PROJEKAT SAUSAU

Treniranje modela koji predviđa ukupan broj iznajmljenih bicikala

Faze izrade:

* Analiza podataka:

DataSet sadrži veliku količinu podataka, ali naravno kao u svakom dataset-u nisu svi podaci bitni za predikciju.Kolone *instant*,*dteday*,*yr*,*casual* i *registred* nisu toliko bitne jer ne sadrže podatke od značaja.Iako *casual* i *registred* djeluju kao bitni podaci nama je najbitnije predvidjeti konačan broj iznajmljenih bicikala-*cnt*.

Podaci od značaja se nalaze u ostalim kolonama i bitno je utvrditi koje moguće vrijednosti imaju i da li možda postoji neka anomalija.Ta cijelokupna procedura će se izvršiti u fajlu **dataImport.py**.

U kodu su odrađene analize,ne postoje nikakve anomalije,utvrđeni su opsezi podataka od značaja,provjere nisu vršene nad *casual* i *registred* jer nisu bitni podaci i samo nas zanima *cnt* konačan broj,koji je bilo koji prirodan broj uključujući i 0.

* Preprocesiranje podataka:

Utvrđene su kolone od interesa,ali se i one razlikuju i možemo ih podijeliti u dvije grupe:

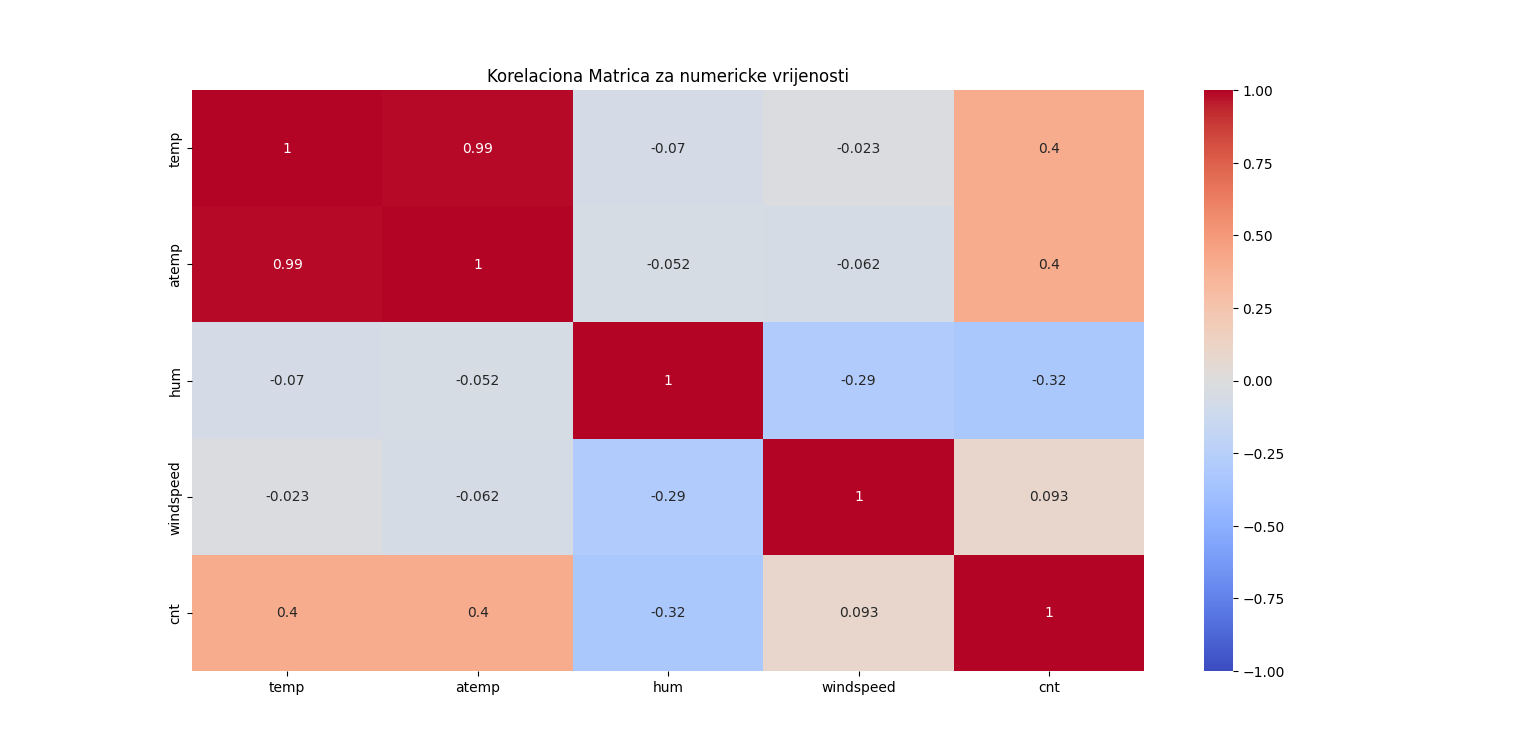
1. „Čisti numerički podaci“, odnosno, podaci koji imaju numeričke podatke za neke vrijednosti i koje kako opadaju ili kako rastu utiču na target.
2. „Kategorijski podaci“, podaci koji jesu brojevi, ali ti brojevi predstavljaju stanja,recimo *season* je kolona koja može imati vrijednosti 1,2,3,4 i to su oznake za godišnja doba predstavljene samo **int** vrijednostima.

U fajlu **dataAnalysis.py** izvršena je analiza podataka,međutim prije toga je izvršeno skaliranje numeričkih podataka kako bi se svi stavili u isti opseg,kao što je već navedeno podaci su podijeljeni u dvije kategorije i na osnovu tih podataka dobili smo sledeće grafike na osnovu kojih možemo vidjeti neke značajne odnose.

U sledećem koraku se vrši eksplorativna analiza skupa.

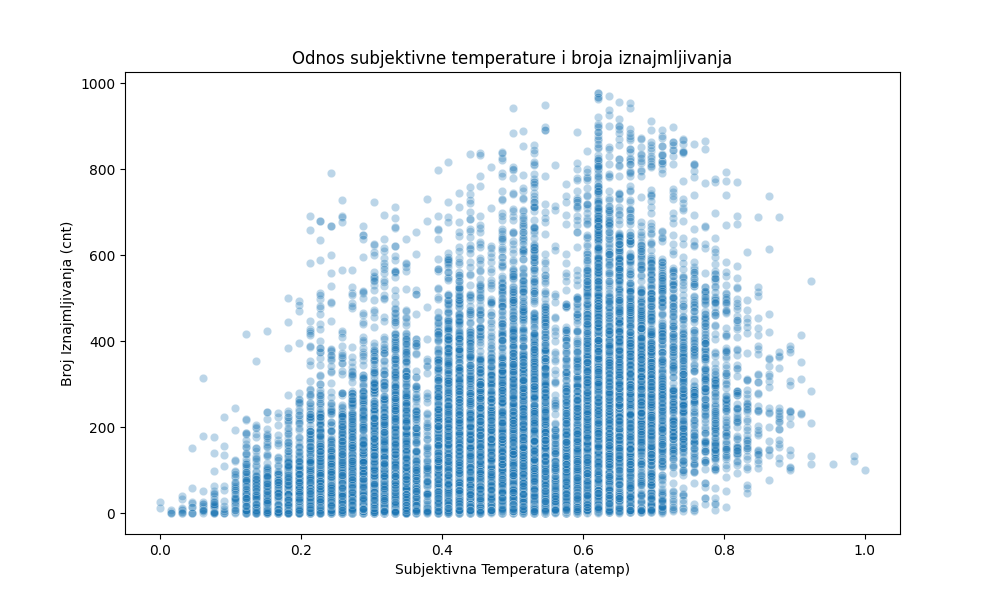
* EDA:

-Korelaciona Matrica:



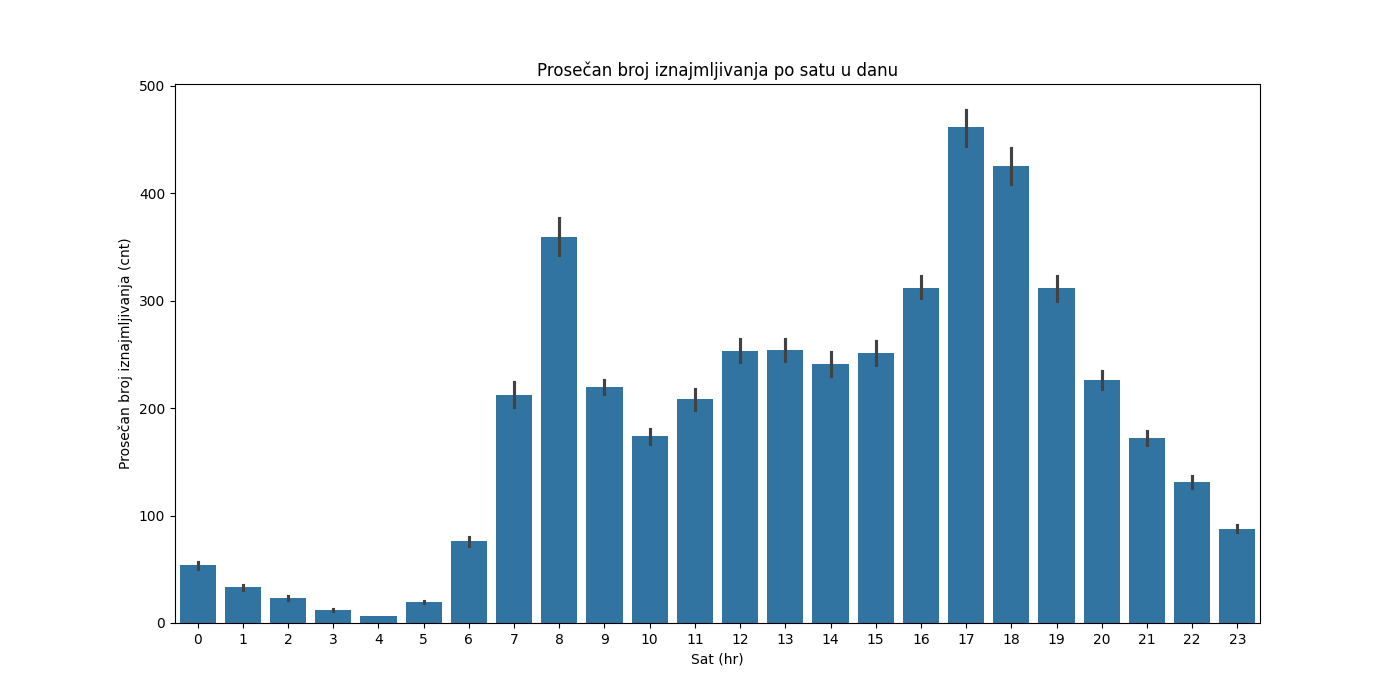
Na osnovu korelacione matrice su izvedeni sledći zaključci:

1. Podaci *temp* i *atemp* imaju skoro savršenu korelaciju što znači da prilikom treniranja modela možemo jednu ovu kolonu izbaciti,i u ovom slučaju izbacićemo kolonu *temp* iz obučavajućeg skupa
2. Podaci a*temp* i *hum* u odnosu na **target** *cnt* imaju vrlo dobar korelacioni odnos,oni se mogu koristiti za korisnu predikciju,jer su proporcionalni tj, obrnuto proporcionalni u odnosu na konačan **target**.

-Scatter Plot(Temperature i iznajmljivanje bicikala): 

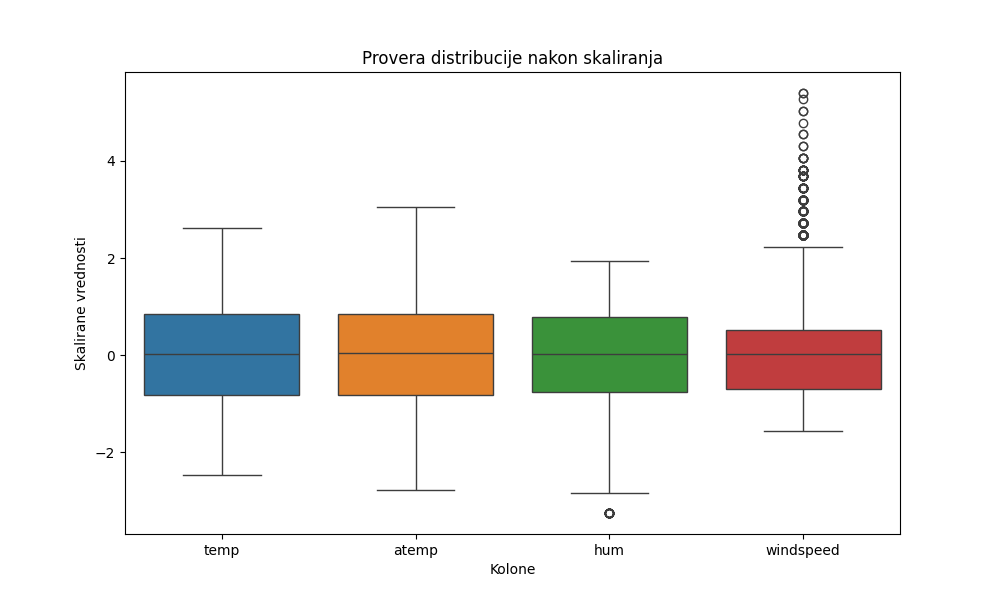
-Iz korelacione matrice smo imali pretpostavku da bi se mogao koristit neki linearan model,ali ipak na osnovu grafika iznad shvatamo da bih mogao,ali on ne bi bio idelan,postoji dosta šumova(oblačića koji bi sigurno odudarali od prave).

-Grafik broja iznajmljivanja u satima:



-Ovaj grafikon nam pokazuje odnos iznajmljivanja po satima,dakle odnos nije linearan ali recimo može se uzeti u obzir da neki vremenski trenuci imaju jaču skalu iznajmljivanja.

-Distribucija nakon sklairanja:



-Sa grafika distribucije vidimo da su sve „čiste numeričke vrijednosti“ skalirane,ali recimo kod podataka *windspeed* i jedan kod *hum* vidimo par izuzetaka, to su zapravo neke ekstremne vrijednosti.

**-ZAKLJUČAK:**

-Na osnovu dosadašnjih podataka možemo zaključiti da prilikom odabira modela za obučavanje ne možemo koristiti linearne modele,oni se ne bi dobro snašli za ovaj skup podataka.Zato se treba napraviti neka kombinacija linearnog modela i stabla odluke,korišćenjem ta 2 modela tj njihove kombinacije mogli bi se ostvariti dobro rezultati.

Formiranje DataSet-a, Treniranje Modela i odabir:

* Pošto je u pitanju regresioni problem,predviđanja nekog broja,velika je vjerovatnoća da će se najbolje pokazati regresioni modeli kao što su **DecisionTree** i **RandomForest**(njihovi modeli **Regressor**).Pored ova 2 modela „šansu“ treba pružiti linearnim modelima **Ridge**  i **Lasso**.**Moja** ideja je bila da se napravi kombinacija 2 modela,jednog koji bi učio na numeričkim podacim, a drugi koji bi radio sa kategorijskim podacima,ali ipak to nije dobra praksa jer sam shvatio da će tako modeli ostati bez bitnih podataka.
* Početna ideja za formiranje DataSet-a je i dalje ista,kao što je prije navedeno izbacuju se određene kolone.Kategorijski podaci su se kodirali, a dok su se numerički skalirali.I nad takvim podacima će se izvršiti trening.Jedino što je ovdje da kažem bitno je hronologija podataka,dakle, ja sam DataSet podijelio na train, test i val skup na sledeći način,train je 80% podataka i to **prvih** 80% koji bi predstavljali „prošlost“ dok val i test imaju po 10% podataka i to tako da val ima 10% nakon traina, „sadašnjost“, i test modela se dešavama na „budućim“ podacima, zadnjih 10%.
* Treniranje modela kreće i program je tako osmišljen da se pamte najbolji parametri treninga i da se po potrebi na grafiku prikažu rezultati testa kako bi se uvidilo o greškama koje je model napravio.U toku treniranja velika je vjerovatnoća da će se DataSet mijenjati da bi se ostvarili što bolji rezultati.Na sledećoj strani se nalazi tabela koja prikazuje kako je za koji DataSet(pošto ih je više formirano) postignut rezultat i koje su greške.

FDS – DataSet čine kategorijski(kodirani) i numerički(skalirani) podaci

FDS2 – DataSet sadrži sin i cos vrijednosti formirane od kolona *hr*,*mnth* i *weekday*(originalne kodirane su izbačene)

FDS3 – DataSet sadrži sin i cos kao i originalne kodirane te kolone

FDS4 – DataSet koji ima parametar drop\_first = False,ne izbacuju se kolone radi redudantnosti, bez sin i cos kolona

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | RIDGE | | | LASSO | | | DT | RF |
| DS\M,E | S | MAE | RMSE | S | MAE | RMSE | R2 | R2 |
| FDS | 0.48 | 96.61 | 135.17 | 0.48 | 96.76 | 135.23 | 0.40 | 0.50 |
| FDS2 | 0.31 | 95.18 | 155.25 | 0.31 | 95.23 | 155.52 | 0.34 | 0.41 |
| FDS3 | 0.48 | 95.61 | 135.17 | 0.48 | 95.61 | 135.17 | 0.33 | 0.41 |
| FDS4 | 0.48 | 96.61 | 135.17 | 0.47 | 95.99 | 135.63 | 0.37 | 0.48 |

* U većini slučajeva nađeni slični parametri modela tokom treninga i oni izgledaju ovako:

-FDS:

Ridge – alpha = 0.001||Lasso – alpha = 0.01

DT – max\_depth = 20, min\_samples\_leaf = 8, min\_samples\_split = 20

RF – max\_depth = None, min\_samples\_leaf = 4, min\_samples\_split = 2, n\_estimators = 200

-FDS2:

Ridge – alpha = 0.1||Lasso – alpha = 0.1

DT – max\_depth = 10, min\_samples\_leaf = 8, min\_samples\_split = 20

RF – max\_depth = 20, min\_samples\_leaf = 8, min\_samples\_split = 2, n\_estimators = 50

-FDS3:

Ridge – alpha = 0.001||Lasso – alpha = 0.001

DT – max\_depth = 15, min\_samples\_leaf = 8, min\_samples\_split = 20

RF – max\_depth = None, min\_samples\_leaf = 8, min\_samples\_split = 20, n\_estimators = 100

-FDS4:

Ridge – alpha = 0.001||Lasso – alpha = 0.1

DT – max\_depth = 15, min\_samples\_leaf = 8, min\_samples\_split = 20

RF – max\_depth = None, min\_samples\_leaf = 8, min\_samples\_split = 20, n\_estimators = 200

-Ovo su do sada najbolja predviđanja modela:

Predvidjanje modela:Lasso(alpha=0.001, max\_iter=25000)

RMSE: 135.1739, MAE: 96.6042, Score: 0.4824

Predvidjanje modela:Ridge(alpha=0.001)

RMSE: 135.1740, MAE: 96.6129, Score: 0.4824

Predvidjanje modela:DecisionTreeRegressor(max\_depth=20, min\_samples\_leaf=8, min\_samples\_split=20, random\_state=42)

RMSE: 117.1231, MAE: 80.0338, Score: 0.6114

Predvidjanje modela:RandomForestRegressor(min\_samples\_leaf=4, n\_estimators=200, random\_state=42)

RMSE: 112.6124, MAE: 77.5154, Score: 0.6407

-Jedan važan detalj, nakon što sam vratio kolonu *yr* rezultati su se značajno poboljšali,to je DataSet5 i on sadrži sin i cos od one 3 pomenute kolone kao i njihove originalne kodirane vrijednosti.I rezultati predikcije su mnogo bolj.

Predvidjanje modela:Lasso(alpha=0.001, max\_iter=25000)

RMSE: 116.1648, MAE: 85.0450, Score: 0.6177

Predvidjanje modela:Ridge(alpha=0.001)

RMSE: 116.1725, MAE: 85.0568, Score: 0.6177

Predvidjanje modela:DecisionTreeRegressor(max\_depth=20, min\_samples\_leaf=8, min\_samples\_split=20, random\_state=42)

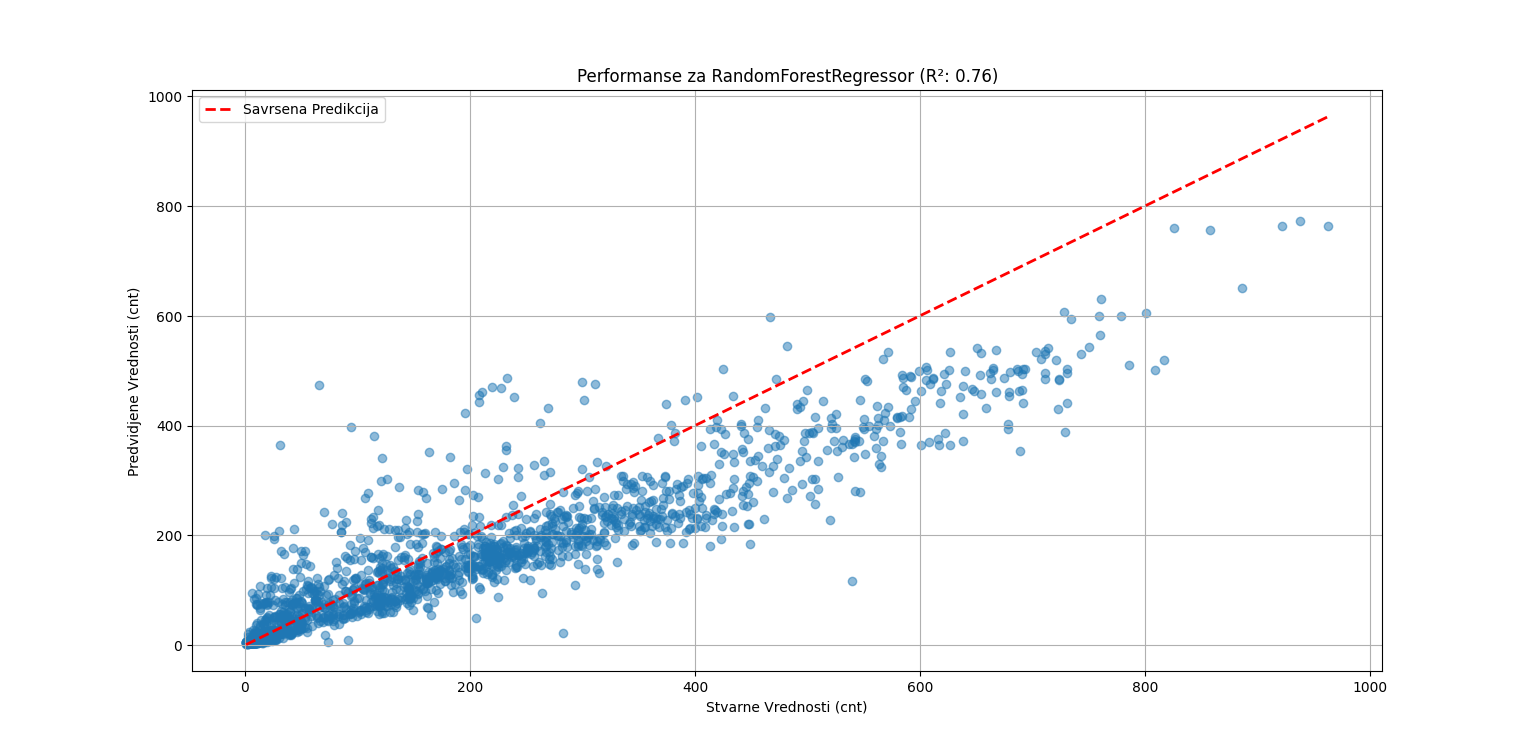
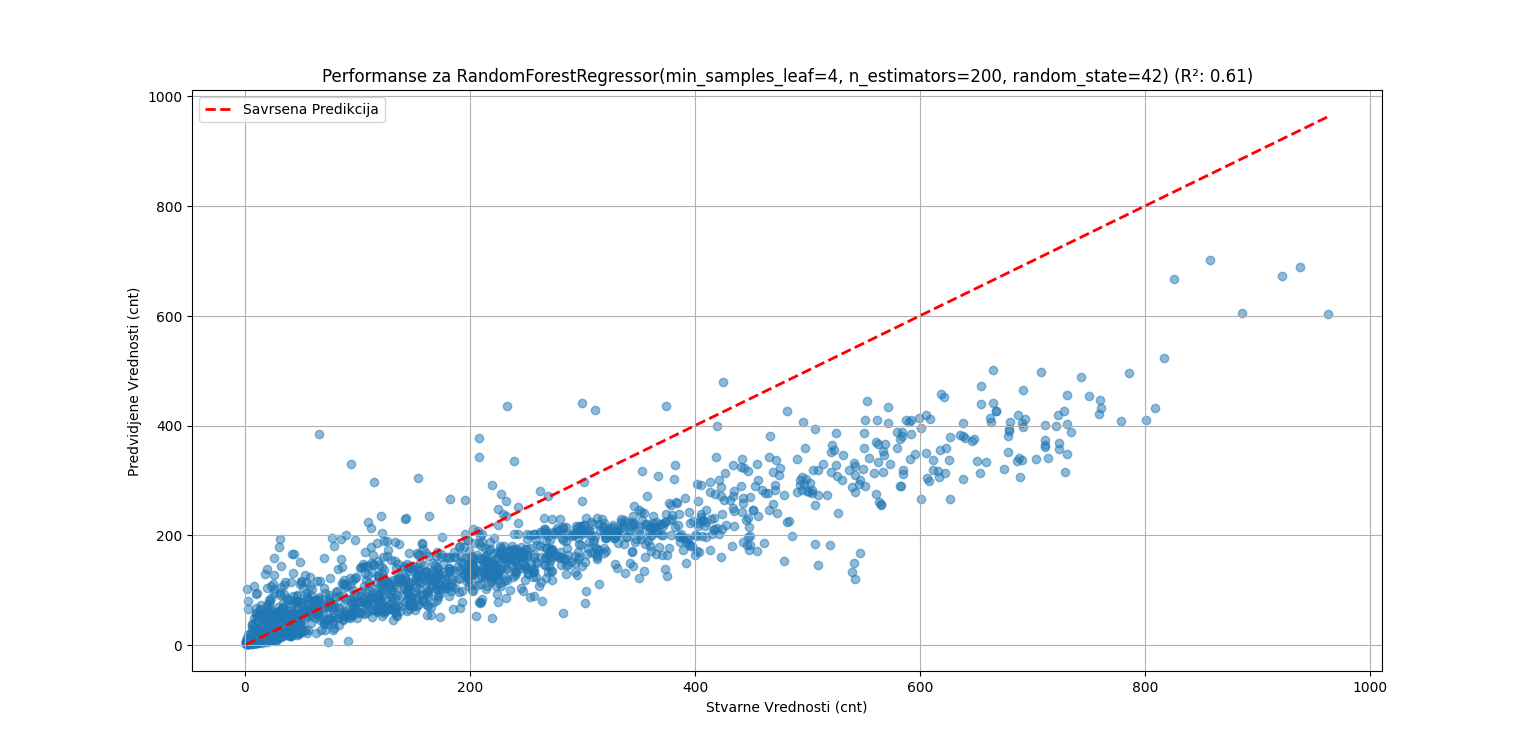
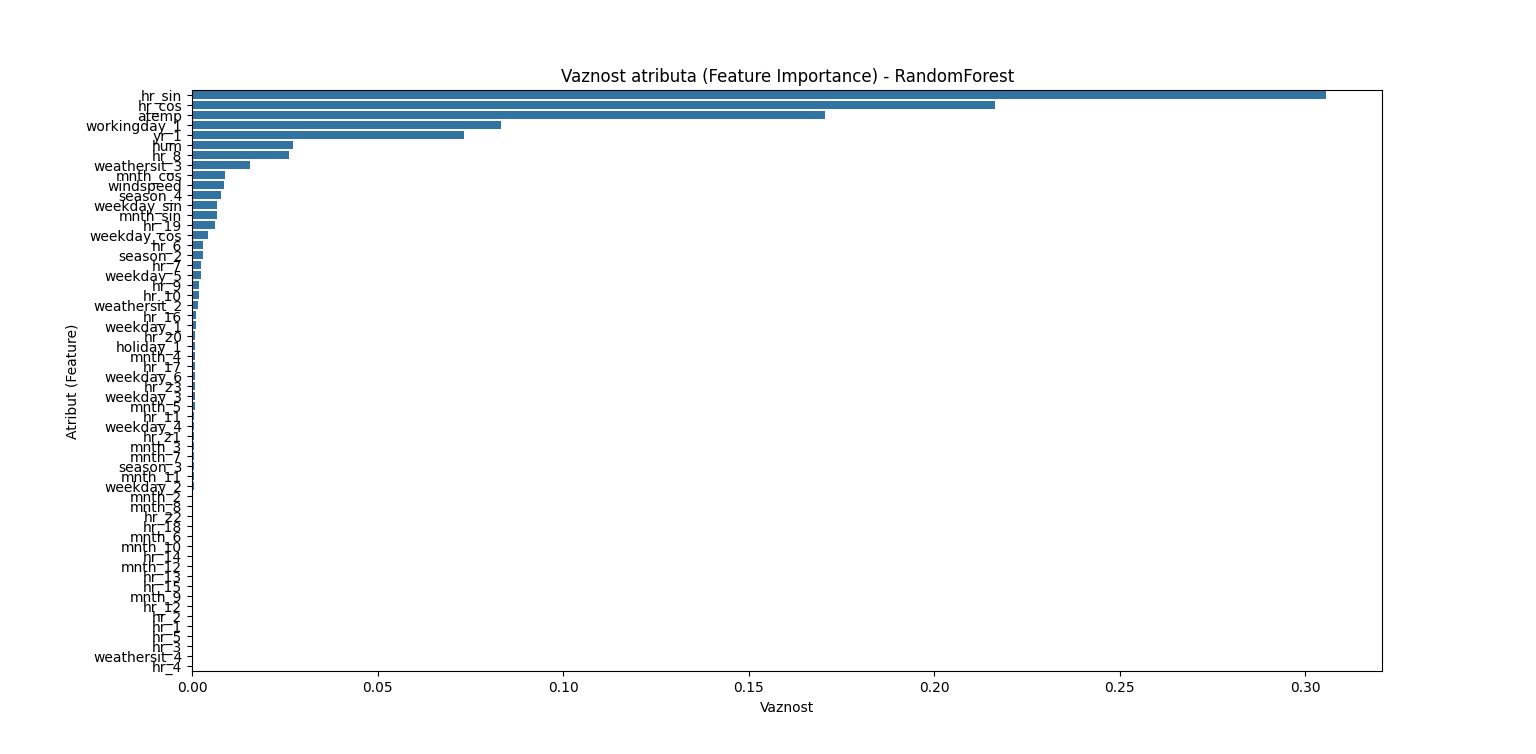
RMSE: 94.7413, MAE: 65.9580, Score: 0.7457

Predvidjanje modela:RandomForestRegressor(min\_samples\_leaf=4, n\_estimators=200, random\_state=42)

RMSE: 91.3947, MAE: 64.7399, Score: 0.7634

I kada bi se pogledali grafici **Feature Importnaces** od svih modela došli bi do zaključka da je mnogo značajan faktor *yr* i ako bismo obratili pažnju na prethodne grafike performansi od modela sa starim DataSet-om vidili bi da baš nedostaje taj trend rasta iz godine u godinu odnosno taj atribut *yr* je malo povećao konačan *cnt* koji je rastao kroz godinu.

-Na sledećoj strani možemo da vidimo grafike gdje jasno možemo da zaključimo koliko se „oblak“ predikcije podigao u novom DataSet-u,isto tako tu je i grafik koji pokazuje važnost novog dodatog atributa.



-Naravno kao što se moglo zaključiti linearni modeli mogu dati pristojne rezultate ali ipak nisu toliko moćni kao DecisionTree i RandomForest,oni su ipak „robusniji“ i od njih se očekivalo da daju kvalitetan rezultat.Odrađeno je i test šta bi se desilo kada bi se izbacili nebitni podaci:

Predvidjanje modela:Lasso(alpha=0.001, max\_iter=25000)

RMSE: 114.3330, MAE: 82.5417, Score: 0.6297

Predvidjanje modela:Ridge(alpha=0.001)

RMSE: 114.3311, MAE: 82.5412, Score: 0.6297

Predvidjanje modela:DecisionTreeRegressor(max\_depth=20, min\_samples\_leaf=8, min\_samples\_split=20,random\_state=42)

RMSE: 94.6893, MAE: 65.8978, Score: 0.7460

Predvidjanje modela:RandomForestRegressor(min\_samples\_leaf=4, n\_estimators=200, random\_state=42)

RMSE: 91.3912, MAE: 64.7032, Score: 0.7634

-Čitav projekat je išao sledećom logikom,za svaki DataSet da se uradi trening modela i nađu njihovi najbolji parametri na osnovu validacionog skupa,u toku samog treniranja i formiranja DataSet-ova dolazilo je do određenih grešaka koje su ispravljene u kasnijim,šta sadrži koji DataSet navedeno je u fajlu **dataSet.py** koji je glavni za formiranje podataka dok se u fajlu **trainModels.py** treniraju modeli i tamo se vrši podjela podataka,metrike su urađene na testnim skupovima i na osnovu toga su birani bolji modeli.I za kraj zadnji DataSet je 7 i on je imao sledeće rezultate

Predvidjanje modela:Lasso(alpha=0.001, max\_iter=25000)

RMSE: 137.3879, MAE: 103.0258, Score: 0.4653

Predvidjanje modela:Ridge(alpha=0.001)

RMSE: 137.3877, MAE: 103.0317, Score: 0.4653

Predvidjanje modela:DecisionTreeRegressor(max\_depth=20, min\_samples\_leaf=8, min\_samples\_split=20,random\_state=42)

RMSE: 91.2457, MAE: 64.0393, Score: 0.7641

Predvidjanje modela:RandomForestRegressor(min\_samples\_leaf=4, n\_estimators=200, random\_state=42)

RMSE: 90.8195, MAE: 64.2951, Score: 0.7663

-Konačan kraj predstavlja sama ideja kombinovanja 2 modela što će se uraditi putem **Stacking-a** a to je da ćemo u **Stacking ubaciti** jedan najbolji linearni(to će biti **Ridge** jer je stabilniji) i **RandomForest** jer je do sad davao najbolje rezultate,i kao glavni **estimator** se uzima **Ridge**.

-U sklopu čitavog procesa je i fajl u kome se nalaze grafici i na kraju u nazivu grafika broj predstavlja za koji se DataSet radio trening i predikcija,i jedino što fajl **finalDataSet2.1** predstavlja pravu verziju jer se u verziji 2 dogodila greška u kodu prilikom formiranja **sin** i **cos** podataka.

