Analiza podataka – snage betona

Ognjen Poznanović

I. Uvod

Pojava modernog betona koji nam je poznat i koji se uz manje modifikacije koristi i danas datira iz 1824. Njegova nova, poboljšana verzija podrazumevala je korišćenje cementa koji se dobijao na potpuno novi način. Masovna proizvodnja kreće 1828. u Irskoj i tada je predstavljao materijal sa najboljim karakteristikama potrebnim za građevinske projekte. Neki od prvih značajnijih projekata na kojima je korišćen jesu izgradnja mosta 1850. i proširenje luke 1875. Glavna karaktetistika betona jeste njegova snaga koja se meri kao mogućnost betona da izdrži određeni pritisak pre neko što izgubi svoje prvobitne karakteristike. Naš cilj jeste da napravimo model koji će računati snagu betona (*Concrete compressive strength*) sa dostupnim podacima. Rešavanje ovog problema podrazumeva veoma nelinearnu funkciju starosti i sastojaka.

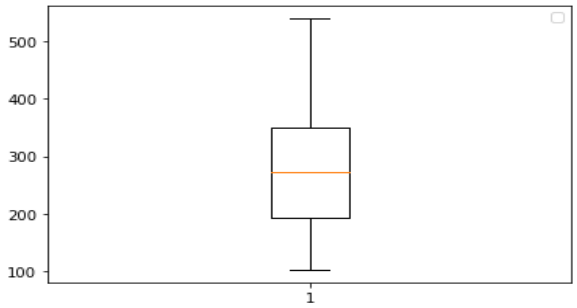
II. Baza podataka

Baza podataka sadrži snagu betona kao i sve informacije potrebne da se ona odredi. Ona sadrži 1030 uzoraka gde je svaki opisan sa 9 obeležja. Obeležja su: *Cement*, *Blast Furnace Slag*, *Fly Ash*, *Water*, *Superplasticizer*, *Coarse Aggregate*, *Fine Aggregate*, *Age*, *Concrete compressive strength*. Sva obeležja su numerička.

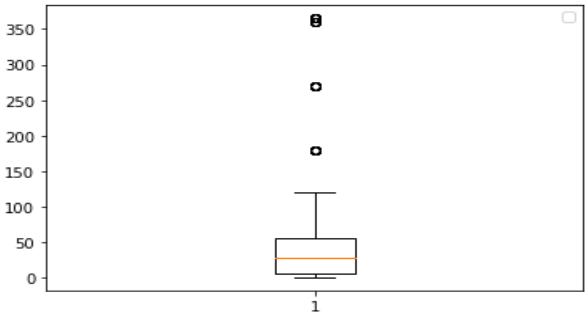
III. Analiza podataka

Prvi zadatak bio je provera broja uzoraka i obeležja. Baza podataka nema nedostajućih vrednosti. Svih 1030 uzoraka sadrži određene vrednosti u svakom od svojih obeležja. Zatim su skraćeni nazivi kolona tako da se ne izgubi smisao ali da olakša dalji rad u analizi i rukovanju sa podacima.

Sledeći korak bio je iscrtavanje histograma i *boxplot*-ova kako bi se stekao utisak o raspodeli i primetile pojedine nelogičnosti. Nakon analize histograma i *boxplot*-ova primećeno je da uzorci kod obeležja: *Cement*, *Fly Ash*, *Coarse Aggregate* nemaju *outlier*-e (Slika 1). Kod ostalih postoje *outlier*-i ali oni u proseku čine manje od 1% ukupnih podataka pa su ostavljeni. Jedini izuzetak su bila odstupanja kod *Age* kolone. Međutim svi podaci su ostavljeni jer ovo čini jedno od najvažnijih osobina koje formira snagu betona. Pretpostavljeno je da su inženjeri želeli da testiraju kako što duži interval stajanja utiče na snagu (Slika 2).



Sl. 1. Prikaz *boxplot*-a za kolonu *Cement*.



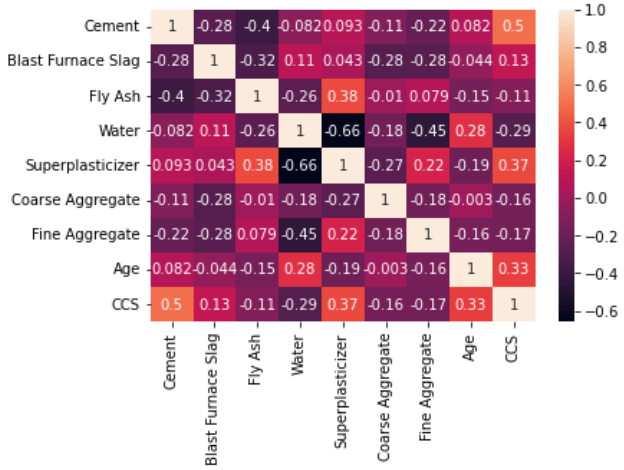
Sl. 2. Prikaz *boxplot*-a za kolonu *Age*.

Iz korelacije *Concrete compressive strength* sa ostalim obeležjima vidi se niska pozitivna korelacija sa kolonama *Cement* (0.5), *Superplasticizer* (0.37) i *Age* (0.33) kao i niska negativna korelacija sa kolonom *Water* (-0,29) (Tabela 1).

Kod korelacije svih obeležja primećena je niska pozitivna korelacije između *Superplasticizer* i *Fly Ash* (0.38) kao i umerena negativna korelacija između *Superplasticizer* i *Water* (-0.66)

KORELACIJA		
OBELEŽJE1	OBELEŽJE2	VREDNOST
CONCRETE COMPRESSIVE STRENGTH	*CEMENT*	0,50
	BLAST FURNACE SLAG	0,13
	FLY ASH	-0,11
	WATER	-0,29
	SUPERPLASTICIZER	0,37
	COARSE AGGREGATE	-0,16
	FINE AGGREGATE	-0,17
	AGE	0,33

Tabela 1. Prikaz korelacije *CCS* sa ostalim obeležjima.



Sl. 3. Korelacija svih obeležja.

IV. LINEARNA REGRESIJA

Linearna regresija predstavlja metodu nadgledanog učenja koja ima za cilj da predviđa određenu vrednost obeležaja kao rezultat zbira proizvoda parametara i vrednosti obeležaja koji se koriste za treniranje. Način na koji se model obučava podrazumeva menjanje vrednosti parametara tako da se model posle svake iteracije približava tačnoj vrednosti. Neki od načina na koji se može meriti uspešnost linearne regresije su: Mean Squared Error (MSE), Root Mean Squared Error (RMSE) i R-squared. Što su manje vrednosti kod prva dva načina to je model bolji. Rezulata R-squared se kreće u opsegu od 0 do 1 i što je broj bliži jedinici to je model bolji. Najveći problemi kod obučavanja modela jesu pristrasnost i varijansa. U slučaju male pristrasnosti to govori da model nije uspeo da razume šta mu je potrebno i važno od podataka sa kojim raspolaže kako bi napravio dobru predikciju. Dok velika vrednost varijanse predstavlja da će model imati probleme sa procenom kada se iskoristi novi skup podataka.

Za rešavanje ovog problema koristili su se modeli iz klase *LinearRegression*, *Ridge*, *Lasso*. Modeli su trenirani sa datim podacima, standardizovanim kao i podacima koji nastaju kao rezultat primene *PolynomialFeatures*. Takođe modeli su trenirani na smanjenom broju dimenzionalnosti.

Polynomial regression je jedna vrsta linearne regresije koja pravi nova obeležja od kombinacije postojećih i ima mogućnost podizanja obeležja na stepen. Za formiranje novih obeležja koristi se funkcija *transform*. Prilikom inicijalizacije podešavaju se parametri kao što su: *degree*, *interaction_only i include_bias*. Prvi predstavlja ceo broj i zadužen je za odlučivanje do kog stepena će podizati obeležja. Drugi može imati vrednosit *True i False*, ako je označeno *True* neće biti dozvoljeno podizanje obeležja na stepen. Poslednji parametar uvodi konstantu kao obeležje.

Kod dva modela iz klase *LinearRegression* menjana je vrednost parametra *fit_intercept* kao i korišćenje datih podataka naspram standardizovanih u drugom modelu.

Kod modela koji koristi podatke iz obeležja koji nastaju nakon primene *Polynomial regression*, testirano je koji će najmanji broj *degree* dati najbolji rezultat.

Kod modela iz klase *Ridge* i *Lasso* menjana je vrednost parametra *alpha* kao i podaci iz obeležja koji nastaju nakon primene *Polynomial regression*.

	LinearRegression		
	dati podaci	standardizovani	*Polynomial regression*
MSE	109	1384	28
R2 score	0.62	-3.7	0.90

Tabela 2. Prikaz rezultata modela klase *LinearRegression*

	Ridge		
	dati podaci	standardizovani	*Polynomial regression*
MSE	109	1384	36
R2 score	0.62	-3.7	0.87

Tabela 3. Prikaz rezultata modela klase *Ridge*

	Lasso		
	dati podaci	standardizovani	*Polynomial regression*
MSE	109	1384	34
R2 score	0.62	-3.7	0.88

Tabela 4. Prikaz rezultata modela klase *Lasso*

Selekcija odabira obeležja urađena je metodom odabira unapred. U ovom slučaju gde postoji mali broj obeležja dodatno smanjenje ne doprinosi poboljšanju rezultata. Zapravo smanjenje dimenzionalnosti u ovom slučaju dovodi do pogoršanja konačnih rezultata.

V. Zaključak

Najbolja predviđanja daje model iz klase *LinearRegression* prilikom korišćenja podataka iz obeležja koji nastaju nakon primene *Polynomial regression*. Identični rezultati se ostvaruju kada parametar *degree* ima vrednost između tri i šest. Za konačan model odabrana je vrednost tri zbog smanjenja kompleksnosti. Za postizanje boljih rezultata potrebna je baza podataka sa više uzoraka.

VI. Literatura

https://concrete.ie/about-concrete/history-of-concrete/#:~:text=The%20birth%20of%20modern%20concrete,stone%20in%20the%20hardened%20state

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

https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.PolynomialFeatures.html

Mast. inž. Tijana Nosek, Doc. dr Branko Brkljač, Mast. inž. Danica Despotović, Prof. dr Milan Sečujski, Prof. dr Tatjana Lončar-Turukalo: Paktikum iz mašinskog učenja, Fakultet tehničkih nauka, Univerzitet u Novom Sadu