Çoklu Lineer Regresyon Analizi

Mehmet Emin Sahin

2023-11-25

Veri Hakkında

Kaynak: https://www.kaggle.com/datasets/mohansacharya/graduate-admissions

Bu veri seti, mezuniyet notları, TOEFL notları ve diger faktörlerle ögrencilerin yüksek lisansa kabul olasılıklarını içerir. Toplam 400 gözlem ve 9 degisken içerir.

- Bağımlı Değişken: chance of admit (kabul olasılığı)
- Açıklayıcı Degisken 1: GREScore - Mezuniyet notu
- Açıklayıcı Degisken 2: TOEFLScore - TOEFL notu
- Açıklayıcı Degisken 3: UniRating - Üniversite degerlendirmesi (5 üzerinden)
- Açıklayıcı Degisken 4: SOP - Amaç mektubu notu
- Açıklayıcı Degisken 5: CGPA - Agırlıklı ortalama not
- Açıklayıcı Degisken 6: research - Arastırma (tez) yapma durumu
- Açıklayıcı Degisken 7: LOR - Referans mektubu notu

Veri Setinin İçe Aktarılması

```
knitr::opts_chunk$set(echo = TRUE)
desktop_path <- file.path("C:", "Users", "mehmet", "Desktop")
file_name <- "Admission_Predict.csv"
file_path <- file.path(desktop_path, file_name)
data <- read.csv(file_path, stringsAsFactors = FALSE)
```

Gerekli Kütüphanelerin YÜklenmesi

```
knitr::opts_chunk$set(echo = TRUE)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.2.3
```

```r
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.2.3
```

```
## corrplot 0.92 loaded
```

```r
library(gridExtra)
```

```
## Warning: package 'gridExtra' was built under R version 4.2.3
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```r
library(MASS)
```

```
## Warning: package 'MASS' was built under R version 4.2.3
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```r
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.2.3
```

```
## Zorunlu paket yükleniyor: lattice
```

```r
library(car)
```

```
## Warning: package 'car' was built under R version 4.2.3
```

```
## Zorunlu paket yükleniyor: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```r
library(lmtest)
```

```
## Warning: package 'lmtest' was built under R version 4.2.3
```

```
## Zorunlu paket yükleniyor: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

Veri Setinden İlk Sütunun Çıkarılması

- İlk degisken olan "Serial no" değişkeni veri hakkında herhangi bir bilgi vermediği için ve degisken seçimi ve yorumunda kafa karışıklığı yaratmaması için veriden çıkartacagım

```
knitr::opts_chunk$set(echo = TRUE)
data <- data[, -1]
```

Veri Setinin Kısa Özeti

```
knitr::opts_chunk$set(echo = TRUE)
summary(data)
```

```
##    GRE.Score      TOEFL.Score    University.Rating      SOP
##  Min.   :290.0   Min.   : 92.0   Min.   :1.000     Min.   :1.0
##  1st Qu.:308.0   1st Qu.:103.0   1st Qu.:2.000     1st Qu.:2.5
##  Median :317.0   Median :107.0   Median :3.000     Median :3.5
##  Mean   :316.8   Mean   :107.4   Mean   :3.087     Mean   :3.4
##  3rd Qu.:325.0   3rd Qu.:112.0   3rd Qu.:4.000     3rd Qu.:4.0
##  Max.   :340.0   Max.   :120.0   Max.   :5.000     Max.   :5.0
##       LOR             CGPA          Research      Chance.of.Admit
##  Min.   :1.000   Min.   :6.800   Min.   :0.0000   Min.   :0.3400
##  1st Qu.:3.000   1st Qu.:8.170   1st Qu.:0.0000   1st Qu.:0.6400
##  Median :3.500   Median :8.610   Median :1.0000   Median :0.7300
##  Mean   :3.453   Mean   :8.599   Mean   :0.5475   Mean   :0.7244
##  3rd Qu.:4.000   3rd Qu.:9.062   3rd Qu.:1.0000   3rd Qu.:0.8300
##  Max.   :5.000   Max.   :9.920   Max.   :1.0000   Max.   :0.9700
```

```
glimpse(data)
```

```
## Rows: 400
## Columns: 8
## $ GRE.Score         <int> 337, 324, 316, 322, 314, 330, 321, 308, 302, 323, 32~
## $ TOEFL.Score       <int> 118, 107, 104, 110, 103, 115, 109, 101, 102, 108, 10~
## $ University.Rating <int> 4, 4, 3, 3, 2, 5, 3, 2, 1, 3, 3, 4, 4, 3, 3, 3, 3, 3~
## $ SOP               <dbl> 4.5, 4.0, 3.0, 3.5, 2.0, 4.5, 3.0, 3.0, 2.0, 3.5, 3.~
## $ LOR               <dbl> 4.5, 4.5, 3.5, 2.5, 3.0, 3.0, 4.0, 4.0, 1.5, 3.0, 4.~
## $ CGPA              <dbl> 9.65, 8.87, 8.00, 8.67, 8.21, 9.34, 8.20, 7.90, 8.00~
## $ Research          <int> 1, 1, 1, 1, 0, 1, 1, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 1~
## $ Chance.of.Admit   <dbl> 0.92, 0.76, 0.72, 0.80, 0.65, 0.90, 0.75, 0.68, 0.50~
```

Çıkarımlar

- Gre Skoru: En düşük GRE skoru 290, en yüksek 340. Bu, başvuranların çoğunun bu sınavda iyi yaptığını gösteriyor. Ortalama ve ortanca skorlar da yüksek, bu yüzden sınıfın genel olarak GRE'de güçlü olduğunu söyleyebiliriz.
- TOEFL Skoru:TOEFL skorları 92 ile 120 arasında değişiyor. Ortalama ve ortanca skorlar 100'ün üzerinde, yani öğrencilerin genel İngilizce seviyesi iyi gibi görünüyor.
- Üniversite Derecesi: Dereceler 1 ile 5 arasında değişiyor ve çoğu öğrenci orta seviyede bir üniversiteden geliyor, çünkü ortalama derece 3'ün biraz üzerinde.

- Sop: SOP puanları da 1 ile 5 arasında ve çoğunluk 3 ve 4 arasında bir yerde. Yani öğrenciler genellikle SOP yazmayı biliyorlar.
- LOR: LOR puanları da benzer bir şekilde 1 ile 5 arasında ve çoğu öğrenci için ortanca puan 3.5 Bu, öneri mektuplarının genellikle iyi olduğunu gösteriyor.
- CGPA: CGPA 6.8 ile 9.92 arasında değişiyor ve ortalama çok yüksek. Yani sınıftaki öğrencilerin akademik performansları oldukça iyi gibi duruyor.
- Research: Araştırma değişkeni, öğrencinin araştırma deneyimi olup olmadığını gösteriyor ve ortalamaya baktığımızda sınıfın yarısından fazlasının bu deneyime sahip olduğunu görüyoruz.
- Kabul Şansı: Kabul şansı için en düşük puan 0.34, en yüksek puan 0.97 ve çoğu öğrenci 0.7 civarında bir şansa sahip. Bu da başvuranların büyük bir kısmının kabul edilme ihtimalinin oldukça yüksek olduğunu gösterir.

Veri Seti Kayıp Gözlem Kontrolü

```
knitr::opts_chunk$set(echo = TRUE)
head(is.na(data))
```

```
##      GRE.Score TOEFL.Score University.Rating   SOP   LOR  CGPA Research
## [1,]     FALSE       FALSE             FALSE FALSE FALSE FALSE    FALSE
## [2,]     FALSE       FALSE             FALSE FALSE FALSE FALSE    FALSE
## [3,]     FALSE       FALSE             FALSE FALSE FALSE FALSE    FALSE
## [4,]     FALSE       FALSE             FALSE FALSE FALSE FALSE    FALSE
## [5,]     FALSE       FALSE             FALSE FALSE FALSE FALSE    FALSE
## [6,]     FALSE       FALSE             FALSE FALSE FALSE FALSE    FALSE
##      Chance.of.Admit
## [1,]           FALSE
## [2,]           FALSE
## [3,]           FALSE
## [4,]           FALSE
## [5,]           FALSE
## [6,]           FALSE
```

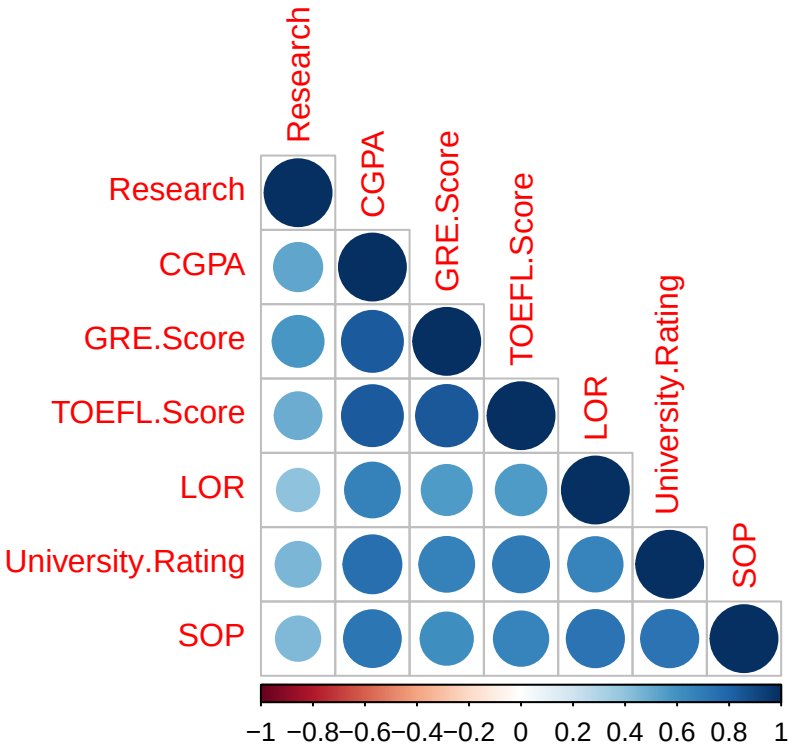
- Veri seti herhangi bir kayıp gözlem içermemektedir.

Modele Geçmeden Önce Veri Görselleştirme İle Değişkenlerin İncelenmesi

Korelasyon Matrisi

- Korelasyon matrisi ile degiskenler arasındaki iliskileri inceleyelim.

```
knitr::opts_chunk$set(echo = TRUE)
cor_matrix <- round(cor(data[,1:7]), 2)
corrplot(cor_matrix, method = "circle", type = "lower", order = "hclust")
```

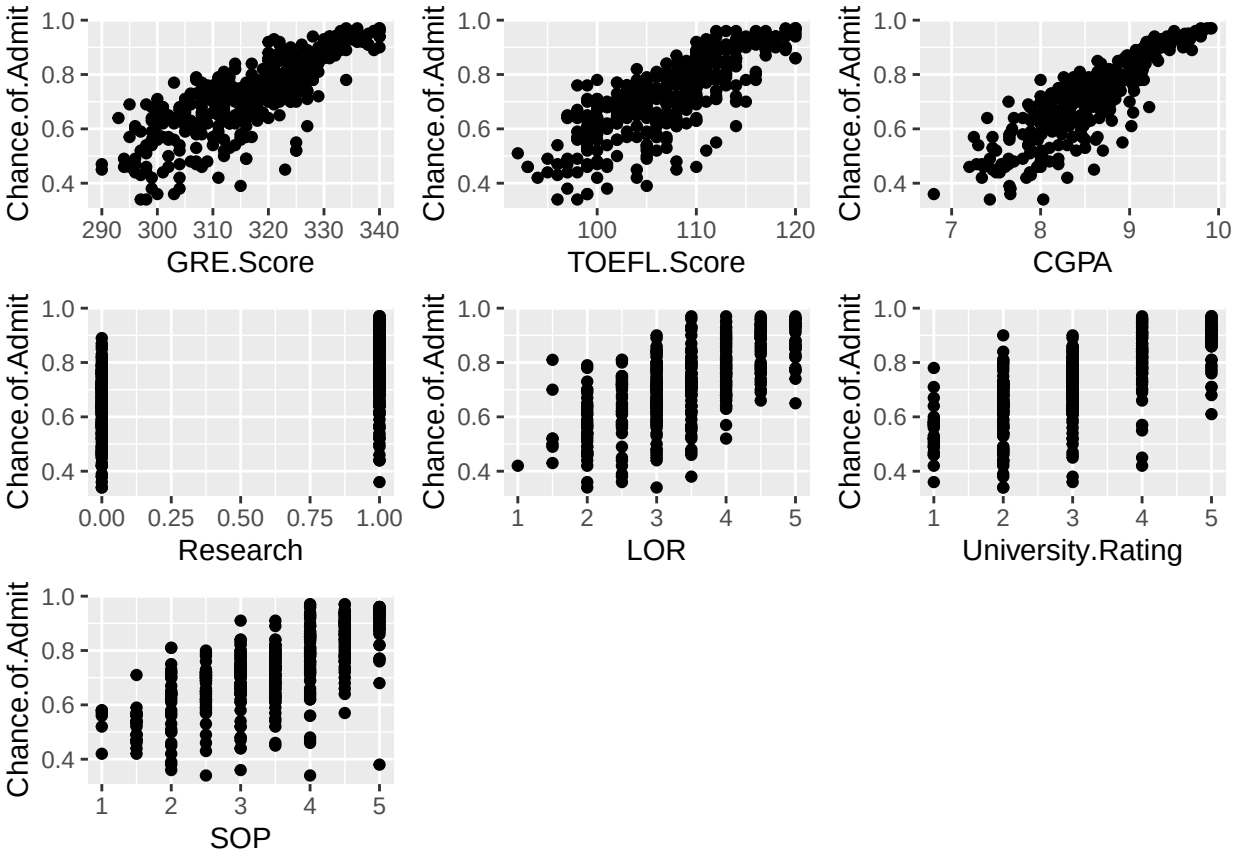


Çıkarımlar

- CGPA ve GRE Score arasındaki korelasyon oldukça yüksek, bu da daha yüksek bir CGPA'nın, daha yüksek bir GRE puanı ile iliskili oldugunu gösterir.
- TOEFL Score da GRE Score ve CGPA ile güçlü pozitif bir iliskiye sahip, bu da dil yeterliliğinin genel akademik yetkinlik ve test performansıyla ilişkili olduğunu gösterir.
- Research değiskeni, diger akademik ölçütlerle karsılaştırıldığında daha düşük bir korelasyona sahip, bu da araştırma deneyiminin diğer akademik ölçütlerle güçlü bir iliskisi olmadığını gösterir.
- Diğer değişkenler (SOP, LOR, University.Rating) akademik puanlar (CGPA, GRE, TOEFL) ile orta derecede pozitif korelasyona sahipken, birbirleri ile de benzer orta seviyede pozitif korelasyonlar sergiliyorlar.

Dağılım Grafikleri

```
knitr::opts_chunk$set(echo = TRUE)
plot1 <- ggplot(data, aes(x = GRE.Score, y = Chance.of.Admit)) + geom_point()
plot2 <- ggplot(data, aes(x = TOEFL.Score, y = Chance.of.Admit)) + geom_point()
plot3 <- ggplot(data, aes(x = CGPA, y = Chance.of.Admit)) + geom_point()
plot4 <- ggplot(data, aes(x = Research, y = Chance.of.Admit)) + geom_point()
plot5 <- ggplot(data, aes(x = LOR, y = Chance.of.Admit)) + geom_point()
plot6 <- ggplot(data, aes(x = University.Rating, y = Chance.of.Admit)) + geom_point()
plot7 <- ggplot(data, aes(x = SOP, y = Chance.of.Admit)) + geom_point()
grid.arrange(plot1, plot2, plot3, plot4, plot5, plot6, plot7, ncol = 3)
```



Çıkarımlar

- GRE skoru ile kabul sansı arasında pozitif bir iliski görünüyor. GRE skoru arttıkça kabul sansının da arttığı görülüyor.
- TOEFL skoru ile kabul sansı arasında da pozitif bir iliski var. Yüksek TOEFL skorları, daha yüksek kabul şanslarıyla ilişkili görünüyor. Bu ilişki de doğrusal gibi görünüyor.
- CGPA ile kabul sansı arasındaki ilişki oldukça güçlü ve pozitif. Bu, CGPA'nın kabul şansını önemli ölçüde etkilediğini ve yüksek CGPA'ların genellikle daha yüksek kabul şanslarıyla iliskili olduğunu gösteriyor.
- Research değişkeni etkisi diger niceliksel değişkenler kadar açık değil.
- Hem LOR hem de SOP için, derecelendirme arttıkça kabul sansının arttıgı görülüyor, ancak bu artıs diğer değiskenler kadar güçlü değil.
- Daha yüksek dereceli üniversiteler için kabul şansının genellikle daha yüksek olduğu görülüyor.

Veri Setinin Eğitim ve Test Olarak Ayırma

```
knitr::opts_chunk$set(echo = TRUE)
set.seed(53)
train_index <- sample(1:nrow(data), size = floor(0.8 * nrow(data)))
```

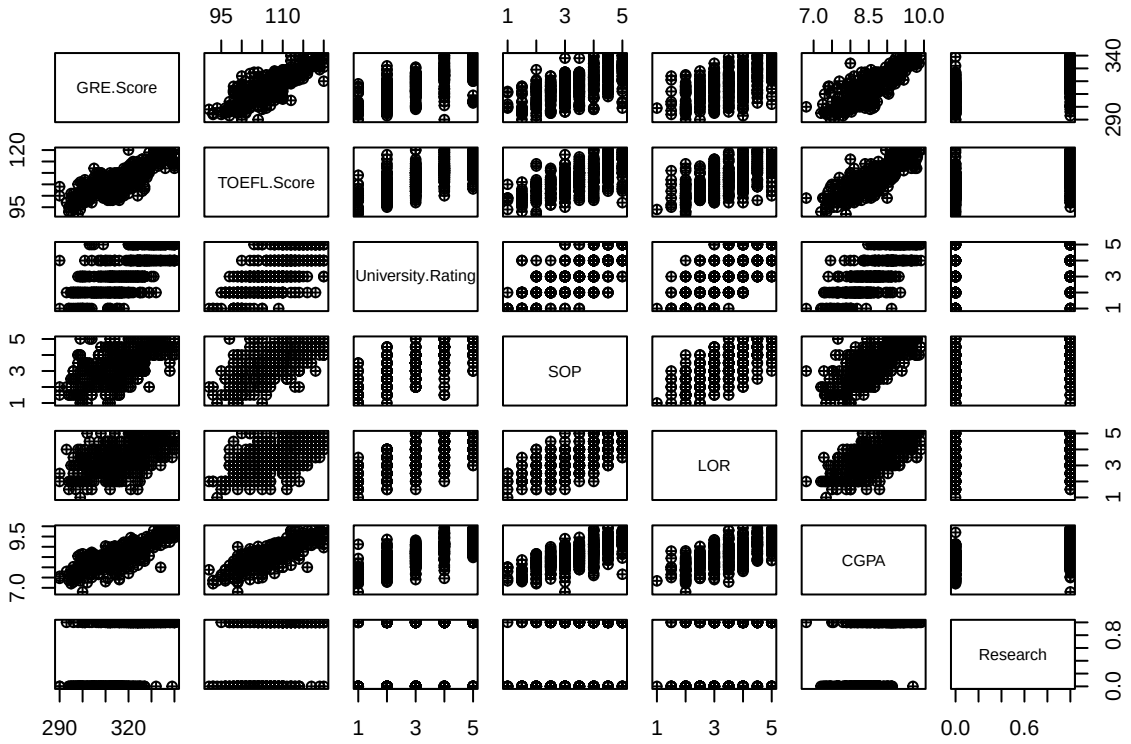
```r
train_data <- data[train_index, ]
test_data <- data[-train_index, ]
```

Sayısal Olarak Korelasyon Matrisi Ve Görselleştirilmesi

```r
knitr::opts_chunk$set(echo = TRUE)
data_names <- data[, c("GRE.Score", "TOEFL.Score", "University.Rating", "SOP", "LOR", "CGPA", "Research
correlation_matrix <- cor(data_names)
print(correlation_matrix)
```

```
##                   GRE.Score TOEFL.Score University.Rating       SOP       LOR
## GRE.Score         1.0000000   0.8359768         0.6689759 0.6128307 0.5575545
## TOEFL.Score       0.8359768   1.0000000         0.6955898 0.6579805 0.5677209
## University.Rating 0.6689759   0.6955898         1.0000000 0.7345228 0.6601235
## SOP               0.6128307   0.6579805         0.7345228 1.0000000 0.7295925
## LOR               0.5575545   0.5677209         0.6601235 0.7295925 1.0000000
## CGPA              0.8330605   0.8284174         0.7464787 0.7181440 0.6702113
## Research          0.5803906   0.4898579         0.4477825 0.4440288 0.3968593
##                       CGPA  Research
## GRE.Score         0.8330605 0.5803906
## TOEFL.Score       0.8284174 0.4898579
## University.Rating 0.7464787 0.4477825
## SOP               0.7181440 0.4440288
## LOR               0.6702113 0.3968593
## CGPA              1.0000000 0.5216542
## Research          0.5216542 1.0000000
```

```r
pairs(data_names, pch = 10)
```



Çıkarımlar

- Korelasyon matrisinde önceden yaptığım çıkarımlara ek olarak şunu söyleyebilirim: Sonuçta, bu korelasyonlar bağımsız değişkenlerin birbiriyle ilişkili olduğunu göstermektedir. Bu durum, çoklu doğrusal bağlantı (multicollinearity) sorununa işaret etmektedir.

Model Oluşturma

```
knitr::opts_chunk$set(echo = TRUE)
model <- lm(Chance.of.Admit ~ ., data = train_data)
summary(model)
```

```
##
## Call:
## lm(formula = Chance.of.Admit ~ ., data = train_data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.260789 -0.024256  0.008853  0.035849  0.158602
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -1.2012949  0.1313772  -9.144  < 2e-16 ***
```

```
## GRE.Score          0.0012915  0.0006344   2.036  0.04260 *
## TOEFL.Score        0.0029216  0.0011662   2.505  0.01274 *
## University.Rating  0.0012980  0.0051713   0.251  0.80198
## SOP                0.0004325  0.0058748   0.074  0.94136
## LOR                0.0243917  0.0058156   4.194 3.57e-05 ***
## CGPA               0.1280001  0.0131977   9.699  < 2e-16 ***
## Research           0.0259691  0.0083962   3.093  0.00216 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06129 on 312 degrees of freedom
## Multiple R-squared:  0.8158, Adjusted R-squared:  0.8117
## F-statistic: 197.5 on 7 and 312 DF,  p-value: < 2.2e-16
```

Çıkarımlar

- Modelin Genel Uyumu R-kare değeri yaklaşık %81.6, yani model, kabul şansındaki varyansın %81.6'sını açıklıyor. Bu oldukça iyi bir uyum anlamına geliyor, yani modelimiz, öğrencilerin kabul şansını oldukça iyi tahmin edebiliyor.
- Modelde GRE.Score, TOEFL.Score, LOR, CGPA ve Research değişkenlerinin p-değerleri düşük, yani bu değişkenlerin kabul şansı üzerinde anlamlı bir etkisi var. Diğer taraftan University.Rating ve SOP değişkenlerinin p-değerleri yüksek, yani bu değişkenlerin etkisi anlamlı değil.
- F-İstatistiği Modelin genel anlamlılığını test eder. Burada çok düşük bir p-değeri ($< 2.2e\text{-}16$) ile çok yüksek bir F-istatistiği var, bu da modelin değişkenlerinin kolektif olarak kabul şansını anlamlı bir şekilde tahmin ettiğini gösteriyor.
- Sonuç olarak modelimize göre, not ortalaması (CGPA) ve araştırma yapmış olmak çok önemli. GRE ve TOEFL sınav sonuçları da kabul şansını etkiliyor ama not ortalaması daha önemli bir etkiye sahip gözüküyor. Diğer taraftan, üniversitenin derecesi ve SOP pek bir etkisi yok gibi. Yani master başvurusu yapacak olanlar, özellikle notlarını ve araştırma deneyimlerini ön plana çıkarmalılar diyebiliriz.

Eğitim Veri Setindeki Sayısal Değişkenler için Shapiro-Wilk Normallik Testi

```r
knitr::opts_chunk$set(echo = TRUE)
shapiro_test_results <- lapply(train_data[, sapply(train_data, is.numeric)], shapiro.test)
shapiro_test_results
```

```
## $GRE.Score
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.98633, p-value = 0.003995
##
##
## $TOEFL.Score
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
```

```
## W = 0.98439, p-value = 0.001516
##
##
## $University.Rating
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.90414, p-value = 2.277e-13
##
##
## $SOP
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.95482, p-value = 2.293e-08
##
##
## $LOR
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.95798, p-value = 5.928e-08
##
##
## $CGPA
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.99118, p-value = 0.0525
##
##
## $Research
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.63188, p-value < 2.2e-16
##
##
## $Chance.of.Admit
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.97693, p-value = 5.144e-05
```

Çıkarımlar

- GRE Skoru: p-değeri 0.003995, yani %0.4. Bu değer 0.05'ten küçük, dolayısıyla GRE skorlarının normal dağılım göstermediğini söyleyebiliriz.
- TOEFL Skoru: p-değeri 0.001516, yani %0.15. Bu da TOEFL skorlarının normal dağılım göstermediğini gösteriyor.
- Üniversite Derecesi: p-değeri çok küçük, neredeyse sıfır. Bu, üniversite derecelerinin normal dağılmadığını gösteriyor.
- SOP: p-değeri yine çok düşük. Bu da SOP değerlerinin normal dağılmadığını gösteriyor.
- LOR: p-değeri çok düşük, normal dağılım göstermediğini gösteriyor.
- CGPA: p-değeri 0.0525, yani %5.25. Bu değer 0.05'e çok yakın, bu nedenle CGPA'nın neredeyse normal dağılım gösterdiğini söyleyebiliriz ama tam olarak diyemeyiz.
- Araştırma: p-değeri çok küçük, bu da araştırma değişkeninin kesinlikle normal dağılım göstermediğini gösteriyor.
- Kabul Şansı: p-değeri kabul şansının da normal dağılım göstermediğini gösteriyor.
- Sonuç olarak optimum modeli kurmaya çalışırken bunlara göre hareket edeceğim.

Modelin Performansı

```
knitr::opts_chunk$set(echo = TRUE)
train_predictions <- predict(model, train_data)
train_actuals <- train_data$Chance.of.Admit
train_mse <- mean((train_predictions - train_actuals) ** 2)
train_rmse <- sqrt(train_mse)
train_r_squared <- summary(model)$r.squared


test_predictions <- predict(model, test_data)
test_actuals <- test_data$Chance.of.Admit
test_mse <- mean((test_predictions - test_actuals) ** 2)
test_rmse <- sqrt(test_mse)

cat("Egitim Seti Performansı:\n")
```

```
## Egitim Seti Performansı:
```

```
cat("MSE:", train_mse, "\n")
```

```
## MSE: 0.003662268
```

```
cat("RMSE:", train_rmse, "\n")
```

```
## RMSE: 0.06051668
```

```
cat("R-Squared:", train_r_squared, "\n\n")
```

```
## R-Squared: 0.8158445
```

```
cat("Test Seti Performansı:\n")
```

```
## Test Seti Performansı:
```

```r
cat("MSE:", test_mse, "\n")
```

```
## MSE: 0.005428367
```

```r
cat("MSE:", test_mse, "\n")
```

```
## MSE: 0.005428367
```

```r
cat("RMSE:", test_rmse, "\n")
```

```
## RMSE: 0.07367745
```

Çıkarımlar

- Eğitim Seti Performansı:
- MSE (Eğitim Seti): 0.003662268. MSE, modelin hatalarının karesinin ortalamasıdır. Düşük bir MSE değeri, modelin tahminlerinin gerçek değerlere yakın olduğunu gösterir. Eğitim setindeki MSE değeri oldukça düşük, bu da modelin eğitim verilerine iyi uyduğunu gösteriyor.
- RMSE (Eğitim Seti): 0.06051668. RMSE, MSE'nin kareköküdür ve hataların ölçeğini daha iyi anlamamıza yardımcı olur. RMSE de düşük, bu da modelin eğitim setindeki performansının iyi olduğunu gösteriyor.
- R-kare (Eğitim Seti): 0.8158445. Bu, modelin eğitim setindeki bağımlı değişkenin varyansının %81.58'ini açıklayabildiğini gösteriyor. Yani model, verilerin çoğunu iyi bir şekilde yakalıyor.
- Test Seti Performansı:
- MSE (Test Seti): 0.005428367. Test setindeki MSE, eğitim setine göre biraz daha yüksek. Bu, modelin test verilerinde biraz daha fazla hata yaptığını gösteriyor.
- RMSE (Test Seti): 0.07367745. Test setindeki RMSE de eğitim setine kıyasla daha yüksek. Bu, modelin test verilerindeki tahminlerinin biraz daha az doğru olduğunu gösteriyor.
- R-kare (Test Seti): 0.7519239. Test setindeki R-kare değeri eğitim setine göre daha düşük, bu da modelin bağımsız verilerde biraz daha az etkili olduğunu gösteriyor. Ancak yine de %75.19 gibi kabul edilebilir bir seviyede.
- Genel Yorum: Modeliniz eğitim setinde oldukça iyi performans gösteriyor ve test setinde de kabul edilebilir bir performans sergiliyor gibi duruyor. Ancak, eğitim ve test setleri arasında bir miktar performans farkı var.Bu, modelin biraz aşırı uyum yapmış olabileceğine işaret edebilir
- !!! Hiçbir aykırı değer taraması yapmadan veya değişkenler üzerinde herhangi bir düzenleme yapmadan oluşturduğum modeli ve sonuçlarını inceledim ve yorumladım. Şimdi bu bilgilere dayanarak optimum bir model kurmaya çalışacağım.

Optimum Model Oluşturma

Eğitim ve Test Olarak Yeniden Ayırma

```r
knitr::opts_chunk$set(echo = TRUE)
set.seed(53)
train_index_1 <- sample(1:nrow(data), size = floor(0.8 * nrow(data)))
train_data_1 <- data[train_index_1, ]
test_data_1 <- data[-train_index_1, ]
```

Mahalanobis Aykırı Değer Kontrolü

```
knitr::opts_chunk$set(echo = TRUE)
# Mahalanobis mesafesini hesaplama fonksiyonu
mahalanobis_distance <- function(data, cov_matrix, center) {
  mahalanobis(data, center = center, cov = cov_matrix)
}

# Egitim veri seti icin kovaryans matrisi ve ortalamalari hesapla
cov_matrix <- cov(train_data_1[, sapply(train_data_1, is.numeric)])
center <- colMeans(train_data_1[, sapply(train_data_1, is.numeric)])

# Her gozlem icin Mahalanobis mesafesini hesapla
train_data_1$mahalanobis <- mahalanobis_distance(train_data_1[, sapply(train_data_1, is.numeric)], cov_

# Aykiri degerleri tanimla
threshold <- qchisq(0.95, ncol(train_data_1) - 1)
outliers <- which(train_data_1$mahalanobis > threshold)

# Aykiri degerleri cikar
train_data_clean_1 <- train_data_1[-outliers, ]

# University.Rating ve SOP degişkenlerini cikar
# train_data_final_1 <- train_data_clean_1 %>% select(-University.Rating, -SOP)
# test_data_final_1 <- test_data_1 %>% select(-University.Rating, -SOP)

train_data_final_1 <- train_data_clean_1[, !colnames(train_data_clean_1) %in% c("University.Rating", "S
test_data_final_1 <- test_data_1[, !colnames(test_data_1) %in% c("University.Rating", "SOP")]


# Son durumu kontrol et
str(train_data_final_1)
```

```
## 'data.frame':    294 obs. of  7 variables:
##  $ GRE.Score      : int  328 314 334 327 336 326 317 326 316 299 ...
##  $ TOEFL.Score    : int  115 107 117 109 112 111 107 113 106 100 ...
##  $ LOR            : num  4 4 4.5 4 5 4 3 4 4 3.5 ...
##  $ CGPA           : num  9.16 8.27 9.07 8.77 9.76 9.23 8.28 9.4 8.32 7.88 ...
##  $ Research       : int  1 0 1 1 1 1 0 1 0 0 ...
##  $ Chance.of.Admit: num  0.78 0.72 0.89 0.79 0.96 0.88 0.66 0.91 0.72 0.68 ...
##  $ mahalanobis    : num  3.9 7.81 7.03 2.99 7.88 ...
```

```
str(test_data_final_1)
```

```
## 'data.frame':    80 obs. of  6 variables:
##  $ GRE.Score      : int  322 314 321 327 328 318 328 300 338 316 ...
##  $ TOEFL.Score    : int  110 103 109 111 112 110 116 97 118 105 ...
##  $ LOR            : num  2.5 3 4 4.5 4.5 3 5 3 4.5 2.5 ...
##  $ CGPA           : num  8.67 8.21 8.2 9 9.1 8.8 9.5 8.1 9.4 8.2 ...
##  $ Research       : int  1 0 1 1 1 0 1 1 1 1 ...
##  $ Chance.of.Admit: num  0.8 0.65 0.75 0.84 0.78 0.63 0.94 0.65 0.91 0.49 ...
```

Gerekli Değişkenler İçin Log Dönüşümü

```
knitr::opts_chunk$set(echo = TRUE)
train_data_final_1$log_GRE <- log(train_data_final_1$GRE.Score)
train_data_final_1$log_TOEFL <- log(train_data_final_1$TOEFL.Score)
test_data_final_1$log_GRE <- log(test_data_final_1$GRE.Score)
test_data_final_1$log_TOEFL <- log(test_data_final_1$TOEFL.Score)
```

Eğitim Seti Üzerinden Model Kurma ve Değerlendirme

```
knitr::opts_chunk$set(echo = TRUE)
model2 <- lm(Chance.of.Admit ~ log_GRE + log_TOEFL + CGPA + Research + LOR, data = train_data_final_1)
summary(model2)
```

```
##
## Call:
## lm(formula = Chance.of.Admit ~ log_GRE + log_TOEFL + CGPA + Research +
##     LOR, data = train_data_final_1)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.210407 -0.024249  0.005489  0.029540  0.116691
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.388313   0.892462  -6.038 4.81e-09 ***
## log_GRE      0.563115   0.182414   3.087  0.00222 **
## log_TOEFL    0.379109   0.117814   3.218  0.00144 **
## CGPA         0.117379   0.012070   9.725  < 2e-16 ***
## Research     0.014588   0.007477   1.951  0.05202 .
## LOR          0.025788   0.004455   5.788 1.86e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05121 on 288 degrees of freedom
## Multiple R-squared:  0.8633, Adjusted R-squared:  0.8609
## F-statistic: 363.7 on 5 and 288 DF,  p-value: < 2.2e-16
```

Çıkarımlar

- Modelin F-istatistiği oldukça yüksek (363.7) ve ilgili p-değeri çok küçük (2.2e-16'dan küçük). Bu, modelin anlamlı olduğunu, yani bağımsız değişkenlerin setinin, bağımlı değişkenin varyansını anlamlı bir şekilde açıkladığını gösterir.
- Modelin R-kare değeri 0.8633'tür, yani model, bağımlı değişkenin varyansının yaklaşık %86.33'ünü açıklıyor. Adjusted R-squared değeri ise %86.09, bu da modelin bağımsız değişken sayısına göre düzeltilmiş bir açıklama gücüne sahip olduğunu gösterir.
- Modelin sabiti negatif ve anlamlıdır.
- GRE puanlarının logaritması, kabul şansı üzerinde pozitif ve anlamlı bir etkiye sahiptir. GRE puanı yükseldikçe, kabul şansı artmaktadır.

- TOEFL puanlarının logaritması da kabul şansı üzerinde pozitif ve anlamlı bir etkiye sahiptir. TOEFL puanı yükseldikçe, kabul şansı artmaktadır.
- CGPA'nın katsayısı pozitif ve oldukça anlamlıdır, yani CGPA'daki artışlar kabul şansını önemli ölçüde artırmaktadır.
- Research, kabul şansını pozitif yönde etkilemektedir.
- LOR kalitesi de kabul şansını pozitif ve anlamlı bir şekilde etkilemektedir.
- Modelin hatalarının standart sapması 0.05121'dir. Bu, modelin tahminlerinin gerçek değerlerle olan ortalama sapmasını gösterir ve düşük bir değerdir, bu da modelin iyi bir uyum sağladığını gösterir.
- Modelin hataları için min, 1Q , medyan, 3Q ve max değerler, residualların dağılımı hakkında bilgi verir. Bu değerler oldukça sınırlı bir aralıkta toplanmıştır (-0.21'den 0.116'ya), bu da hataların büyük olmadığını ve modelin tahminlerinin çoğu zaman gerçek değerlere yakın olduğunu gösterir.
- Ancak, her zaman olduğu gibi, modeli test veri seti üzerinde de değerlendirmeleyiz

Test Seti Üzerinde Tahmin Yapma ve Değerlendirme

```
knitr::opts_chunk$set(echo = TRUE)
# Test setinde tahmin yapma
test_predictions <- predict(model2, newdata = test_data_final_1)

# Gerçek degerlerle tahminleri karsilastirma ve performans metriklerini hesaplama
test_data_final_1$predicted_Chance.of.Admit <- test_predictions
mean_squared_error <- mean((test_data_final_1$Chance.of.Admit - test_predictions)^2)
root_mean_squared_error <- sqrt(mean_squared_error)

# Performans metrikleri
print(paste("MSE:", mean_squared_error))
```

```
## [1] "MSE: 0.00550305140002735"
```

```
print(paste("RMSE:", root_mean_squared_error))
```

```
## [1] "RMSE: 0.0741825545531249"
```

```
# AIC ve BIC degerlerini hesapla
model_aic <- AIC(model)
model_bic <- BIC(model)

# AIC ve BIC degerleri
print(paste("AIC:", model_aic))
```

```
## [1] "AIC: -868.974559753318"
```

```
print(paste("BIC:", model_bic))
```

```
## [1] "BIC: -835.059670791174"
```

- Metrikler, modelimizin ne kadar iyi performans gösterdiğini anlamanız için yararlıdır. AIC ve BIC, modelin karmaşıklığını ve uyumunu birlikte değerlendirirken; MSE ve RMSE, model tahminlerinin gerçek değerlere ne kadar yakın olduğunu gösterir. Genellikle, daha düşük AIC ve BIC değerleri tercih edilir, çünkü bu değerler modelin verileri daha iyi yakaladığını ve aynı zamanda gereğinden fazla karmaşık olmadığını gösterir. MSE ve RMSE'nin düşük olması, modelin tahminlerinin gerçek değerlere yakın olduğunu gösterir.

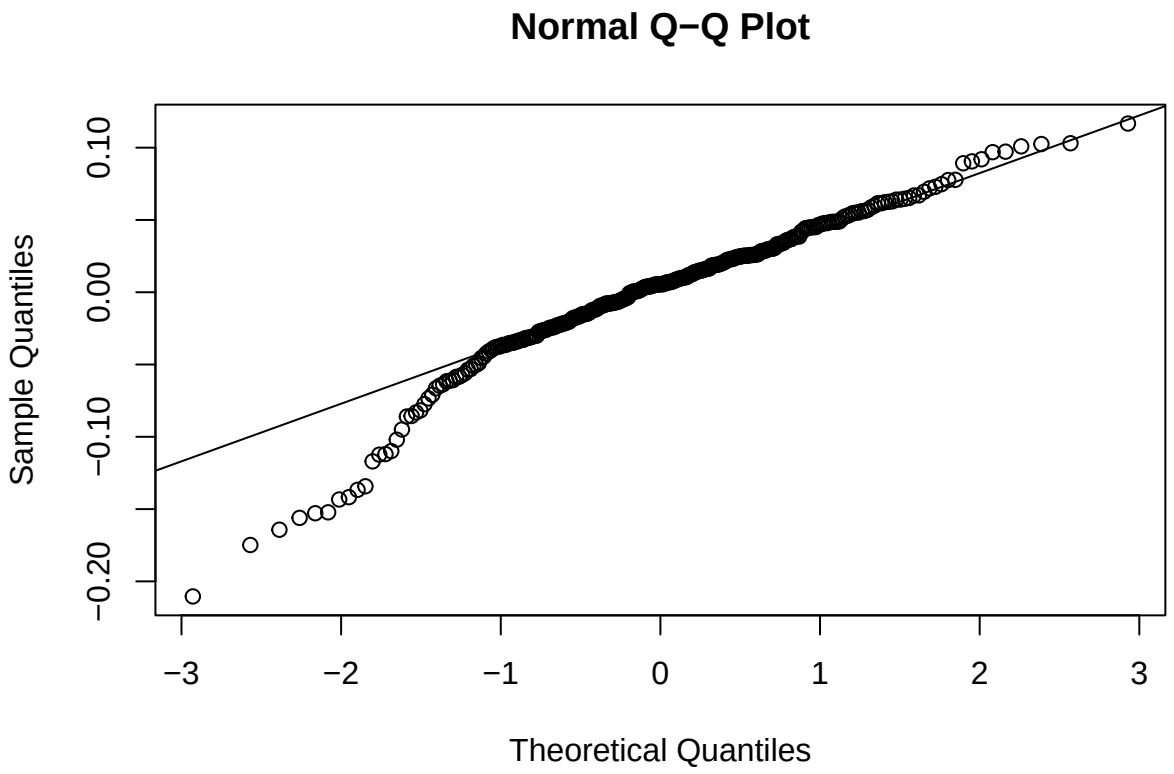
Çıkarımlar

- Genel olarak, kurduğum modelin AIC ve BIC değerlerinin negatif ve büyük bir negatif değer olması, modelinizin iyi bir uyum sağladığını ve aşırı uyum (overfitting) olmadığını gösterir. RMSE ve MSE değerleri de oldukça düşük, bu da modelin test setindeki verileri iyi tahmin ettiğini gösteriyor.

Son Kurduğum Model İçin Varsayım Kontrolü

V1:Hata Terimlerinin Normal Dagılımını Kontrol Etme

```
knitr::opts_chunk$set(echo = TRUE)
qqnorm(residuals(model2))
qqline(residuals(model2))
```



V2:Hata Terimlerinin Bagımsızlıgını Kontrol Etme

- Hata terimleri arasındaki oto-korelasyonu kontrol etmek için Durbin-Watson testi

```
knitr::opts_chunk$set(echo = TRUE)
dwtest(model2)
```

```
##
##  Durbin-Watson test
##
## data:  model2
```

```
## DW = 1.8305, p-value = 0.07284
## alternative hypothesis: true autocorrelation is greater than 0
```

V3:Eşit Varyans Varsayımını Kontrol Etme

```
knitr::opts_chunk$set(echo = TRUE)
# Breusch-Pagan testi
bptest(model2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model2
## BP = 25.502, df = 5, p-value = 0.0001114
```

V4:Çoklu Doğrusal Bağlantı Varsayımını Kontrol Etme

```
knitr::opts_chunk$set(echo = TRUE)
vif(model2)
```

```
##   log_GRE log_TOEFL      CGPA  Research       LOR
##  4.829254  4.912565  5.650141  1.537736  1.840158
```

Çıkarımlar

- Modelimde residualları genel olarak normal dağılıma yakın, ancak bazı potansiyel sapmalar var.
- Residuallar arasında düşük otokorelasyon riski olabilir, ancak bu durum çok kritik görünmüyor.
- V3 varsayımı ihlal edilmiş gibi görünüyor, bu da modelin hata terimlerinin bağımsız değişkenlerin değerlerine bağlı olarak farklılık gösterdiği anlamına gelmektedir.
- CGPA'yı dışarıda tutarsak, doğrusal bağlantı ile ilgili bir problem görünmüyor..
- Genel olarak modelimin varsayımlarını daha yakından incelemek ve potansiyel olarak modelimi iyileştirebilecek adımları göz önünde bulundurmak faydalı olacaktır.