Predikcija stope nasilnog kriminala primenom naprednih regresionih modela i ansambala

Vladimir Čajka SV55/2024

19.1.2026

1. Opis problema

Cilj projekta je izgradnja i poređenje modela za **predikciju stope nasilnog kriminala** u lokalnim zajednicama na osnovu dostupnih socio-ekonomskih i demografskih karakteristika. Podaci sadrže velik broj ulaznih promenljivih koje opisuju populaciju, obrazovanje, zaposlenost, strukturu domaćinstava, prihode, stanovanje i policijske resurse, dok su izlazne varijable povezane sa kriminalitetom.

Ovakav problem je izazovan iz više razloga. Prvo, skup sadrži **značajan procenat nedostajućih vrednosti**, što zahteva pažljiv izbor strategije imputacije i obradu nedostajućih podataka bez curenja informacija između trening i test skupa. Drugo, prisutna je **multikolinearnost** i potencijalna redundantnost među atributima, što otežava klasične linearne modele i može dovesti do nestabilnih procena parametara. Treće, moguće je postojanje **outlajera** i heterogenosti u zajednicama, zbog čega modeli treba da budu dovoljno robusni.

Zbog navedenih osobina, projekat se fokusira na primenu **savremenih regresionih metoda** i ansambala koji se u praksi pokazuju najuspešnijim na tabelarnim podacima, uz strogu metodologiju evaluacije (podela na skupove / k-fold validacija i podešavanje hiperparametara). Pored tačnosti predikcije, razmatra se i **interpretabilnost** modela kroz analizu važnosti atributa i objašnjavanje doprinosa najuticajnijih promenljivih, kako bi rezultati bili korisni i u kontekstu razumevanja faktora povezanih sa kriminalitetom.

2. Ciljevi projekta

Ciljevi projekta su:

- **izgradnja i evaluacija modela** za predikciju ciljne varijable na tabelarnim podacima sa velikim brojem obeležja i nedostajućim vrednostima,
- **poređenje naprednih regresionih algoritama** nad istim protokolom obrade podataka,
- **robustnost modela** na autlajere i šum,
- **analiza značaja obeležja i interpretabilnost** radi identifikacije faktora koji najviše doprinose predikciji.

3. Skup podataka

Skup podataka koji se analizira i nad kojim će se primenjivati regresioni modeli je **Communities and Crime (Unnormalized)** skup podataka sa **UCI Machine Learning Repository** (https://archive.ics.uci.edu/dataset/211/communities+and+crime+unnormalized).

Skup podataka sadrži podatke o **2215 redova (lokalne zajednice)** i obuhvata **147 kolona (atributa)**. Podaci opisuju demografske, socio-ekonomske, stambene i bezbednosne karakteristike zajednica, kao i statistike kriminala.

Kolona čiju vrednost želimo da predvidimo je **violentPerPop**, koja predstavlja **stopu nasilnog kriminala na sto hiljada stanovnika**.

U nastavku su navedeni primeri nezavisnih promenljivih koje su intuitivno relevantne za modelovanje kriminaliteta, konačan značaj obeležja biće utvrđen kroz feature importance i SHAP analizu:

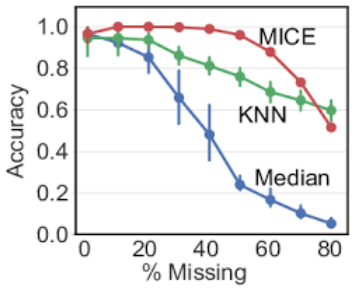
1. **pctKids2Par**
    - Opis: Procenat male dece koja žive u porodicama sa oba roditelja.
    - Tip podataka: Numerički.
2. **pctKidsBornNeverMarr**
    - Opis: Procenat dece rođene van braka.
    - Tip podataka: Numerički.
3. **pct2Par**
    - Opis: Procenat familija u kojima deca imaju oba roditelja.
    - Tip podataka: Numerički.
4. **pctBlack**
    - Opis: Procenat stanovništva afroameričkog porekla.
    - Tip podataka: Numerički.
5. **kidsBornNeverMarr**
    - Opis: Broj dece rođene van braka.
    - Tip podataka: Numerički.
6. **pctPopDenseHous**
    - Opis: Procenat stanovništva u gusto naseljenim domaćinstvima.
    - Tip podataka: Numerički.
7. **State**
    - Opis: Država u Sjedinjenim Američkim Državama.
    - Tip podataka: Kategorički.
8. **pctWhite**
    - Opis: Procenat stanovništva bele rase.
    - Tip podataka: Numerički.

4. Metodologija

1. **Pretprocesiranje podataka**
   Prvi korak u realizaciji projekta predstavlja pretprocesiranje podataka iz skupa *Communities and Crime (unnormalized)*. Skup podataka sadrži nedostajuće vrednosti koje su označene simbolom ?, kao i atribute sa različitim skalama i značajnim rasponom vrednosti.
   U okviru pretprocesiranja biće sprovedeni sledeći koraci:

- zamena simboličkih oznaka nedostajućih vrednosti odgovarajućim numeričkim oznakama,
- izbacivanje atributa sa velikim procentom nedostajućih vrednosti,
- imputacija preostalih nedostajućih vrednosti,
- kodiranje kategorijske promenljive **state**,
- uklanjanje potencijalnih **leakage** atributa vezanih za kriminal (sve "crime" kolone osim ciljne),
- skaliranje obeležja tamo gde je potrebno (StandardScaler / RobustScaler – zavisno od modela).

   Interpolacija se ne koristi, budući da redovi u skupu podataka ne poseduju prirodan redosled, te interpolirane vrednosti ne bi imale statističko niti semantičko značenje. **Median imputacija kao baseline**, uz poređenje sa **KNN/MICE** imputacijom, pri čemu se parametri imputacije računaju isključivo na trening skupu (sprečavanje leakage-a).
   Nakon pripreme, podaci se dele **na trening/validacioni/test** skup u odnosu **60/20/20**, dok se izbor hiperparametara vrši **k-fold cross-validacijom** na trening delu, a test skup se koristi isključivo za finalno izveštavanje.



Slika 1: Imputacija

2. **Bazni regresioni model (OLS)**
   Kako bi se uspostavila jasna referentna tačka za poređenje, u prvoj fazi se trenira bazni model višestruke linearne regresije korišćenjem metode najmanjih kvadrata (Ordinary Least Squares – OLS). OLS predstavlja najjednostavniji regresioni pristup i služi isključivo kao **baseline** za procenu dobitaka koje donose naprednije metode.

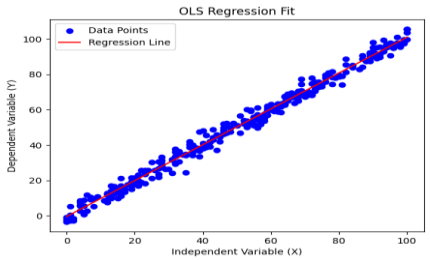
Model se trenira na trening skupu, dok se performanse proveravaju na validacionom skupu, a test skup se koristi samo za finalno izveštavanje. Posebna pažnja se posvećuje uticaju skaliranja i numeričke stabilnosti, pa se OLS ispituje u dve varijante:

- bez prethodnog skaliranja ulaznih obeležja,
- sa standardizacijom obeležja (z-score).

U okviru ove faze posmatraju se:

- greške predikcije (RMSE, R^2),
- stabilnost procenjenih koeficijenata,
- indikatori numeričkih problema (kondicioni broj, singularnost dizajn matrice).

Dobijeni rezultati služe kao polazna osnova za naredne faze projekta u kojima se uvode stabilizacioni i napredniji regresioni pristupi.

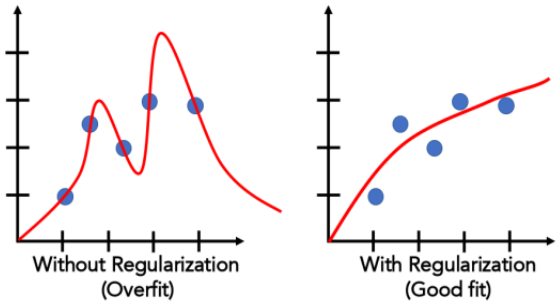


Slika 2: OLS regresioni model

3. **Regularizovani linearni modeli**

Primenjuju se linearni modeli sa regularizacijom koji smanjuju varijansu procene i stabilizuju rešenje u prisustvu korelisanih obeležja:

- **Ridge regresija (L2)** – stabilizuje koeficijente i poboljšava numeričku stabilnost.
- **Lasso regresija (L1)** – vrši selekciju obeležja (sparse rešenje) i može smanjiti prekomerno prilagođavanje.
- **Elastic Net (L1+L2)** – kombinuje prednosti Ridge i Lasso i često je stabilniji kada postoje grupe korelisanih atributa.
- Hiperparametri (α za Ridge/Lasso, kao i α i l1_ratio za Elastic Net) biraju se GridSearch procedurom u okviru k-fold cross-validacije na trening delu.

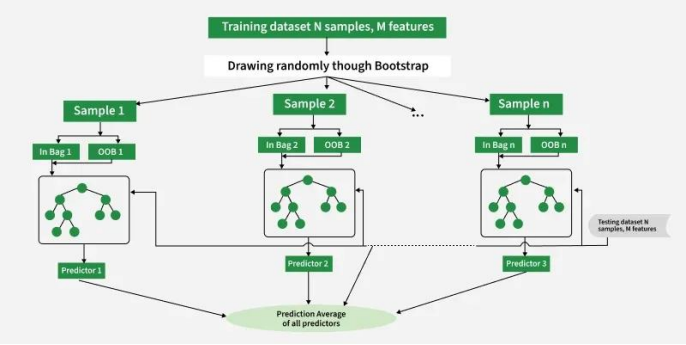


Slika 3: Regularizacija

4. **Nelinearni modeli i ansambli za tabelarne podatke**
   Kako bi se postigla veća tačnost predikcije i obuhvatile nelinearne zavisnosti između socio-ekonomskih obeležja i stope nasilnog kriminala, u ovoj fazi se primenjuju moderni ansambl modeli zasnovani na gradijentnom pojačavanju stabala (Gradient Boosting Decision Trees – GBDT). Kao reprezentativan i posebno pogodan model za tabelarne podatke koristi se **CatBoostRegressor**.
   Za razliku od linearnih modela, GBDT modeli mogu da modeluju složene interakcije između obeležja i nelinearne efekte, pri čemu su robustni na multikolinearnost i često daju najbolje rezultate na tabelarnim skupovima podataka. Dodatno, CatBoost se pokazuje stabilnim u prisustvu nedostajućih vrednosti i šuma u podacima, što je važno za analizirani skup. Za linearne modele state se kodira (one-hot), dok se kod CatBoost-a tretira kao kategorijska promenljiva (native handling).
   U okviru ove faze sprovodi se:

- treniranje CatBoost modela nad trening skupom,
- izbor hiperparametara pomoću k-fold cross-validacije,
- evaluacija performansi na validacionom skupu,
- finalna procena na test skupu.

   Rezultati se upoređuju sa baznim OLS modelom i regularizovanim linearnim modelima (Ridge/Lasso/ElasticNet) u pogledu metrika RMSE i R^2, kao i u pogledu generalizacije. Dodatno, analizira se značaj obeležja (feature importance) kako bi se identifikovali faktori koji najviše doprinose predikciji stope nasilnog kriminala.
   Tokom treniranja koristi se early stopping (na osnovu validacionog dela u okviru cross-validacije) kako bi se sprečilo preprilagođavanje i odabrao optimalan broj iteracija.



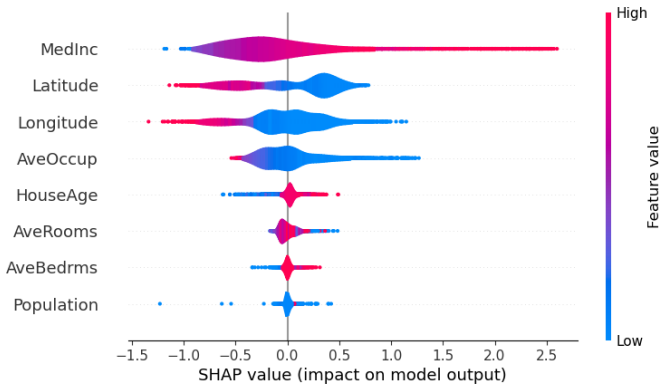
Slika 4: CatBoostRegressor

5. **Interpretabilnost i objašnjavanje predikcija (Feature importance + SHAP)**
   Iako ansambl modeli poput CatBoost-a postižu visoku tačnost, često se smatraju "black-box" pristupima. Zbog toga je u projektu važan deo analiza interpretabilnosti, kako bi se razumelo koja obeležja najviše utiču na predikciju stope nasilnog kriminala i u kom smeru.
   U okviru ove faze sprovodi se:

- **Globalna važnost obeležja (feature importance)** – analiza ukupnog doprinosa svakog atributa modelu,

- **Permutation importance** – robusnija procena važnosti obeležja kroz merenje pada performansi kada se pojedinačno obeležje permutuje (na validacionom skupu).
- **SHAP analiza (SHapley Additive exPlanations)** – objašnjenje predikcija kroz:
    - globalne SHAP vrednosti (najuticajnije promenljive u proseku),
    - lokalna objašnjenja (za pojedinačne zajednice – zašto je model dao baš tu vrednost),
    - vizuelizacije (summary plot, dependence plot) za identifikaciju nelinearnih efekata i interakcija između promenljivih.

  Cilj ove faze je da se, pored same tačnosti, dobije i **kvalitativno objašnjenje** koji socio-ekonomski i demografski faktori najviše doprinose stopi nasilnog kriminala, čime rezultati postaju korisni i u kontekstu interpretacije problema, a ne samo predikcije.



Slika 5: SHAP analiza

5. Evaluacija performansi

Evaluacija se sprovodi konzistentno za sve modele korišćenjem istog protokola podele podataka i iste metrike. Kao primarne metrike koriste se RMSE, MAE i koeficijent determinacije R^2, dok se dodatno prati stabilnost generalizacije kroz razliku između trening i validacionih/test rezultata. Pored numeričkih metrika, radi se i vizuelna analiza predikcija i reziduala, kako bi se identifikovali sistematski obrasci greške.

6. Očekivani rezultati

Očekuje se da ansambl modeli zasnovani na gradijentnom pojačavanju stabala (CatBoostRegressor) ostvare najbolju tačnost na ovom tabelarnom skupu podataka, nadmašujući bazni OLS i regularizovane linearne modele. Regularizovani modeli (Ridge/Lasso/ElasticNet) treba da pokažu poboljšanje u odnosu na OLS u pogledu generalizacije i stabilnosti, posebno u prisustvu korelisanih obeležja. Dodatno, očekuje se da interpretabilnost (feature importance, permutation importance i

SHAP) izdvoji ključne socio-ekonomske faktore povezane sa stopom nasilnog kriminala i pruži razumljiva globalna i lokalna objašnjenja predikcija.

7. Literatura

https://archive.ics.uci.edu/dataset/211/communities+and+crime+unnormalized

https://www.researchgate.net/figure/MICE-imputed-data-have-shown-to-yield-superior-and-robust-classification-accuracy_fig1_358919438

https://scikit-learn.org/stable/modules/linear_model.html

https://en.wikipedia.org/wiki/Ridge_regression

https://www.youtube.com/watch?v=FgakZw6K1QQ

https://www.datacamp.com/tutorial/ols-regression

https://towardsdatascience.com/using-shap-values-to-explain-how-your-machine-learning-model-works-732b3f40e137/

https://www.geeksforgeeks.org/data-science/ordinary-least-squares-ols-using-statsmodels/

https://www.geeksforgeeks.org/machine-learning/methods-for-dealing-with-outliers-in-regression-analysis/

https://www.geeksforgeeks.org/machine-learning/catboost-ml/

https://thaddeus-segura.com/lasso-ridge/