Modelo de clasificación para la predicción del voto en las elecciones presidenciales en los Estados Unidos de América

UANL
UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

HERNÁNDEZ-SEMPORINI Héctor Jesús
jesus.hernandezs@uanl.edu.mx
Universidad Autónoma de Nuevo León
Facultad de Ciencias Físico Matemáticas

FCFM
FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

Introducción

La predicción del sentido del voto ha sido un problema que se ha abordado desde el nacimiento de la democracia misma. Este es un problema multifacético, debido al constante y cambiante carácter de la sociedad y la infinidad de diferencias que construyen la individualidad de las personas.
La importancia de la elección de un presidente es una que impacta múltiples aspectos de la vida política, económica y social de un país. Es por esto que la previsión del resultado electoral sirve como principio importante de planificación en muchos sentidos. Uno de estos sentidos sería el económico ya que la diferencia en corrientes entre candidatos puede poner el riesgo la operación de industrias enteras que pueden afectar el rumbo económico.

Metodología

La metodología a utilizar es se construye en tres partes. La primera parte es sobre la definición de de los datos a utilizar, siendo la limpieza de la base de datos y la selección de variables lo más relevante, la segunda parte consiste en el uso de la metodología de *Extreme Gradient Boosting* (*XGBoost*) o refuerzo de gradientes extremo para el entrenamiento del modelo en base a los datos de 2016 y por último la tercera parte consiste en la generación de resultados al utilizar el modelo de datos entrenado con los datos de 2016 para predecir el voto en la elección presidencial de 2020 en Estados Unidos de América.

La primera parte de la metodología consiste en hacer una selección importante de las variables a utilizar y a hacer un preprocesamiento general de los mismos datos. Los datos se toman de una serie de encuestas de corte longitudinal a un muestreo de personas votantes en Estados Unidos. La encuesta consiste de un banco de preguntas de diferentes características que consisten en aspectos sociales, demográficos, políticos, económicos, de creencias personales y sobre temas considerados relevantes para la vida política y social en el país [1]. Una de estas preguntas es si se votó en las elecciones de ese año y por qué candidato se votó. Cada una de estas personas participa de este estudio en cada año por lo que se tiene ese corte longitudinal, lo que permite el estudio de cambios a lo largo del tiempo en sus comportamientos y creencias. En base a un estudio de la literatura existente sobre las características consideradas relevantes para la predicción del voto por otros autores [3-7] se hace la selección de preguntas a tomar en cuenta de las más de 700 preguntas individuales que se realizan a cada uno de los candidatos.

Se termina con la siguiente selección de características:

- presvote_2016: Por quién se votó en las elecciones del 2016*
- obamaapp_2016: Que tanto aprueban del gobierno del presidente Obama (presidente anterior)
- fairsociety_2016: Que tanto consideran que la sociedad americana es justa con todos
- beliefinmedia_2016: Que tanto creen en lo que dicen en los medios de comunicación
- persfinretro_2016: Situación económica mejor o peor que hace un año
- trustgovt_2016: Confianza y creencia en los políticos de Washington
- immi_contribution_2016: Los inmigrantes contribuyen a la sociedad americana
- view_deathpen_2016: Creen en la pena de muerte como un castigo adecuado
- envwarm_2016: Creen en el calentamiento global
- govtreg_business_2016: Creen que el gobierno debe regular más fuertemente a los mercados
- sexism_equality_2016: Creen en la búsqueda de la igualdad para las mujeres
- gunowner_2016: Son dueños de un arma de fuego
- educ_2016: Nivel educativo
- marstat_2016: Estatus marital
- employment_2016: Estatus de empleo
- newsint_2016: Nivel de interés en las noticias
- religion_2016: Cual es su religión si es que se identifican con alguna

*Como parte de la limpieza de datos solo se dejaron observaciones donde el voto fuera estrictamente a favor de los candidatos demócrata o republicano debido a las mínimas observaciones para otros candidatos.

Ya con el set de datos trabajados se procede a hacer uso de la metodología *XGBoost*. La metodología está basada en la siguiente función objetivo, en la que se busca agregar un t-ésimo *booster* a nuestro modelo [2].

$$L^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i)) + \Omega(f_t)$$

Para esta función objetivo se le busca minimizar utilizando el método de Newton, lo cual se puede lograr utilizando la aproximación de series de Taylor de segundo orden en el punto $\hat{y}_i^{(t-1)}$

$$l(y_i, \hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i)) \approx l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t(\mathbf{x}_i)^2$$

Donde la gradiente y el hessiano se definen como

$$g_i = \frac{\partial}{\partial \hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}) \qquad h_i = \frac{\partial}{\partial^2 \hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$$

Con lo que se obtiene la siguiente función de pérdida

$$\tilde{L}^{(t)} = \sum_{i=1}^{n} [g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t(\mathbf{x}_i)^2] + \Omega(f_t)$$

Para calcular el valor de predicción óptimo para cada rama en ft se usa

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda}$$

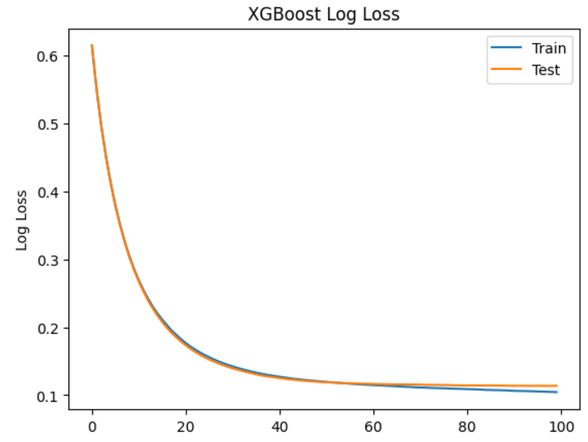
Finalmente, para ir calculando y determinando si cada rama debe detenerse porque no logra una ganancia se utiliza

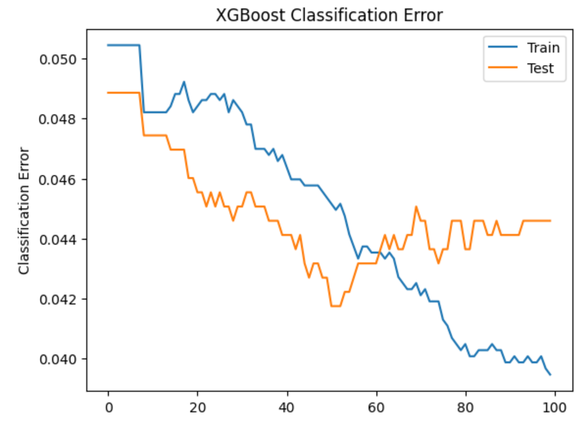
$$\Delta L = \frac{1}{2} \left[\frac{(\sum_{i \in I_L} g_i)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{(\sum_{i \in I_R} g_i)^2}{\sum_{i \in I_R} h_i + \lambda} - \frac{(\sum_{i \in I} g_i)^2}{\sum_{i \in I} h_i + \lambda} \right] - \gamma$$

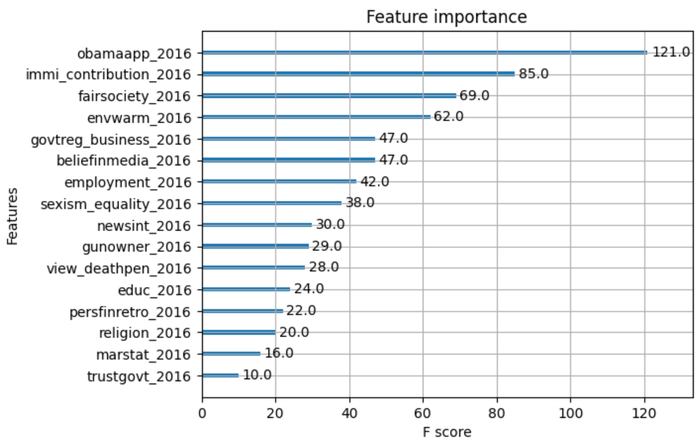
Resultados Preliminares

Primero viendo los resultados del entrenamiento del modelo realizado a los datos de las elecciones de 2016.

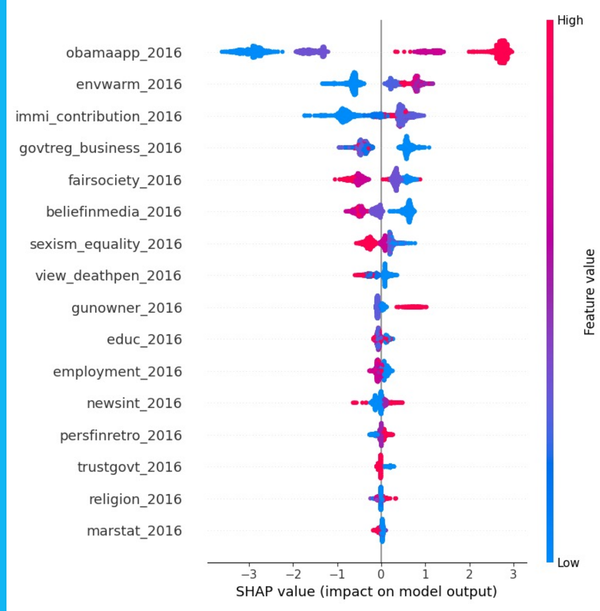
$$M\ de\ Conf.1 = \begin{bmatrix} 1024 & 47 \\ 47 & 990 \end{bmatrix} \quad M\ de\ Conf.2 = \begin{bmatrix} 981 & 90 \\ 80 & 957 \end{bmatrix}$$

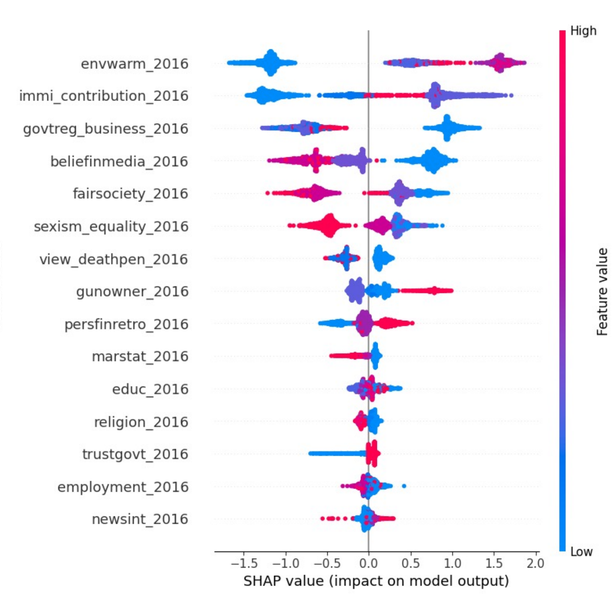






Indicador	Resultado 1	Resultado 2
Exactitud (Acc.)	95.54%	91.94%
Precisión	95.47%	91.40%
Recall	95.46%	92.29%
f1	95.46%	91.84%





Según los resultados del primer entrenamiento es excelente en cuestión de las métricas de evaluación tradicionales, viendo las gráficas de importancia de las características y de los valores de Shapley se observaba que la característica de aprobación de Obama es muy alta e importante por lo que se decidió comprobar los resultados sin esa característica en particular, para evitar que esa característica fungiera como una variable proxy para nuestra variable dependiente. Lo que se observó en el segundo entrenamiento es no era el caso y aún con sin esa característica en particular se obtenía buen resultado.

Conclusiones Preliminares

De manera preliminar en base a los resultados se observa que las características selecionadas sí son de alta importancia para la definición del voto y se observa que varios de los temas centrales en los puntos de vista de cada uno de los candidatos de las elecciones tanto de 2016 como del 2020 son altamente relevantes para la predicción, como lo son el calentamiento global, los inmigrantes, una sociedad justa, los medios de comunicación y el sexismo. Por otro lado, varias de los temas típicamente polarizantes como el empleo, la economía, religión y educación tuvieron un impacto menos fuerte al esperado.
Los siguientes pasos son ver la efectividad del entrenamiento del voto del 2016 y aplicarlo al 2020 para ver si los patrones de voto cambiaron y el modelo deja de funcionar predictivamente.

Referencias

[1] Democracy Fund Voter Study Group. Views of the Electorate Research Survey, December 2016 (2017). Washington DC: Democracy Fund Voter Study.
[2] Bowers, M. (2022). XGBoost Explain. Random Realizations.
[3] Lewis-Beck, M. & Steigmaier M. (2000). Economic Determinants of Electoral Outcomes. Annual Review of Political Science (3). pp. 183-219.
[4] Markus, G. (1988). The Impact of Personal and National Economic Conditions on the Presidential Vote: A Pooled Cross-Sectional Analysis. American Journal of Political Science 32: 137-154.
[5] Molina J. & Pérez C. (2004). Radical Change at the Ballout Box: Causes and Consequences of Electoral Behavior in Venezuela's 2000 Elections. Latin American Politics and Society 46 (1): 103-134.
[6]Morris, P. (1978). Economic Retrospective Voting in American National Elections: A Micro-Analysis. American Journal of Political Science 22: 426-443.
[7]Settle R. & Abrams B. (1973). The Determinants of voter participation: a more general model. Public Choice 27: 81-89.