Извештај

Анализа података и обука модела линеарне регресије - концентрација PM2.5 честице

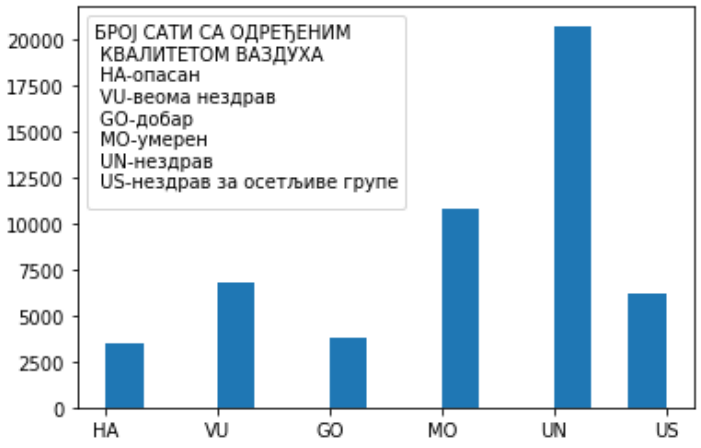
Аутор: Гашпар Јован

1.Опис базе података

Овај извештај се бави анализом података везаних за концентрацију PM2.5 честица у ваздуху. Један узорак садржи информације о концентрацији PM2.5 честица у једном сату. Узорци су забележени у станици 'US PM Post' у Пекингу у периоду од 2010 до 2015 године. База садржи 52578 узорака. Сваки од узорака има следеће атрибуте: концентрацију PM2.5 честица, температуру ваздуха,брзину и правац ветра влажност ваздуха, количину падавина, температуру кондензације, ваздушни притисак као и информације о томе када је прикупљен узорак(година,годишње доба месец, дан и сат). Атрибути који се односе на време када је прикупљен податак као и правац ветра су категорички, док су остали атрибути нумерички.

2.Анализа података

Приликом анализирања улазних вредности и додатног упознавања са самом темом закључио сам да је неопходно додати још један атрибут који се користи за означавање квалитета самог ваздуха. Тај атрибут сам назвао AQ(eng. *Air quality*) и додао сам га у жељи да прикажем расподелу улазних вредности.AQ представља категорички атрибут.

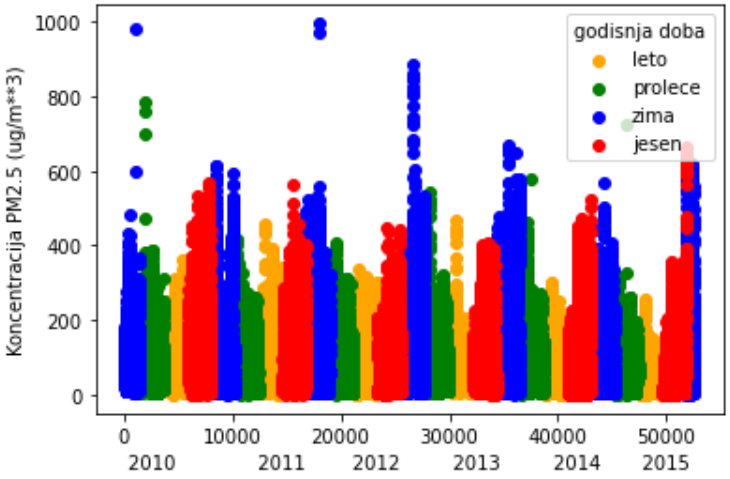


Слика1: Приказ расподеле квалитета ваздуха

У каснијој анализи атрибут AQ ће се показати као користан при одређивању концентрације PM2.5 честица јер директно упућује на опсег вредности из кога се узима

сама вредност PM2.5 по следећој скали: GO(0,12), MO(12,35.4), US(35.5,55.4), UN(55.5,150.4), VU(150.5,250.4),HA(>250.5)

Приликом одабира узорака за рад одбачено је занемариво мало узорака због недозвољених вредности на више од 20% обележја(свега 6 узорака).Код осталих узорака је ради попуњавања недостајућих и нелогичних вредности примењена медијана, претходно их групишући водећи се идејом да ће бити приближније вредности обележја које се односе на исто годишње доба.

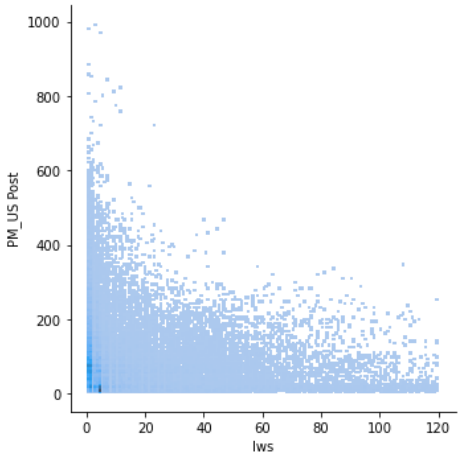


Слика2: Приказ концентрацијеPM2.5 честица у зависности од годишњег доба

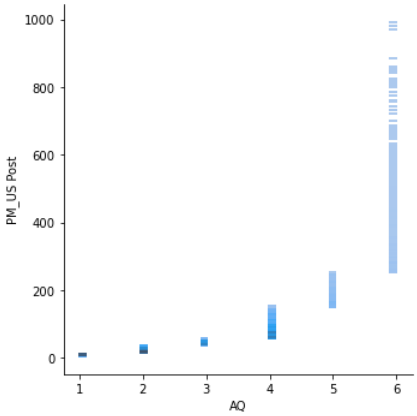
Јасно се види да је просечна концентрација PM 2.5 честица мања током лета у односу на друга годишња доба.

Коришћењем матрице корелације издвојио сам најизраженије зависности од обележја 'PM_US Post' то су зак ”квалитет ваздуха - *AQ*”: 0.84; ”влажност ваздуха - *HUMI*”:0.4; “правац ветра - *cbwd*”=0.25;”годишње доба - *season*”:0.128; ”топлоту кондезације - *DEWP*”:0.12; ”брзина ветра - *Iws*”:-0.19

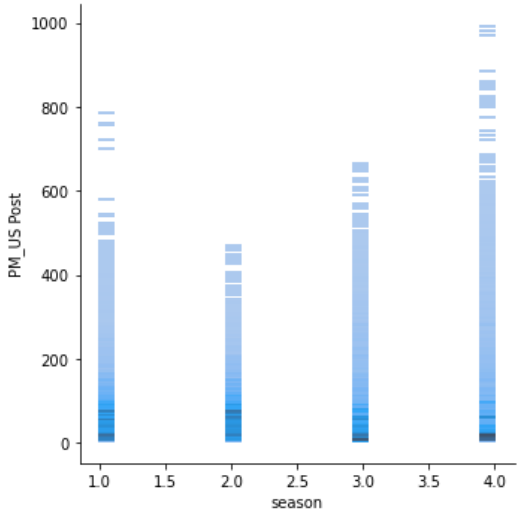
Испод су наведени различити графици расподеле концентрације PM2.5 честица у зависности од одређеног обележја као и базне функције које су коришћене приликом обуке модела.



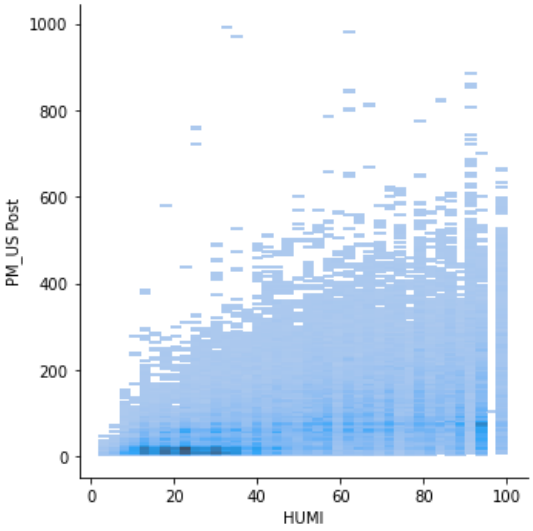
Слика3: Приказ зависности концентрације PM 2.5 честица у зависности од брзине ветра,базна функција коришћена над обележјем „lws“ је: Y=(e^-x)+20



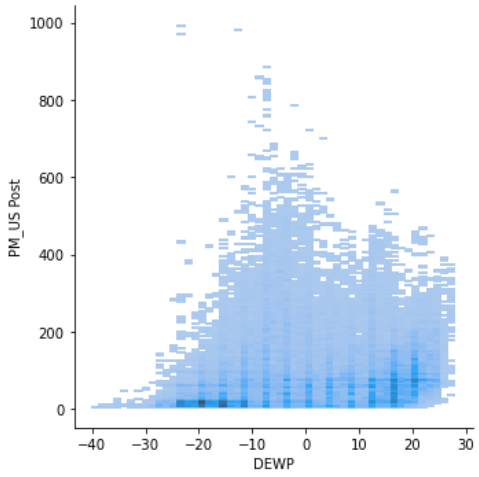
Слика4: Приказ зависности концентрације PM 2.5 честица у зависности од квалитета ваздуха, базна функција која је коришћена над обележјем „AQ“ је: y=e^x



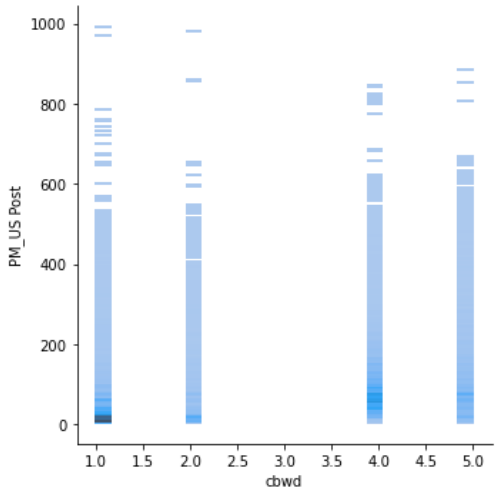
Слика5: Приказ зависности концентрације PM 2.5 честица у зависности од годишњег доба



Слика6: Приказ зависности концентрације PM 2.5 честица у зависности од влажности ваздуха, базна функција која је примењена у овом случају је : Y= (e^sqrt(x))/32



Слика7: Приказ зависности концентрације PM 2.5 честица у зависности од температуре кондензације



Слик8: Приказ зависности концентрације PM 2.5 честица у зависности од правца ветра

На основу вредности обележја датих у бази извршена је обука модела линеарне регресије у три случај 1.без регуларизације 2. Са Ridge регуларизацијом 3. Са Lasso регуларизацијом.

У почетку су коришћена искључиво обележја која су била дата у бази. Резултат успешности модела након обуке је био испод 0.4. Након тога је уврштено и обележје „AQ“. Добијен је знатно бољи модел за сва три начина обуке(око 0.7). После прелиставања литературе у жељи да побољшам прецизност модела додајем и вештачка обележја која се односе на производе постојећих обележја са одговарајућом корелацијом. На крају су додате и базне функције добијене анализирајући претходно исцртане графике. Резултат је био задивљујућ. Иако је модел постао комплекснији проширивањем са 11 нових обележја али је због тога прецизност порасла нешто мало

изнад 0.9. Упоредивши међусобно моделе закључујем да се модел са Ridge регуларизацијом незнатно боље показао.

3.Резиме

Приликом обуке регресионог модела је важно уочити која су обележја међусобно корелисана, поготово је важно увидети која су обележја корелисана са обележјем које предвиђа модел линеарне регресије.Да би се боље искористили улазни подаци додају се нова обележја.Када модел постане превише комплексан, тада треба одбацити обележја са најмањом корелисаношћу. За додатну прецизнност се користе базичне функције чијим се увршташем боље описује зависност неког обележја од интересног.