دانشگاه صنعتی شریف
دانشکده علوم ریاضی

آمار و کابرد آن

گزارش پروژه

امیرعباس افضلی ۴۰۰۱۰۰۶۶۲

مدرّس: دکتر جواد ابراهیمی

نیم سال دوم ۱۴۰۲-۱۴۰۱

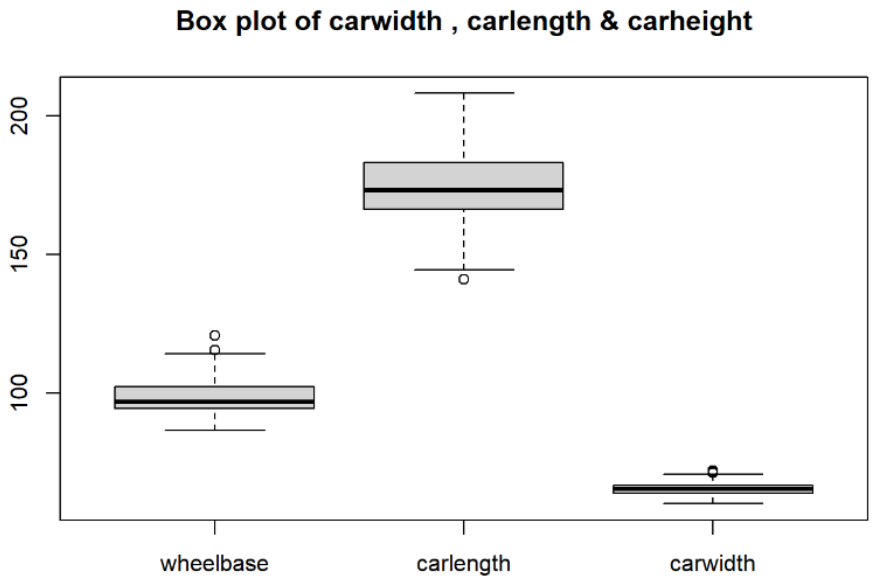
در این فایل ، کمی توضیحات بیشتر درباره ی پروژه داده شده است

در ابتدا دیتاست را لود می کنیم و چند ستون اول آن را نمایش می دهیم

"X"	"car_ID"	"symboling"	"wheelbase"	"carlength"	"carwidth"	"carheight"	"curbweight"	"enginesize"	"boreratio"	"stroke"	"compressionratio"	"horsepower"	"peakrpm"	"citympg"	"highwaympg"	"price"
0	1	3	88.6	168.8	64.1	48.8	2548	130	3.47	2.68	9.0	111	5000	21	27	13495.0
1	2	3	88.6	168.8	64.1	48.8	2548	130	3.47	2.68	9.0	111	5000	21	27	16500.0
2	3	1	94.5	171.2	65.5	52.4	2823	152	2.68	3.47	9.0	154	5000	19	26	16500.0
3	4	2	99.8	176.6	66.2	54.3	2337	109	3.19	3.4	10.0	102	5500	24	30	13950.0
4	5	2	99.4	176.6	66.4	54.3	2417	136	3.19	3.4	8.0	115	5500	18	22	17450.0
5	6	2	99.8	177.3	66.3	53.1	2507	136	3.31	3.4	8.5	110	5500	19	25	15250.0

شکل ۱

حال سه بار پلات مربوط به سه ستون $carlength, carwidth, carheight$ را نمایش می دهیم و آن ها را تحلیل می کنیم.

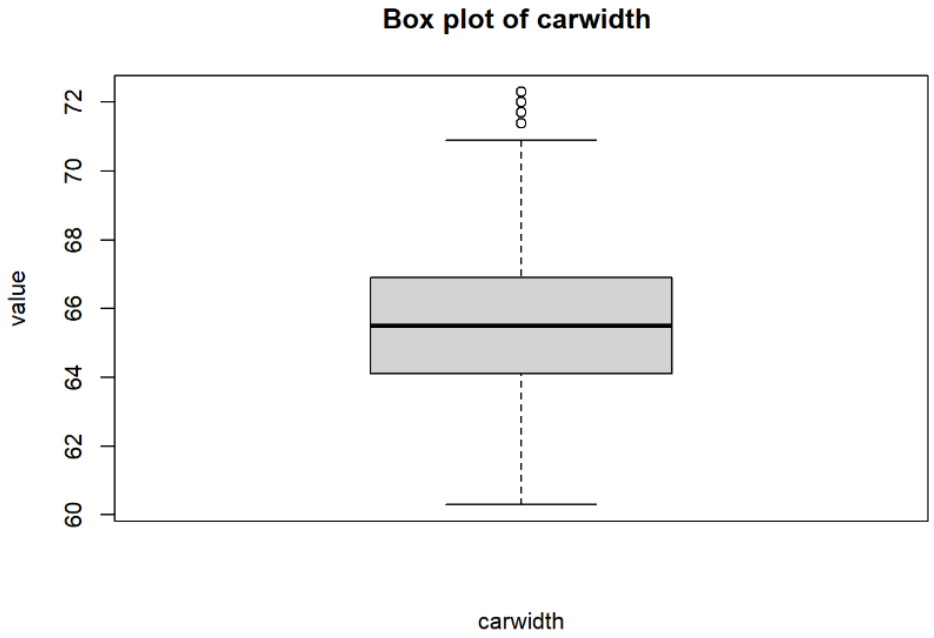


شکل ۲

با توجه به نمودار بالا مشخص است که پراندگی نسبی داده های ستون $carlength$ نسبت دو ستون دیگر بیشتر است . همچنین در ستون $carlength$ تجمع داده ها در مقادیر پایین تر است و همچنین در این ستون داده پرت $(outlier)$ نیز که مقدار آن از سایر دیتا ها کمتر است نیز وجود دارد.

مشابه توضیحات بالا برای ستون $wheelbase$ نیز وجود دارد ، با این تفاوت که داده های پرت این ستون ، زا سایر داده ها بزرگتر هستند.

در ستون $carwidth$ که پراکندگی این داده ها نسبت به دو ستون دیگر کمتر است ، برای مشاهده و تحلیل بهتر آن را بصورت جداگانه پلات می کنیم :



شکل ۳

مشاهده می شود که توزیع داده ها تقریبا متقارن است و فاصله میانه با $Quantile\ 25$ و $Quantile\ 75$ برابر می باشد و همچنین چند داده پرت که مقادیرشان بزرگتر از سایر داده ها است نیز در داده های این ستون وجود دارد.

توجه شود که به طور دیفالت در اینجا داده های بزرگتر از $Q3 + 1.5IQR$ و کوچکتر از $Q1 - 1.5IQR$ داده پرت در نظر گرفته شده اند.

Data wrangling and remove missing value

در این قسمت مشکل داده گم شده $(missing\ value)$ را برطرف می کنیم در ستون های $numeric$ ، $missing\ value$ ها را با میانه ستون و در ستون های $categorical$ ، $missing\ value$ ها را با مد جایگزین می کنیم .

$missing\ value$ ها در بعضی ستون ها بصورت استرینک خالی و در بعضی بصورت NA می باشند .

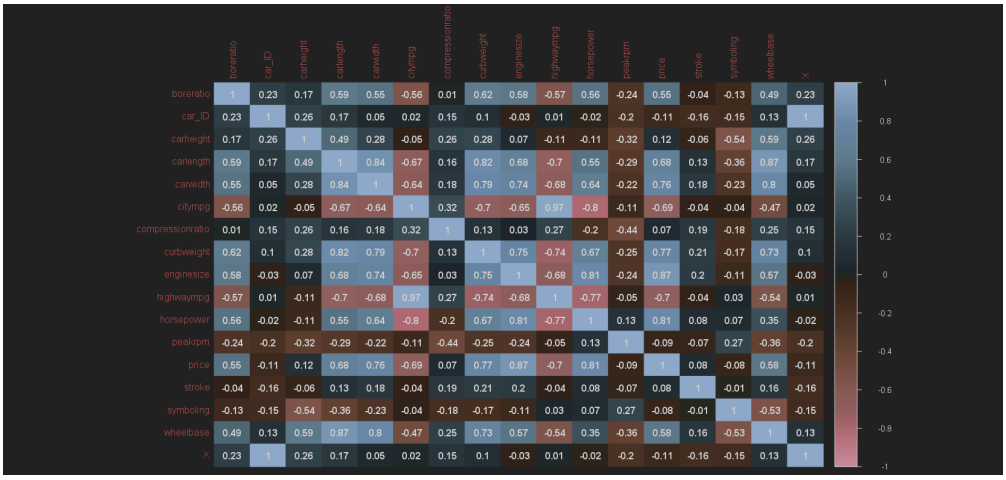
مشاهده می شود که در ستون های $curbweight, boreratio, carbody, cylindernumber$ داده گم شده وجود دارد.

داده های گم شده در $curbweight$ و $boreratio$ را با میانه و داده های گم شده در $carbody$ و $cylindernumber$

را با مد ستون خود جایگزین می کنیم .
دلیل استفاده از مد و میانه هم در گزارش فایل *rmd*. آورده شده است.

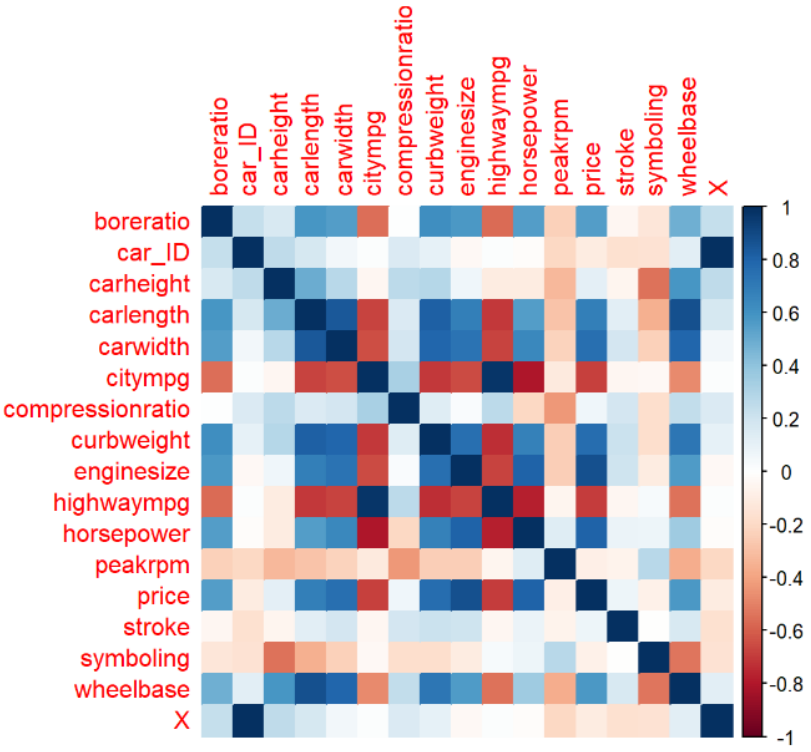
Correlation map

در این بخش با دستور *corr* ، همبستگی ویژگی های *numeric* دادگان را می یابیم و آن را نمایش می دهیم.



شکل ۴:

Correlation map



شکل ۵:

Correlation map

بررسی اثرات علی :

با توجه به نقشه همبستگی بالا می توان گفت که که متغیر قیمت که متغیر پاسخ مساله رگرسیون است ، همبستگی مثبت میان قیمت خودرو و متغیر های زیر تا حدی قابل توجه می باشد :

$$Enginesize, curbweight, horsepower, borreratio, carwidth, carlength, wheelebase$$

همچنین متغیر های $Citymap$ و $highwaympg$ با یکدیگر کورلیشن منفی دارند.

حال ۴ آزمون فرض طراحی می کنیم و با سطح اطمینان ۵ درصد برای آن تصمیم گیری می کنیم.

اگر متغیر پاسخ که در اینجا $Price$ می باشد ، به صورت تابعی خطی از متغیر های گفته شده در نظر بگیریم ، یعنی فرض کنیم یک مساله رگرسیون $MultipleLinearRegression(MLR)$ داریم ، رابطه متغیر پاسخ با

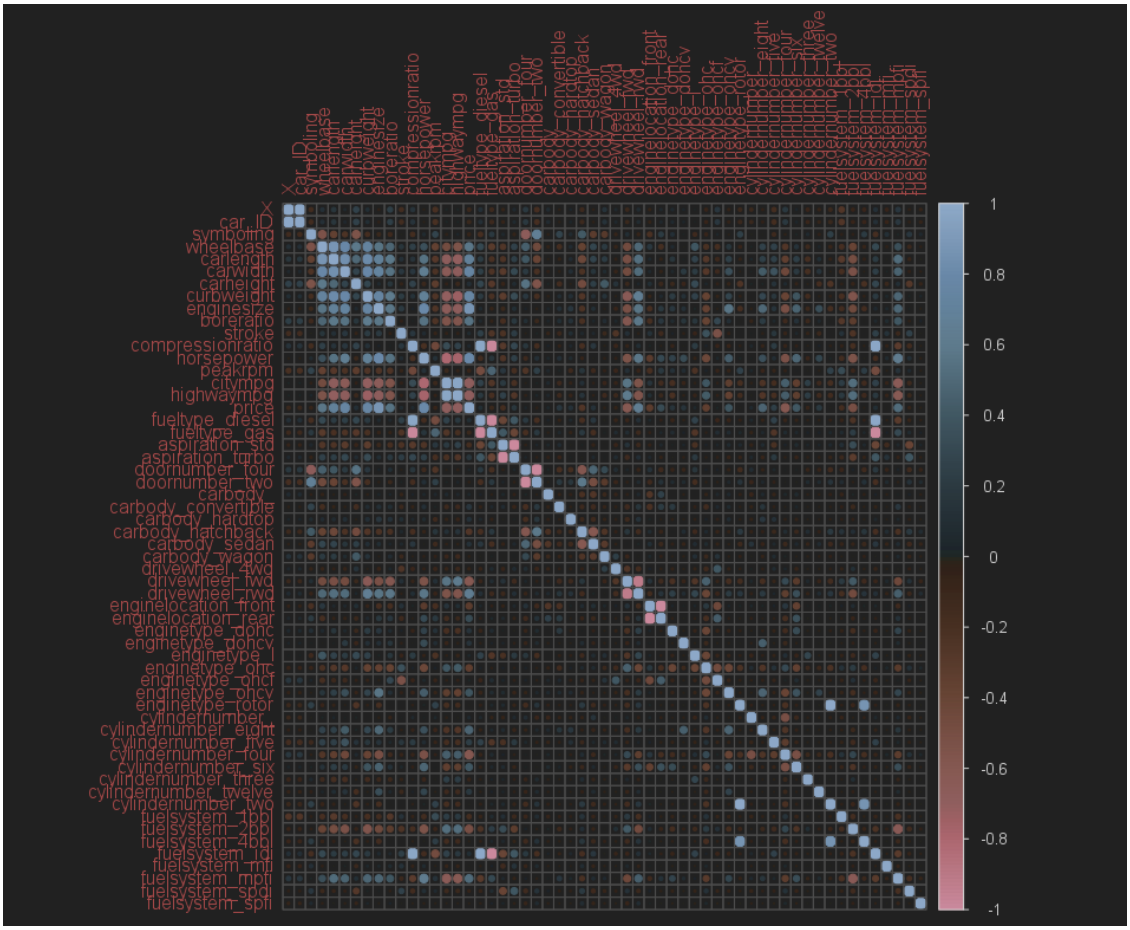
متغیر های پیشگو به این صورت خواهد بود :

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n$$

بخش طراحی آزمون فرض در فایل rmd. توضیح داده شده است .

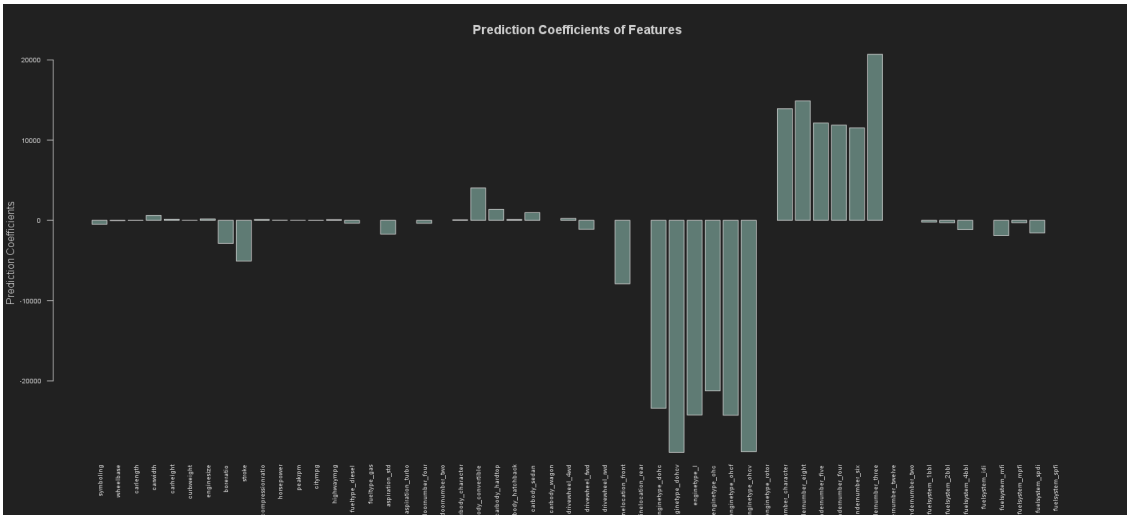
define dummy variable

طبق توضیحاتی که در فایل گزارش *rmd*. داده شده است ، برای ستون های کتگوریکال ، متغیر های جدید تعریف می کنیم و آن ستون ها را حذف می کنیم . و در آخر برای دیتاست جدید که حدود ۶۰ ستون دارد ، *correlation map* رسم می کنیم .



شکل ۶:

New Correlation map



شکل ۷

در نمودار بالا که نشان دهنده جدول ضرایب می باشد، مقدار $Prediction$ هر ضریب مشخص شده است. صرفا بزرگتر بودن میزان یک ضریب ، به معنای اهمیت آن در مدل نمی باشد . برای مثال مقدار مقیاس $(scale)$ دادگان می تواند ب مقدار ضریب اثر بگزارد.همانطور که در تصویر بالا نیز مشاهده می شود، براخی از ضرایب بزرگ متناظر با ویژگی هایی هستند که رنج بزرگتری دارند.

حال اگر مقیاس ضزایب برای همه ی ویژگی ها در مدل رگرسیون یکی باشد ، بزرگی ضرضرایب می تواند، می تواند نشانه ای نسبی از اهمیت آن ویژگی ها باشد.

در این ویژگی هایی با ضرایب بزرگتر می توانند ، اثر بیشتری بر متغیر پاسخ داشته باشند (اثر مثبت یا منفی).

راهکار های بهبود عملکرد هم به طور مفصل در فایل rmd. توضیح داده شده است.

Feature Selection

ابتدا فیچر سلکشن را با استفاده از آزمون t و روش $backward$ ، همانطور که در فایل $.rmd$ به طور مفصل توضیح داده شده است ، انجام می دهیم.
ستون های انتخاب شده در شکل زیر نمایش داده شده است.

	Coefficients	P_Value
1	(Intercept)	0.100180
2	symboling	0.0164736
3	carwidth	0.000392687
4	enginesize	2.78343e-26
5	boreratio	0.00671621
6	stroke	4.75959e-09
7	compressionratio	0.0294139
8	peakrpm	5.31453e-06
9	aspiration_std	0.000127084
10	carbody_convertible	0.000355109
11	carbody_sedan	0.0705188
12	drivewheel_fwd	0.00748450
13	enginelocation_front	3.95500e-05
14	enginetype_dohc	1.80154e-08

	Coefficients	P_Value
13	enginelocation_front	3.95500e-05
14	enginetype_dohc	1.80154e-08
15	enginetype_dohcv	2.50871e-09
16	enginetype_l	5.02607e-09
17	enginetype_ohc	1.51065e-07
18	enginetype_ohcf	1.41614e-08
19	enginetype_ohcv	1.63658e-11
20	cylindernumber_character	2.11927e-06
21	cylindernumber_eight	1.01156e-08
22	cylindernumber_five	5.38912e-05
23	cylindernumber_four	4.08693e-05
24	cylindernumber_six	6.11248e-06
25	cylindernumber_three	7.27658e-08
26	fuelsystem_spdi	0.100942

شکل ۸:

Selected features

همانطور که مشاهده می شود ، همه ی پی ولیو ها کمتر از ۵ درصد می باشند.
با مشاهده نتایج معیار های خطا در مدل فیت شده ی جدید (با ویژگی های کمتر) ، $TSS, MSE, R-Squared$ تغییر زیادی نکرده اند . اما درمورد $adjust\ R\ squared$ ، همانطور که در فایل $.rmd$ هم ذکر شد ، این ویژگی تا حدی وابسته به تعداد ویژگی ها نیز می باشد و به همین دلیل تغییر تعداد ویژگی تغییر زیادی در آن مشاهده می شود.

حال فیچر سلکشن را با استفاده از $ANOVA$ و $F-statistics$ انجام می دهیم.
و الگوریتم آن به این صورت است که از بین مثلا ۵۰ ویژگی از داده ترین ، یک ستون ویژگی انتخاب میکنیم و $F-score$ آن را با متغیر پاسخ که در اینجا $price$ می باشد اندازه می گیریم .
در انتهای این عملیات ما ۵۰ $F-score$ خواهیم داشت و ۱۰ ستونی که بیشترین $F-score$ را دارند را به عنوان ویژگی های انتخاب شده $(Selected\ feature)$ بر می گزینیم.
ستون های انتخاب شده در شکل زیر نمایش داده شده است.

	Coefficients	F_Value
1	enginesize	546.991
2	horsepower	329.591
3	curbweight	248.189
4	carwidth	190.300
5	highwaym...	140.345
6	citympg	136.750
7	carlength	127.651
8	drivewheel...	109.894
9	drivewheel...	89.7064
10	boreratio	88.6903
11	cylindernu...	81.8387
12	fuelsystem...	63.3388
13	wheelbase	63.2889

	Coefficients	F_Value
13	wheelbase	63.2889
14	fuelsystem...	49.9723
15	cylindernu...	48.9134
16	cylindernu...	36.9274
17	enginetype...	27.9672
18	enginetype...	27.6638
19	engineloca...	23.2610
20	engineloca...	23.2610
21	carbody_h...	9.95162
22	carbody_c...	8.43683
23	cylindernu...	8.03881
24	carbody_h...	6.22126
25	enginetype...	5.01834

شکل ۹:

$Selected\ features$

$$Synergy\ \ effect$$

برای انتخاب و بررسی ۱۰ جفت ویژگی ، از بین ۱۰ ستونی که در فیچر سلکشن با $ANOVA$ بدست آوردیم ، ۲ ستون را انتخاب می کنیم و ستون جدیدی که حاصل ضرب این دو ستون می باشد را به دیتای خود اضافه می کنیم و برای دیتای جدید(با ۱۱ متغیر پیشگو) ، رگرسیون فیت می کنیم و در انتها اگر پی ولیو متناظر ستون جدید ، از $\frac{\alpha}{2}$ کمتر شد ، یعنی تاثیر این ویژگی زیاد است و می توانیم این ویژگی را به ویژگی های دادگان اضافه کنیم و در غیر این صورت این ویژگی را به ویژگی های دادگان اضافه نمی کنیم.
با استفاده از آزمون t و $P-Value$ نیز می توان سطح معنی داری را مشاهده کرد.

$Other\ Models$:

بر روی چند مدل دیگر داده ترین را فیت می کنیم و مقادیر MSE و $R-Squares$ مدل را بر روی داده تست گزارش می کنیم.

1.$Decision\ Tree$:

مدل $Decision\ Tree$ یک مدل درختی است که با استفاده از تقسیم دادهها بر اساس مجموعه از شرایط ویژگی ($feature$)، تصمیمات گرفته و پیش بینی میکند.
این مدل بسیار قابل فهم است و میتواند قوانین قابل فهمی را از دادهها استخراج کند.
$Decision\ Tree$ به طور طبیعی با تبدیلات غیرخطی و تعامل بین این ویژگیها سازگار است.
با این حال، $DecisionTree$ به شدت تمایل به بالا/پایین رفتن از حدود تقسیم است و ممکن است به بیش برازش ($overfit$) برای دادهها منجر شود.

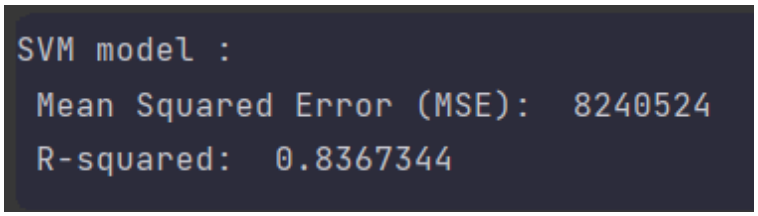
```
Decision tree model :
 Mean Squared Error (MSE):  10712526
 R-squared:  0.7877578
```

شکل ۱۰:

Decision tree model

2.$SVM(SupportVectorMachine)$:

SVM یک مدل ماشین لرنینگ تمیز کننده است که در تعیین یک طرح تصمیم بهینه برای دادههای دو دستهبندی مشکلات به نظر میرسد.
با استفاده از تابع هسته $(Kernel)SVM$ قادر به ایجاد طرح تصمیم مستقل از ابعاد داده است و میتواند با دادههای خطی و غیرخطی کار کند.
SVM استفاده از بردارهای حاشیهای $(SupportVector)$ برای تصمیمگیری استفاده میکند و برای دادههایی که انتقال هستند مقاوم است.

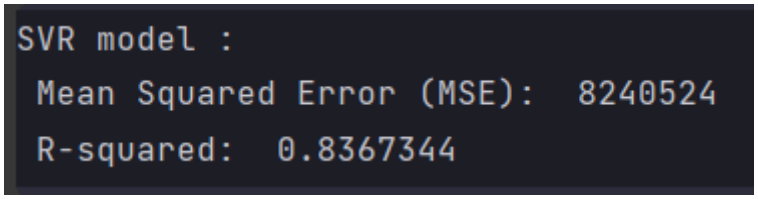


شکل ۱۱:

$SVM\ model$

$3.SVR(SupportVectorRegression)$:

SVR مدلی است که برای مسائل پیشبینی و رگرسیون به کار میرود. - این مدل، بردارهای حاشیهای $(SupportVector)$ را به عنوان نقاط مرجع برای مدلسازی عملکرد پیشبینیکننده استفاده میکند.
SVR همچنین با استفاده از تابع هسته $(Kernel)$، میتواند با دادههای غیرخطی کار کند و روابط پیچیدهتری را با دادهها نمایش دهد.



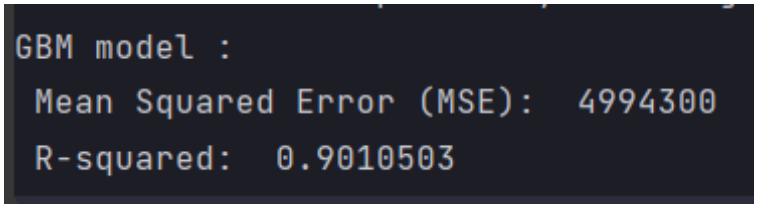
شکل ۱۲:

$SVR\ model$

$4.GBM(GradientBoostingMachine)$:

GBM یک مدل ترکیبی است که مجموعه از مدلهای ضعیف (مثلاً درختان تصمیم کوچک) را بهبود میدهد و به یک مدل قویتر تبدیل میکند. - این مدل با بهبود مداوم در تابع هزینه $(CostFunction)$ از طریق یک روش گرادیان بهینهسازی میکند.

GBM قادر به مقابله با دادههای پرت و اثربخشی در حضور دادههای پراکنده است. - اما نکته مهمی که باید به آن توجه کنید این است که GBM ممکن است به خاطر مقدار زمان و منابع مورد نیاز اجرا، پیادهسازی مستعد خطا باشد.

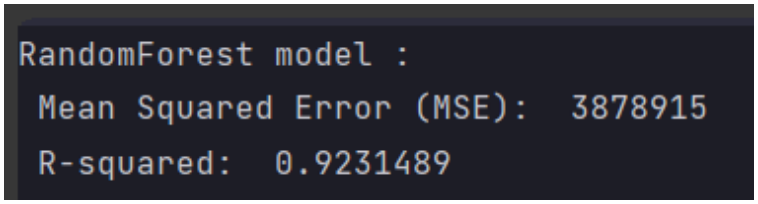


شکل ۱۳:

GBM model

5.$RandomForest$:

$RandomForest$ یک مدل گراف تصادفی است که از مجموعه بزرگی از$DecisionTree$ تصمیم میگیرد. // این مجموعه از درختان تصمیم توسط روند تصادفی از دادهها و ویژگیها ساخته میشود. // $RandomForest$ میتواند با دادههای پرت و نویزی کار کند و از بیشبرازش $(overfitting)$ میتواند غلبه کند. // همچنین،$RandomForest$ قادر به ارزیابی اهمیت ویژگیها در پیشبینی است



شکل ۱۴:

Random Forest model