Uticaj kršenja pretpostavki linearne regresije na tačnost i stabilnost regresionih modela

Kristian Zhou Gubić SV25/2024 i Miloš Vukić SV22/2024

DEFINICIJA PROBLEMA

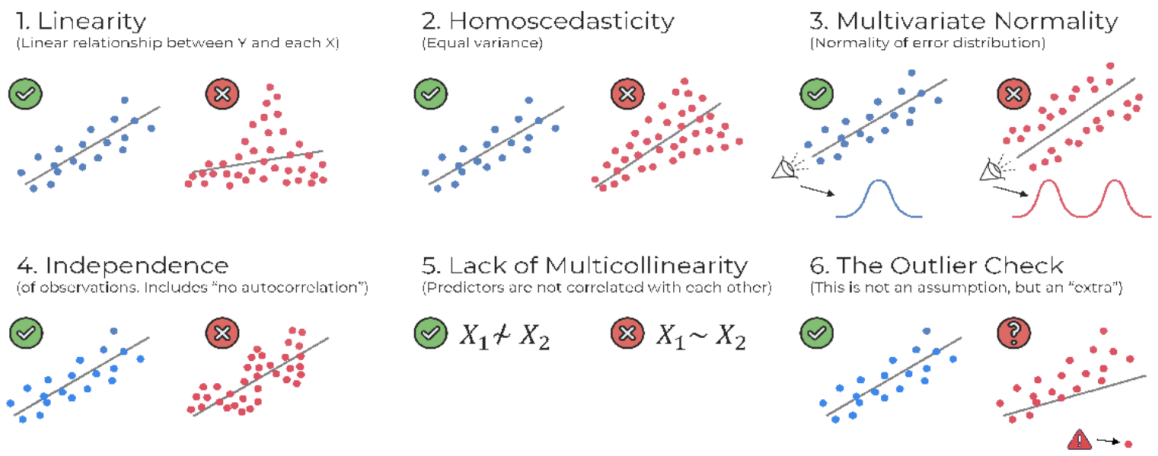
Linearna regresija jeste jedna od najrasprostanjenijih i najkorišćenijih metoda za modelovanje zavisnosti između numeričkih promenljivih vrednosti.

Teorijska osnova linearne regresije ogleda se u sledećim pretpostavkama:

- Linearne zavisnosti između ulaznih i izlaznih promenljivih
- Nezavisnosti grešaka
- Normalnosti grešaka
- Jednake varijanse
- Nepostojanja jake kolinearnosti

U realnim skupovima podataka često dolazi do situacije da su ove pretpostavke narušene, što može dovesti do nestabilnih regresionih koeficijenata, povećane greške predikcije kao i smanjenje pouzdanosti modela.
Upravo zbog toga, važno je ispitati kako se standardni regresioni modeli ponašaju u uslovima koji odstupaju od teorijskih pretpostavki, kao i u kojoj meri alternativni, robusni i regularizovani modeli mogu ublažiti negativne posledice tih odstupanja.

Cilj projekta nije samo procena tačnosti regresionih modela, već i analiza njihove **stabilnosti** i **robusnosti** u uslovima koji odstupaju od teorijskih pretpostavki linearne regresije, kroz kontrolisane numeričke eksperimente i ponovljene evaluacije.



ANALIZA STABILNOSTI MODELA

Stabilnost modela u ovom projektu predstavlja osetljivost modela na promene u trening podacima i narušavanje pretpostavki linearne regresije.

U suštini, kao stabilnost podrazumeva se osobina da **male** promene u ulaznim podacima (npr. drugačiji random split train/test skupa) ili u pretpostavkama modela dovode do relativno **malih** promena u predikcijama i parametrima modela.
Ako **male** promene u ulaznim podacima ili u pretpostavkama dovode do **velikih** promena u predikcijama ili parametrima modela, to nam pokazuje da naš model nije stabilan i podložan je nepredviđenim ponašanjima.

Stabilnost će biti kvantifikovana kroz:

- varijabilnost regresionih koeficijenata kroz ponovljene eksperimente
- varijansu greške predikcije (MSE/RMSE) između različitih podela skupa
- poređenje degradacije performansi usled kontrolisanog narušavanja pretpostavki

Primer kvantifikacije:
Bitno je razlikovati da RMSE sam po sebi predstavlja tačnost, dok praćenje promene RMSE-a predstavlja stabilnost modela. Koristeći RMSE kao kvantifikator stabilnosti, prvo izmerimo RMSE baznog modela bez promene, zatim uradimo malu promenu podataka ili pretpostavke (npr. drugačiji random split train/test skupa), a onda izmerimo RMSE izmenjenog modela. Model je nestabilan ako mu je RMSE skočio iako je promena bila sitna, a stabilan je ako je RMSE ostao sličan. Ako uzmemo dva različita regresiona modela, i uporedimo njihovu relativnu promenu, možemo zaključiti koji model je stabilniji.

$$\Delta RMSE = \frac{RMSE_{novo} - RMSE_{bazno}}{RMSE_{bazno}}$$

Relativna promena RMSE u odnosu na bazni scenario, koristi se kao indikator stabilnosti
Vrednosti bliže nuli ukazuju na veću stabilnost modela

Za XGBoost stabilnost koeficijenata se ne razmatra, već kroz varijabilnost performansi i degradaciju pri narušavanju pretpostavki.

U okviru analize ponašanja modela, stabilnost i robusnost modela biće ispitane kroz sledeće konkretne scenarije:

- različite slučajne podele skupa podataka na trening i test skup (različiti random shuffle/split)
- ponovljene k-fold podele skupa podataka
- dodavanje skoro-linearno zavisnih (kolinearnih) atributa radi pojačavanja multikolinearnosti
- dodavanje heteroskedastičnog šuma izlaznoj promenljivoj
- dodavanje šuma koji ne prati normalnu raspodelu

Za svaku od navedenih situacija biće analizirane promene performansi i stabilnosti modela u odnosu na bazni scenario.
Bazni scenario podrazumeva originalni skup podataka bez veštačkih perturbacija uz standardnu podelu na trening i test skup.

SKUP PODATAKA

U projektu će se koristiti **Energy Efficiency Dataset**, koji je dostupan u okviru UCI Machine Learning Repository baze podataka (Literatura i Izvori)

Skup podataka sadrži ukupno 768 instanci i 8 ulaznih promenljivih, koje opisuju geometrijske i energetske karakteristike zgrada:

- relativna kompaktnost
- površina
- površina zidova
- površina krovova
- ukupna visina
- orijentacija
- površina ostakljenja
- raspodela površine ostakljenja

Izlazne promenljive predstavljaju:

- potreba za grejanjem
- potreba za hlađenjem

Iako skup podataka sadrži dve izlazne promenljive, projekat će se fokusirati isključivo na **potrebu za grejanjem**, kako bi analiza stabilnosti i ponašanja regresionih modela bila metodološki jasnija i fokusiranija.
Na ovaj način omogućeno je detaljnije ispitivanje uticaja kršenja pretpostavki linearne regresije, bez uvođenja nepotrebnih komplikacija.

Zbog prisutne kolinearnosti između ulaznih promenljivih, kao i potencijalnih odstupanja od normalnosti grešaka i jednake varijanse, ovaj skup podataka je pogodan za analizu ponašanja regresionih modela u realnim uslovima.

METODOLOGIJA

Radi dobijanja pouzdanijih zaključaka i smanjenja uticaja slučajnosti podele skupa podataka, evaluacija regresionih modela biće sprovedena kroz veći broj ponovljenih eksperimenata sa različitim podelama na trening i test skup. Za svaki scenario biće zabeležene vrednosti relevantnih metrika greške, kao i vrednosti regresionih koeficijenata.

Na osnovu dobijenih rezultata, analiziraće se stabilnost modela, kako u pogledu predikcije, tako i u pogledu osetljivosti regresionih koeficijenata. Posebna pažnja će se obratiti poređenju stabilnosti standardnog **OLS modela** u odnosu na **regularizovane i robusne** regresione modele.

Pored klasičnih i regularizovanih linearnih regresija, u analizu će biti uključen i **eXtreme Gradient Boosting (XGBoost) regresioni model**, kao predstavnik nelinearnih "ensemble" metoda, sa ciljem poređenja robusnosti i stabilnosti u odnosu na linearne metode.

Izrada projekta će obuhvatiti sledeće korake:

1. Eksplorativna analiza podataka
    - analiza osnovnih statističkih karakteristika
    - vizualizacija distribucija promenljivih
    - analiza kolinearnosti između atributa

2. Provera pretpostavki linearne regresije
    - ispitivanje linearne zavisnosti između promenljivih
    - analiza nezavisnosti grešaka
    - pregled normalnosti grešaka
    - ispitivanje jednake varijanse
    - detekcija multikolinearnosti pomoću VIF (Variance Inflation Factor)

3. Izgradnja regresionih modela
    - standardna višestruka linearna regresija (**OLS**)
    - regularizovani modeli (**Ridge i Lasso regresija**)
    - robusni regresioni modeli (**Huber regresija, RANSAC**)
    - **XGBoost regresioni model**, predstavnik nelinearnih ensemble algoritama

4. Kontrolisano narušavanje pretpostavki linearne regresije
    - Heteroskedastični šum
    - Nenormalne greške
    - dodavanje skoro-linearno zavisnih atributa

5. Analiza stabilnosti modela
    - poređenje vrednosti regresionih koeficijenata

- analiza osetljivosti modela na narušene pretpostavke
- poređenje ponašanja svih modela (standardnih, regularizovanih, robusnih i XGBoost modela)

6. Poređenje rezultata i interpretacija
    - kvantativno poređenje performansi
    - kvalitativna analiza dobijenih rezultata
    - interpretacija krajnjih rezultata i donošenje zaključka

KONTROLISANO NARUŠAVANJE PRETPOSTAVKI MODELA

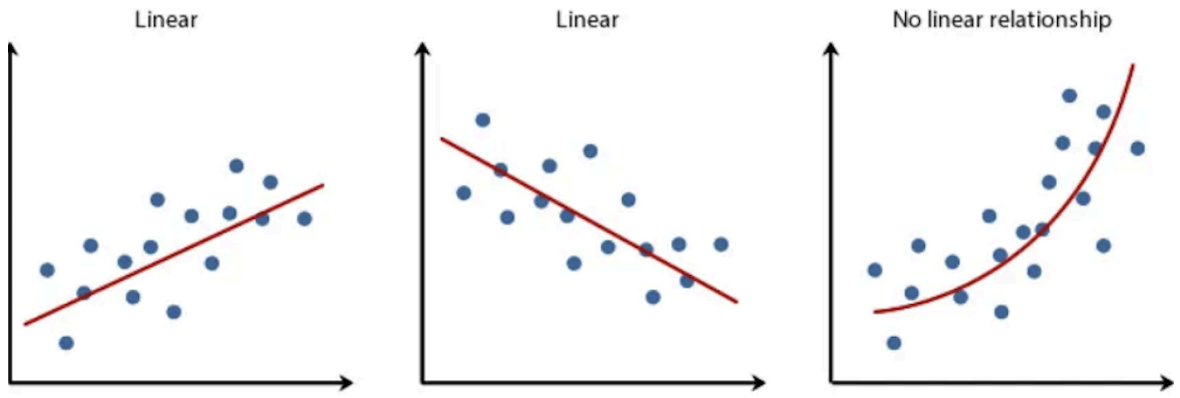
Pored analize realnog skupa podataka, u okviru projekta biće sprovedeni kontrolisani numerički eksperimenti u kojima se pojedine pretpostavke linearne regresije namerno narušavaju. Cilj ovih eksperimenata jeste da se ispita kako različiti regresioni modeli reaguju na sistematska odstupanja od teorijskih pretpostavki, kao i u kojoj meri alternativni modeli mogu da ublaže negativne posledice tih odstupanja.

Narušavanje pretpostavki biće realizovano dodavanjem veštačkog šuma različitih karakteristika na izlazne promenljive, kao i modifikovanjem ulaznih atributa sa ciljem pojačavanja multikolinearnosti. Biće ispitani slučajevi heteroskedastičnih grešaka, kao i grešaka koje ne prate normalnu raspodelu.

VIZUELIZACIJA I ANALIZA REZULTATA

U cilju boljeg razumevanja ponašanja regresionih modela i ispitivanja pretpostavki linearne regresije, u okviru projekta biće korišćene odgovarajuće vizuelizacije, poput matplotlib metoda.
Obuhvataće prikaz matrice kolinearnosti ulaznih promenljivih radi analize jake kolinearnosti, grafički prikaz reziduala u odnosu na predikcije radi ispitivanja homoskedastičnosti i ponašanja grešaka, kao i QQ-plot reziduala u cilju provere normalnosti grešaka.

Rezultati eksperimenata biće predstavljeni i grafički. Biće korišćeni boxplot grafici za poređenje distribucija grešaka predikcije kroz ponovljene eksperimente, kao i grafički prikazi regresionih koeficijenata radi analize njihove varijabilnosti i stabilnosti.
Ovakav način prikaza omogućava intuitivno poređenje ponašanja različitih modela i dodatno ilustruje uticaj kršenja pretpostavki linearne regresije na tačnost i pouzdanost rezultata.



NAČIN EVALUACIJE

Modeli će biti trenirani i evaluirani za izlaznu promenljivu koja predstavlja **potrebu za grejanjem.**

Evaluacija modela će se vršiti pomoću sledećih metrika:

Srednja kvadratna greška, eng. Mean Squared Error (MSE)

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Koren srednje kvadratne greške, eng. Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{(\hat{y}_i - y_i)^2}{n}}$$

Koeficijent determinacije, eng. R Squared

Prilagođeni koeficijent determinacije, eng. Adjusted R Squared

TEHNOLOGIJE

Za realizaciju projekta biće korišćene sledeće tehnologije:

- programski jezik Python
- biblioteke: NumPy, Pandas, Scikit-learn, Matplotlib

PODELA RADA

Projekat će se raditi sa sledećom podelom zadataka:
Kristian Zhou-Gubić, SV25/2024

- eksplorativna analiza podataka
- provera pretpostavki linearne regresije
- implementacija i analiza standardnog OLS modela
- generatori slučajeva kršenja pretpostavki
- agregacija metrika i boxplotovi

Miloš Vukić, SV22/2024

- implementacija regularizovanih (Ridge, Lasso) i robusnih (Huber, RANSAC) i XGBoost regresionog modela
- analiza stabilnosti i osetljivosti tih modela
- poređenje performansi i interpretacija rezultata
- ponovljeni eksperimenti (N split, k-fold)

Zajednički zadatak

- donošenje zaključka na osnovu krajnjih rezultata

LITERATURA I IZVORI

UCI Machine Learning Repository - Energy Efficiency Dataset
https://archive.ics.uci.edu/dataset/242/energy+efficiency

The Consequences of Violating Linear Regression Assumptions - TowardsDataScience
https://towardsdatascience.com/the-consequences-of-violating-linear-regression-assumptions-4f0513dd3160/

Linear Regression with Ordinary Least Squares - Coding Train
https://www.youtube.com/watch?v=szXbuO3bVRk&list=PLRqwX-V7Uu6bCN8LKrcMa6zF4FPtXyXYj&index=5

Introduction to Linear Regression Analysis - Montgomery, Peck, Vining
https://www.kwcsangli.in/uploads/3--Introduction_to_Linear_Regression_Analysis__5th_ed._Douglas_C._Montgomery__Elizabeth_A._Peck__and_G._.pdf

XGBoost - Wikipedia
https://en.wikipedia.org/wiki/XGBoost