évaluation du modèle de prédiction du choix de vote fédéral. L'objectif est d'évaluer la robustesse et la fiabilité du modèle, notamment sa capacité à prédire correctement le premier ou le deuxième choix de vote.

Le modèle a été entraîné sur des données provenant de deux sources principales: une étude pilote et une application de collecte de données. Les évaluations ont été effectuées sur un ensemble de validation représentant 30% des données disponibles.

2.1 Spécification du modèle

Le modèle évalué dans ce rapport est basé sur la formule suivante, qui comprend de nombreuses variables explicatives en interaction avec les régions (Québec vs. Reste du Canada):

```
dv_voteChoice
```

Cette formule montre que le modèle:

1. Utilise une approche différenciée entre le Québec et le reste du Canada (ROC)
2. Intègre des variables socio-démographiques (âge, genre, éducation, revenu)
3. Inclut des variables de style de vie (transport, alimentation, loisirs)
4. Incorpore des prédictions issues d'autres modèles pour chaque parti

3 Méthodologie

3.1 Préparation des données

Les données ont été préparées en suivant ces étapes:

1. Combinaison des données enrichies de l'étude pilote et de l'application
2. Filtrage pour ne conserver que les observations avec un choix de vote valide
3. Sélection des variables pertinentes selon le modèle
4. Suppression des observations avec des valeurs manquantes

L'ensemble final pour la validation comprend 17 591 observations, réparties comme suit:

Table 1: Distribution des votes dans l'ensemble de validation

Parti	Application	Pilote	Total
BQ	4197	42	4239
CPC	4248	101	4349
LPC	5686	71	5757
NDP	2825	55	2880
GPC	353	13	366

3.2 Méthodes d'évaluation

Pour évaluer la robustesse du modèle, nous avons utilisé plusieurs métriques:

- **Accuracy globale**: Proportion des prédictions correctes
- **Matrice de confusion**: Distribution des prédictions par rapport aux valeurs réelles
- **Précision, Recall et F1-Score**: Métriques par parti
- **Top-N Accuracy**: Proportion des cas où la vraie classe est parmi les N premières prédictions
- **Analyse régionale**: Performance du modèle par région
- **Analyse de stabilité**: Estimation de la variabilité de la performance via bootstrap
- **Analyse de calibration**: Évaluation de la fiabilité des probabilités prédites

4 Résultats

4.1 Performance globale

Le modèle atteint une accuracy globale de 54,19%, ce qui est considérable dans un contexte multipartite à 5 options. Lorsqu'on considère le premier ou deuxième choix (top-2 accuracy), la performance s'élève à 82,72%.

Table 2: Métriques de performance globale

Métrique	Valeur
Accuracy globale	0.5419
Accuracy premier ou deuxième choix	0.8272
Recall moyen (macro)	0.4509
Recall pondéré (weighted)	0.5419

4.2 Performance par parti

Table 3: Métriques par parti politique

Parti	Précision	Recall	F1_Score	Support	Top2_Accuracy
BQ	0.5459	0.8205	0.6556	4239	0.9571
CPC	0.5539	0.5061	0.5289	4349	0.7705
LPC	0.5177	0.4341	0.4722	5757	0.9024
NDP	0.5600	0.4667	0.5091	2880	0.6712
GPC	0.5263	0.0273	0.0519	366	0.0437

La performance varie considérablement selon les partis:

- Le **Bloc Québécois (BQ)** obtient le meilleur recall (82,05%) et une excellente top-2 accuracy (95,71%)
- Le **Parti Conservateur (CPC)** et le **NPD** montrent des performances équilibrées
- Le **Parti Libéral (LPC)** a un bon taux quand on considère le premier ou deuxième choix (90,24%)
- Le **Parti Vert (GPC)** présente des performances plus faibles, probablement dues au nombre limité d'observations

4.3 Matrice de confusion

Table 4: Matrice de confusion

	BQ	CPC	LPC	NDP	GPC
BQ	3478	818	1624	372	79
CPC	222	2201	1046	428	77
LPC	392	1092	2499	732	112
NDP	147	234	587	1344	88
GPC	0	4	1	4	10

La matrice de confusion révèle plusieurs tendances intéressantes:

- Le modèle prédit bien le BQ pour les électeurs du BQ (3478 bonnes prédictions)
- Il existe une confusion significative entre les principaux partis, notamment entre LPC et CPC
- Le GPC est rarement prédit, même pour les électeurs réels du GPC

4.4 Performance par région

Table 5: Performance par région

Région	Observations	Accuracy	Top2_Accuracy
Prairies	1890	0.5799	0.8402
Colombie-Britannique	1223	0.4816	0.8029
Atlantique	787	0.6010	0.8818
Ontario	3900	0.5487	0.8556
Québec	9762	0.5348	0.8122
Territoires	29	0.4483	0.7586

L'analyse régionale montre que:

- Les meilleures performances sont obtenues dans la région **Atlantique** (60,10% d'accuracy)
- Les **Prairies** suivent de près avec 57,99%
- La **Colombie-Britannique** et les **Territoires** présentent les performances les plus faibles
- Le **Québec**, qui représente le plus grand nombre d'observations, atteint une accuracy de 53,48%

4.5 Comparaison Québec vs Reste du Canada (ROC)

Table 6: Comparaison Québec vs Reste du Canada

Région	Accuracy	Top2_Accuracy
Québec	0.5348	0.8122
Reste du Canada (ROC)	0.5506	0.8460

Le modèle performe légèrement mieux dans le reste du Canada (ROC) comparativement au Québec, tant pour l'accuracy simple que pour l'accuracy du premier ou deuxième choix.

4.6 Stabilité du modèle

L'analyse bootstrap avec 100 répétitions sur un échantillon de 1000 observations révèle:

Table 7: Résultats de l'analyse bootstrap

Métrique	Valeur
Accuracy moyenne	0.5177
Écart-type	0.0138
Borne inférieure IC-95%	0.4845
Borne supérieure IC-95%	0.5404

Cette faible variation indique une bonne stabilité du modèle malgré les différences régionales et partisanes.

4.7 Analyse de calibration

L'analyse de calibration examine si les probabilités prédites correspondent aux fréquences observées. Voici un aperçu pour le BQ:

Table 8: Calibration pour le Bloc Québécois

Bin_Probabilité	Fréquence_Observée	Nombre_Observations	Prob_Moyenne_Prédite
[0,0.1]	0.0070	8468	0.0151
(0.1,0.2]	0.1597	745	0.1523
(0.2,0.3]	0.2572	1112	0.2533
(0.3,0.4]	0.3360	1494	0.3532
(0.4,0.5]	0.4706	1887	0.4512
(0.5,0.6]	0.5787	2065	0.5493
(0.6,0.7]	0.6432	1522	0.6434
(0.7,0.8]	0.7061	296	0.7273
(0.8,0.9]	1.0000	2	0.8112

Cette analyse montre que le modèle est généralement bien calibré pour le BQ, avec des probabilités prédites proches des fréquences observées, surtout dans les bins intermédiaires.

5 Discussion

5.1 Forces du modèle

1. **Performance satisfaisante en contexte multipartite**: Une accuracy de 54,19% est substantielle dans un système à 5 partis
2. **Excellente performance en top-2**: Prédire le premier ou deuxième choix avec 82,72% d'accuracy
3. **Bonne performance régionale**: Particulièrement en Atlantique et dans les Prairies
4. **Stabilité**: Faible variation des performances dans l'analyse bootstrap
5. **Bonne calibration**: Les probabilités prédites correspondent généralement aux fréquences observées

5.2 Limites et points d'amélioration

1. **Performance inégale selon les partis**: Particulièrement faible pour le Parti Vert
2. **Confusion entre partis**: Notamment entre le LPC et le CPC
3. **Performance régionale variable**: Plus faible en Colombie-Britannique et dans les Territoires
4. **Déséquilibre dans les données**: Surreprésentation du Québec dans l'échantillon

6 Conclusion

Le modèle de prédiction électorale démontre une robustesse et une fiabilité satisfaisantes, particulièrement dans sa capacité à identifier le premier ou deuxième choix des électeurs. Avec une accuracy globale de 54,19% et une top-2 accuracy de 82,72%, il offre une base solide pour l'analyse électorale.

Les variations de performance entre partis et régions soulignent l'importance de considérer ces facteurs dans l'interprétation des prédictions. La stabilité du modèle, confirmée par l'analyse bootstrap, renforce sa crédibilité malgré ces variations.

En conclusion, ce modèle constitue un outil précieux pour l'analyse des intentions de vote, tout en reconnaissant ses limites actuelles et les opportunités d'amélioration futures.

7 Annexes

7.1 Détails techniques

- **Seed pour la partition**: 42
- **Seed pour le bootstrap**: 123
- **Proportion d'entraînement/validation**: 70%/30%
- **Nombre de répétitions bootstrap**: 100
- **Taille de l'échantillon bootstrap**: 1000 observations

7.2 Partis politiques

- **BQ**: Bloc Québécois
- **CPC**: Parti Conservateur du Canada
- **LPC**: Parti Libéral du Canada
- **NDP**: Nouveau Parti Démocratique
- **GPC**: Parti Vert du Canada