Nikola Kerleta IN 34/2020
Srđan Petrović IN 47/2020

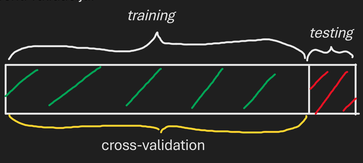
SECOND MILESTONE

1. Obrada nedostajućih i nepotrebnih vrednosti:
U našoj bazi imamo nedostajuće vrednosti za ukupno 3 obeležja.
Međutim, pošto smatramo da ta obeležja ne igraju značajnu ulogu u našem slučaju, možemo samo odbaciti kolone gde se nalaze ta obeležja.
Odbacujemo:
- track_name
- track_artist
- track_album_name

Pored tih obeležja, postoje još njih koja nam ne doprinose proceni jer su ili jedinstvena za svaki uzorak ili nemaju neki preterani značaj, kao što su:
- track_id
- track_album_id
- .....

2. Nakon što smo završili sa odbacivanjem obeležja koja nećemo koristiti nadalje, možemo se pozabaviti i sa direktnim odbacivanjem uzoraka sa nevalidnim vrednostima. Zbog jako dobrog prvobitnog stanja baze, bilo je potrebno samo odbaciti jedan uzorak koji u stvari predstavljao reklamu koja je nekako dospela u bazu.

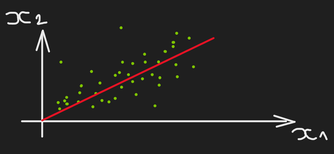
3. Podela podataka biće odrađena na trening i test skup, gde ćemo na trening skupu vršiti unakrsnu validaciju.



4. Manipulacija podacima
Pre procesa regresije, planiramo da prethodno obavimo normalizaciju obeležja kao i pretvaranja određenih obeležja u numerički tip.
Takođe, isprobaćemo selekcije obeležja putem SequentialFeatureSelector funkcije u cilju poboljšanja performansi modela.

5. Odabir modela
S obzirom da predviđamo neku realnu vrednost (popularnost pesme), razumno je da koristimo neki regresioni algoritam.
Neke od opcija koje ćemo razmatrati su:
- Linearna regresija
- SVR (Support Vector Regression)
- Regresija bazirana na stablima odluka

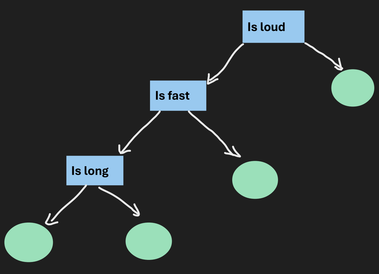
6. Linearna regresija
Testiraćemo više različitih hipoteza kao što su linearna sa i bez interakcija, kao i hipoteza 2. stepena sa interakcijama.
Pored toga, testiraćemo i performanse ridž i laso regularizacije u cilju poređenja istih.



7. SVR
Drugi pristup regresiji, planiramo da obavimo putem SVR modela.
Odnosno, pokušavamo da pronađemo hiperravan koja najbolje prolazi kroz naše uzorke.
Prilikom ovog procesa, testiraćemo performanse sa regularnim SVR pristupom kao i sa NuSVR pristupom, gde uvodimo parametar Nu kao granicu.
Takođe, vršićemo i testiranje sa različitim kernelima kao što su polinomijalni i rbf.

Različiti kerneli regresije

8. Regresija bazirana na stablima odluke
Još jedan pristup koji smo razmatrali je pristup koji koristi stabla odluke. Ovaj pristup je pogodan za interpretaciju jer najčešće koristi binarno stablo zajedno sa kriterijumom deljenja da bi postigao predikciju. Za hiperparametre smo planirali da koristimo max_depth = 3 i min_sample_leaf = 1.



9. Pri kraju procesa predviđanja, testiraćemo performanse svakog od izabranih modela regresija i zabeležiti rešenja i performanse koje nam ti algoritmi pružaju.

10. Ta testiranja obavljaće se pre i nakon primene PCA algoritma za smanjenje dimenzionalnosti.

11. Kao finalni korak, uporedićemo te performanse i rešenja i donositi određene zaključke.

model	error_1	error_2
I		
II		
⋮	⋮	⋮