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

4. januára 2025

Obsah

1 Predstavenie témy

V histórii slovenských volieb sme sa stretli už s viacerými prekvapeniami. Či už to bol výsledok strany Smer SD v roku 2012, ktorý im stačil na zloženie jednofarebnej vlády, nečakané prekonanie hranice zvoliteľnosti stranou ĽSNS v roku 2016, ktorá mala mesiac pred voľbami v prieskumoch verejnej mienky okolo 2%, alebo takisto nečakaná dominantná výhra strany OĽaNO v roku 2020. Tieto javy ukazujú, že prieskumy volebných preferencií sú dobrým odhadom skutočného výsledku volieb, no nezachytávajú v sebe zmenu nálad a názorov v spoločnosti, ktoré nevyhnutne volebný deň prináša.

V našej práci sa budeme venovať práve tomuto. Chceme predikovať výsledky volieb na základe prieskumov verejnej mienky a rôznych informácií o strane či socioekomickej situácie na Slovensku. Výsledný model môže slúžiť ako nový odhad výsledkov z volebného dňa, popri odhadovaní samotným posledným prieskumom či exit pollom.

V nasledujúcich kapitolách sa pozrieme na dáta, s ktorými budeme pracovať. Pokúsime sa opísať nami pozorované správanie v nich a nadizajnujeme klasifikačný a predikčný model. Natrénovaný predikčný model využijeme na odhad výsledku volieb, keby sa konali v decembri 2024, resp. ak by sa konali v máji 2025.

2 Dáta

2.1 Prieskumy volebných preferencií

Na účely tohto projektu potrebujeme mať značnú históriu volebných prieskumov v jednotnom formáte. Budeme sa zaoberať voľbami v rokoch 2012, 2016, 2020 a 2023, čiže potrebujeme prieskumy z obdobia od januára 2010. Nakoľko však agentúra FOCUS nezverejňuje s každým prieskumom jednotný .xlsx alebo .csv súbor a rovnako na svojej stranke nemá zverejnené všetky prieskumy, ktoré kedy vykonala, tak je táto úloha obzvlášť problematická. Väčšina vykonaných prieskumov je však k dispozícii vo formáte .pdf v štýle reportu (Press release). Javí sa, že FOCUS pri týchto reportoch udržuje jednotný formát v priebehu rokov, čo využijeme na extrakciu dát z nich.

Súčasťou každého takéhoto .pdf súboru je aj samotný prieskum, napríklad:

Politická strana	**% rozhodnutých voličov**	**95% interval spoľahlivosti**	% rozhodnutých voličov	% rozhodnutých voličov
	september 2019		august 2019 II.	august 2019 I.
SMER-SD	**21,7%**	18,8% - 24,6%	20,8%	21,8%
koalícia strán Progresívne Slovensko a SPOLU	**13,3%**	10,9% - 15,7%	14,9%	14,0%
Kotleba – Ľudová strana Naše Slovensko	**10,6%**	8,4% - 12,8%	11,4%	12,1%
SME RODINA – Boris Kollár	**7,2%**	5,4% - 9,0%	6,6%	6,3%
KDH	**6,9%**	5,1% - 8,7%	7,3%	7,5%
SNS	**6,8%**	5,0% - 8,6%	7,0%	7,0%
OĽANO	**6,8%**	5,0% - 8,6%	6,1%	6,0%
Za ľudí	**6,5%**	4,7% - 8,3%	5,5%	5,0%
SaS	**6,4%**	4,6% - 8,2%	7,9%	7,0%
MOST-HÍD	**4,1%**	2,7% - 5,5%	4,3%	4,7%
SMK – MKP	**3,3%**	2,1% - 4,5%	3,5%	3,4%
Dobrá voľba	**2,1%**	-	x	x
SZS	**1,0%**	-	0,3%	0,9%
iná strana[3]	**3,3%**	-	4,4%	4,3%

** v tabuľke sú uvedené len subjekty, ktoré dosiahli preferencie 1% a viac*

Automatizovaného sťahovanie z webu sme vykonali pomocou Python knižnice Selenium, ktorá vie interagovať s webovým prehliadačom ako bežný používateľ. Pomocou iteratívneho odosielania query do prehliadača vo forme: "focus prieskum volby month year"+ "filetype:pdf"sme stiahli prvý result vo forme .pdf súboru. Avšak ak v danom mesiaci nebol uskutočnený prieskum, resp. prvý nájdený .pdf súbor obsahoval niečo iné, stiahli sme zbytočný súbor. Takéto situácie sme museli následne ručne identifikovať a odstrániť.

Nakoniec sa nám úspešne podarilo takýchto reportov získať 127.

Ďalej sme tabuľky z reportov agentúry FOCUS automatizovane extrahovali na účel vytvorenia jednotného .csv súboru. Použili sme Python knižnicu Docling, pomocou ktorej sme prekonvertovali tabuľky z .pdf do pd.DataFrame, spojili a exportovali do jednotného `data/raw/polls/focus_polls.csv`.

Tento proces nebol úplne priamočiary. Názvy politických strán sa extrahovali veľmi nekonzistentne. Takisto viaceré politické strany menili meno v priebehu rokov, takže bolo potrebné manuálne vytvoriť mapper, ktorý tieto názvy zjednotil. Ďalej bolo potreba niektoré záznamy aj manuálne opraviť, keďže strany s dlhším názvom (v reporte zabrali dva riadky tabuľky) sa duplikovali v našom extrahovanom datasete.

Prieskumy z obdobia od januára 2010 do novembra 2024, ktoré sme neboli schopní získať automatizovane, sme dopísali ručne. Agentúra Focus však nezverejňuje výsledky každý mesiac, čiže sme tieto údaje doplnili z iných agentúr. Vyskytlo sa však 10 mesiacov, kedy žiadna známa agnetúra prieskum nezverejnila. Tieto dáta sme lineárne interpolovali zo „susedných“ mesiacov. Údaje o zdrojoch jednotlivých prieskumov sú v súbore `data/polls_agencies.csv`. Výsledné dáta z prieskumov sú v súbore `data/polls_data.csv`.

Rozdelenie prieskumov podľa volieb

Pre neskoršie účely regresie a klasifikácie potrebujeme poznať výsledky jednotlivých volieb pre jednotlivé strany. Tieto údaje sme získali zo Štatistického úradu Slovenskej republiky, sú uložené v súbore `data/election_results.csv`. Takisto môže pri klasifkácii/predikovanii pomôcť údaj, či bola daná strana v konkrétnom volebnom období v koalícii/opozícii. Tieto údaje boli vyplnené ručne do súboru `data/elected_parties.csv`. K týmto dátam sme ešte pre každú stranu pridali jej politické preferencie z prieskumov 12 mesiacov pred voľbami, čo môžete nájsť v súbore `data/polls_by_election.csv`.

Tieto dáta chceme využiť v dizajne modelov, preto ich rozdelíme na trénovaciu a testovaciu sadu v pomere $4:1$. Treba však odfiltrovať dáta, ktoré sú podobné typu „volebný výsledok 0%, všetkých 12 prieskumov pred voľbami okolo 0%“, aby sme mohli pozorovať trendy a správanie, nie iba konštantnú nulu. Preto sú v trénovacom a testovacom datasete iba tie strany, ktoré počas 12-tich mesiacov pred konkrétnymi voľbami dosiahli aspoň raz hranicu 1.5%, čo je polovica hranice na štátny príspevok za výsledok vo voľbách. Vo výsledku je teda v datasete `data/polls_by_election_train.csv` 50 dátových bodov, v `data/polls_by_election_test.csv` 13.

2.2 Všeobecné štatistiky o štáte

V modelovaní chceme zistiť, či interakcia hodnoty rôznych indikátorov o Slovenskej republike napríklad s účasťou v koalícii nezachytáva nejaké trendy. Preto sme z portálu DATAcube získali dáta o nasledovných indikátoroch pre roky 2010 až 2023 (vymenované sú iba tie, ktoré sú použité v neskoršej analýze)

- nezamestnanosť v percentách
- HDP na človeka v eurách
- riziko chudoby v percentách
- priemerný príjem v domácnosti v eurách za mesiac
- inflácia v percentách
- priemerná cena benzínu s oktánovým číslom 95 v eurách za liter
- výdavky na vedu a vzdelávane v eurách za rok

Tieto dáta sú uložené v súbore `data/general_data.csv`.

2.3 Politický kompas

Ako posledné sme chceli získať údaje o hodnotách a nastaveniach politických strán. Získali sme nasledovné údaje z ZDROJ AKÝ? pre politické subjekty Progresívne Slovensko, Smer SD, Hlas SD, Slovensko, SaS, KDH, Republika, SNS, Sme rodina, Maďarská aliancia a Demokrati:

- liberalizmus-konzervativizmus: na škále od -1 do 1, kde -1 je najviac liberálny
- ľavica-pravica: na škále od -1 do 1, kde -1 je najviac ľavicový
- životné prostredie: na škále od -1 do 1, kde 1 značí najväčší možný dôraz na témy týkajúce sa životného prostredia

- integrácia v Európskej únii: na škále od -1 do 1, kde 1 značí najväčší možný dôraz na témy týkajúce sa integrácie do EÚ
- internacionalizmus a zahraničná politika: na škále od -1 do 1, kde 1 značí najväčší možný dôraz na témy týkajúce sa zahraničnej politiky a spolupráce so zahraničím

Tieto dáta sú uložené v súbore `data/political_compass_data.csv`.

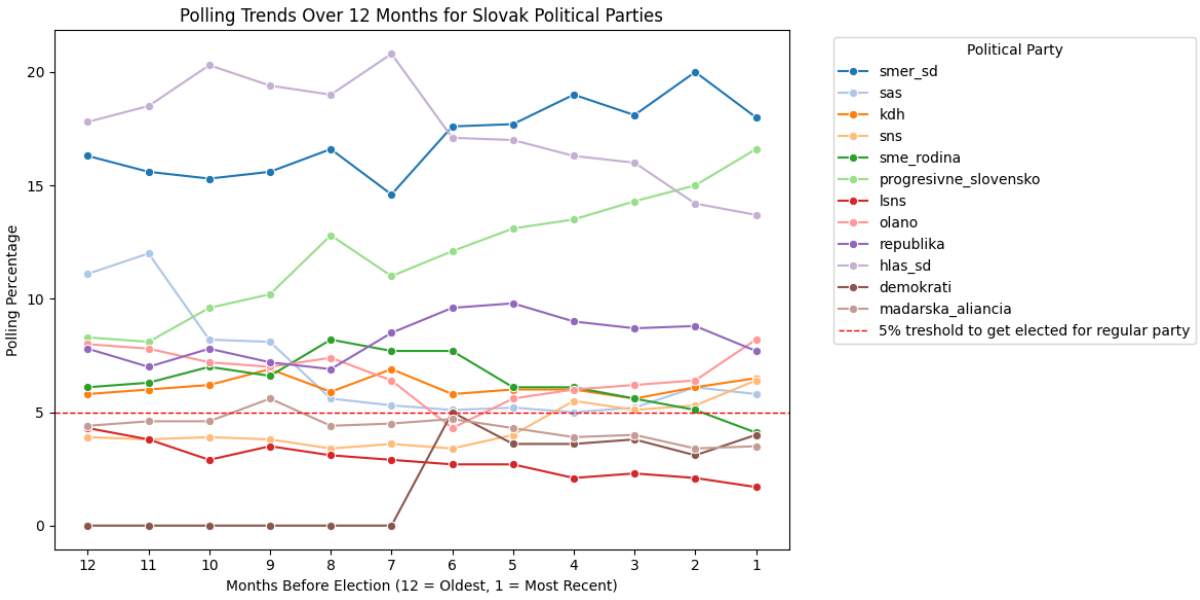
3 Exploratívna analýza

V exploratívnej analýze sme sa zamerali na ukázanie relevantnosti dát, ktoré sme nazbierali a rovnako aj na vytvorenie si očakávaní od predikčného modelu. Teda po nazbieraní a upratání dát sme hlavne riešili to, ako dané dáta budeme kombinovať a či niektoré z nich budú relevantné pre náš predikčný model.

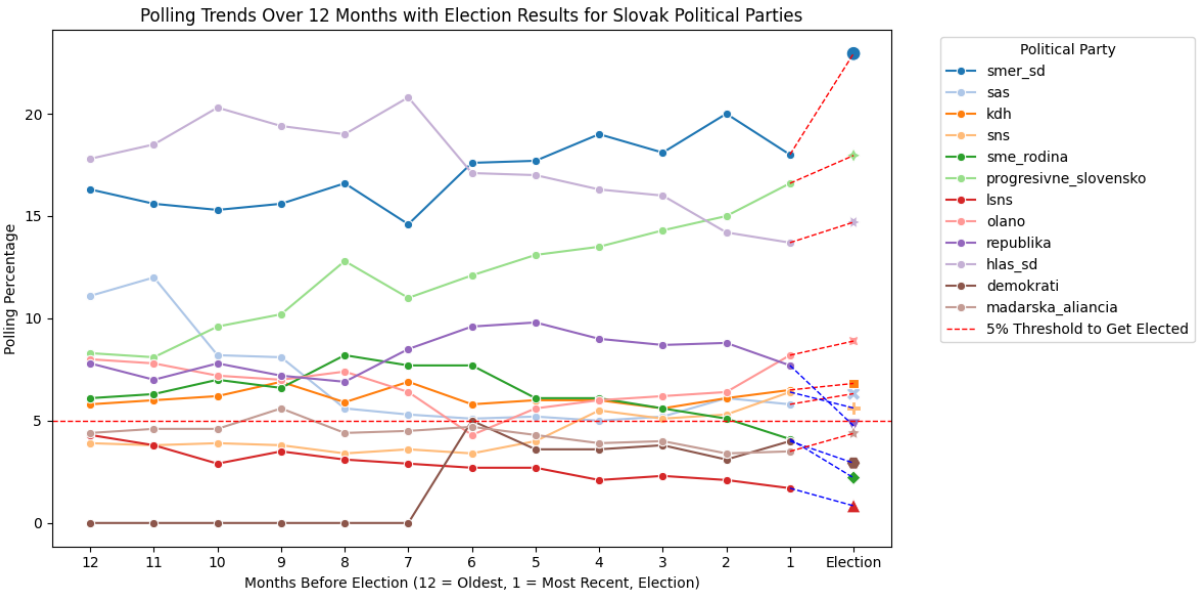
3.1 Náhľad do dát polls_by_election

Hlavný dataset `data/polls_by_election.csv` obsahuje dáta, ktorý máme v pláne v upravenej verzii použiť na predikciu volebných výsledkov. Pre každú stranu dataset obsahuje posledných 12 prieskumov a výsledky volieb spolu s informáciami, či bola strana v parlamente, opozícii, alebo koalícii. Bolo by teda vhodné sa pozrieť na to, ako sa vyvíjali volebné prieskumy 12 mesiacov pred voľbami. Ak by sme našli jasné trendy, kedy strana postupne rastie v čase, chceli by sme očakávať od nášho modelu, že tento rast sa v ňom ukáže.

Na ukážku sme si vybrali dáta z minuloročných volieb a rovnako sme vyfiltrovali strany, čo dosiahli v prieskumoch pred voľbami nenulový výsledok.

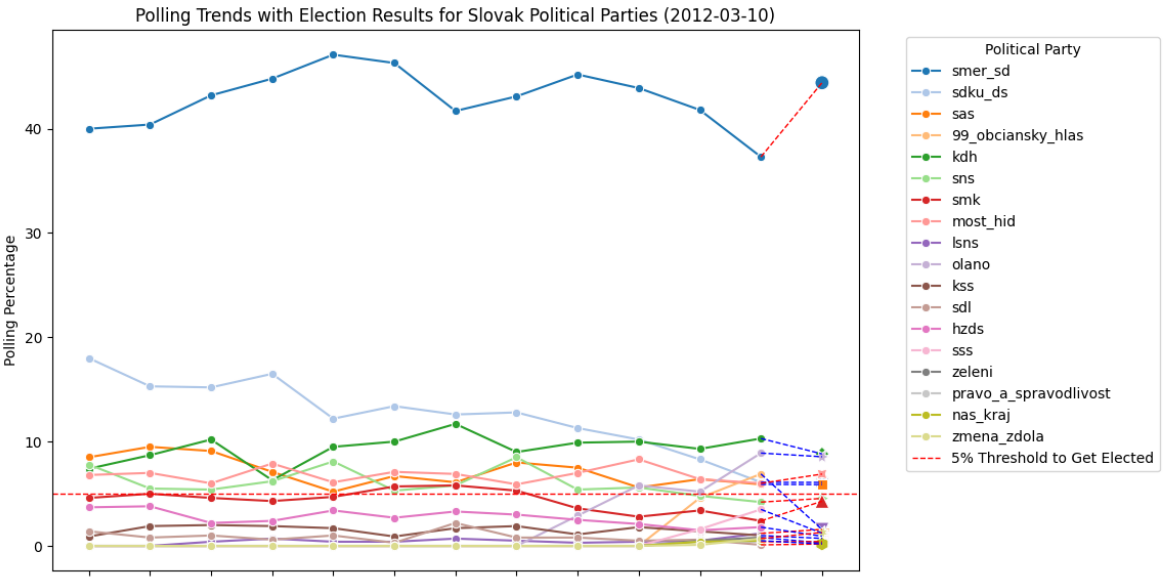


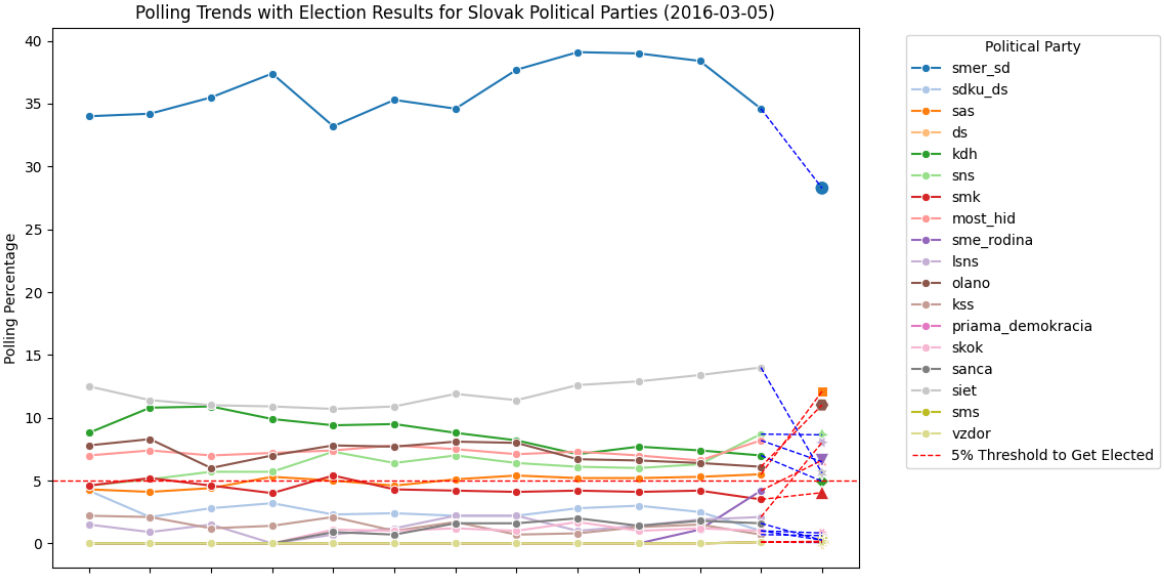
Už môžeme pozorovať ako sa vyvíjali názory voličov pred voľbami, napríklad zrod strany Demokrati, mierny prepad Hlasu, či postupný nárast Progresívneho Slovenska. Pridanie výsledku vo voľbách by malo ešte väčšiu výpovednú hodnotu, keďže budeme vedieť zhodnotiť aj to, ako sa preferencie premietli už aj vo voľbách, a tak urobíme presne to.

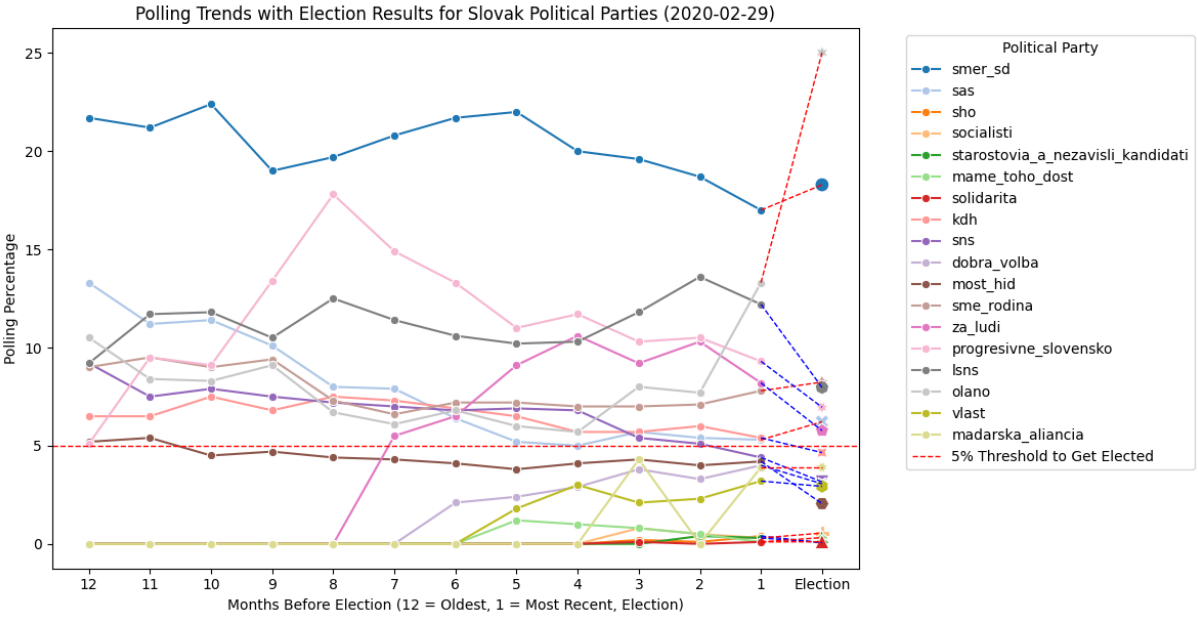


Je jasné, že prieskumy zachytávajú realitu pred voľbami relatívne dobre, aj keď vidíme aj značné skoky pre určité strany. Avšak poradie toho ako strany dopadli vo voľbách sa skoro nezmenilo, až na výrazný prepad strany Republika. Vidíme aj, že trendy v prieskumoch majú vplyv na to, ako strana dopadne vo voľbách. Aj tu sa nájdu výnimky, ako Hlas, ktorý niekoľko mesiacov pred voľbami padal, ale nakoniec dopadol lepšie, ako v posledných prieskumoch. Taktiež, ak sa niektoré strany dostali až pod hranicu zvoliteľnosti v prieskumoch, často ju už neprekonali. Voliči majú prirodzený strach, že im prepadne hlas, a tak práve takýto výsledok v prieskumoch môže veľmi ublížiť strane.

Toto bol pohľad len na najnedávnejšie voľby, ale ako vyzerali aj tie predtým?


Polling Trends with Election Results for Slovak Political Parties (2012-03-10)
Polling Percentage
Political Party
smer_sd
sdku_ds
sas
99_obciansky_hlas
kdh
sns
smk
most_hid
lsns
olano
kss
sdl
hzds
sss
zeleni
pravo_a_spravodlivost
nas_kraj
zmena_zdola
5% Threshold to Get Elected


Polling Trends with Election Results for Slovak Political Parties (2016-03-05)
Polling Percentage
Political Party
smer_sd
sdku_ds
sas
ds
kdh
sns
smk
most_hid
sme_rodina
lsns
olano
kss
priama_demokracia
skok
sanca
siet
sms
vzdor
5% Threshold to Get Elected


Polling Trends with Election Results for Slovak Political Parties (2020-02-29)
Polling Percentage
Political Party
smer_sd
sas
sho
socialisti
starostovia_a_nezavisli_kandidati
mame_toho_dost
solidarita
kdh
sns
dobra_volba
most_hid
sme_rodina
za_ludi
progresivne_slovensko
lsns
olano
vlast
madarska_aliancia
5% Threshold to Get Elected
12
11
10
9
8
7
6
5
4
3
2
1
Election
Months Before Election (12 = Oldest, 1 = Most Recent, Election)

Voľby v roku 2012 mali jasného favorita v strane SMER, ktorá po voľbách aj sama zostavila vládu, čo sa považuje za následok chaotického pádu vlady a následné predčasné voľby. Strana SDKÚ sa dramaticky prepadala v prieskumoch z 18% rok pred voľbami, až na 6% vo voľbách.

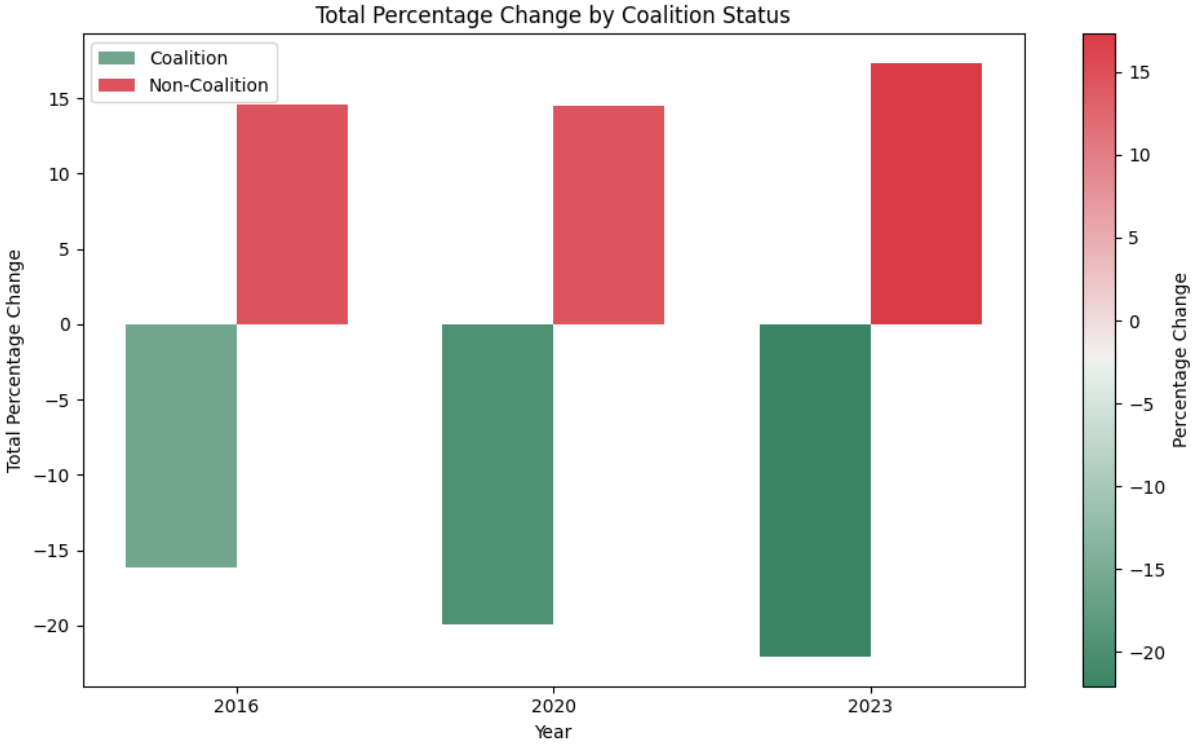
Vo voľbách v roku 2016 už SMER stratil na sile a v posledných mesiacoch kampane utrpel prepad na 28%. Zaujimavý vývoj mala strana Sieť, ktorá sa v prieskumoch dostávala nad 10% celkom konzistentne, avšak v poslednom mesiaci kampane jej voliči pravdepodobne prestúpili k iným stranám ako OĽaNO, SaS či Sme rodina. Sieť sa tak prepadla až na hranicu zvoliteľnosti.

Voľby v roku 2020 sa ukazujú ako najvyrovnanejšie a to v tom zmysle, že viac strán dosahovalo v prieskumoch nad hranicu 10%. Avšak ujať vedenia sa podarilo strane OĽaNO, ktorá rástla v posledných mesiacoch pred voľbami zatiaľ najvýraznejšie zo všetkých pozorovaných strán v prieskumoch, až ich aj nakoniec vyhrala. Graf nezachytáva realitu, toho že koalícii Progresívne Slovensko-Spolu sa nepodarilo dostať do parlamentu, keďže kandidovali ako koalicia dvoch stán a pre takýto typ politického subjektu je potrebné dosiahnuť vo voľbách viac ako 7%.

Trend toho, že vládne strany počas výkonu moci strácajú na podpore v následujúcich voľbach vyzerá ustálený v rámci všetkých volieb, ešte sa naň pozrieme bližšie.

3.2 Dopad predošleho vládnutia na voľby

V našich dátach je informácia o tom či v predošlom volebnom období bola daná strana v parlamente a či bola súčasťou koalície/opozicie. Bolo by teda vhodné sa pozrieť na to, ako sa menila voličská zakladňa strán v koalícii a strán, ktoré v nej neboli.



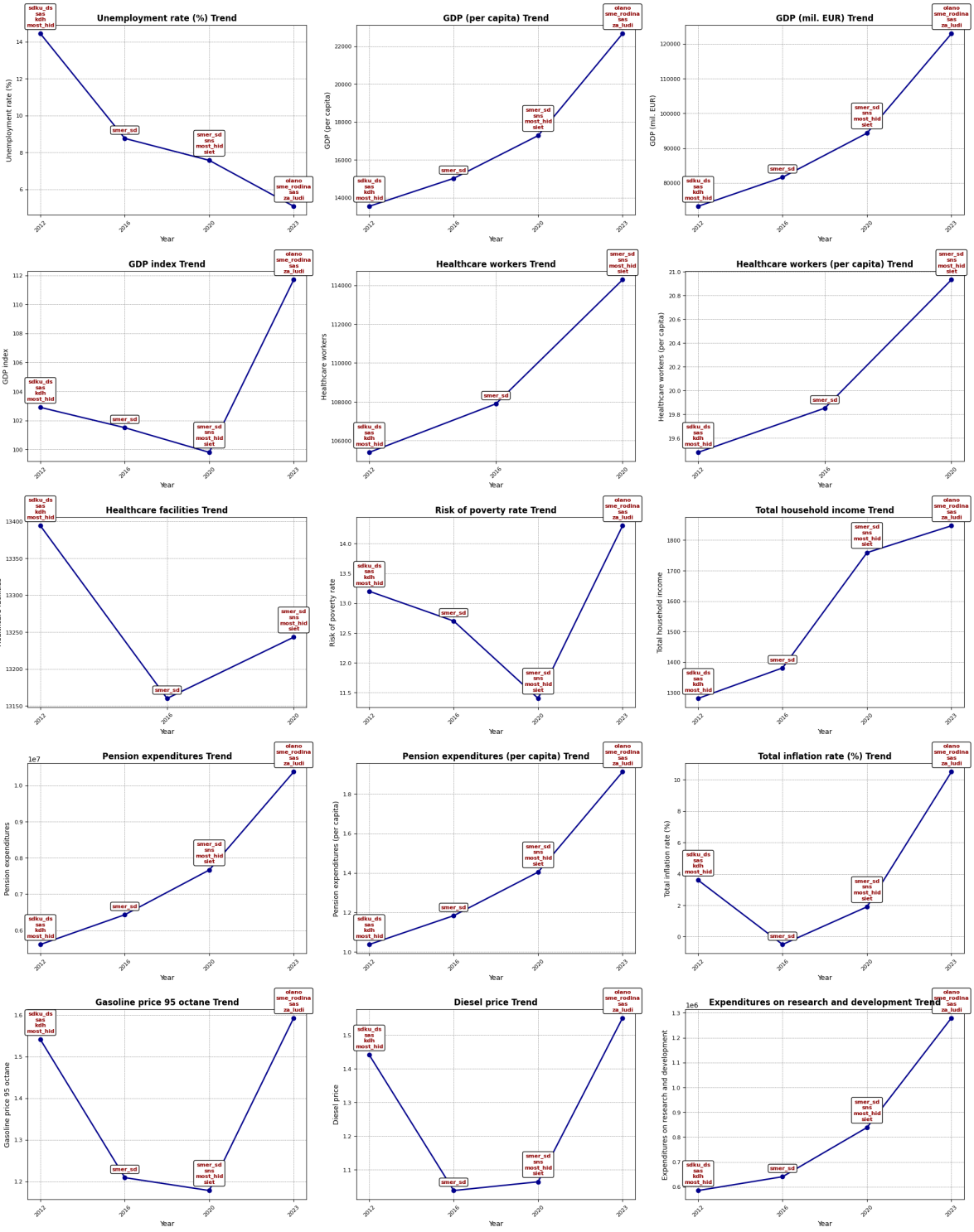
V súčte strany vládnej moci vždy stratili až cez 15 percent v nasledujúcich voľbách, avšak miera straty nebola ustálená, teda chceme veriť, že kvalita vládnutia aspoň do nejakej miery ovplyvnila, koľko by dané strany stratili v následujúcich voľbách. V závislosti od tohto strany mimo vladnej koalície nabrali vo výsledkoch približne v rovnakom

množstve. Rovnako volič mohol od vládnej koalicie prejsť k strane, ktorá má k nej blízko a po voľbách sa k nej pridala. Preto aj keď z roku 2012 na 2016 SMER stratil cez 15 percent komfortne zostavil vládu v roku 2016 za pomoci SNS, MOST - HÍD a Siete.

3.3 Pohlaď na koaličné strany a rozvoj krajiny

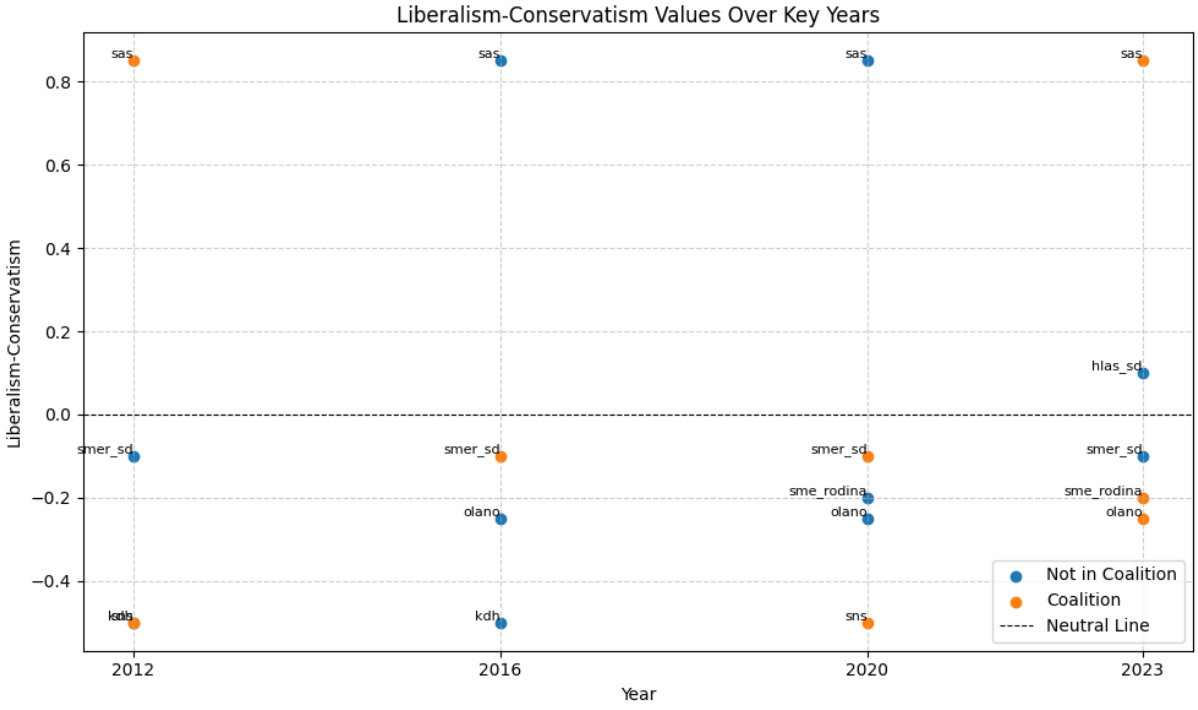
Po zistení z predchádzajúcich analýz, že sa stranám po spolupráci v koaličných vládach znižuje voličské zastúpenie, sme sa rozhodli preskúmať, či by tento jav nemohol byť zapríčinený ekonomicko-sociálnymi faktormi. Po úprave našich datasetov, ktoré zachytávajú vývoj všeobecných faktorov krajiny, ako napríklad HDP, a datasetu spojeného so stranami, ich koaličnými partnermi a obdobiami ich vládnutia, sme zistili, že napriek pravidelným stratám vo voličskom zastúpení jednotlivých koaličných strán nevieme tento úpadok vidieť v trendoch rôznych štatistík o krajine.

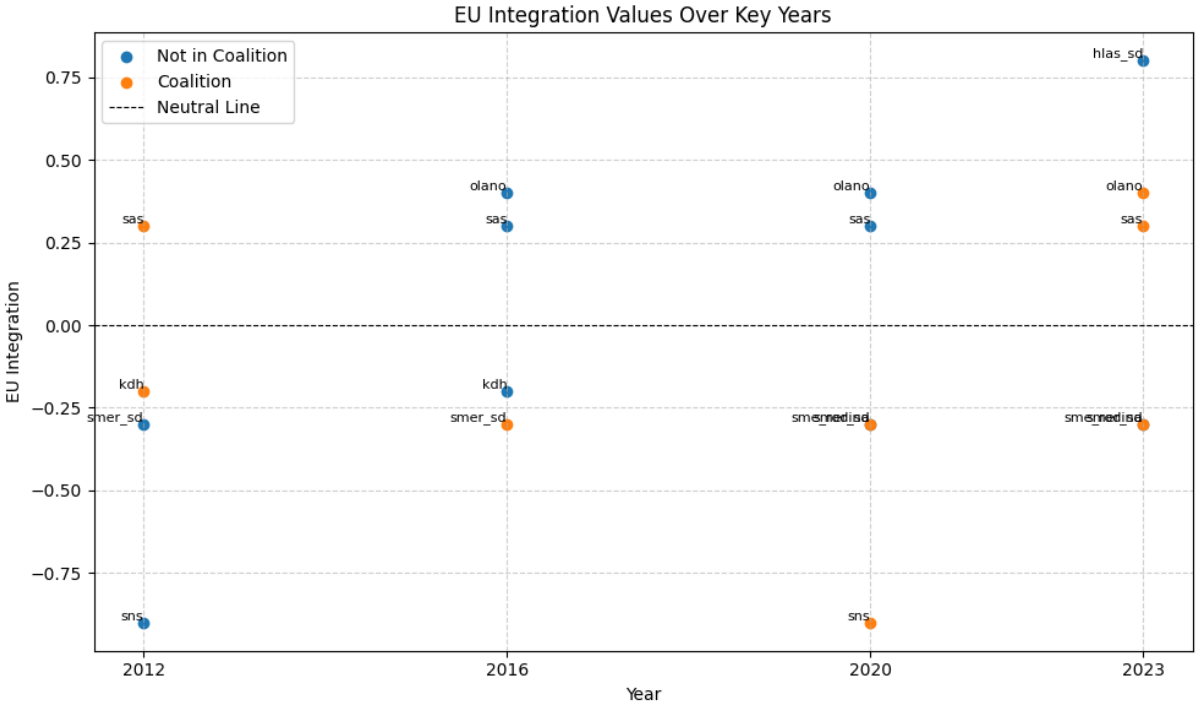
Na grafoch nižšie je zobrazená koalícia a jej celé predošlé obdobie vládnutia. To znamená, že rozhodnutia koaličných strán z roku 2016 mali vplyv na vývoj situácie od roku 2012 až po 2016. Pri mnohých ukazovateľoch, ako je napríklad percento nezamestnanosti, sa podarilo udržať klesajúci trend. Rovnako sa podarilo udržať rastúci trend pri HDP a ďalších ukazovateľoch. V prípade situácie s klesajúcim rizikom chudoby sa tiež prevažne podarilo udržať klesajúci trend. Dá sa teda povedať, že rozhodnutie voličov prestať podporovať danú stranu nemusí byť vždy spôsobené negatívnymi rozhodnutiami strán, ale aj inými faktormi, ako sú napríklad vyjadrenia politikov či iné okolnosti. Avšak, nie je to pravidlom.


Unemployment rate (%) Trend
Unemployment rate (%)
GDP (per capita) Trend
GDP (per capita)
GDP (mil. EUR) Trend
GDP (mil. EUR)
GDP index Trend
GDP index
Healthcare workers Trend
Healthcare workers
Healthcare workers (per capita) Trend
Healthcare workers (per capita)
Healthcare facilities Trend
Healthcare facilities
Risk of poverty rate Trend
Risk of poverty rate
Total household income Trend
Total household income
Pension expenditures Trend
Pension expenditures
Pension expenditures (per capita) Trend
Pension expenditures (per capita)
Total inflation rate (%) Trend
Total inflation rate (%)
Gasoline price 95 octane Trend
Gasoline price 95 octane
Diesel price Trend
Diesel price
Expenditures on research and development Trend
Expenditures on research and development
Year
sdku_ds sas kdh most_hid
smer_sd
smer_sd sns most_hid slet
olano sme_rodina sas za_ludi

Ďalším naším krokom bolo hlbšie sa zamerať na ideologické a štrukturálne faktory vládnucich koalícií a preskúmať, či ich spoločné hodnoty, ako je liberalizmus, orientácia na ľavú alebo pravú politickú stranu, postoj k integrácii do EÚ a ďalšie, zohrávali úlohu pri formovaní rastu krajiny. Porovnaním koalícií s rôznymi ideológiami sa snažíme zistiť, či tieto rozdiely ovplyvnili ich schopnosť podporovať pokrok v kľúčových oblastiach.

Naše dáta boli získané z Politického kompasu, ktoré však nie vždy presne zodpovedajú dátam týkajúcim sa zvolených strán, keďže nie všetky strany boli zahrnuté v Politickom kompase.





Ako si môžeme všimnúť napríklad v politickom období medzi rokmi 2020 a 2023, koalícia bola zložená zo strán SaS, OĽaNO, Sme rodina a Za ľudí. Hoci sa tieto politické subjekty líšia v otázkach integrácie do EÚ a v pozícii na škále liberalizmus – konzervativizmus, napriek tomu dokázali posilniť celkovú hospodársku situáciu v krajine. Výraznejšie napríklad narástlo HDP a zároveň sa podarilo znížiť mieru nezamestnanosti. Na druhej strane však nedošlo k žiadnemu výraznému zlepšeniu v oblasti inflácie a cien plynu, ktoré sa im počas tohto obdobia nepodarilo dostať pod kontrolu. Za niektoré nepriaznivé javy, ktoré sa im nepodarilo odstrániť, však nemusí niesť plnú zodpovednosť iba táto koalícia, keďže do vývoja zasiahli aj iné externé faktory a krízy.

4 Klasifikačné modely

V tejto časti sa budeme venovať binárno-klasifikačným modelom a aplikovať ich na úlohu predpovedania, či sa konkrétna politická strana dostane do parlamentu, alebo nie.

4.1 Výber algoritmov

Rozhodli sme sa porovnať tri základné modely, ktoré poskytujú rôzne prístupy ku riešeniu tejto klasifikačnej úlohy:

- logistická regresia
- rozhodovací strom
- support vector machine

4.2 Predspracovanie dát

Pred trénovaním týchto modelov bolo potrebné získané dáta spracovať do vhodného formátu. Dáta o samotných prieskumoch, ktoré sú už vopred rozdelené na trénovaciu a testovaciu vzorku, majú takúto štruktúru.

Tabuľka 1: Ukážka dát o prieskumoch

political_party	...	**elected_to_parliament**	**1**	**2**	...
olano	...	0	8.2	6.4	...
smer_sd	...	1	34.6	38.4	...
⋮	⋮	⋮	⋮	⋮	

Teda máme informácie o názve strany, či bola v danom roku naozaj zvolená alebo nie (hodnota 1 označuje, že bola) a údaj v percentách z prieskumu z k mesiacov pred voľbami, kde $k \in \{1, \ldots, 12\}$. Takisto máme pre každé pozorovanie dátum volieb, informáciu o tom či strana bola predtým v koalícii alebo v opozícii a skutočný percentuálny výsledok volieb v danom roku.

Okrem týchto „hlavných“ dát, máme dáta aj dáta „vedľajšie“, ktoré zachytávajú všeobecné informácie o sociálno-ekonomickej situácii na Slovensku v danom roku.

Tabuľka 2: Ekonomické ukazovatele

indicator	**2010**	**2011**	...
Unemployment rate (%)	12.46	13.59	
GDP (per capita)	12668	13254	
⋮	⋮	⋮	

Tieto dve tabuľky sme na základe rokov spojili a následne zaviedli aj nové interakčné premenné. Napríklad sme vynásobili binárnu premennú *in_coallition_before* s premennou *Unemployment rate* a ešte niektorými premennými zo všeobecných dát.

Nakoniec sme všetky premenné štandardizovali na tzv. z-skóre, teda každá premenná má priemer 0 a štandardnú odchýlku 1. Z takto naškálovaných premenných sme vybrali podmnožinu, na ktorej budeme trénovať naše klasifikačné algoritmy. Menovite to boli tieto premenné.

```
    final_variables: list[str] = [
        "1", "2", "3", "4", "5", "6", "7", "8", "9", "10", "11", "12",
        "in_coalition_before_Unemployment rate (%)",
        "in_coalition_before_Pension expenditures (per capita)",
        "in_coalition_before_Expenditures on research and development",
        "Risk of poverty rate", "Total household income",
        "Total inflation rate (%)", "Gasoline price 95 octane",
    ]
```

4.3 Trieda Ensemble

Okrem hore-uvedených algoritmov sme naimplementovali aj triedu `Ensemble`, ktorá spája ľubovolné klasifikačné algoritmy, ktoré sa podieľajú štýlom „hlasovania“ na finálnej predikcii. Parametre potrebné na inicializáciu sú tieto.

```
    def __init__(self, models: list[Type[BaseEstimator]],
                 metric: Callable, threshold: float = 0.5,
                 weights: Optional[list[float]] = None) -> None:
```

Argument `models` je zoznam modelov, ktoré budú „hlasovať“ o finálnej predikcii. Parameter `metric` je metrika na základe ktorej sa proporčne pridelí každému modelu „sila hlasovania“. Hyperparameter `threshold` určuje, kedy je pozorovanie klasifikované ako „dostane sa do parlamentu“ a kedy nie. Voliteľným parametrom `weights` môžeme dopredu prideliť silu každému modelu.

Po inicializácii môžeme náš „spojený“ model natrénovať. Teda trénuje sa osobitne inštancia každého zo vstupných modelov. Nakoniec sa vypočítajú sily predikcií podľa danej metriky, ktoré sa napokon znormalizujú tak, aby dali v súčte 1.

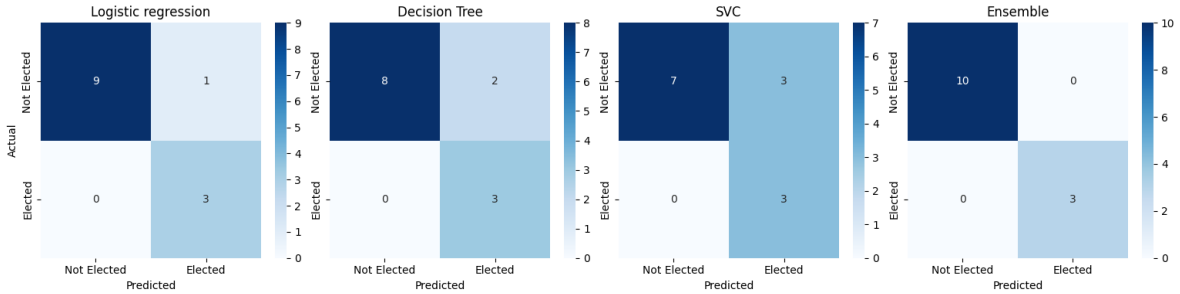
Po úspešnom trénovaní, môžeme predikovať takto.

```
    def predict(self, X: pd.DataFrame):
    predictions: list[np.ndarray] = []

    for classifier in self.fitted_classifiers:
        predictions.append(np.array(classifier.predict(X=X)))

    weighted_sum: np.ndarray = np.dot(
      np.array(predictions).T, np.array(self.weights)
      )

    return (weighted_sum >= self.threshold).astype(int)
```

Vstupom do metódy predict je matica dát a výsledkom je vektor klasifikácií. Po predikcií všetkých modelov sa vypočíta súčin matice predikcií a vektora váh. Nakoniec sa pozložkovo porovná výsledok tohto násobenia so vstupným hyperparametrom `threshold` .

4.4 Porovnanie algoritmov

Pre každý z uvedených prístupov nájdeme doprednou selekciou podmnožinu premenných, ktorá nakoniec použijeme pri finálnom trénovaní. Pre našu triedu `Ensemble` takisto kros-validáciou nájdeme optimálnu hodnotu hyperparametra `threshold` pre túto podmnožinu. Nakoniec už môžme len porovnať tieto natrénové modely na testovacích dátach.



Môžeme vidieť, že model `Ensemle`, v tomto prípade pozostávajúca práve zo všetkých troch modelov, klasifikovala všetky testovacie dáta správne, narozdiel od samotných modelov, kde boli prípady takzvanej falošnej pozitivity. Hoci sa tento výsledok zdá sľubný, testovacie dáta pozostávali len z trinástich pozorovaní. Otázkou teda zostáva, ako by si náš model viedol na väčších dátach.

5 Predikčné modely

Úlohu predikovania volebného výsledku môžeme rozdeliť na dve časti. Prvá je modelovanie vývoja volebných preferencií politických strán. Druhý aspekt je predikovanie skutočného výsledku volieb na základe posledného prieskumu preferencií. V sekcii 3 sme videli, 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 teda stavať na pozorovaniach z exploratívnej analýzy a trénovať model, ktorý bude predikovať tento rozdiel a skúsiť pomocou neho vysvetliť toto netriviálne správanie.

5.1 Predikovanie vývoja volebných preferencií

V tejto časti budeme pracovať s prieskumami volebných preferencií politických strán a hnutí od januára v roku 2010 do novembra v roku 2024. Pre každý z 81 politických subjektov máme 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. Porovnáme 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, ktoré vudú vykazovať lineárny trend.

Teória aj označenie v tejto sekcii boli 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) časového radu musia byť konštantné/nezávislé od času a jeho 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 sa používa diferencovanie – namiesto X_t modelujeme $Y_t = X_t - X_{t-1}$. Tento proces diferenciácie 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 sezónnosťou v časových radoch. Keďže tento trend 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 na výber modelu nepridali, keďže vznikli iba tesne pred voľbami v roku 2023.

Model vyberáme 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 (rok pred voľbami). 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 $\alpha \in [0.1; 1)$ a $\beta \in [0.1; 1)$ budú nastavené z dát okrem posledných 6 mesiacov tak, aby minimizalizovali strednú kvadratickú chybu predikcie posledných 6 mesiacov

Priemerné chyby vyšli nasledovne:

$$\begin{aligned} \mathrm{err}_{\mathrm{ARIMA}} &= 9.55 \\ \mathrm{err}_{\mathrm{Holt,MLE}} &= 8.92 \\ \mathrm{err}_{\mathrm{Holt,MSE}} &= 12.01 \end{aligned}$$

Výsledný predikčný model bude teda predikovať vývoj preferencií politickej strany Holtovým dvojitým exponenciálnym vyrovnávaním. Jej parametre α a β budú nastavené pomocou metódy maximálnej vierohodnosti na preferenciách 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 či 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 vyššie-spomenutými pridanými premennými. Každý model budeme hodnotiť krosvalidáciou na 10 častí, pričom výsledná chyba modelu bude priemerná stredná

kvadratická chyba jednotlivých častí krosvalidácie. Predtým je potrebné však vybrať podmnožinu stĺpcov pre lineárnu regresiu a vybrať hyperparameter λ pre lasso a hrebeňovú regresiu.

Podmnožinu stĺpcov pre lineárnu regresiu budeme vyberať pažravo 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 tá 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 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 teda modelovať lineárnou regresiou s jedinou premennou – účasť politickej strany v opozícii. Otestujeme na testovacích dátach spomenutých v kapitole vyššie. Aby sme vedeli ohodnotiť našu predikciu, pozrieme sa na jej strednú kvadratickú a strednú absolútnu chybu 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 (december 2024) 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 tieto údaje nemáme. Jedná sa teda o výraznú redukciu počtu dát už z aj tak malého počtu, ale napriek tomu môžeme pozorovať správanie modelu.

Znovu 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 modeli používať túto verziu.

6 Predikcie volieb

V tejto sekcií ukážeme výsledky nadizajnovaného modelu zo sekcie 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

Pre Holtovo dvojité exponenciálne vyrovnávanie natrénované parametre vyšli nasledovne:

politická strana	α	β
Progresívne Slovensko	0.909	0.018
Smer SD	0.68	0.1
Hlas SD	0.98	0
SaS	0.57	0.39
KDH	0.4	0.12
Slovensko	0.79	0
Republika	0.89	0.003
Maďarská aliancia	0.5	0
Demokrati	0.71	0.01
Sme rodina	0.82	0
SNS	0.67	0

Medzi zaujímavé výsledky môžeme uviesť vysoké hodnoty parametra α pre Progresívne Slovensko a Hlas SD, čo indikuje vysokú kolísavosť dát a teda pri vyhladzovaní časového radu je dôležitá aktuálna hodnota. Na druhej strane prevažne nízke hodnoty

parametra β naznačujú, že vyhladzovaný trend je upravovaný pomaly, čiže pre inú ako „horizontálnu“ predikciu vývoja preferencií by mala strana vykazovať očividný stúpajúci alebo klesajúci trend pred voľbami.

Pri lineárnej regresii, 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). Tento fakt naznačuje, že opozičné strany by mohli mať skok dohora medzi výsledkami v prieskumoch a vo voľbách vyšší pri zvyšujúcich sa výdavkoch na dôchodky. Takisto z výsledkov vidíme koreláciu medzi konzervatívnosťou strany v koalícii s tým, že vo voľbách získava viac percent, ako predpokladajú prieskumy. Ešte spomenieme, že pre model bez dát z politického kompasu bol predikovaný sklon vplyvu toho, či strana bola v opozícii, pozitívny. Teda môžeme predpokladať, že opozičným stranám sa darí vo voľbách viac, ako predpokladajú prieskumy.

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. Sčítaním dostávame výslednú predikciu.

Môžeme predikovať výsledky volieb, keby sa konali v decembri 2024 a v máji 2025:

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

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

Osobnú interpretáciu a zhodnotenie týchto predikovaných výsledkov nechávame na čitateľa.

7 Zhrnutie

V našej práci sme sa venovali predikovaniu volebných výsledkov za pomoci prieskumov, hodnotového nastavenia strany a socio-ekonomickej situácie štátu. Rozobrali sme si pozorované správanie v dátach, ukázali funkčnosť klasifikovania pokorenia hranice zvoliteľnosti a následne aj predikciu volebného výsledku. Z nie signifikantného zlepšenia testovacej chyby oproti naivnému prediktoru môžeme usúdiť, že voľby sú výrazne komplexnejší proces, ako náš model dokáže zachytiť. Našli sme však prípady, kedy môžeme predpokladať, že strane sa bude dariť lepšie alebo horšie vo voľbách, ako predikujú prieskumy verejnej mienky.

V budúcej práci by sme sa mohli venovať napríklad predikovaniu volieb v iných štátoch, resp. vo viacerých štátoch naraz. Počas celého pracovania na projekte nás limitoval fakt, že dát máme príliš málo na to, aby natrénovaný model zachytil všeobecné správanie. Tiež by sme mohli implementovať do predikcie vývoja preferencií vzájomné ovplyvňovanie sa medzi stranami. Je známy fakt, že voliči sa často medzi podobnými stranami „prelievajú“, k čomu sú aj dostupné údaje o tzv. druhej voľbe. Takisto pokles popularity koaličných strán by mohol spustiť nástup opozičných strán.

Na zhodnotenie relevancie našej práce nám ostáva iba zbierať údaje z prieskumov a čakať na najbližšie voľby. Môžeme pozorovať, či sa naše predikcie s blížiacimi sa voľbami ustáľujú alebo kolíšu, a nakoniec ako ďaleko od skutočného výsledku budú. Veríme v pozitívny výsledok.