به نام خدا

حل مسئله‌ی رگرسیون به وسیله درخت تصمیم

"درس مبانی علم داده‌ها"

استاد مربوطه: دکتر موسی گلعلی‌زاده

پژوهشگر: ساجده لشگری

آبان ۹۸

داده‌های مورد استفاده در این مسئله، مربوط به اطلاعات اتومبیل‌ها می‌باشد. تعداد کل اتومبیل‌ها ۱۱۷ است که ٪۹۰ از آن‌ها به عنوان نمونه آموزش و باقی‌مانده به عنوان نمونه آزمایش قرار گرفته شده‌اند.

هدف پیش‌بینی مسافت طی شده توسط اتومبیل‌ها با توجه به صفت‌های مربوط به آن و استفاده از الگوریتم درخت تصمیم می‌باشد.

این صفت‌ها عبارتند از:

قیمت (متغیر عددی بین ۵۸۶۶ تا ۴۱۹۹۰)، کشور سازنده (متغیر رسته‌ای)، میزان قابلیت اطمینان به اتومبیل‌ها از نظر شرکت سازنده (متغیر رسته‌ای ۵ سطحی)، نوع (متغیر رسته‌ای ۶ سطحی) و در آخر مسافت طی شده توسط اتومبیل (متغیر عددی بین ۱۸ تا ۳۷) که به عنوان متغیر پاسخ در نظر گرفته شده است.

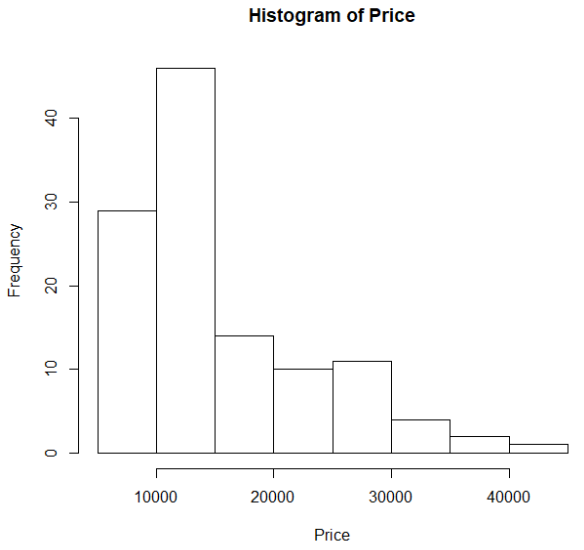
خلاصه‌ای از داده‌ها به صورت

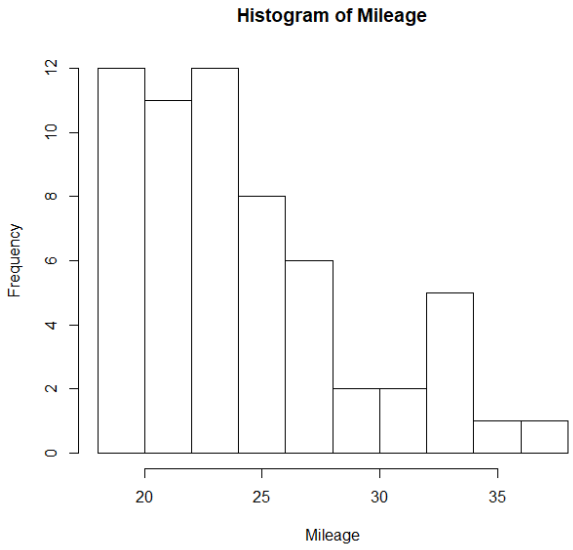
```
     Price          Country      Reliability    Mileage         Type
 Min.   : 5866   USA      :49   Much worse :18   Min.   :18.00   Compact:22
 1st Qu.:10125   Japan    :31   worse      :12   1st Qu.:21.00   Large  : 7
 Median :13150   Germany  :11   average    :26   Median :23.00   Medium :30
 Mean   :15743   Japan/USA: 9   better     : 8   Mean   :24.58   Small  :22
 3rd Qu.:18900   Korea    : 5   Much better:21   3rd Qu.:27.00   Sporty :26
 Max.   :41990   Sweden   : 5   NA's       :32   Max.   :37.00   Van    :10
                 (Other)  : 7                    NA's   :57
```

می‌باشد.

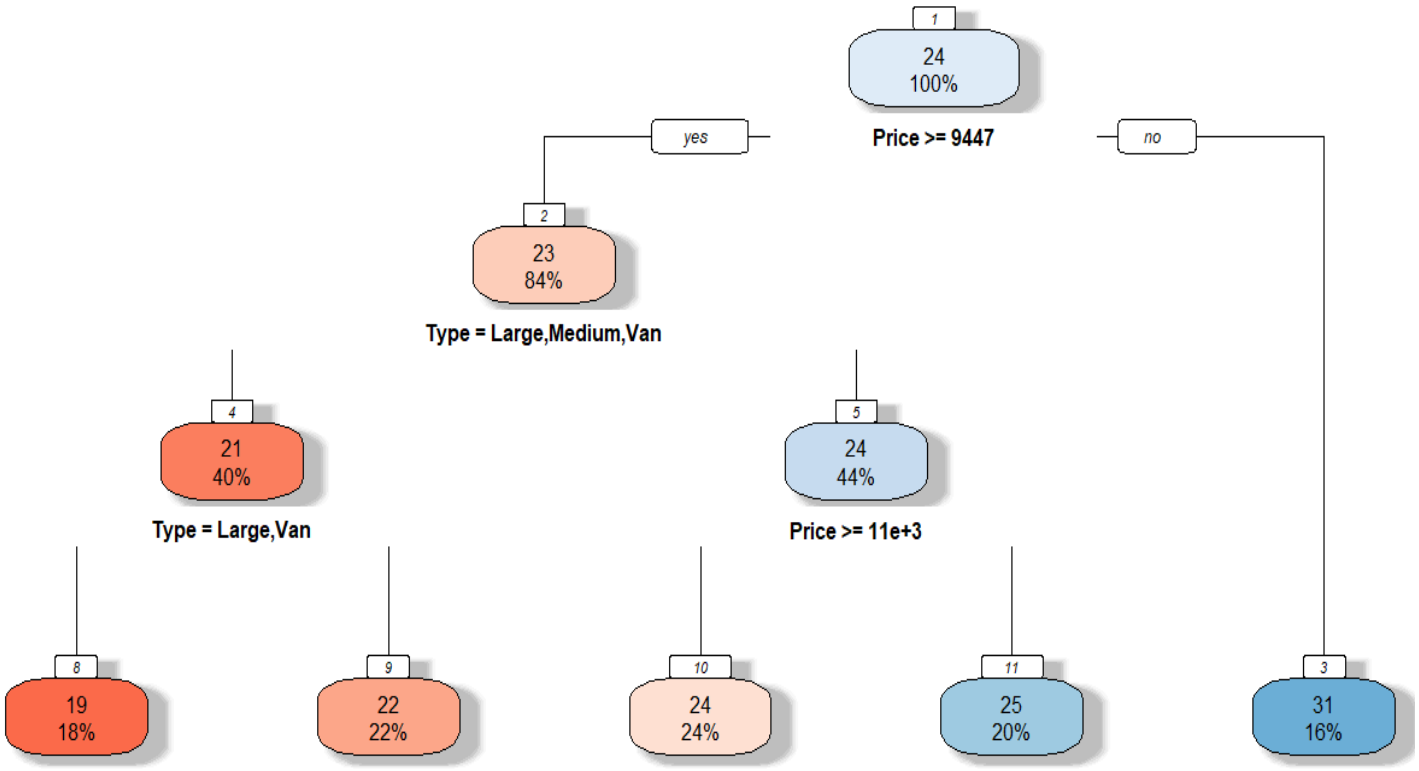
همانطور که در خلاصه‌ی بالا مشاهده می‌شود، در صفت‌های قابلیت اطمینان و مسافت طی شده، داده گمشده وجود دارد که باتوجه به مدل انتخابی (الگوریتم CART) نیازی به پیش پردازش وجود ندارد و با وجود داده‌های گمشده مدل ساخته و پیش‌بینی انجام می‌شود.

همچنین باتوجه به نمودار هیستوگرام داده‌های عددی که در زیر مشاهده می‌شود، می‌توان گفت توزیع آن‌ها نرمال نیست.





در مرحله بعد با استفاده از درخت تصمیم، مدل‌سازی انجام شده، سپس به وسیله‌ی این مدل، برای متغیر پاسخ (مسافت طی شده توسط اتومبیل) پیش‌بینی انجام شده است.

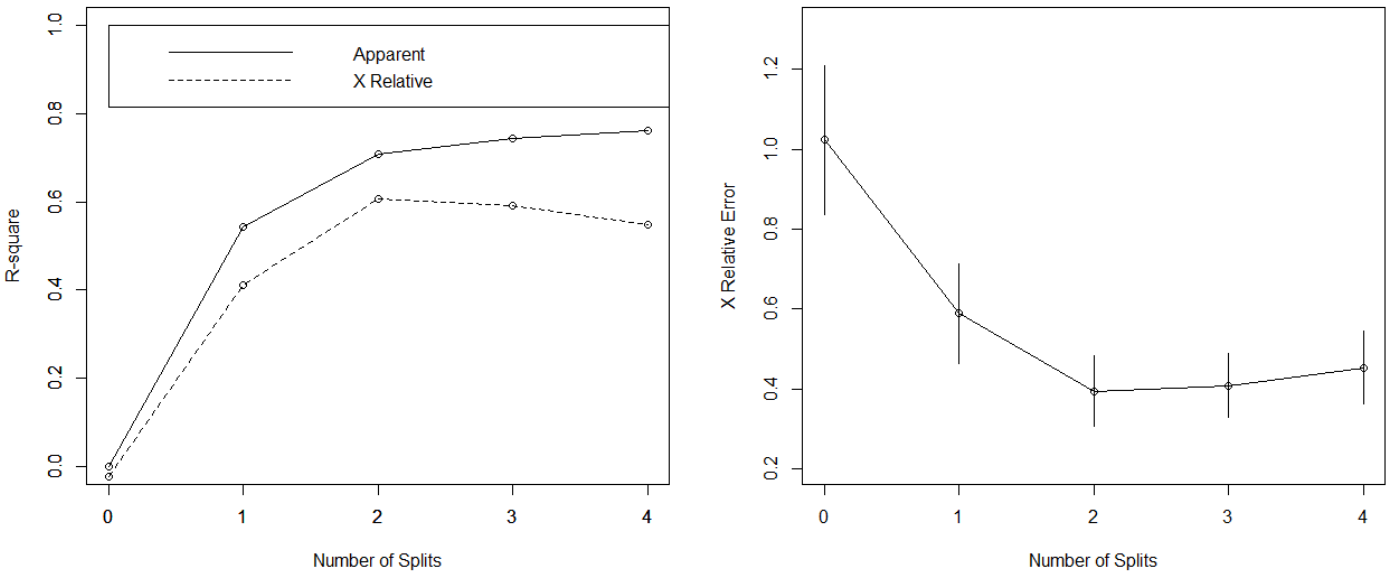


نمودار ۱

در بین 4 صفت نام برده شده در بالا، تنها 2 صفت (قیمت و نوع اتومبیل) در ساختن مدل استفاده شده اند (انتخاب ویژگی (feature selection) انجام شده) که می‌توان این نتایج را در جدول زیر مشاهده کرد.

```
Regression tree:
rpart(formula = Mileage ~ ., data = train, method = "anova")

Variables actually used in tree construction:
[1] Price Type

Root node error: 1011.7/55 = 18.395

n=55 (50 observations deleted due to missingness)

        CP nsplit rel error  xerror     xstd
1 0.543679      0   1.00000 1.02298 0.186812
2 0.165399      1   0.45632 0.58864 0.124258
3 0.034601      2   0.29092 0.39453 0.088583
4 0.018289      3   0.25632 0.40837 0.080547
5 0.010000      4   0.23803 0.45249 0.090802
```

جدول ۱



نمودار ۲

با توجه به نمودار اول، تفسیرهای زیر حاصل می‌شوند.

- ٪۱۶ از اتومبیل‌هایی که قیمتشان کمتر از $۹۴۴۷ بوده، ۳۱ مایل طی کرده‌اند.
- ٪۲۰ از اتومبیل‌هایی که قیمتشان بین $۹۴۴۷ تا $۱۱۰۰۰ بوده و نوعشان sporty یا small یا compact است، ۲۵ مایل طی کرده‌اند.
- ٪۲۴ از اتومبیل‌هایی که قیمتشان بیشتر از $۱۱۰۰۰ بوده و نوعشان sporty یا small یا compact است، ۲۴ مایل طی کرده‌اند.
- ٪۲۲ از اتومبیل‌هایی که قیمتشان بیشتر از $۹۴۴۷ بوده و نوعشان medium است، ۲۲ مایل طی کرده‌اند.
- ٪۱۸ از اتومبیل‌هایی که قیمتشان بیشتر از $۹۴۴۷ بوده و نوعشان large یا van است، ۱۹ مایل طی کرده‌اند.

همانطور که در نمودار۲ (سمت چپ) مشاهده می‌شود، مقدار ضریب تعیین برای split چهارم نزدیک به ۰,۷۶ است و مقدار X Relative در دومین split، بیشترین مقدار را دارد پس می‌توان گفت تعداد split مناسب برابر با ۲ است.

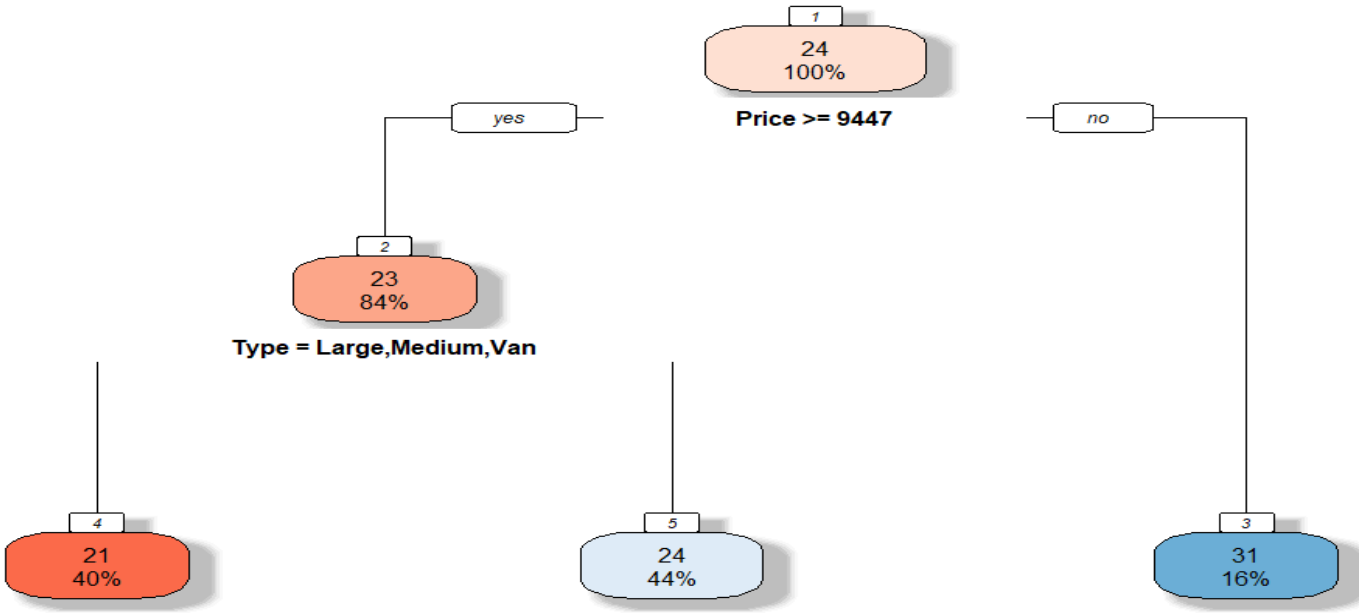
همچنین در مرحله بعد با استفاده از نتایج به‌دست آمده در جدول۱ و نمودار۲ تعداد splitهای مناسب تعیین می‌شود و post-pruning به وسیله آن انجام می‌شود.

با توجه به کوچکترین مقدار xerror در جدول یادشده و نقطه‌ی شکستگی یا زانو در نمودار۲ (سمت راست)، Split مناسب برابر با ۲ می‌باشد و درخت حاصل از این مدل به صورت نمودار۳ است.

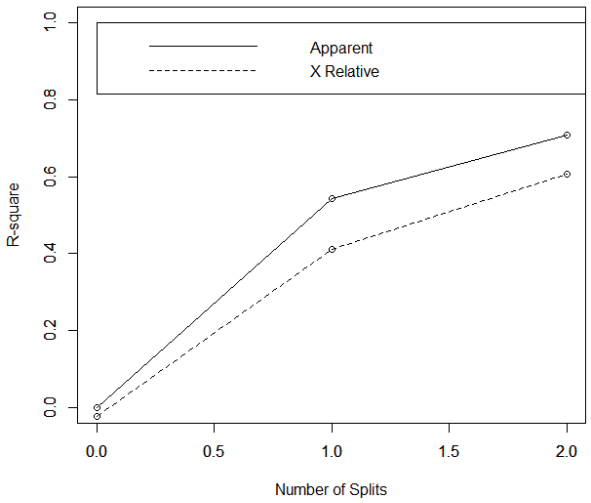
```
Regression tree:
rpart(formula = Mileage ~ ., data = train, method = "anova")

Variables actually used in tree construction:
[1] Price Type

Root node error: 1011.7/55 = 18.395

n=55 (50 observations deleted due to missingness)

        CP nsplit rel error  xerror     xstd
1 0.543679      0   1.00000 1.02298 0.186812
2 0.165399      1   0.45632 0.58864 0.124258
3 0.034601      2   0.29092 0.39453 0.088583
```

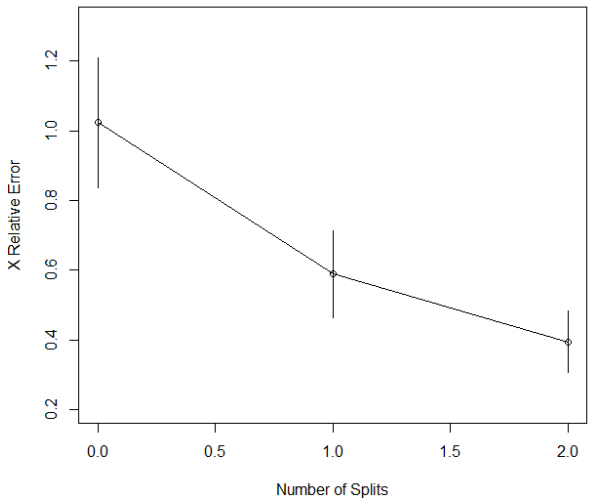
همچنین نتایج و تفسیرهای حاصل از آن به صورت زیر می‌باشد.

- ٪۱۶ از اتومبیل‌هایی که قیمتشان کمتر از $۹۴۴۷ بوده، ۳۱ مایل طی کرده‌اند.
- ٪۴۴ از اتومبیل‌هایی که قیمتشان بیشتر از $۹۴۴۷ بوده و نوعشان sporty یا small یا compact است، ۲۴ مایل طی کرده‌اند.
- ٪۴۰ از اتومبیل‌هایی که قیمتشان بیشتر از $۹۴۴۷ بوده و نوعشان large یا medium یا van است، ۲۱ مایل طی کرده‌اند.



نمودار۳





همچنین مقدار ضریب تعیین برای ۲ تقسیم‌بندی تقریبا برابر ۰,۷۲ است و می‌توان گفت مدل تقریبا عملکرد خوبی دارد.

و به ترتیب مقادیر پیش‌بینی شده و مقادیر واقعی برای نمونه‌های جدید (نمونه‌های آزمایش) به صورت زیر است:

```
 [1] 31.22222 31.22222 31.22222 24.50000 24.50000 24.50000 24.50000 31.22222 24.50000 20.68182
[11] 20.68182 20.68182
```

مقادیر پیش‌بینی شده

```
[1]  33  37  34  NA  NA  NA  NA  NA  26  NA  21  NA
```

مقادیر واقعی

همانطور که مشاهده می‌شود مقادیر تقریبا نزدیک به هم هستند و برای مقادیر گمشده نیز پیش‌بینی انجام شده است که این یکی از مزیت‌های استفاده از این مدل است.

همچنین مقدار MSE برگ‌ها (مقادیر پیش بینی شده) در جدول زیر نمایش داده شده، که باتوجه به کم بودن مقدار آن می‌توان گفت مدل خوبی به داده‌ها برازش داده شده است.

مدل اول	مدل دوم (هرس شده)
Node number 3: 9 observations mean=31.22222, MSE=7.506173 Node number 8: 10 observations mean=19.3, MSE=2.21 Node number 9: 12 observations mean=21.83333, MSE=0.8055556 Node number 10: 13 observations mean=23.69231, MSE=7.905325 Node number 11: 11 observations mean=25.45455, MSE=3.520661	Node number 3: 9 observations mean=31.22222, MSE=7.506173 Node number 4: 22 observations mean=20.68182, MSE=3.035124 Node number 5: 24 observations mean=24.5, MSE=6.666667