Estimación de voto a partir del CIS - Electomedia

A lo largo de la práctica fin de bootcamp hemos hecho análisis sobre resultados ya conocidos de las elecciones. Nos podemos preguntar si sería posible hacer una estimación de voto futuro a partir de un sondeo. Para ello nos basamos en los microdatos del sondeo del CIS antes de las elecciones de noviembre de 2019. El CIS es la única firma de sondeos que hace público sus microdatos, y en el caso de los sondeos preelectorales se trata de una muestra de casi 18.000 encuestas.

Objetivo del modelo

La previsión de voto la hemos desarrollado en el cuaderno ***CIS_N19_def***, y realmente no ha resultado excesivamente complicado. La variable objetivo a modelizar es la intención directa de voto de los encuestados; aproximadamente un 30% de éstos, en cambio no dice lo que van a votar, y ello es precisamente lo que hay que modelar y después estimar.

Preprocesado

El dataset cuenta con 146 columnas, que desde luego no nos parecen todas relevantes; nos centramos en aquellas que contienen valoraciones políticas de partidos o líderes, así como escalas de ubicación política tanto de los partidos como de los propios encuestados. Se han considerado algunos datos personales de estos últimos, tales como la edad. En total nos quedamos con unas 42 columnas. Como se trata de una prueba de concepto, hemos seleccionado las encuestas de solo una de las provincias, en este caso Málaga, por lo que consideramos unas 480 filas.

Pasamos a modelizar la intención de voto declarada, presente en unas 370 filas, incluido aquellos que declaran que no irán a las urnas. Las 110 filas restantes se tendrán que estimar con el modelo ya definido. El preprocesado consiste principalmente en un mapeo de columnas categóricas a valores numéricos, un paso en su mayor parte bastante sencillo.

Modelo de random forest

Separamos las 370 filas entre train y test, y aplicamos un modelo *random forest classifier*, que ajustamos mediante *grid search*. Conseguimos un modelo con un *accuracy* en el dataset de test superior al 90%, evitando en principio pues el *overfitting*.

A continuación, aplicamos el modelo a las columnas que no declararon su intención de voto y así la estimamos. Ya teniendo la intención de voto de todos los encuestados, la comparamos con su recuerdo de voto en las elecciones de abril de 2019. Así podemos ver los cambios de voto respecto a esas elecciones de cara a noviembre de 2019, siendo esta nuestra predicción del resultado.

Resultado y conclusiones

La estimación no es muy precisa cuando se la compara con los resultados reales. Existe un sesgo a favor del PSOE y Ciudadanos, y en contra de PP y Vox. Con todo, esta misma desviación, casi en las mismas magnitudes, se dio en el sondeo del CIS para el conjunto de toda España, por lo que no pensamos que sea para estar insatisfecho. Quizás el punto más satisfactorio es que conseguimos modelizar la intención declarada de voto de una manera muy precisa. Los detalles del proceso de estimación se pueden consultar en el propio cuaderno.