**The Caret Package**

**Library:** Yüklenmiş olan paket programların kütüphanelerini çağıran komuttur.

**Names:** Veri içindeki değişkenlerin adlarını alma ve ayarlama komutudur.

**Attach:** Bir veri setini değerlendirme sırasında veri tabanında R tarafından aranmasını sağlayan komuttur. Attach fonksiyonu kullanılan veri setinde değişkenlerin çağırılma işlemi sadece adları kullanılarak yapılabilmektedir.

**Str:** Veri setindeki değişkenlerin karakterini ve sayısını ve içindeki gözlemleri getirir.

**Summary:** Veri setinin veya içindeki değişkenlerin özet istatistiklerini çağırdığımız komuttur.

Caret paketi (Classification And Regresion Training'in kısaltması), tahmine dayalı modeller oluşturma sürecini kolaylaştırmaya çalışan bir dizi işlevdir. Caret, model oluşturma ve değerlendirme sürecinin yanı sıra özellik seçimi ve diğer teknikleri düzenlemeye çalışan çeşitli işlevlere sahiptir.

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

İlk önce, createDataPartition kullanarak verileri bölmeyi test ediyoruz. Burada verilerin %75'i eğitim setine, %25'i test setine ayrılmış bir veri seti oluşturmak için kullanıyoruz. createDataPartition(), her iki veri kümesinde de sağlanan değişkenin benzer dağılımına sahip bir bölme sağlamaya çalışır.

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

xxx değişkenine eğitim setini ve test setini alt alta kümelemek için kullanabileceğim bir gösterge işlevi atanır.

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

Oluşturduğumuz trainingx içindeki değişkenlerin isimleri ise;

metin içeren bir resim

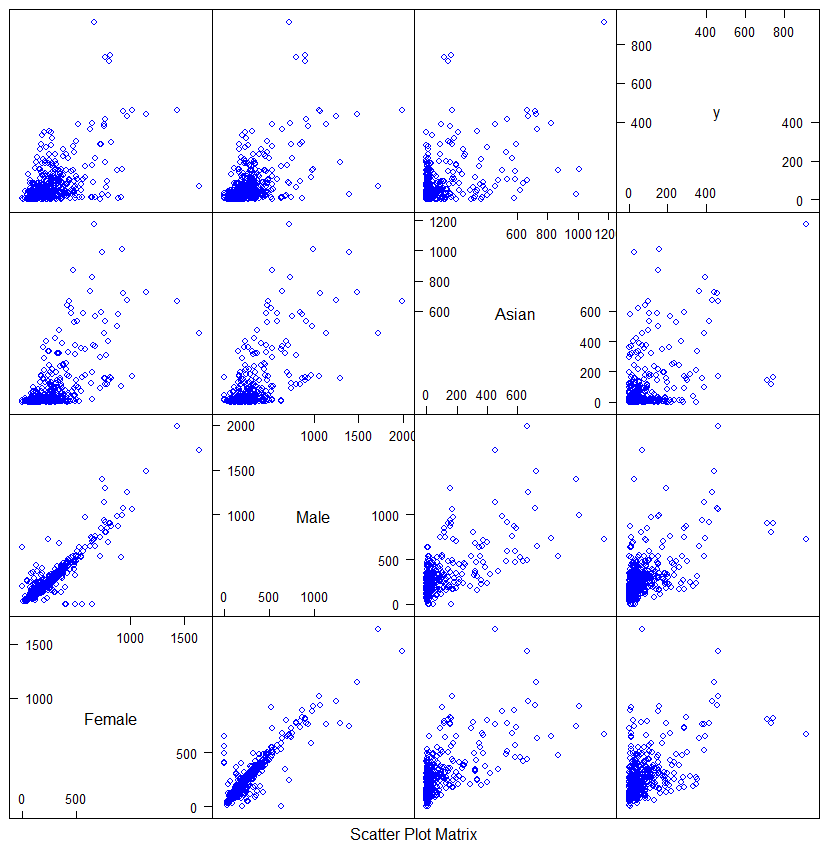
Açıklama otomatik olarak oluşturuldu

şeklindedir.

metin içeren bir resim

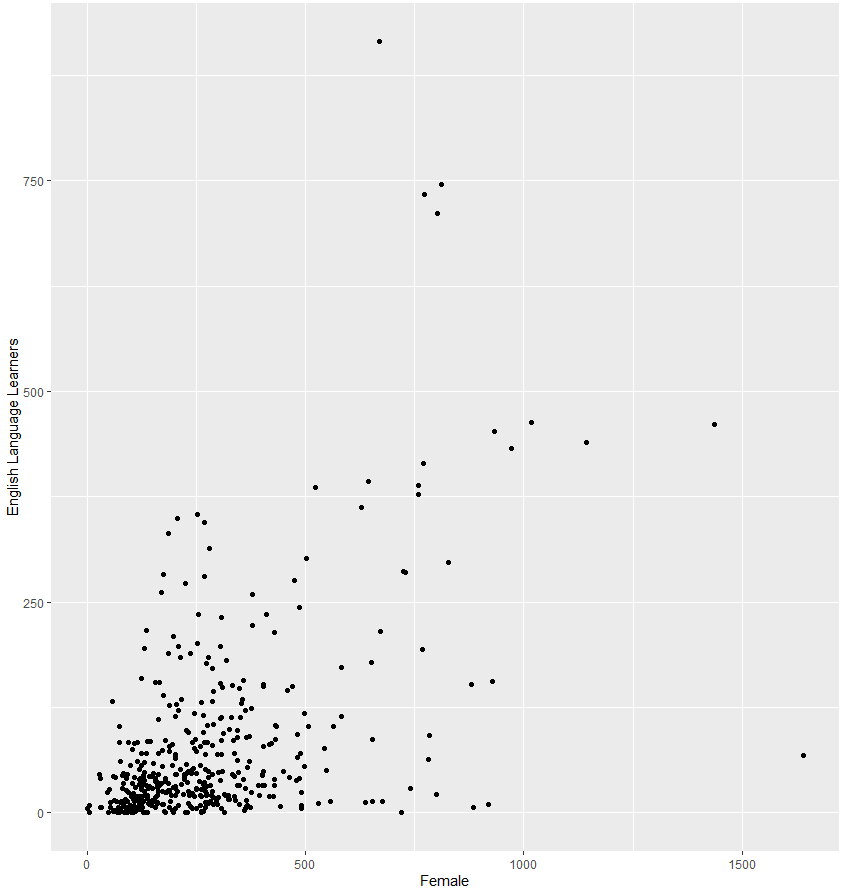
Açıklama otomatik olarak oluşturuldu

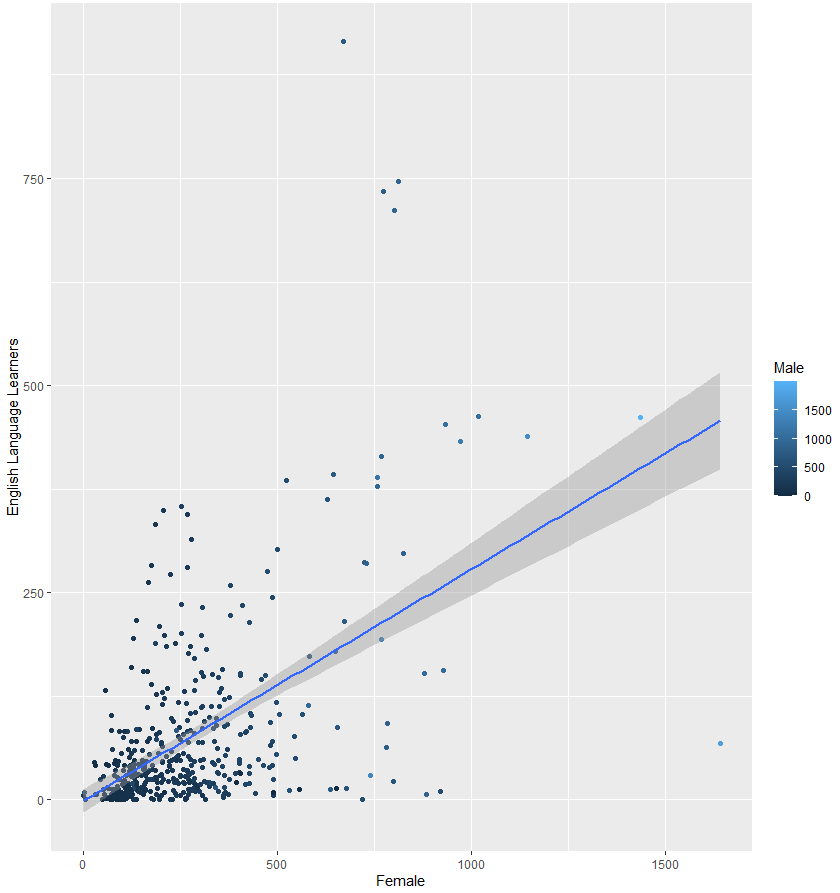
**featurePlot()**, kafes grafiği olarak da bilinir amaç sahip olduğumuz değişkenleri karşılıklı olarak çizerek gösterir. Amaç y değişkenine karşı çizilen değişkenlerin her birine karşılık gelen grafikleri görmektir.



metin içeren bir resim

Açıklama otomatik olarak oluşturuldu





**Ggplot2():** ggplot2, bir veri çerçevesindeki verilerden karmaşık grafikler oluşturmak için yardımcı komutlar sağlayan bir çizim paketidir. Hangi değişkenlerin çizileceğini, nasıl görüntüleneceğini ve genel görsel özellikleri belirlemek için daha programlı bir ara yüz sağlar. Bu nedenle, yalnızca temel veriler değişirse veya bir çubuk grafiğinden dağılım grafiğine geçmeye karar verirsek minimum değişikliklere ihtiyacımız var. Bu, minimum miktarda ayarlama ve ince ayar ile yayın kalitesinde grafikler oluşturmaya yardımcı olur.

**Qplot():** Hızlı çizimin kısaltması qplot, genellikle tek bir satırda diğer çizim sistemlerini kullanan birkaç kod satırı gerektiren karmaşık çizimler üretmeyi kolaylaştırır.

**Geom\_smooth():** Üst çizimin varlığında desenleri görmede göze yardımcı olur.

metin içeren bir resim

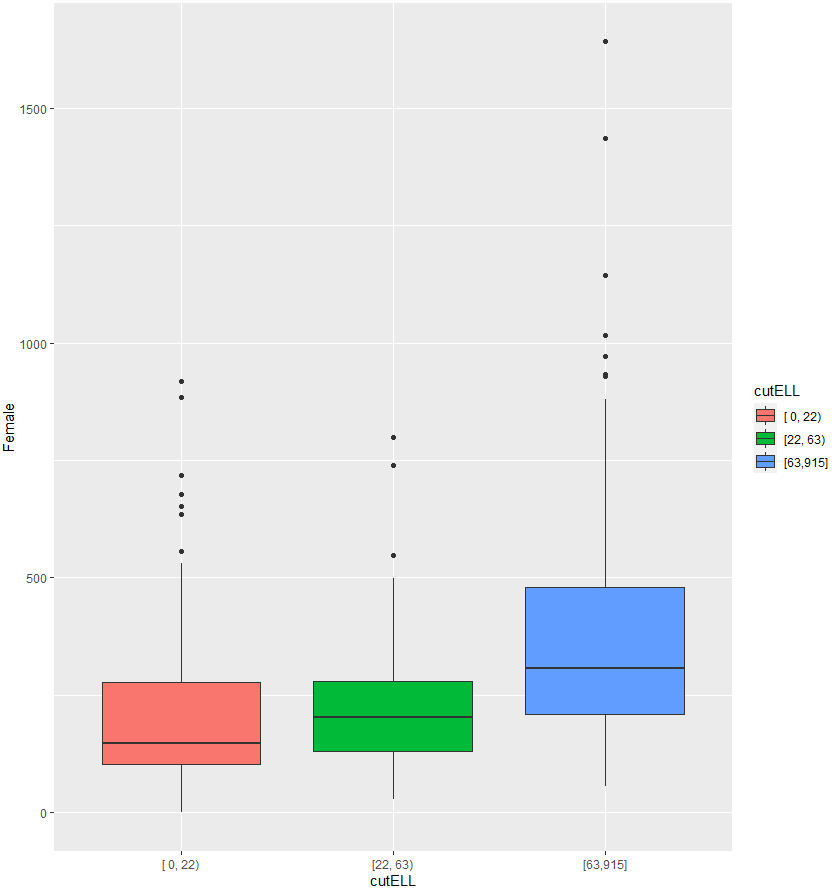
Açıklama otomatik olarak oluşturuldu

**Cut**gibi işlevler ancak sol uç noktalar kapsayıcıdır ve etiketler [alt, üst] biçimindedir, ancak son aralığın [alt, üst] şeklinde olması önemlidir. Kesimler verilirse, varsayılan olarak kesimlerin tüm x aralığını içerdiğinden emin olunmalı fakat kesintiler verilmezse, x'i nicel gruplara veya belirli bir minimum gözlem sayısına sahip gruplara böler. cut bir kategori nesnesi yaratırken, cut2 bir faktör nesnesi yaratır.

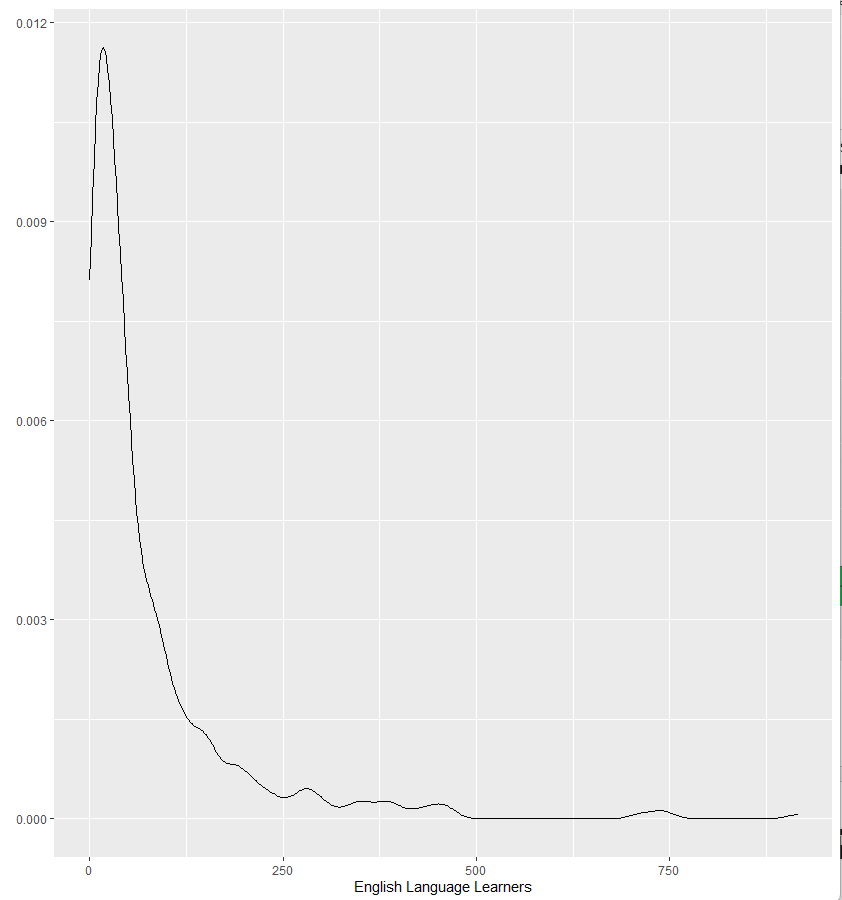
Yukarıda verilen çıktıya göre ‘English Language Learners’ değişkeni 3 farklı gruba ayrılmıştır.

**Hmisc;** veri analizi, üst düzey grafikler, yardımcı işlemler, örnek boyutu ve gücü hesaplama işlevleri, simülasyon, veri kümelerini içe aktarma ve açıklama ekleme, eksik değerleri atama, gelişmiş tablo oluşturma, değişken kümeleme, karakter dizisi işleme, R nesnelerinin dönüştürülmesi için yararlı birçok işlev içerir.



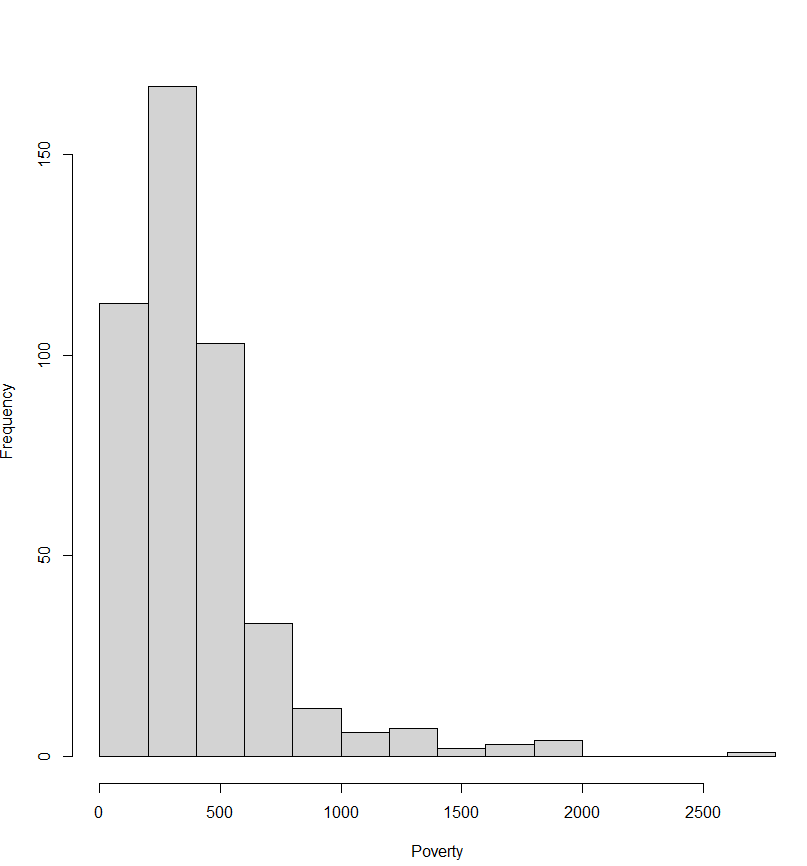






metin içeren bir resim

Açıklama otomatik olarak oluşturuldu



**preProcress()**, eğitim verilerinden tahmin edilebilir ve aynı değişkenlerle herhangi bir veri kümesine uygulanabilir. İşlev x’in ikiden az benzersiz değeri olan herhangi bir değişkene hata atar

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

X: bir matris veya veri çerçevesi

Method: işleme türünü belirten bir karakter vektörü. Olası değerler "center", "scale", "pca" ve "spartialSign"

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

**predict():** R'deki tahmin işlevidir, giriş verilerine dayalı değerleri tahmin etmek için kullanılır.

object: tahminlerin istendiği bir model nesnesi.<br><br>

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

**dummyVars():** dummyVars işlevi, yapay değişkenler oluşturmanıza olanak tanır - başka bir deyişle, metin verilerini modelleme amacıyla sayısal verilere dönüştürür.

**Formula:** Uygun bir R modeli formülü

**Data:** İlgilenilen tahmin edicilere sahip bir veri çerçevesi

**head():** Bir vektörün, matrisin, tablonun, veri çerçevesinin veya işlevin ilk parçalarını almak için kullanılır.

**!!** Yukarıda verilen çıktıya göre 1 değeri ait olduğunuz sınıfı, 0 değeri ise ait olmadığınız sınıf değerini gösterir. Bu nedenle her satırda sadece bir tane 1 değeri vardır.

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

**nearZeroVar():** Benzersiz bir değere sahip (yani sıfır varyanslı tahmin ediciler) tahmin edicileri teşhis eder. checkConditionalX, y düzeylerine göre koşullandırılmış x sütunlarının dağılımına bakar ve y grupları içinde seyrek olan x sütunlarını tanımlar.

**X:** sayısal bir vektör veya matris veya tüm sayısal verileri içeren bir veri çerçevesi

**saveMetrics:** Yanlışsa, sıfır veya sıfıra yakın tahmin edicilerin konumları döndürülür. Doğruysa, tahmin bilgisine sahip bir veri çerçevesi döndürülür.

**NearZeroVar:** saveMetrics = FALSE ise, sorunlu öngörücülerin sütun konumlarına karşılık gelen bir tamsayı vektörü. saveMetrics = TRUE ise, sütunları olan bir veri çerçevesi

**freqRatio:** En yaygın değer için frekansların ikinci en yaygın değere oranı

**percentUnique:** Toplam veri noktası sayısından benzersiz veri noktalarının yüzdesi

**ZeroVar:** Tahmin edicinin yalnızca bir ayrı değere sahip olup olmadığına ilişkin bir mantık vektörü

**nzv:** Öngörücünün sıfıra yakın bir varyans öngörücü olup olmadığına ilişkin bir mantıksal vektör

Verileri böldükten sonra bir dizi modeli eğitmeye başlanır. Basit bir eklemeli lojistik regresyonla;

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

**Train():** Bir dizi sınıflandırma ve regresyon rutini için bir ayar parametreleri ızgarası kurar, her modele uyar ve yeniden örneklemeye dayalı bir performans ölçüsü hesaplar.

**X:** örneklerin satırlarda ve özelliklerin sütunlarda olduğu eğitim verilerini içeren bir veri çerçevesi.

**Data:** Formülde belirtilen değişkenlerin tercihen alınması gereken veri çerçevesi

train, optimal yeniden örnekleme istatistikleriyle ilişkili karmaşıklık parametrelerini seçerek modelleri ayarlamak için kullanılabilir. Belirli bir model için, bir parametreler ızgarası (varsa) oluşturulur ve model, ayar parametrelerinin her bir aday kombinasyonu için biraz farklı veriler üzerinde eğitilir. Her veri setinde, dışarıda tutulan örneklerin performansı hesaplanır ve her kombinasyon için ortalama ve standart sapma özetlenir

Burada, caret paketinden train() işlevine dört argüman sağladık. Bunlar:

**form;** yanıt olarak varsayılan değişkeni belirtir. Ayrıca, mevcut olan tüm tahmin edicilerin kullanılması gerektiğini belirtir.

**data;** eğitimin kullanılan veriye kapalı olacağını belirtir.

**trControl;** 5 kat çapraz doğrulama kullanacağımızı belirtir.

**method:** glm, genelleştirilmiş linear modele uyacağımızı belirtir.

**RMSE:** Karekök ortalama hatası

**Rsquared:** Regresyon modelinden R2

Yöntem esas olarak hem modeli (ve daha özel olarak söz konusu modele R'ye sığdıracak işlevi) hem de kullanılacak paketi belirtir. train() işlevi, esasen seçtiğimiz yöntemin etrafındaki bir sarmalayıcıdır. Bu durumda işlev, temel R işlevi glm()'dir, bu nedenle ek paket gerekmez.

trControl argümanına ilettiğimiz liste, caret'ten trainControl() işlevi kullanılarak oluşturulur. trainControl() işlevi, özellikle yeniden örnekleme şeması olmak üzere, train() tarafından gereken bir dizi eğitim seçeneğini belirtmek için güçlü bir araçtır.

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

**trainControl():** Train işlevinin hesaplama nüanslarını kontrol eder.

**Number**: Ya kat sayısı ya da yeniden örnekleme yineleme sayısı

**classProbs:** mantık; Her yeniden örneklemede sınıf olasılıkları (tahmin edilen değerlerle birlikte) sınıflandırma modelleri için hesaplanmalı mı? Sorusuna cevap verir.

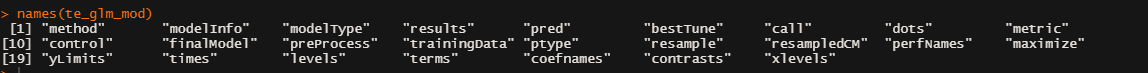
Yukarıda verilen R çıktısında yeniden örneklemenin nasıl yapılacağı ile ilgili bu listenin sadece ilk üç öğesini verilmektedir. Bunlar bizim en çok ilgileneceğimiz üç unsurdur.

**Method**, yeniden örneklemenin nasıl yapılacağını belirtir.

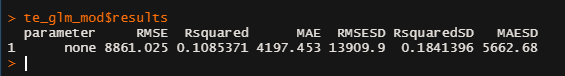
**number,** cv ve boot gibi yeniden örnekleme gerektiren yöntemler için yeniden örneklemenin kaç kez yapılması gerektiğini belirtir.

**repeats,** tekrarlanan cv gibi yöntemler için yeniden örneklemenin tekrarlanma sayısını belirtir

* te\_glm\_model çıktısına göre 38 tahmin edicisi olan 451 gözlem kullanıldığını görüyoruz. Herhangi bir veri ön işlemesi yapılmamıştır ve 5 kat çapraz doğrulama kullanılmıştır.



Train() tarafından döndürülen listede saklanan çok sayıda bilgi olduğunu görüyoruz. Sıklıkla ilgileneceğimiz iki unsur, results ve finalModel'dir.



**RMSESD,** doğruluk tahminindeki belirsizliğin tahminini verir.

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

**RMSE:** Modelin test verisindeki tahmin hatası

**Rsquared:** Modelin test verisi üzerindeki r kare değeri

**MAE:** Modelin mutlak hata değeri

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

**Call:** Modeli kurmak için kullanılan fonksiyonun ismini verir

**Residuals:** Artıkların dağılım istatistiklerini gösterir

Coefficients: katsayıların istatistiklerini gösterir

**Estimate:** veri setindeki her bir bağımsız değişkenin katsayıları

**Std. Error:** tahmin edilen katsayıların standart hataları

**t\_value:** katsayıların t değerleri

**p\_value:** katsayıların p değerleri

**Residual standard error:** Artık standart hata

**Multiple R-squared:** R2 değeri

**F-statistic:** F istatistik ölçümleri