**ML\_31\_07\_2023**

**Decision Tree Regression**

Decision tree regression’in tek farki target’in continous veri olmasi.

Decision tree classification gini ve entrophy’e gore hesaplama yapiyordu.

Decision tree regression ise min squarred error kullanacak.

Variance’in en dusuk oldugu yere gore bolumleme yaparak calisir.

Variance’in daha dusuk oldugu yer daha tutarli bir noktadir.

**Describe:**

Owner’da std, mean’den daha fazla. Outlier var denebilir.

**Car\_name.value\_counts-**->kac tane unique deger var ona bakiyoruz. Cogundan birer tane var. Boyle bir dataseti geldiginde, napariz?

-daha anlamli siniflar olusturacak sekilde toplulastirmaya calisirim.

Herbir feature icin en az 15-20 gozlem isteriz deriz ya. Class sayisi fazla ise, 15-20 sekilde gozlem olacak sekilde onlari birlestiririm yani. Ilgili sinifi temsil edebilecek kadar gozlem olmali yani.

-Gozlem sayisi az olanlari drop etmek de mumkun. Cunku 1 gozlemle ne egitebilirim ne model kurabilirim.Extra bilgi alabilirsen bilgi alip, gozlem sayisini artir illa kalsin diyorsan.

**Feature Engineering:**

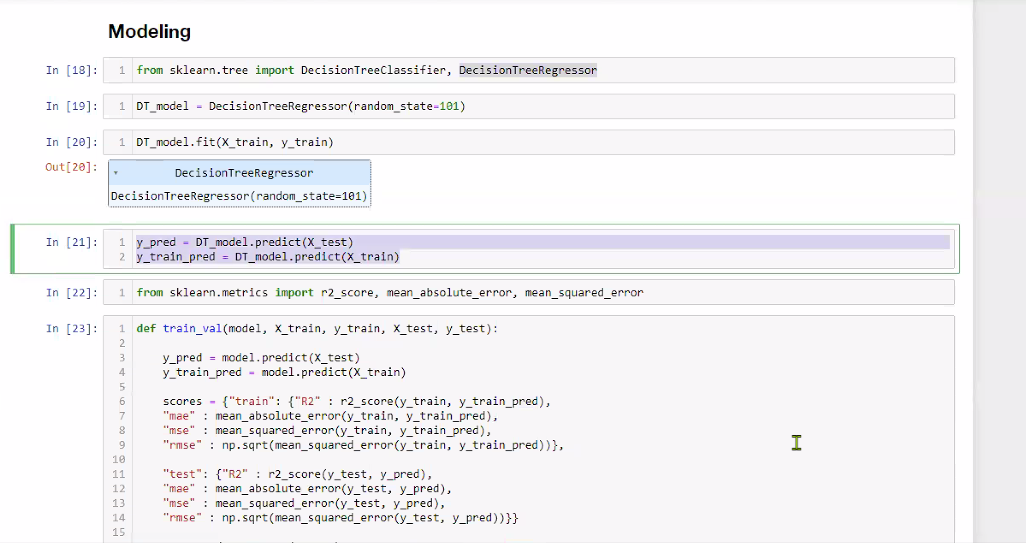
Model yilini arabanin yasi olarak guncelleyebilirim.

Yil da yuksekse km de yuksekse ayni anlama gelebilir

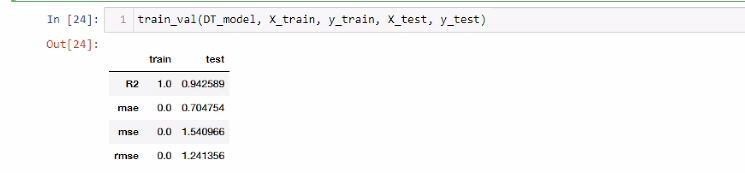
Vehicle\_age yaptim year’i dropladim.

**Get\_dummies**

Str olan featurelari numeric hale getirdim, drop=first dedim

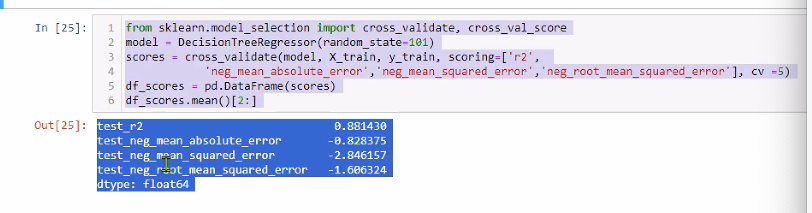


Regression modeli oldugu icin regression skorlari geliyor

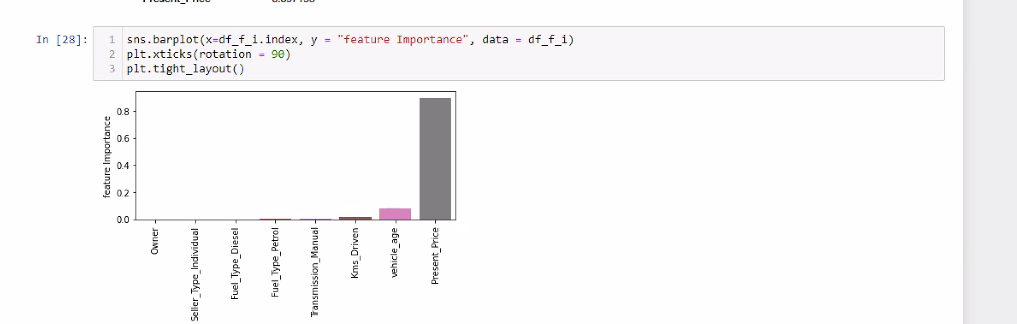


R2 train’de 1 olmasi fazla olmayan bir sey. testte 1’e yakin thats’ fine.

Mae, mse, rmse 0 olmasi cok imkansiz. Ozellikle regression modellerde boyle skor gormek zor. Boyle bir sey gorunce Overfitting var denebilir rahatlikla.



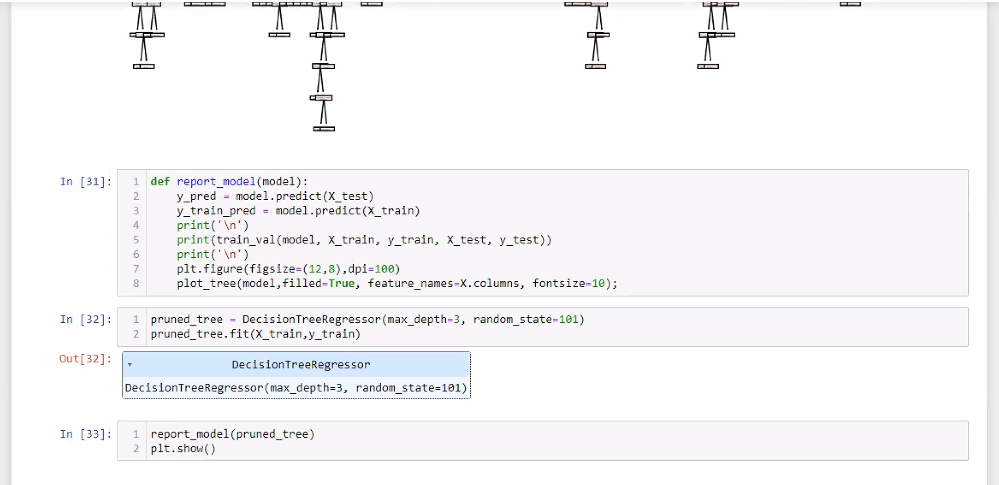
**Feature Importance**

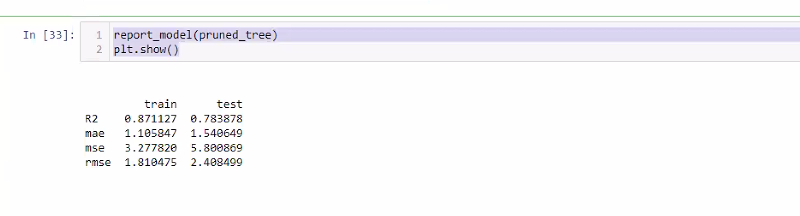


DT’de feature importance ozellikle guvenilir, ozellikle de random forest kisminda.

DT modelimiz en onemli feature’u verdikten sonra digerlerine o kadar prim vermez.

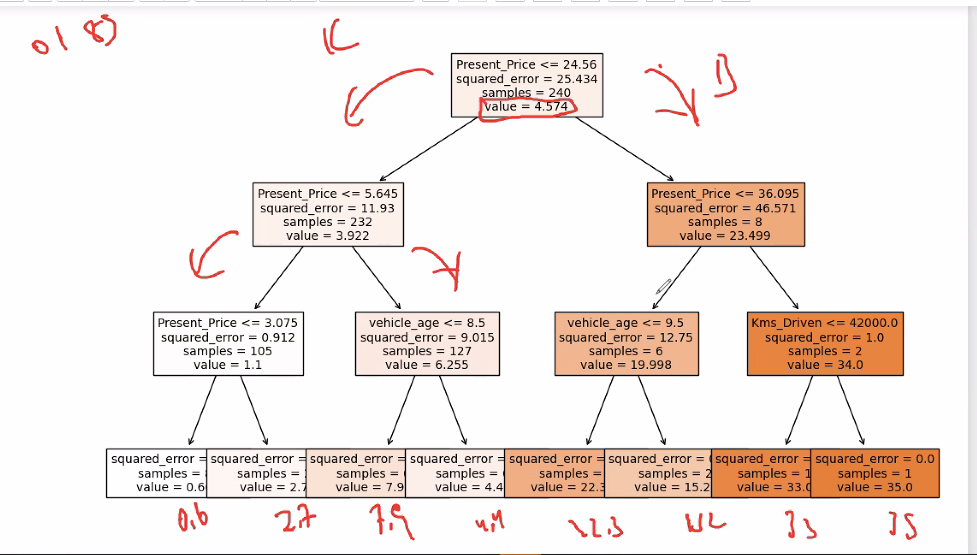
Bu yuzden en yuksek feature bir dusurulerek tekrar skor alinmasi gerekir.





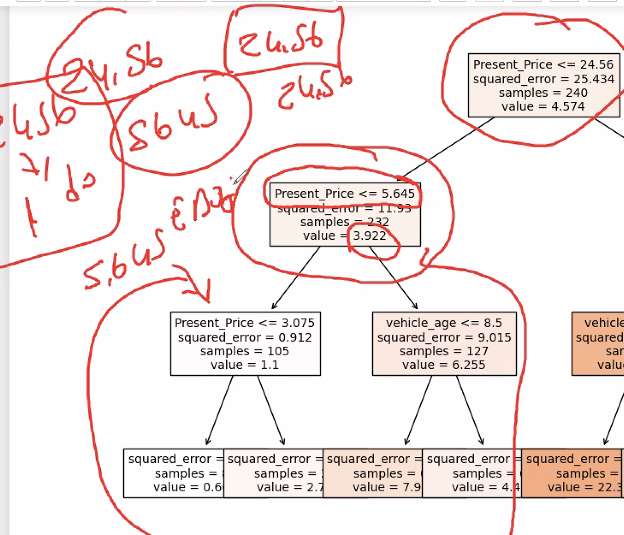
Regresyon modelime sunu diyorum:

Train datasinda bozulma saglayarak test datamda duzelme saglamis oldum.



En iyi bolunmeyi present\_price’a gore yapiyormus. Ondan basliyor. Once Present\_price’a gore sorular soruyor, buyuk mu kucuk mu sorusu uzerinden dallanmaya devam ediyor. Aldigi skorlara gore satis fiyati belirlenecek.

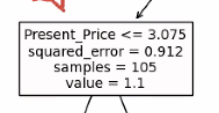
Burda bir tek value var.



Standard variance’e gore optimum bolunme noktasini belirliyor.

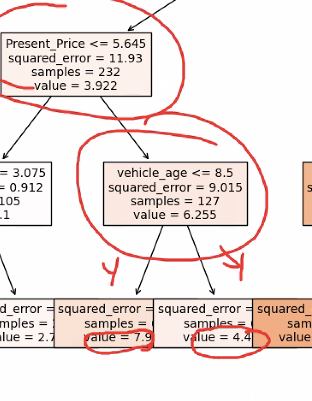
Solda 24.566’dan kucuk olanlar var. Buraya dusenlerin ortalamasi 3.922 imis.

Bunlarin bolunmek icin en ideal noktasi, yani variancein en dusuk oldugu bolunme noktasi 5.645 imis present\_price acisindan.



Yani bu kutuda en yuksek degeri 5.645 olanlar varmi. Buraya dusen gozlemlerin de ortalamasi 1.1’mis.

Bir sonraki bolunmeye bakacak olursak:

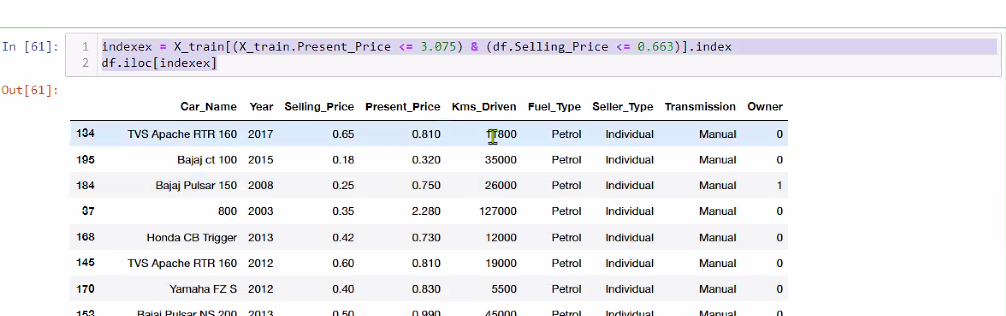


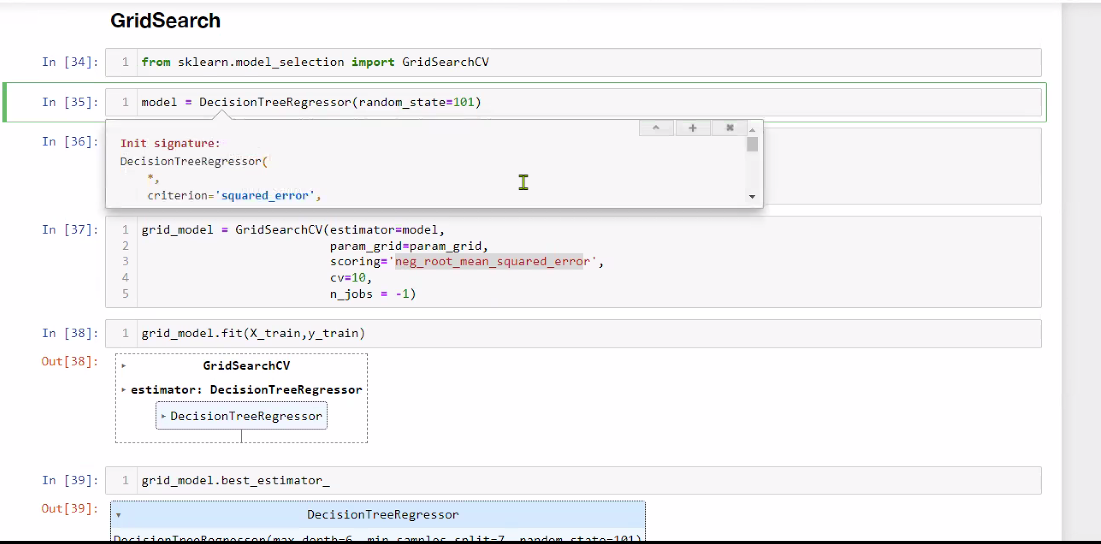
Present\_price’i 5.645 olanlarin ortalamasi 6.255 imis. Bunlara standard sapma en dusuk olacak sekilde vehicle\_age<=8.5 sorusu sorulmus.

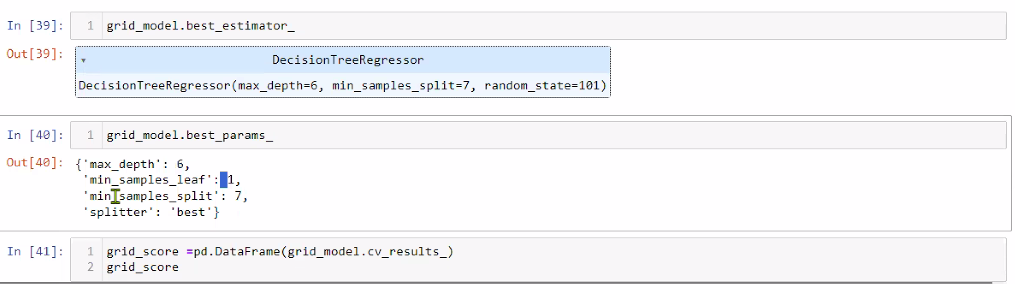
vehicle\_age<=8.5 olanlardan fiyat ortalamasi 7.9 iken

vehicle\_age>8.5 olanlardan fiyat ortalamasi 4.4 imis.

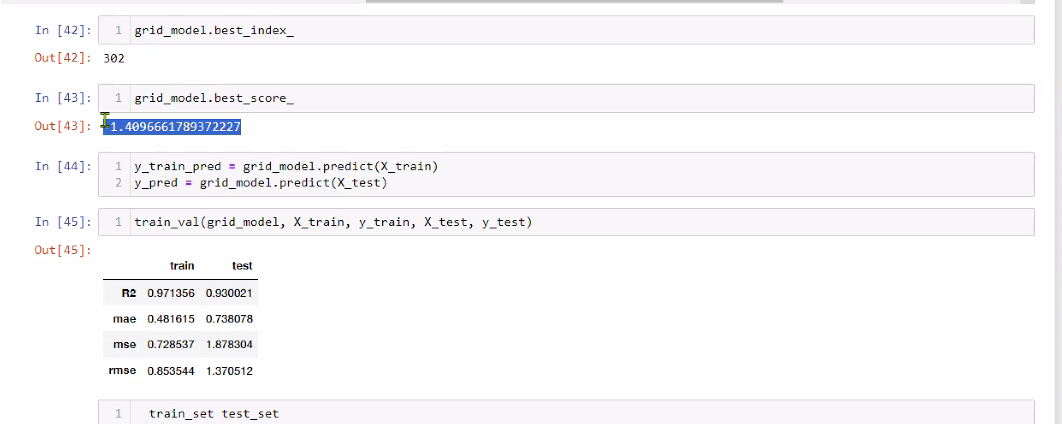
**Yani value’lar hep target’in ortalamasi**



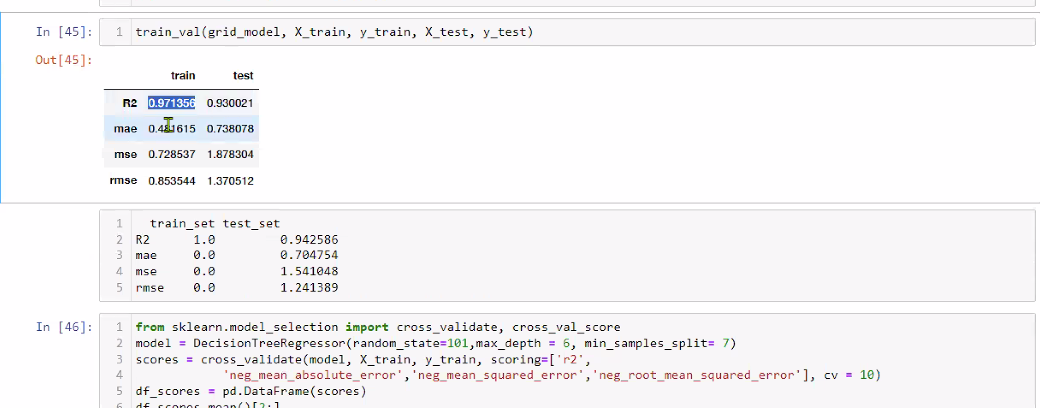




Max\_depth’i 6’ya dusurdu.



Best\_score - rmse skorum

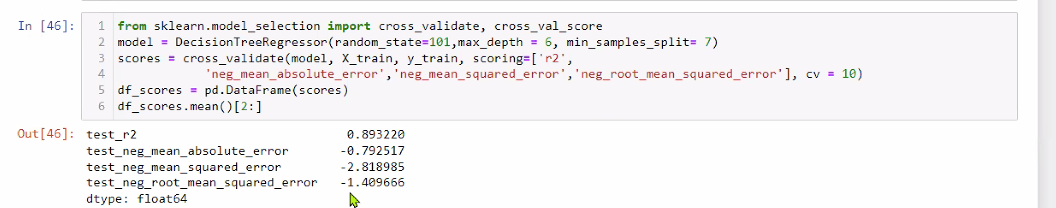


Overfit durumu ortadan kalkmis.

0.97’lere kadar train datam gelmis.

**Butun skorlari gormek icin tekrar manuel CV ‘ye bakarsam:**

Manuel kod bu:



Grid search yaptim sonucu begenmedim diyelim o zaman ancak hyperparametrelerle oynayabilirim ama bilerek yapmam lazim. Feature engineering’e geri donerim.

