Dobar dan! Ja sam Valentina i tema mog rada je prediktivno modeliranje stopa nataliteta i mortaliteta u RH na temelju demografskih i ekonomskih varijabli.

* Natalitet i mortalitet jedne su od ključnih demografskih značajki države te njihove procjene mogu pomoći u izgradnji populacijskih politika te sustava poput obrazovnog, zdravstvenog i mirovinskog.
* Postoje mnogi matematički modeli za ovakve procjene no oni su ograničeni na manje skupove podataka te teže nalaze nelinearne odnose značajki.
* Stoga se posljednjih godina ovo područje sve više okreće strojnom učenju koji mogu prepoznati različite složenije obrasce i odnose između značajki

Prije same izrade modela napravljena je analiza korelacije među varijablama.

Prvi graf pokazuje korelacije između broja rođenih te drugih varijabli. Imamo očekivane korelacije poput pozitivne sa ruralnim i mlađim stanovništvom te negativne sa urbanim i starim stanovništvom. Zanimljiva je negativna korelacija sa BDP-om i izvozom, no to se može objasniti kao posljedica veće urbanizacije, veće uključenosti žena na tržište rada, više obrazovanje i slično što doprinosi BDP-u i izvozu, ali ne i broju rođenih.

A chart with different colored lines

Description automatically generated with medium confidence

Sljedeći graf pokazuje korelacije između broja umrlih te drugih varijabli. Ponovo imamo neke očekivane korelacije poput pozitivne sa starijim stanovništvom, očekivanim trajanjem života te negativne sa mlađim stanovništvom. Od neočekivanih imamo pozitivnu korelaciju s urbanim stanovništvom, BDP-om i izvozom te negativnu sa ruralnim i sa stopom nezaposlenosti, što opet možemo pripisati općenitom razvoju države s kojim dolazi do urbanizacije, veće zaposlenosti i moguće bolje zdravstvene skrbi no višeg mortaliteta i starenja stanovništva.

A chart with different colored lines

Description automatically generated with medium confidence

Modele koje sam odabrala su linearna regresija, holt, holt-winters, arimax te extreme gradient boosting modeli. Podatci nad kojima su izgrađeni su podijeljeni na dva skupa, prvi (veći) skup od 1960. do 2021. godine sa manjim brojem varijabli, od ekonomskih samo BDP te (manji) skup od 1986. do 2021. godine sa većim brojem varijabli, posebice ekonomskih.

Prvi su modeli **linearne regresije**, izgrađeni na 70% podataka za treniranje te 30% podataka za testiranje.

Model za broj rođenih na prvom skupu podataka je uključivao sljedeće varijable:

* Godina
* Stanovništvo najvećeg grada (% urbanog stanovništva)
* Godišnji % rast ruralnog stanovništva
* Stanovništvo (15-64)
* Omjer ovisnosti o dobi, mladi
* BDP po stanovniku

Na ovom grafu, kao i na svim sljedećim ovog tipa, plava linija predstavlja podatke za učenje, zelena podatke za testiranje, a crvene crtkaste linije predikcije modela.

Te na ovom grafu možemo vidjeti da je model uspio uhvatiti generalni trend no ima dosta značajno odstupanje za 2021. godinu.

A graph with lines and numbers

Description automatically generated

Model na drugom skupu podataka je uključivao dvije dodatne varijable:

* Tečaj valute
* Indeks potrošačkih cijena

Rezultate tog modela možemo vidjeti na sljedećem grafu, možemo vidjeti da model malo bolje od prijašnjeg prati trend te nema toliko odstupanje u 2021. godini.

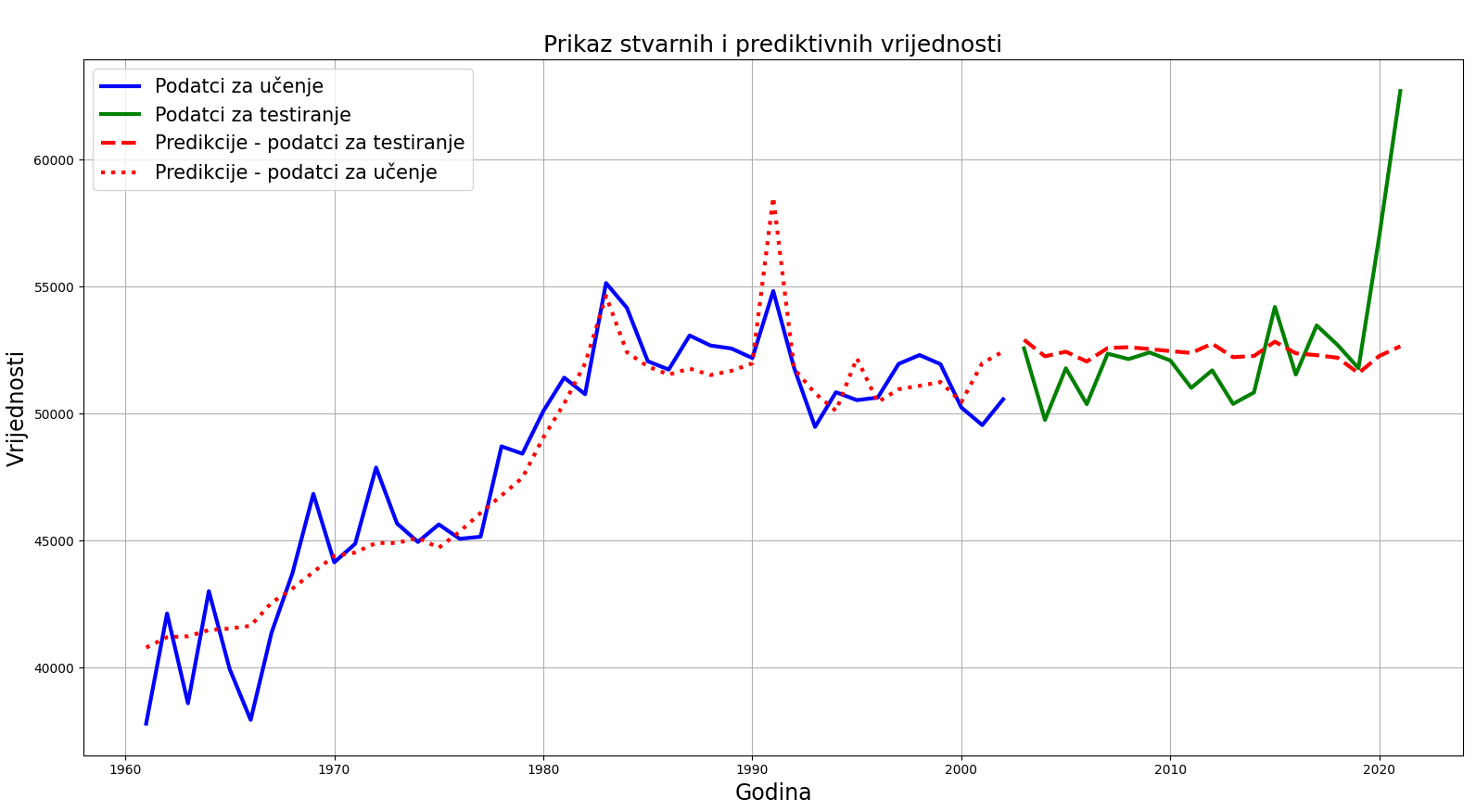
A graph with lines and numbers

Description automatically generated

Modeli linearne regresije za broj umrlih su postigle dosta zadovoljavajuće performanse sa prosječnom apsolutnom postotnom pogreškom manjom od 3%.

Na prvom skupu podataka korištene su sljedeće varijable:

* Godina
* Omjer ovisnosti o dobi, stari
* Očekivano trajanje života
* Preživljavanje do 65. godine, muškarci

Te na sljedećem grafu vidimo da model odlično prati opći trend iako ne i same oscilacije. 

Što se tiče modela na drugom skupu podataka, njemu su dodane varijable indeksa potrošačkih cijena te uklonjena varijabla omjera ovisnosti o dobi. Vidimo da model također prati opći trend, ali ne i oscilacije.

A graph with different colored lines

Description automatically generated

**HOLT**

Kod modela zaglađivanja odnosno Holt i Holt-Winters modela, kao ulaz se koriste samo prijašnji podatci ciljne varijable te su stoga izgrađeni samo na prvom skupu podataka.

Modeli za stope rođenih su izgrađeni sa 50% podataka za treniranje odnosno testiranje. Same predikcije možemo vidjeti na ovom grafu te vidimo da oba modela prate trend.

A graph with different colored lines

Description automatically generated

Oba modela koriste prigušenje što znači da će se s vremenom trend prigušiti u ravnu liniju. Trend tog prigušenja je 0.9 za Holtov te 0.82 za Holt-Winters model što znači snažno prigušenje trenda, odnosno utjecaj prijašnjih promjena se brzo smanjuje.

Parametar zaglađivanja je za oba modela 0.2 što sugerira da model sporije reagira na promjene u trendu, što je korisno kod odstupanja. Te razina zaglađivanja je za oba modela jednaka 1. Takav iznos pokazuje da modeli brzo reagiraju na promjene u osnovnoj razini odnosno koriste samo posljednju vrijednost.

A graph with different colored lines

Description automatically generated

Model za stope umrlih je koristio 60% podataka za treniranje, 40% za testiranje.

Kod modela za stope umrlih, bolje performanse su ostvarene bez prigušenja te stoga imamo ravne linije. Za Holta level zaglađivanja je 0.9, a trend zaglađivanja 0.25, što znači veliki utjecaj neposredno prethodnih iznosa vremenske serije te mali utjecaj promjena u trendu. Holt-Winters ima level 0.35, a trend skoro nula što ukazuje na umjereno zaglađivanje osnovne razine te vrlo sporo prilagođavanje promjenama u trendu. Na grafu možemo vidjeti da modeli otprilike prate osnovni trend, te Holt-Winters to radi bolje.

A graph with different colored lines

Description automatically generated

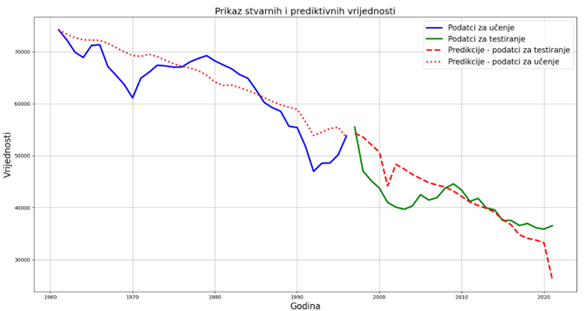
**ARIMAX**

Svi ARIMAX modeli su izgrađeni na 60% podataka za treniranje te 40% za testiranje.

AR i MA parametri objasnit?

Model za broj rođenih na prvom skupu podataka koristio je varijable:

* Godina
* Neto migracije
* Godišnji rast ruralnog stanovništva
* Stanovništvo najvećeg grada

Model uočava trend što možemo vidjeti i na ovome grafu, no ima značajno odstupanje za 2021. godinu. 

Utjecaji AR (autoregresivnih) te MA(pokretnog prosjeka) parametara na model su vrlo mali dok je varijanca greške vrlo visoka i statistički značajna. To ukazuje na nesigurnost modela i sugerira da možda nije optimalan za analizirane podatke.

Model na drugom skupu podataka ne koristi varijablu neto migracija, te iako ima malo manje prosječne i postotne pogreške i dalje ima značajna odstupanja što se više približavamo 2021. godini. Kod ovog su modela također zanemarivi AR i MA parametri.

A graph with lines and numbers

Description automatically generated with medium confidence

Za broj umrlih na prvom skupu podataka korištene su varijable:

* Godina
* Očekivano trajanje života
* Omjer ovisnosti o dobi, stari
* Preživljavanje do 65. godine, muškarci

Na grafu možemo vidjeti da model odlično prati trend pa čak i oscilacije na podatcima za testiranje, iako na podatcima za treniranje ima veća odstupanja, sam model ima postotnu pogrešku od 1.7%. Kod ovog modela AR i MA parametri su vrlo značajni.

A graph with lines and a red line

Description automatically generated with medium confidence

Model na drugom skupu podataka koristi dodatne varijable za indeks potrošačkih cijena te ne koristi broj muškaraca preživjelih do 65 godine. Ovaj model ima postotku pogrešku manju od 3% te sama predviđanja možemo vidjeti na grafu. AR i MA parametri imaju malu značajnost.

A graph with lines and a red line

Description automatically generated with medium confidence

**XGBoost**

Svi Extreme Gradient Boosting modeli imaju stopu učenja od 0.1, broj treniranih stabala je 100 te kao funkciju gubitka koriste kvadratnu grešku.

Modeli na prvom skupu podataka koriste 70% podataka za treniranje, a na drugom skupu 60%.

Model za broj rođenih na prvom skupu podataka ima varijable:

* Ruralno stanovništvo kao % ukupnog stanovništva
* Stanovništvo (15-64)
* Žensko stanovništvo (20-24)
* Godišnji rast stanovništva

Iako model ima zadovoljavajuću postotnu i apsolutnu pogrešku na grafu možemo vidjeti na ne prati trend već su predikcije osim 2005. konstantne.

A graph with lines and dots

Description automatically generated with medium confidence

---- za prenaučenost idk

Model na drugom skupu podataka ima dvije dodatne varijable, inflacija te urbano stanovništvo, no također ima sličan problem kao prijašnji model, odnosno ne prati trend već su predviđanja više-manje konstantna.

A graph with lines and numbers

Description automatically generated with medium confidence

Model za broj umrlih na prvom skupu podataka koristi sljedeće varijable:

* Očekivano trajanje života
* Stanovništvo iznad 80.-te godine
* Neto migracije
* Godišnji rast stanovništva
* Stanovništvo najvećeg grada (% urbanog stanovništva)
* Omjer ovisnosti o dobi, stari

Na grafu možemo vidjeti da model prati trend osim značajnog odstupanja za 2021. godinu.

A graph with lines and numbers

Description automatically generated with medium confidence

Model za broj umrlih na drugom skupu podataka je uz svaku kombinaciju parametara i varijabli pokazivao prenaučenost, što možemo vidjeti na sljedećem grafu. Znači vidimo da se prosječna apsolutna pogreška na podatcima za testiranje smanjuje samo malo na početku te nakon toga ponovno raste i ostaje visoka.

A graph of a person with a blue line

Description automatically generated with medium confidence

Usporedbom svih modela, za natalitet je najbolje performanse imala linearna regresija na manjem skupu podataka te Holt na prvom skupu podataka. Najčešće korištene varijable i najbitnije su stanovništvo najvećeg grada, ruralno stanovništvo te kalendarska godina.

Sveukupno najbolji model je ARIMAX za broj umrlih na prvom skupu podataka, te općenito modeli za mortalitet imaju bolje performanse. Tu su najčešće i najvažnije varijable preživljavanje do 65 godine za muškarce te očekivano trajanje života.

Rad je pokazao da su manje složeni modeli poput linearne regresije i eksponencijalnog zaglađivanja zadovoljavajući za ovakve probleme, dok za složenije modele je potrebno više podataka.

Naravno dalje postoji potencijal za bolja predviđanja uključivanjem većih skupova podataka sa dodatnim varijablama te možda i drugim tipovima modela poput neuronskih mreža.