Fakulta matematiky, fyziky a informatiky Univerzity Komenského, Bratislava

Predikcie volieb podľa prieskumov verejnej mienky

projekt z predmetu Princípy dátovej vedy

Tomáš Antal
Erik Božík
Teo Pazera
Andrej Špitalský
Tomáš Varga

3. januára 2025

Obsah

1 Predstavenie témy

2 Spracovanie dát

3 Exploratívna analýza

4 Klasifikačné modely

5 Predikčné modely

Úlohu predikovania volebného výsledku môžeme rozdeliť na dve časti. Prvá je modelovanie vývoja volebných preferencii politických strán. Druhý aspekt je predikovanie skutočného výsledku volieb na základe posledného prieskumu preferencií. Je očakávateľné, že percentuálny zisk vo voľbách je väčšinou blízky percentám v prieskume mesiac pred voľbami, no história slovenských volieb ukázala, že to nie vždy platí. Budeme sa teda snažiť nájsť relevantné charakteristiky, pomocou ktorých budeme modelovať tento rozdiel.

5.1 Predikovanie vývoja volebných preferencií

V tejto časti budeme pracovať s priekumami volebných preferencií politických strán a hnutí od januára v roku 2010 do novembra v roku 2024. Pre každý z 81 politických subjektov máme teda 179 údajov v percentách o ich preferenciách.

Budeme sa teda na tieto dáta pozerať ako na časové rady, kde rozostupy jednotlivých údajov sú 1 mesiac. Tiež budeme modelovať vývoj preferencií pre každú stranu samostatne, teda náš model predpokladá, že preferencie strán sú vzájomne nezávislé. Porvnáme prístupy k modelovaniu časových radov, konkrétne Holtovo dvojité exponenciálne vyrovnávanie a ARIMA model.

5.1.1 Teória – Holtovo dvojité exponenciálne vyrovnávanie

Exponenciálne vyrovnávanie je používané na „vyhladzovanie“ časových radov. Využíva na to predpoklad, že hodnota časového radu v čase $t+1$ závisí najviac od hodnoty v čase t, menej od hodnoty v čase $t-1$ atď. Môžeme sformulovať model jednoduchého exponenciálneho vyrovnávania

$$\begin{aligned} s_0 &= X_0 \\ s_t &= \alpha X_t + (1-\alpha)s_{t-1} && \text{pre } t > 0 \end{aligned}$$

Hodnota s_t je „vyhladený“ časový rad v čase t, α je jediný koeficient modelu. Keby týmto modelom chceme predikovať budúce hodnoty časového radu $\widehat{X_{t+i}} = s_{t+i}$, vidíme, že by predikcie boli konštantné. Pre zachytenie trendu v dátach sa používa dvojité (Holtovo lineárne) exponenciálne vyrovnávanie.

$$\begin{aligned} s_0 &= X_0 \\ b_0 &= X_1 - X_0 \\ s_t &= \alpha X_t + (1-\alpha)(s_{t-1} + b_{t-1}) && \text{pre } t > 0 \\ b_t &= \beta(s_t - s_{t-1}) + (1-\beta)b_{t-1} && \text{pre } t > 0 \end{aligned}$$

Takýto model s parametrami α a β vieme využiť na predikovanie budúcich hodnôt. Tieto predikcie budú mať lineárny trend.

Teória aj označenie čerpané z: Exponential smoothing.

5.1.2 Teória – ARIMA

Model pre predikciu časových radov ARIMA je nadstavbou pre model ARMA – autoregressive moving-average. Autoregresívna časť indikuje, že predikujeme premennú

na základe jej hodnôt v minulosti. Tento prístup však vyžaduje stacionaritu, teda, neformálne, stredná hodnota a autokovariancia (kovariancia časového radu samého so sebou) a časového radu musia byť konštantné/nezávislé od času a variancia musí byť všade konečná. Moving-average model stupňa q predikuje hodnotu v čase t naslovne:

$$X_t = \mu + \varepsilon_t + \sum_{i=1}^{q} \beta_i \varepsilon_{t-i}$$

Hodnota μ je stredná hodnota časového radu, $\beta_1, \ldots, \beta_2$ sú parametre modelu a ε_i reprezentuje šum v čase i (zdroj Beáta Stehlíková: ARIMA modely, časť 1). $\mathrm{ARMA}(p, q)$ vyzerá teda nasledovne:

$$X_t = c + \sum_{i=1}^{p} \alpha_i X_{t-i} + \varepsilon_t + \sum_{i=1}^{q} \beta_i \varepsilon_{t-i}$$

Parametre $\alpha_1, \ldots, \alpha_p$ sú parametre pre autoregresívnu časť a parameter c „slúži“ na vertikálny posun. Ako sme spomínali, takýto model vyžaduje stacionaritu časového radu. Na eliminovanie trendu (a teda na dosiahnutie stacionarity) v dátach slúži diferencovanie časového radu – namiesto X_t modelujeme $Y_t = X_t - X_{t-1}$. Tento proces opakujeme d-krát, čo nám dáva integrovaný model $\mathrm{I}(d)$.

Spojením týchto troch modelov dostávame model $\mathrm{ARIMA}(p, d, q)$ – autoregressive integrated moving-average.

Spomenieme ešte nadstavbu na tento model s názvom SARIMA, ktorá pracuje aj so sezónalitou v časových radoch. Keďže sezónnosť sme v našich dátach nepozorovali, rozhodli sme sa nepredikovať týmto modelom.

5.1.3 Výber modelu

Metodiku výberu modelu sme volili s ohľadom na to, že chceme predikovať volebný výsledok, keby sú voľby v blízkej budúcnosti. Preto sme chceli, aby čo najlepšie zachytil vývoj preferencií pred poslednými voľbami v roku 2023 pre strany, ktoré sú „relevantné“ na terajšej politickej scéne. Konkrétne sú to Progresívne Slovensko, Smer-SD, Hlas-SD, Slovensko, SaS, KDH, Republika, SNS, Sme rodina a Maďarská aliancia. Politickú stranu Demokrati sme do tejto skupiny nepridali, keďže vznikli iba tesne pred voľbami v roku 2023.

Model vyberáme teda na základe toho, ako dobre predikuje vývoj volebných preferencií pre 10 spomenutých subjektov. Každý z týchto časových radov rozdelíme na trénovaciu a testovaciu časť, kde testovacia časť pozostáva z posledných 12 zložiek. Následne, pre trénovací časový rad spočítame rôznymi modelmi predikciu pre 12 ďalších mesiacov a spočítame strednú kvadratickú chybu vzhľadom na testovací „rok“. Model s najnižšou priemernou chybou použijeme na predikciu volieb.

Modely budeme testovať tri:

1. ARIMA, ktorej parametre nastavíme pomocou funkcie `pmdarima.auto_arima`
   - parameter d je nastavený pomocou Kwiatkowski–Phillips–Schmidt–Shin testu stacionarity
   - parametre (p, q) nastavíme tak, aby minimalizovali Akaikeho informačné kritérium

2. Holtovo dvojité vyrovnávanie, ktorého parametre α a β budú nastavené metódou maximálnej vierohodnosti (funkciou `statsmodels.tsa.api.Holt.fit`)

3. Holtovo dvojité vyrovnávanie, ktorého parametre $\alpha \in [0.1; 10]$ a $\beta \in [0.1; 10]$ budú nastavené z dát okrem posledných 6 mesiacov tak, aby minimizalizovali strednú kvadratickú chybu predikcie posledných 6 mesiacov

Priemerné chyby vyšli nasledovne:

$$\begin{aligned} \mathrm{err}_{\mathrm{ARIMA}} &= 9.55 \\ \mathrm{err}_{\mathrm{Holt,MLE}} &= 8.92 \\ \mathrm{err}_{\mathrm{Holt,MSE}} &= 11.99 \end{aligned}$$

Výsledný predikčný model bude teda predikovať vývoj preferencií politickej strany Holtovým dvojitým exponenciálnym vyrovnávaním. Parametre α a β budú nastavené pomocou metódy maximálnej vierohodnosti na dátach preferencií danej strany v období od januára 2010 do novembra 2024.

5.2 Predikovanie rozdielu volebného výsledku a volebných preferencií v prieskume mesiac pred voľbami

V tejto časti opíšeme našu metodiku za výberom modelu na predikciu rozdielu volebného výsledku a volebných preferencií v prieskume mesiac pred voľbami. Budeme porovnávať obyčajnú lineárnu regresiu a dve jej penalizované verzie – lasso a hrebeňová regresia.

Budeme pracovať s podobnými dátami ako v kapitole o klasifikácii, ale vynecháme volebné preferencie z prieskumov, keďže rozdiel chceme predikovať na základe iných informácií o strane, ako jej podpora. Pre každú stranu a každé voľby od roku 2012 máme údaje, či bola v korešpondujúcom volebnom období v koalícii alebo opozícii (nadobúdajúce hodnoty 0 alebo 1). Pridáme stĺpce údajov rôznych indikátorov o štáte v období volieb spolu s rozdielmi hodnôt týchto indikátorov medzi koncom a začiatkom korešpondujúceho volebného obdobia. Takisto pridáme stĺpce interakcií medzi týmito premennými (aby sme zistili napríklad to, že strana bola v koalícii a počas toho volebného obdobia stúplo HDP na človeka, nemá vplyv na rozdiel volebného výsledku a posledného predvolebného prieskumu).

5.2.1 Výber modelu

Výber modelu uskutočníme na rovnakej trénovacej vzorke ako v kapitole vyššie, iba s niekoľkými pridanými premennými. Každý model budeme hodnotiť krosvalidáciou na 5 častí, pričom výsledná chyba modelu bude priemerná stredná kvadratická chyba. Predtým ale musíme vybrať podmnožinu stĺpcov pre lineárnu regresiu, resp. vybrať hyperparametre pre lasso a hrebeňovú regresiu.

Podmnožinu stĺpcov pre lineárnu regresiu budeme vyberať iteratívne vzhľadom na priemernú strednú kvadratickú chybu krosvalidácie. Najprv ohodnotíme model bez premenných, potom pre všetky premenné spočítame chybu a medzi vybrané premenné pridáme tú s minimálnou. Následne budeme modelovať s dvomi premennými – prvá bude vybratá v predchádzajúcom kroku a ako druhú premennú vyskúšame všetky ostatné. Do ďalšieho kroku pôjde dvojica premenných s minimálnou priemernou krosvalidačnou chybou. Takto budeme pokračovať, kým nepoužijeme všetky premenné.

Výsledná podmnožina stĺpcov bude tá, ktorá počas celého behu algortimu dosiahla najnižšiu chybu. Týmto algoritmom sme pre lineárnu regresiu vybrali iba jednu premennú a to indikátor, či strana bola vo volebnom období pred voľbami v opozícii.

Pre lasso a hrebeňovú regresiu je to jednoduchšie. Pre obe skúšame sto ekvidištančných hodnôt pre parameter $\lambda \in [0.1; 10]$, vyberieme tú s najnižšou priemernou krosvalidačnou chybou vzhľadom na strednú kvadratickú chybu. Pre lasso algoritmus vrátil $\lambda = 0.3$, pre hrebeňovú regresiu $\lambda = 10$.

Nakoniec, pre finálne modelovanie spomedzi týchto troch modelov vyberieme ten s najnižšou krosvalidačnou chybou:

$$\begin{aligned} \mathrm{err}_{\mathrm{LR}} &= 8.14 \\ \mathrm{err}_{\mathrm{lasso}} &= 8.59 \\ \mathrm{err}_{\mathrm{ridge}} &= 9.32 \end{aligned}$$

Budeme modelovať lineárnou regresiou s jedinou premennou – účasť politickej strany v opozícii. Otestujeme na testovacích dátach spomenutých v kapitolách vyššie. Aby sme vedeli ohodnotiť našu predikciu, pozrieme sa na jej strednú kvadratickú a strednú absolútnu predikciu v porovnaní s naivným modelom, ktorý modeluje rozdiel volebného výsledku a posledného predvolebného prieskumu konštantnou nulou:

$$\begin{aligned} \mathrm{MSE}_{\mathrm{LR}} &= 7.24 \\ \mathrm{MSE}_{\mathrm{naive}} &= 7.99 \\ \mathrm{MAE}_{\mathrm{LR}} &= 1.74 \\ \mathrm{MAE}_{\mathrm{naive}} &= 1.71 \end{aligned}$$

Z výsledkov vidíme, že náš model lepšie zachytáva odľahlé dáta (teda veľký „skok“ medzi posledným prieskumom a voľbami) ako model, ktorý predpokladá, že tam žiadny „skok“ nie je. V absolútnej chybe je však mierne horší.

Využijeme fakt, že pre „relevantné“ strany v tejto dobe (január 2025) máme k dispozícii dáta z politického kompasu, ktoré hovoria veľa o profile a prioritách strany. Vytvoríme teda rovnaký model ako vyššie, len z trénovacej a testovacej sady odfiltrujeme strany, pre ktoré údaje nemáme. Jedná sa teda o výraznú redukciu počtu dát už z aj tak malého počtu, ale napriek tomu môžeme pozorovať správanie modelu.

Znova nám vyšlo, že najnižšiu krosvalidačnú chybu dosiahla lineárna regresia, ale tentokrát s tromi premennými:

1. interakcia medzi účasťou v koalícii a hodnotou liberálnosti/konzervatívnosti
2. interakcia medzi účasťou v opozícii a hodnotou výdavkov na dôchodky (na človeka)
3. interakcia medzi nezamestnanosťou v percentách a HDP na človeka

Spočítame znova chyby takéhoto a naivného modelu na (zredukovaných) testovacích dátach

$$\begin{aligned} \mathrm{MSE}_{\mathrm{LR}} &= 6.53 \\ \mathrm{MSE}_{\mathrm{naive}} &= 11.29 \\ \mathrm{MAE}_{\mathrm{LR}} &= 1.72 \\ \mathrm{MAE}_{\mathrm{naive}} &= 2.22 \end{aligned}$$

Vidíme, že pri použití politického kompasu má náš model lepšiu aj MSE aj MAE oproti modelu, ktorý predpokladá, že volebný výsledok bude rovnaký, ako posledný prieskum. Keďže nás zaujíma najmä predikcia budúcich volieb a pre v tomto období relevantné strany máme údaje o politickom kompase, budeme teda vo finálnom modely používať túto verziu.

6 Predikcie volieb

V tejto sekcií ukážeme výsledky nadizajnovaného modelu v sekcií 5. Chceme predikovať výsledky volieb, keby sa konajú o mesiac a o 6 mesiacov od novembra 2024 (mesiac posledných nami zozbieraných údajov z prieskumov). Treba poznamenať, že budeme modelovať volebný výsledok iba pre nasledovné „relevantné“ strany (v prieskumoch agentúry Focus majú za november aspoň 3%) – Progresívne Slovensko, Smer-SD, Hlas-SD, Slovensko, SaS, KDH, Republika, SNS, Sme rodina, Maďarská aliancia a Demokrati.

Náš model sa skladá z dvoch častí:

- na predikciu vývoja preferencií politickej strany používa Holtovo dvojité exponenciálne vyrovnávanie. Parametre α a β sú nastavené pre každú stranu samostatne metódou maximálnej vierohodnosti z ich preferencií v prieskumoch od januára 2010

- rozdiel skutočného volebného výsledku oproti preferenciám strany v prieskume mesiac pred voľbami budeme modelovať lineárnou regresiou s troma premennými:

  1. interakcia medzi účasťou v koalícii a hodnotou liberálnosti/konzervatívnosti
  2. interakcia medzi účasťou v opozícii a hodnotou výdavkov na dôchodky (na človeka)
  3. interakcia medzi nezamestnanosťou v percentách a HDP na človeka

  Parameter $\hat{\beta} \in \mathbb{R}^4$ pre lineárnu regresiu natrénujeme zo všetkých dostupných dát (ich tvar je opísaný vyššie) pre vyššie spomenutých 10 politických strán

Môžeme uviesť, že natrénovaný parameter $\hat{\beta}$ na naškálovných dátach (nulový priemer, jednotková variancia) mal nasledovný tvar:

$$\hat{\beta} = (0.806, 0.305, 1.88, 0.192)$$

Vidíme teda že všetky tri premenné majú pozitívny vplyv na predikciu, čiže ich zvýšenie predpokladá, že strana dosiahne vo voľbách vyšší výsledok, ako mala v prieskume pred voľbami. Najstrmší vplyv má interakcia medzi účasťou v opozícii a hodnotou výdavkov na dôchodky (na človeka).

Pre predikciu volieb o n mesiacov modelujeme vývoj preferencií pre $n-1$ mesiacov exponenciálnym vyrovnávaním a pre n-tý mesiac predikujeme rozdiel medzi preferenciami a prieskumom v $(n-1)$-om mesiaci. Keď tieto dva vektory sčítame, dostaneme výslednú predikciu.

Môžeme teda predikovať výsledky volieb, keby sa konali v decembri 2024:

politická strana	volebný výsledok 12-2024	poslancov v parlamente 12-2024
Demokrati	3.51	0
Hlas SD	11.29	19
KDH	9.68	16
Maďarská aliancia	3.69	0
Slovensko	7.38	12
Progresívne Slovensko	24.15	42
Republika	7.04	12
SaS	10.12	17
Sme rodina	2.36	0
Smer SD	18.68	32
SNS	2.11	0

Takisto môžeme predikovať výsledky volieb, keby sa konajú v máji 2025:

politická strana	volebný výsledok 05-2025	poslancov v parlamente 05-2025
Demokrati	3.77	0
Hlas SD	10.98	19
KDH	9.65	16
Maďarská aliancia	3.28	0
Slovensko	7.31	12
Progresívne Slovensko	24.90	43
Republika	6.87	11
SaS	10.37	17
Sme rodina	2.37	0
Smer SD	18.46	32
SNS	2.03	0

Osobnú interpretáciu a zhodnotenie týchto predikovaných výsledkov nechávame na čitateľa.

7 Zhrnutie