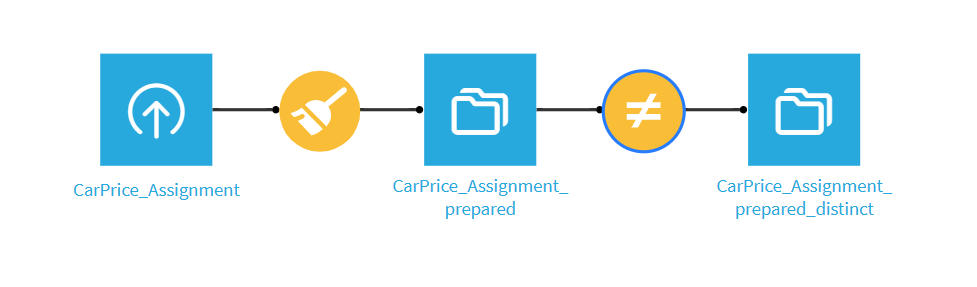
|  |
| --- |
|  |
|  |
|  |

# ÖDEV KONUSU

|  |
| --- |
| Ödev2 Sizlerle Hafta 3 klasörü içerisinde paylaşılan **CarPrice\_Assignment.csv** veri setini kullanarak öncelikle veri setindeki düzenleme işlemlerini yapmanız, ardından da düzenlediğiniz veri setini kullanarak araba fiyatını tahmin eden bir Regresyon Modeli oluşturmanız beklenmektedir. Hem **veri düzenlemesinde yapmış olduğunuz işlem adımlarını** hem de **elde ettiğiniz model sonuçlarını** ekran görüntüleri aracıyla **yorumlarla da destekleyerek** bir **pdf dosyası** **formatında** 5 Ocak 2024 tarihine kadar sisteme yüklemeniz beklenmektedir. |
|  |
|  |

1. CarPrice\_Assignment.csv  veri setinin düzenlenmesi
   1. Proje oluşturma ve veri setinin Import edilmesi

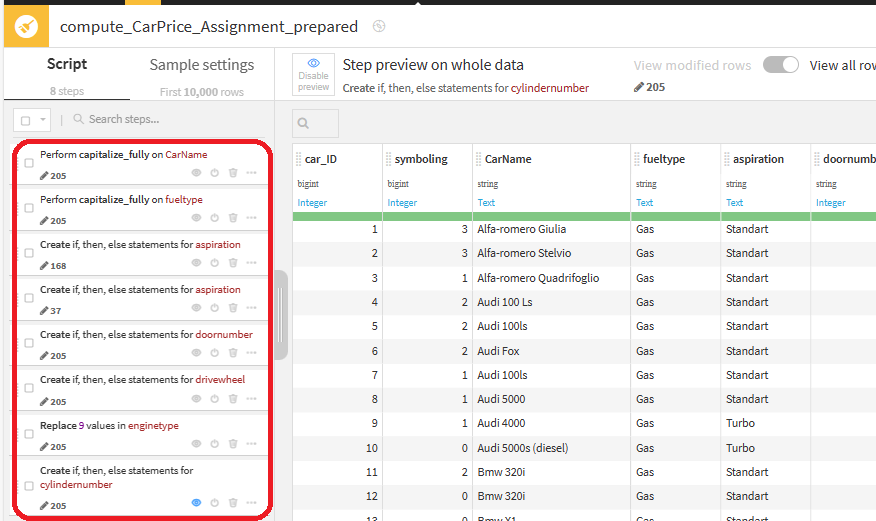
Regresyon isminde bir proje oluşturularak CarPrice Assignment csv dosyayı projeye import edilmiştir.



* 1. Verinin Hazırlanması

Visual Recipeslerden Prepare Recipe ile veride düzenlenmesi gereken alanla düzenlenmiştir. Bunlar boş veriler,bazı text dataların işlenebilir sayısal veriye dönüştürülmesi görsel okunurluk vs.

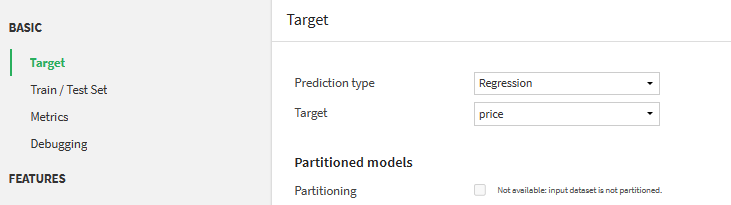
Ardından Distinct Recipe ile tekrar eden veriler ayıklanmıştır.



1. Visual Analyses – ML Regresyon Modelinin uygulanması [Train &Test]
   1. ML Target Regresyon seçimi:

"Prediction" (tahmin), gelecekte olabilecek olayları veya durumları önceden belirleme veya hesaplama sürecidir. Bu terim, genellikle istatistik, veri bilimi, yapay zeka ve diğer analitik alanlarda kullanılır. Prediction işlemi, mevcut veri setlerine dayanarak gelecekteki bir durumu tahmin etmeye çalışır.

Tahmin tipi olarak regresyon ve tahmin edeceğimiz alan Fiyat olduğu için gerekli seçimleri yapıyoruz.



* 1. Train/Test Set: Modelin eğitim:

Eğitim Seti (Train Set):

Modelin eğitildiği veri setidir.

Model, bu set üzerindeki verilere maruz kalarak öğrenir.

Eğitim seti, genellikle veri setinin büyük bir kısmını oluşturur.

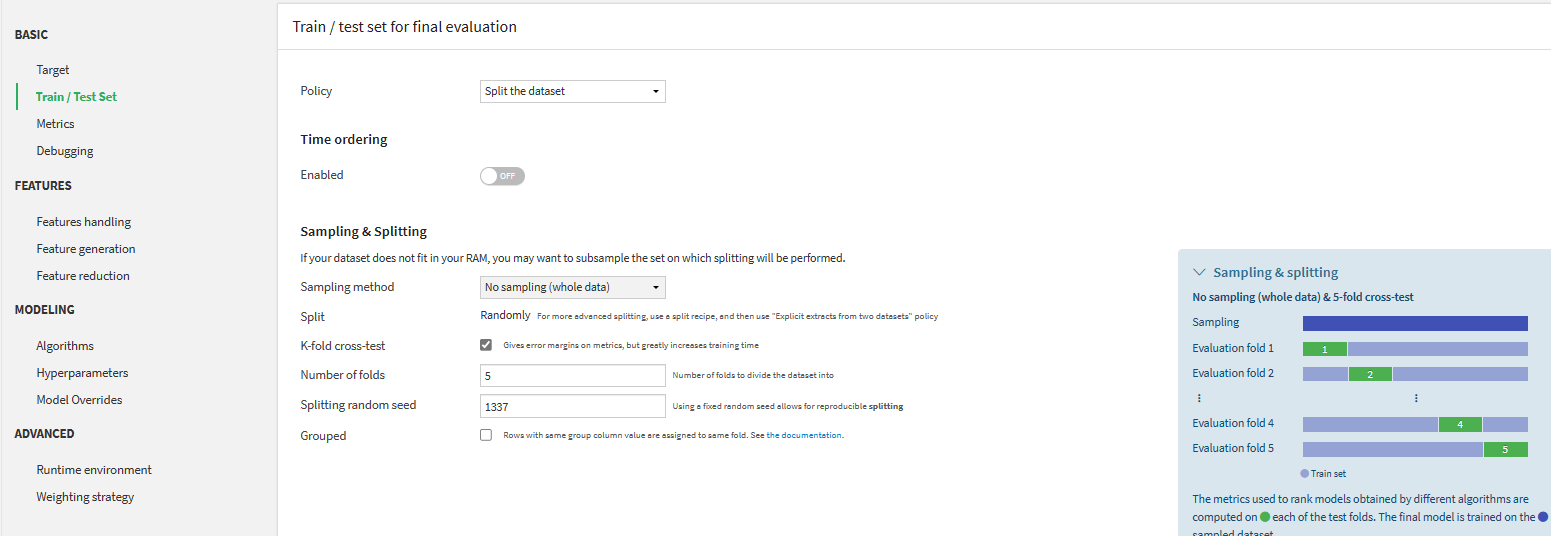
Test Seti (Test Set):

Modelin performansının değerlendirildiği veri setidir.

Model, eğitildikten sonra test setindeki verilere uygulanır ve tahminler yapar.

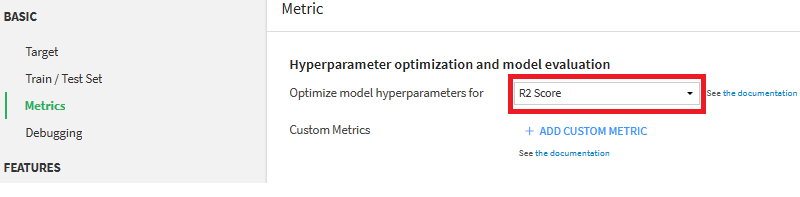
Test seti, genellikle veri setinin küçük bir kısmını oluşturur.

Sampling method alanında tüm datayı seçtil ve Validasyon için K-Folds Croos-Test işaretleyip fold sayısı defult 5 seçildi. Buradaki seed veri gruplamayı ifade ediyor aynı sonucu elde etmek için bu sayı farklı 2 node de kullanıldığında aynı sonuç alınır.



* 1. Metric-R2 Scrore seçimi

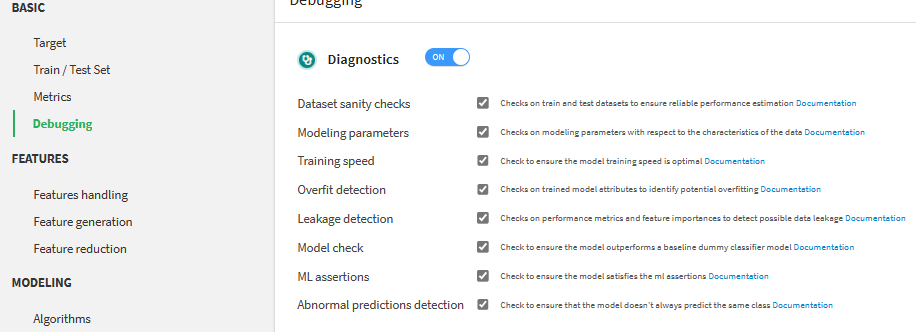
R2 score (R-kare skoru veya R-kare değeri), bir regresyon modelinin ne kadar iyi uyum sağladığını ölçen bir istatistiksel ölçüdür. Bu değer genellikle modelin bağımsız değişkenlerle bağımlı değişken arasındaki varyansın yüzde kaçını açıkladığını ifade eder.



Kısaca, R2 score, bir regresyon modelinin veri setindeki değişkenliği açıklama yeteneğini gösterir. Değer 1'e ne kadar yakınsa, modelin veriyi o kadar iyi açıkladığı anlamına gelir. Ancak, 0'a ne kadar yaklaşırsa, modelin veriyi açıklama yeteneği o kadar düşüktür.

* 1. Hata Ayıklama ve İzleme:

"Debugging" alanı, kullanıcılara projelerinin farklı aşamalarında ortaya çıkan hataları ayıklama ve çözme yetenekleri sunar. Hataların nedenini belirleme, kayıtları inceleme ve gerekirse kod veya konfigürasyon düzenlemeleri yapma imkanı sağlar.

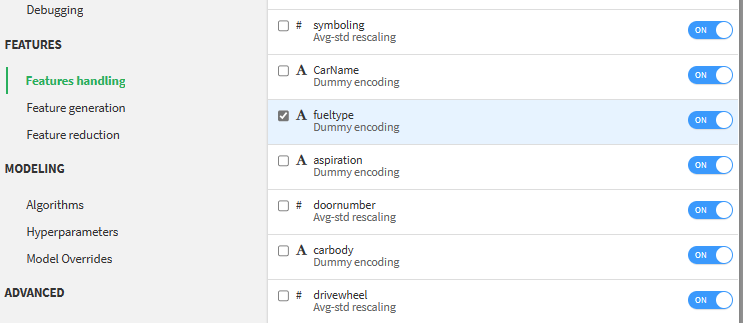


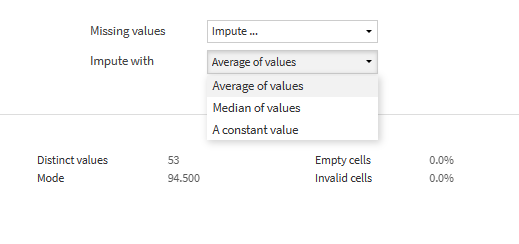
* 1. Features Handling[Özellik işleme]

Değişken tipimiz Categorical veri olanlar için price değerinin değişişmine katkı sağlayacaksa ordinary seçilip diğer durumlar için Ordinal encoding seçilerek ilerlenmiştir. Detay seçimlerde ise aşağıdaki hedefler gözetilmiştir.

Özellik işleme, veri bilimi projelerinde önemli bir adımdır çünkü kullanılan özellikler (features), modelin performansını doğrudan etkiler. Özellik işleme aşamaları arasında şunlar yer alabilir:

1. **Eksik Verilerin İşlenmesi:** Veri setindeki eksik veya boş değerlerin belirlenmesi ve uygun bir şekilde işlenmesi.
2. **Outlier (Aykırı Değer) İşleme:** Aykırı değerlerin belirlenmesi ve bu değerlerin model üzerindeki etkilerini azaltmak için işlenmesi.
3. **Özellik Mühendisliği:** Var olan özelliklerin kullanılarak yeni özelliklerin oluşturulması veya mevcut özelliklerin dönüştürülmesi.
4. **Kategorik Değişken İşleme:** Kategorik değişkenlerin sayısal formata dönüştürülmesi (örneğin, One-Hot Encoding veya Label Encoding).
5. **Standartlaştırma ve Normalizasyon:** Özelliklerin ölçeklendirilmesi ve farklı ölçeklerdeki özelliklerin benzer bir ölçekte olacak şekilde düzenlenmesi.
6. **Özellik Seçimi:** Modelin performansını artırmak için gereksiz veya düşük bilgi içeren özelliklerin çıkarılması.





Missign alanlar için Dataikunun alanları nasıl dlduracağına ait seçimler yapılır.

1. Dummy (One-Hot Encoding):
   * Ne İşe Yarar: Kategorik değişkenleri "Dummy" seçeneğiyle işlediğinizde, bu değişkenler "One-Hot Encoding" yöntemiyle temsil edilir. Her kategori, bir dizi ikili (0 veya 1) değişken olarak ifade edilir. Seçilen kategorinin varlığı durumunda ilgili ikili değişken 1'e, diğerleri ise 0'a eşit olur.
   * Kullanım Örneği: Bir renk kategorik değişkenini ele alalım (örneğin, "Kırmızı", "Mavi", "Yeşil"). Dummy encoding ile her renk için ayrı bir ikili sütun oluşturulur: "Renk\_Kırmızı", "Renk\_Mavi", "Renk\_Yeşil".
2. Labeling (Ordinal):
   * Ne İşe Yarar: Kategorik değişkenleri "Labeling" seçeneğiyle işlediğinizde, bu değişkenlerin kategorilerine sıralı sayısal değerler atanır. Kategori sıralaması, belirli bir sıralamaya göre yapılır ve sayısal değerlerle temsil edilir. Bu, kategoriler arasındaki ilişkiyi belirlemek için kullanışlıdır.
   * Kullanım Örneği: Kullanıcı memnuniyeti kategorik bir değişkense ("Düşük", "Orta", "Yüksek"), Labeling ile "Düşük"e 1, "Orta"ya 2, ve "Yüksek"e 3 gibi sayısal değerler atanabilir.
   1. Algoritma seçimi:

Aşağıdaki Algoritmalar seçilmiştir.

Random Forest:

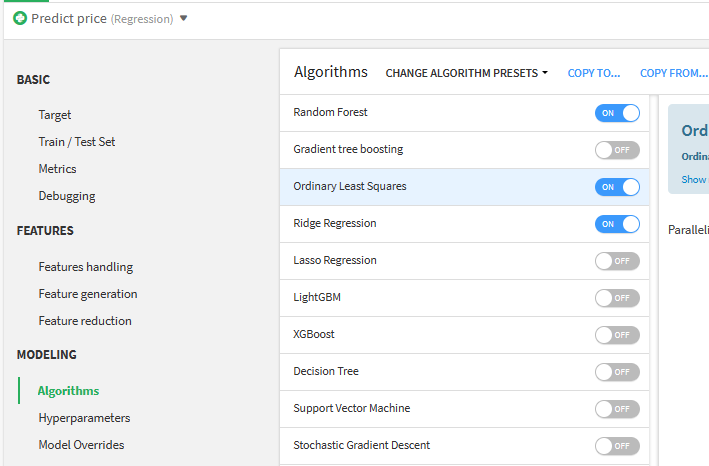
* Ensemble öğrenme algoritmasıdır.
* Yüksek doğruluk ve dirençli model performansı sunar.
* Sınıflandırma ve regresyon görevlerinde kullanılabilir.

Ordinary Least Squares (OLS):

* Doğrusal regresyon yöntemidir.
* Basit, yorumlanabilir ve hesaplama açısından hızlıdır.
* Doğrusal ilişkileri ifade etmek için kullanılır.

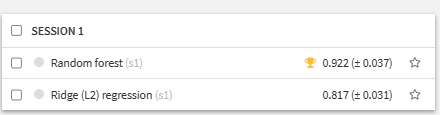
Ridge Regresyon:

* Doğrusal regresyonun bir türüdür.
* L2 düzenlemesi ekleyerek modelin katsayılarını sınırlar.
* Çoklu korelasyon ve aşırı uymaya karşı dirençlidir.

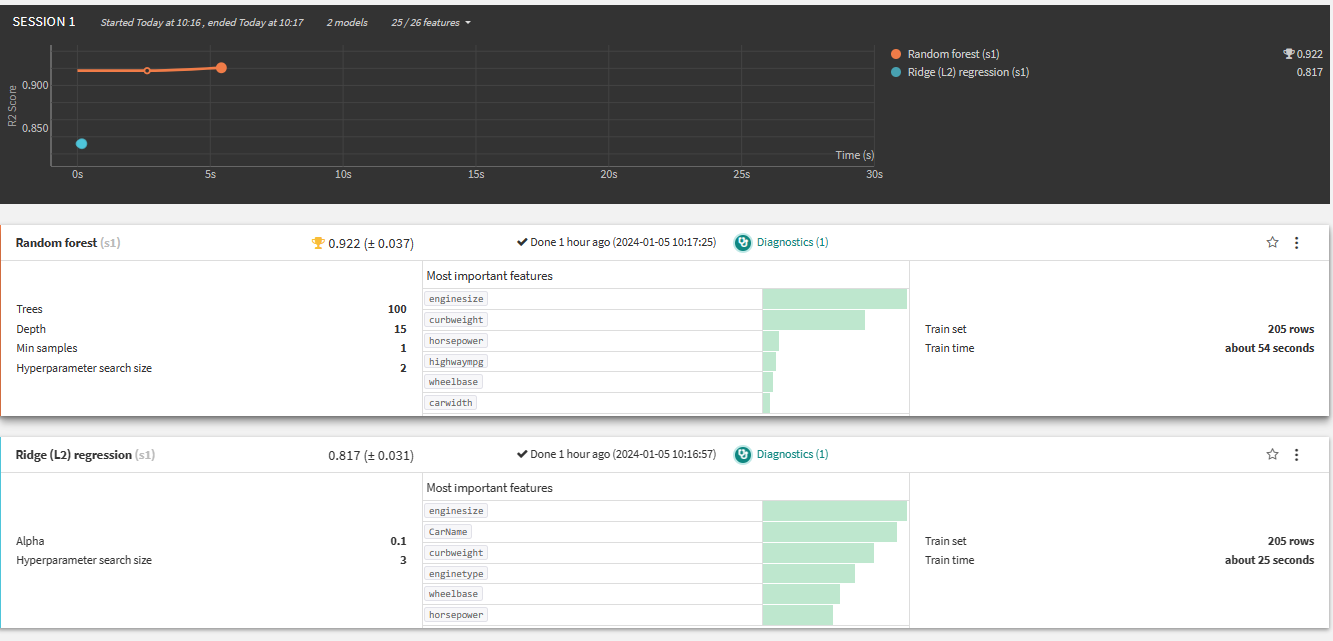


* 1. Train ve Yorumlar

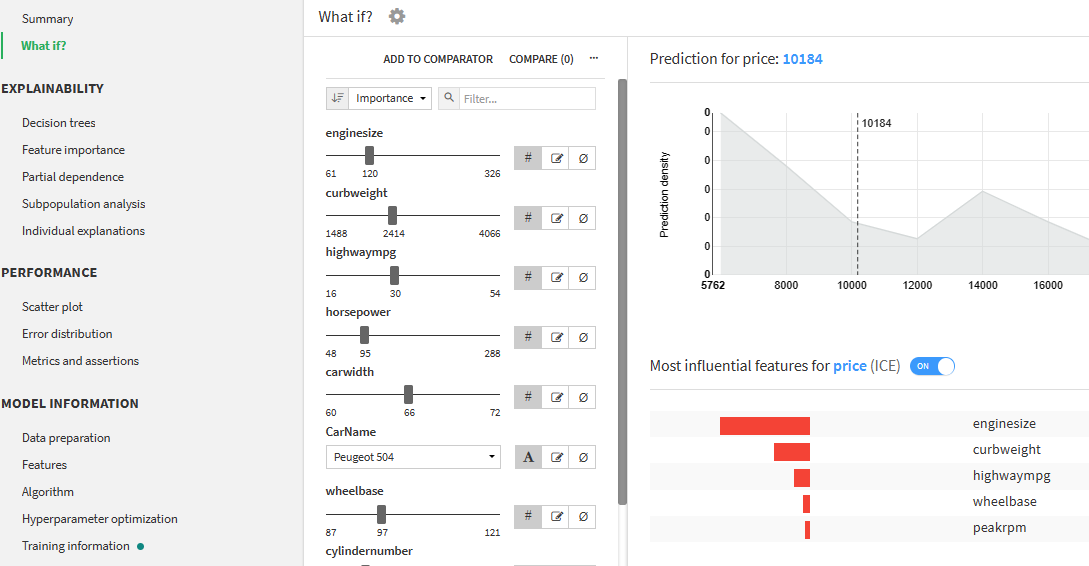
Aşağıda seçilen algoritmalar mevcut.

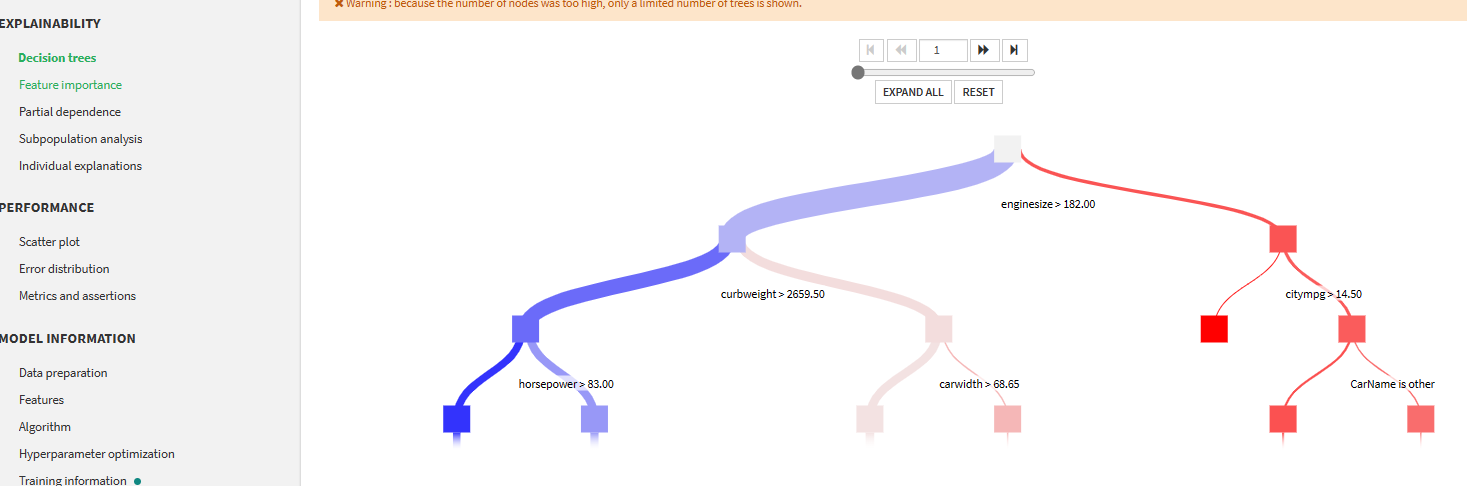


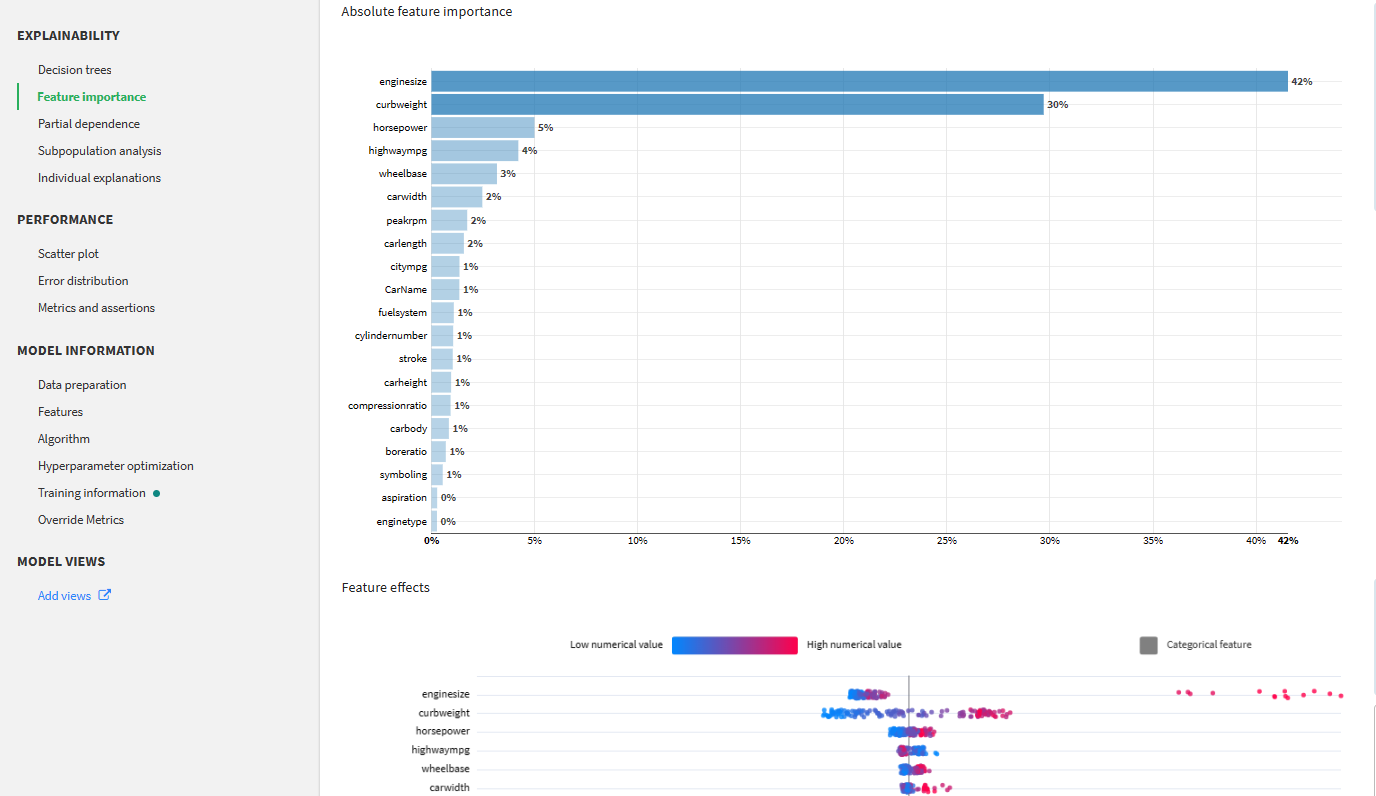
Seçilen algoritma sonuçlarında önemli features lar neler gözlemleyebiliyoruz.

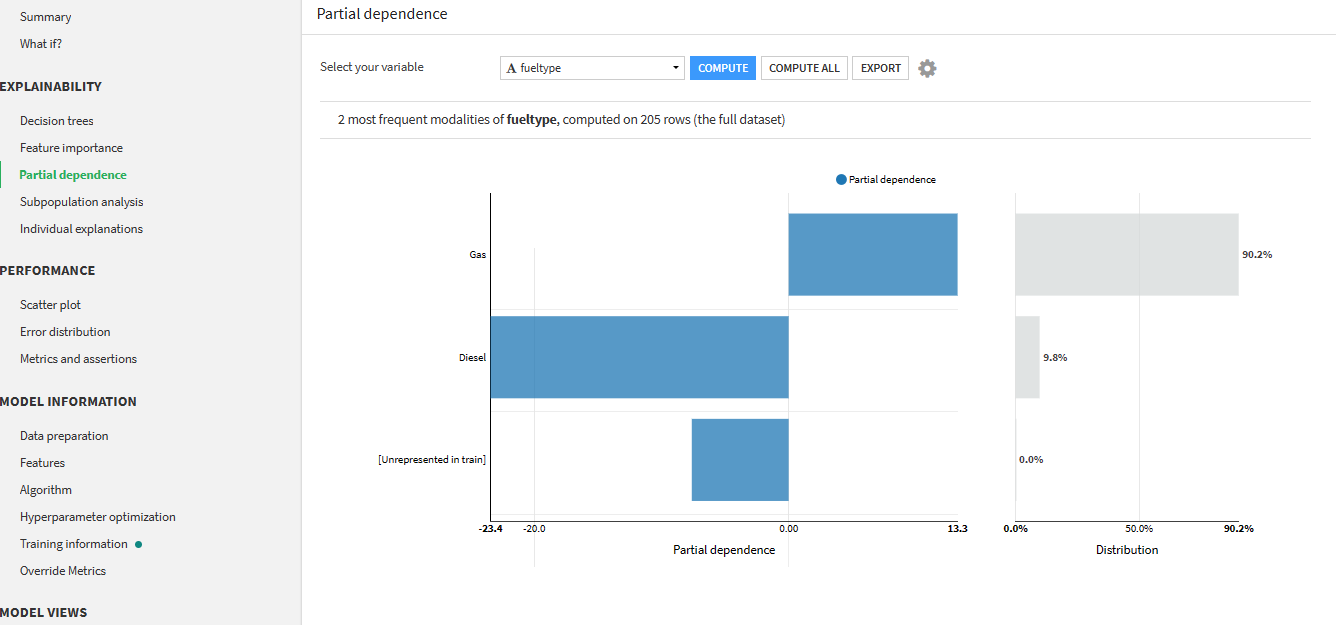


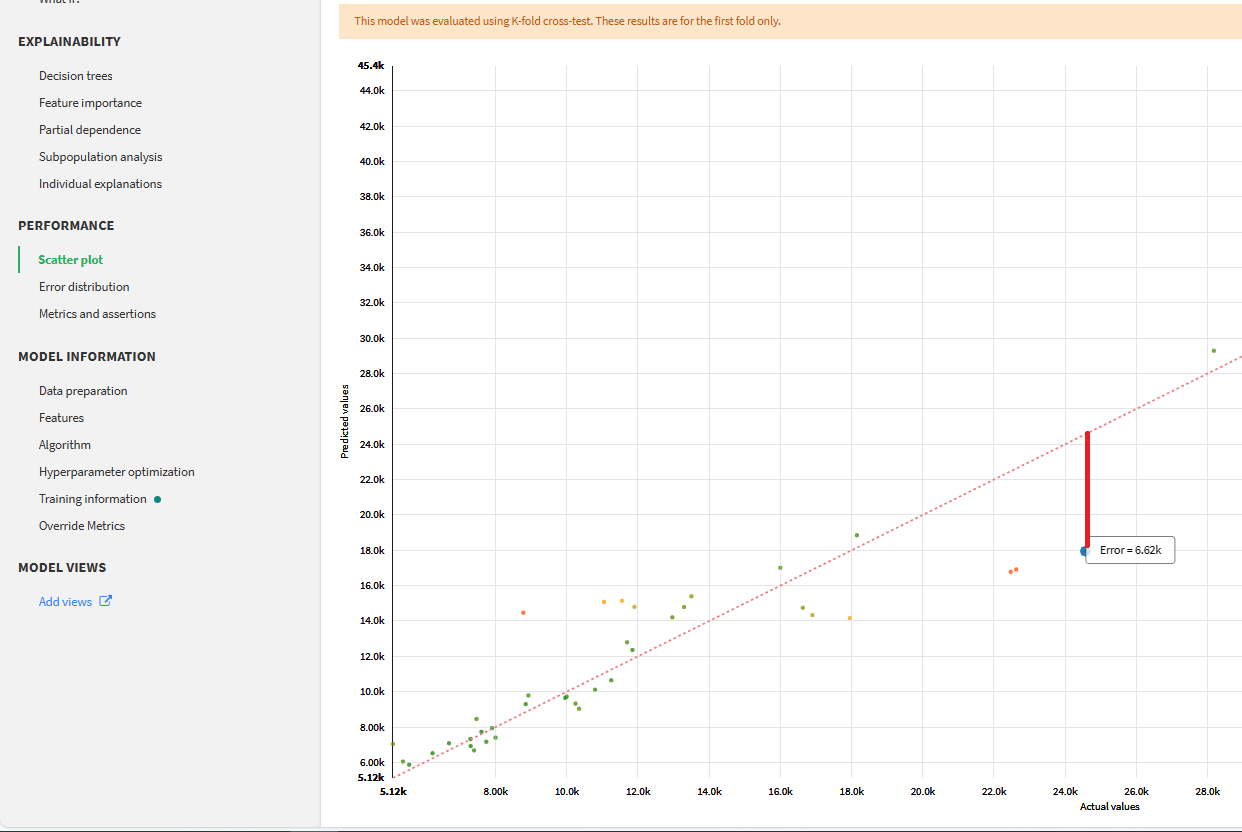
* 1. Özellikleri değiştirdiğimizde ücretin değişimini gözlemleyebiliriz.

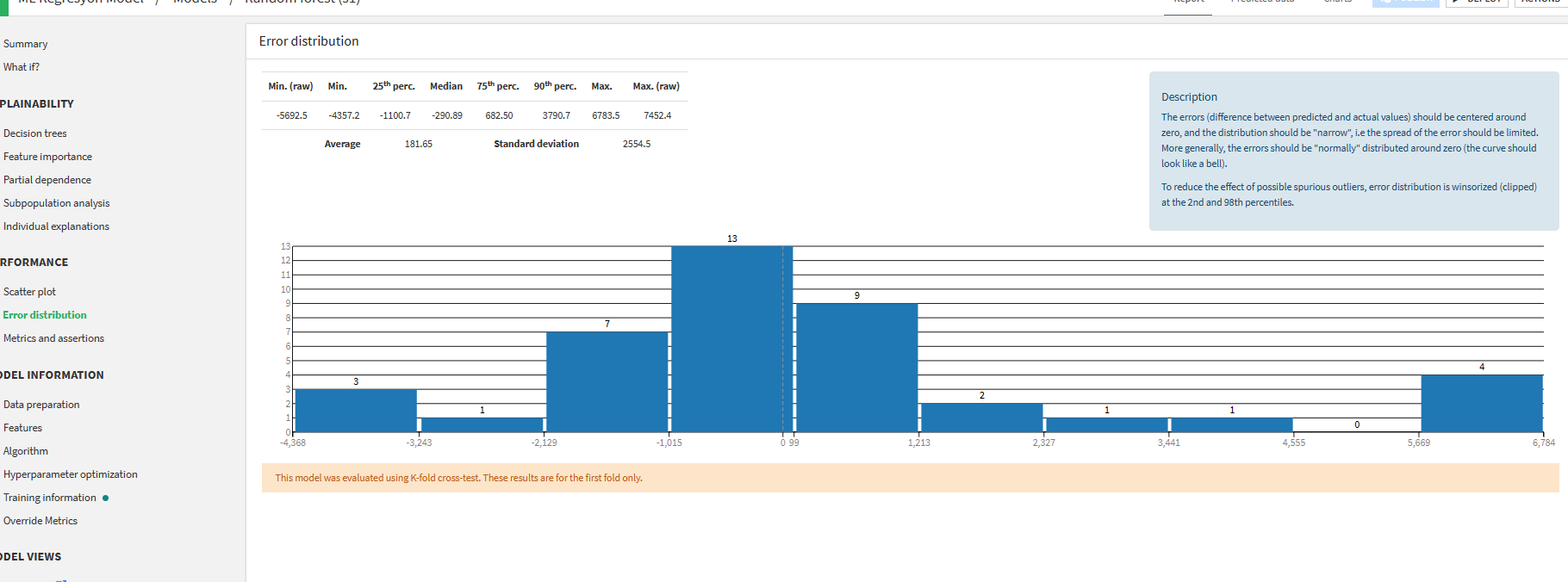


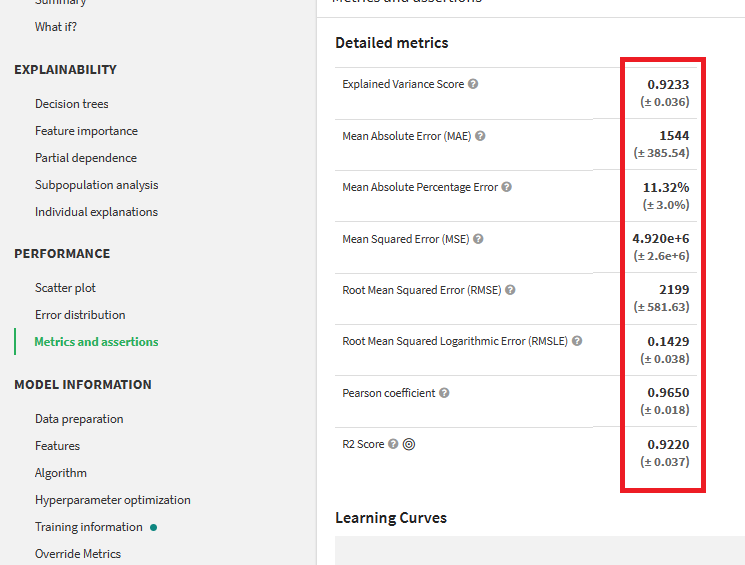


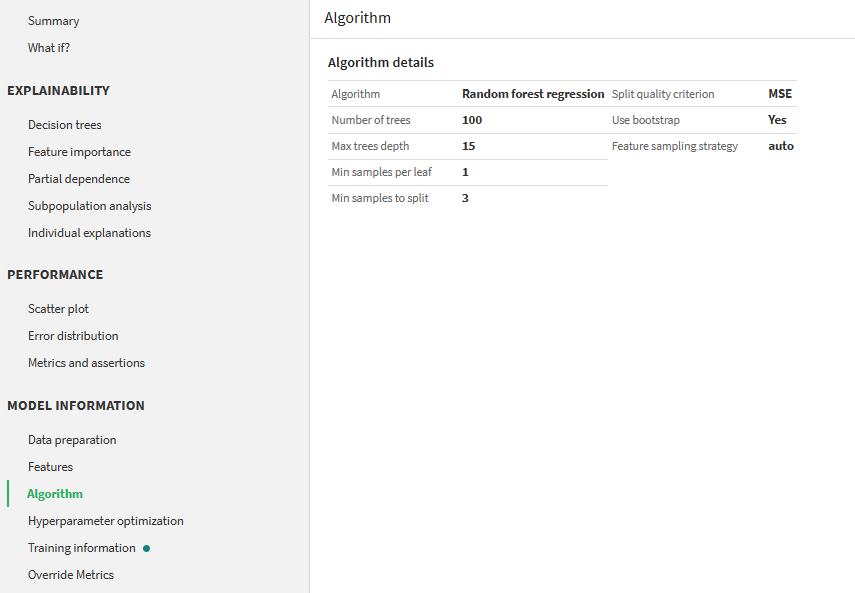


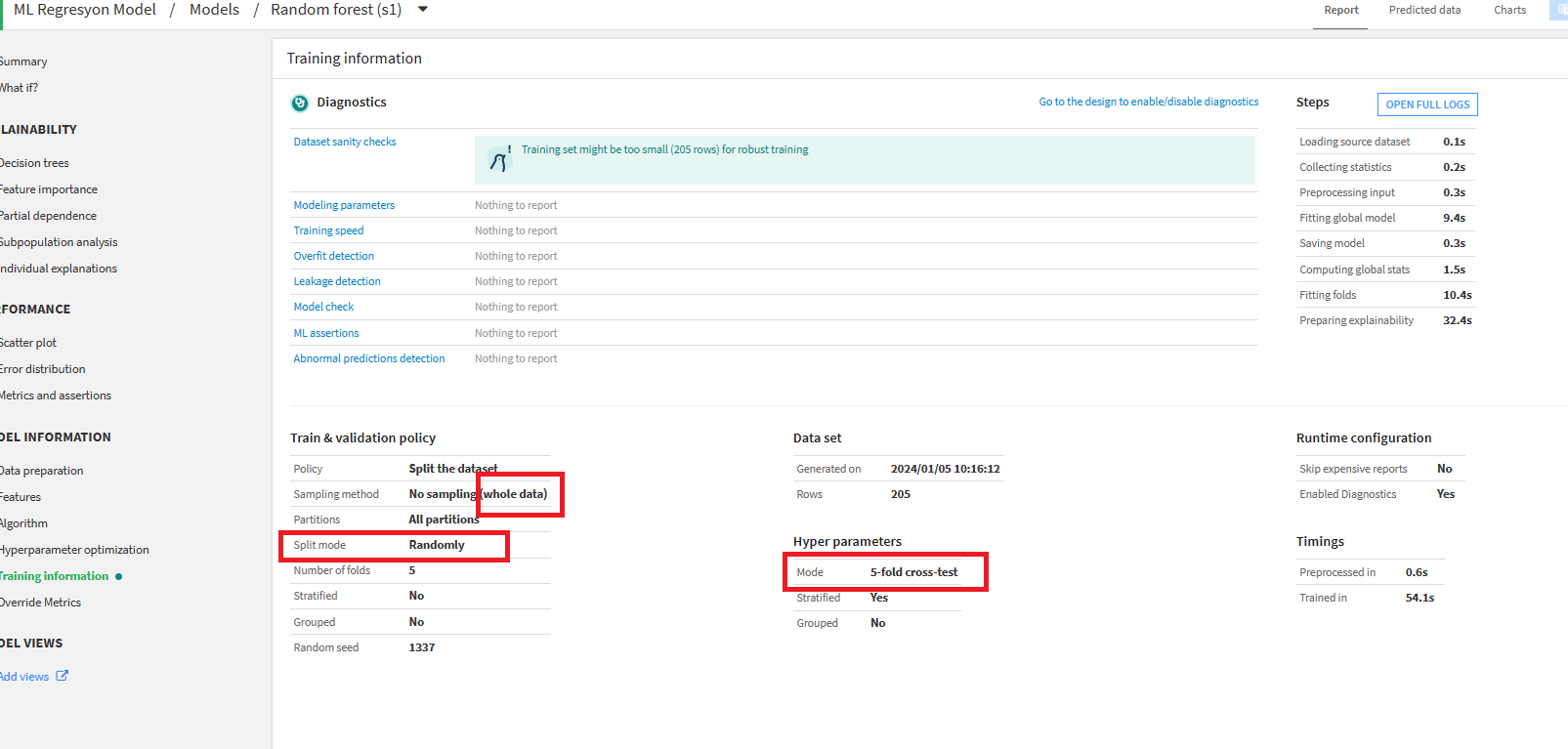












* 1. as
  2. as

1. SONUÇLAR VE YORUMLAR

R2 skoru, bir regresyon modelinin ne kadar iyi uyum sağladığını gösteren bir ölçüdür. Bu skor, 0 ile 1 arasında bir değer alır; 1'e ne kadar yakınsa, modelin veriyi o kadar iyi açıkladığı anlamına gelir. İşte bu R2 skorlarına göre yorumlar:

1. Random Forest (s1):
   * R2 Skoru: 0.922 (± 0.037)
   * Yorum: Oldukça yüksek bir R2 skoru, Random Forest modelinin veri setini çok iyi açıkladığını gösterir. Bu, modelin genel performansının yüksek olduğunu ve veri setindeki varyansın büyük bir kısmını açıkladığını gösterir. ± 0.037 değeri, skorun belirli bir belirsizlikle tahmin edildiğini gösterir. Modelinizin tutarlılık gösterdiği ve performansının genellikle bu aralıkta olduğu anlamına gelir.
2. Ridge (L2) Regression (s1):
   * R2 Skoru: 0.817 (± 0.031)
   * Yorum: İyi bir R2 skoru, Ridge Regresyon modelinin de veri setini iyi açıkladığını gösterir, ancak Random Forest'a kıyasla biraz daha düşüktür. Ridge regresyon, lineer ilişkileri ifade etmek için daha uygundur, bu nedenle veri setinizde daha doğrusal bir yapı varsa başarılı olabilir. ± 0.031 değeri, skorun belirli bir belirsizlikle tahmin edildiğini ve modelin tutarlı bir performans gösterdiğini gösterir.

Sonuç olarak, her iki model de başarılıdır, ancak Random Forest modeli daha yüksek bir R2 skoruna sahiptir, bu da genel performansının biraz daha üst düzeyde olduğunu gösterir. Her iki model de belirtilen belirsizlik aralıklarında tutarlılık gösteriyor gibi görünüyor.