
بسم الله الرحمن الرحيم

دانشگاه تهران

دفتر آموزش‌های حرفه‌ای و تخصصی

پیش بینی قیمت مسکن بوستون

پروژه نهایی دوره مفاهیم علم داده

نام دانشجو:

تینا شاهدی

نام استاد:

مهندس محمدرضا محتاط

شهریور ۱۴۰۰

چکیده

در این پژوهش گام‌های مربوط به داده‌کاوی و یادگیری ماشین در پروژه پیش‌بینی قیمت مسکن در شهر بوستون پیاده سازی شده است. این پروژه دارای دیتاستی با تعداد ۵۰۶ مشاهده، ۱۳ ستون ورودی و یک ستون خروجی می‌باشد. با توجه به اطلاعاتی که از این دیتاست در وب سایت‌ها در اختیار هست، میزان خطای متوسط[1] ۹.۲۱ هزار دلار برای پیش بینی قیمت مسکن گزارش شده است. بنابراین هدف از این پژوهش پیمودن گام های داده‌کاوی یعنی شناخت کسب وکار، درک داده، آماده سازی داده‌ها، مدلسازی، ارزیابی و توسعه و مقایسه خطای مدل بدست آمده با مقدار خطای فوق الذکر می‌باشد. باتوجه به ماهیت تحقیق از گام‌های اول و آخر یعنی شناخت کسب و کار و توسعه مدل می‌توان چشم پوشی کرد. بنابراین این پژوهش شامل چهار فصل است، در فصل اول می‌کوشیم تا شناختی نسبت به ویژگی‌های ورودی و همچنین ستون خروجی یا هدف را بدست آوریم. در فصل دوم پا پیمودن ۹ مرحله اساسی برای مدلسازی، دیتاست پرژه را آماده سازی می‌کنیم. در فصل سوم با استفاده از الگوریتم‌های مرتبط با مسائل کمی، مدل حاصل از دیتاست آماده شده را انجام می‌دهیم و در فصل آخر مدل ساخته شده را در حالت‌های مختلف ارزیابی می‌کنیم. در نهایت خطای بهترین مدل را بدست آورده و دقت انجام پروژه را برآورد می‌کنیم.

[1] RMSE

فهرست مطالب

عنوان **صفحه**

فهرست جداول

عنوان **صفحه**

فهرست اشکال

عنوان **صفحه**

فصل ۱: درک داده

۱-۱ مقدمه

پروژه پیش بینی قیمت مسکن بوستون با دیتاست Boston House Price شامل پیش بینی قیمت خانه‌ها و همسایه‌های آن بر اساس هزار دلار می باشد. این دیتاست یک مسئله رگرسیون با تعداد مشاهدات ۵۰۶، ورودی ها ۱۳ و یک ستون خروجی می باشد. در ادامه به تشریح اسامی ستون های ویژگی می‌پردازیم.

۱-۲ شناسایی داده‌ها

قبل از شروع کار در ابتدا باید هر یک از ویژگی‌های ورودی و همچنین ستون خروجی یا هدف را بشناسیم.

۱-۲-۱ ستون اول؛ CRIM

این ستون یکی از ستون‌های بردار ویژگی است که نشان دهنده نرخ سرانه تبه کاری است. این ویژگی تاثیر به سزایی در تعیین قیمت مسکن مناطق می‌گذارد و بنابراین در نظر گرفتن آن حائز اهمیت است.

۱-۲-۲ ستون دوم؛ ZN

این ویژگی ورودی نشان دهنده میزان مناطق زمین‌های مسکونی با بیش از ۲۵۰۰۰ فوت مربع[۲]است.

۱-۲-۳ ستون سوم؛ INDUS

این ویژگی ورودی نشان دهنده میزان زمین‌های مشاغل غیر خُرد است.

۱-۲-٤ ستون چهارم؛ CHAS

این ویژگی نشان دهنده متغیر مصنوعی باینری مرتبط با رودخانه Charles واقع در شهر بوستون است. برای هر رکورد اگر این رودخانه مسیرهای سطح شهر را محدود کند مقدار یک به خودش می‌گیرد و در غیر اینصورت صفر خواهد بود.

۱-۲-۵ ستون پنجم؛ NOX

این ستون نشان دهنده غلظت اکسیدهای نیتریک (واحد در ۱۰ میلیون) است. تعداد ذرات در میلیون (PPM) روشی برای تعیین مقدار غلظت مواد است. برای مثال، 1 PPM معادل با ۱ میلی گرم از یک ماده در یک لیتر از یک مایع (mg/L)، یا ۱ میلی گرم از یک ماده در یک کیلوگرم از یک ماده جامد، است(mg/kg). این ویژگی می‌تواند تاثیر به سزایی در قیمت مسکن مناطق با غلظت اکسیدهای نیتریک بالا بگذارد.

۱-۲-٦ ستون ششم؛ RM

این ویژگی نمایانگر متوسط تعداد اتاق‌های هر مسکن در هر یک از رکوردها است.

[۲] یک فوت مربع (Square Foot) معادل ۰.۰۹۲۹۰۳۰٤ مترمربع (Square Meter) است.

۱-۲-۷ ستون هفتم؛ AGE

این ویژگی نمایانگر نسبت تعداد واحدهای مسکونی که مالک در آنها زندگی می‌کند و قبل از سال ۱۹٤۰ ساخته شده‌اند.

۱-۲-۸ ستون هشتم؛ DIS

این ویژگی نمایانگر میانگین وزنی فاصله تا پنج مرکز شغلی شهر بوستون می‌باشد که برای هر رکورد از طریق فرمول زیر محاسبه شده است:

i اندیس مراکز شغلی

n تعداد مراکز شغلی

v_i ضریب وزنی مرکز شغلی i

d_i فاصله تا مرکز شغلی i

$\hat{v}$ میانگین وزنی فاصله رکورد

$$\hat{v} = \frac{\sum_{i=1}^{n} \frac{1}{d_i} v_i}{\sum_{i=1}^{n} \frac{1}{d_i}}$$

۱-۲-۹ ستون نهم؛ RAD

این ویژگی نمایانگر شاخص دسترسی هر یک از مشاهدات به بزرگراه‌های محوری سطح شهر بوستون است.

۱-۲-۱۰ ستون دهم؛ TAX

این ویژگی نمایانگر نرخ کل مالیات بر دارایی هر یک از مشاهدات به ازای هر ۱۰۰۰۰ دلار است.

۱-۲-۱۱ ستون یازدهم؛ PTRATIO

این ویژگی نمایانگر نسبت دانش آموز به معلم بر اساس منطقه است.

۱-۲-۱۲ ستون دوازدهم؛ B

این ویژگی بر اساس فرمول زیر محاسبه می‌شود که در آن B_k نمایانگر نسبت سیاه پوستان در هر منطقه (رکورد یا مشاهده) است: $B = 1000(B_k - 0.63)^2$

۱-۲-۱۳ ستون سیزدهم؛ LSTAT

این ویژگی نمایانگر درصد قشر طبقه پایین جمعیت (در هر مشاهده) است.

۱-۲-۱٤ ستون چهاردهم؛ MEDV

این ستون، همان ستون هدف هست که متوسط قیمت خانه‌های دارنده مالک (یعنی خانه‌هایی که مالک در آنجا زندگی می‌کند) بر اساس هزار دلار را نمایش می‌دهد.

۱-۳ تشخیص خطا یا نویز

با توجه به تعریف داده‌ها، بطور کلی مقادیری که داده‌ها **نمی‌پذیرند** را در جداول زیر مشخص کردیم، همچنین در فصل بعدی به طور مفصل تری به این مورد می‌پردازیم.

جدول ۱-۱ مقادیری که ویژگی‌ها نمی‌پذیرند

CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS
صفر یا منفی	منفی	منفی	غیر از 0 و 1	صفر، منفی یا بیشتر از یک	منفی	منفی	منفی

RAD	TAX	PTRATIO	B	LSTAT	MEDV
منفی، اعشاری یا صفر	منفی یا صفر	منفی یا صفر	منفی یا صفر	غیر از 0 تا 100	منفی یا صفر

۱-٤ ساختار کلی پروژه

در این فصل داده‌ها را شناسایی و درک کردیم، متوجه شدیم این داده‌ها نمایانگر چه چیزی هستند، چه مقادیری را نمی‌پذیرند و همچنین از چه جنس و در چه مقیاسی می‌باشند. قبل از انجام مدلسازی داده‌ها که در فصل سوم به آن پرداخته شده است، در فصل دوم داده‌ها را آماده سازی می‌کنیم. با توجه به اطلاعاتی که از این دیتاست در وب سایت‌ها در اختیار هست، میزان خطای متوسط ۹.۲۱ هزار دلار برای پیش بینی قیمت مسکن گزارش شده است. بنابراین در فصل چهارم مدل ساخته شده خود را ارزیابی می‌کنیم، خطای مدل را بدست آورده با مقدار خطای فوق مقایسه می‌کنیم.

فصل ۲: آماده‌سازی داده

۲-۱ مقدمه

وقتی دیتاست را از سایت Kaggle دریافت می‌کنیم، یک فایل csv به ما می‌دهد که داده های آن نیاز به جداسازی و آماده سازی قبل از Import کردن آن در IBM دارد. ۱۰ رکورد از این فایل را می‌توانید در تصویر زیر مشاهده کنید:

جدول ۲-۱ ده رکورد از فایل csv دیتاست

	A
1	**Boston House Price**
2	0.00632 18.00 2.310 0 0.5380 6.5750 65.20 4.0900 1 296.0 15.30 396.90 4.98 24.00
3	0.02731 0.00 7.070 0 0.4690 6.4210 78.90 4.9671 2 242.0 17.80 396.90 9.14 21.60
4	0.02729 0.00 7.070 0 0.4690 7.1850 61.10 4.9671 2 242.0 17.80 392.83 4.03 34.70
5	0.03237 0.00 2.180 0 0.4580 6.9980 45.80 6.0622 3 222.0 18.70 394.63 2.94 33.40
6	0.06905 0.00 2.180 0 0.4580 7.1470 54.20 6.0622 3 222.0 18.70 396.90 5.33 36.20
7	0.02985 0.00 2.180 0 0.4580 6.4300 58.70 6.0622 3 222.0 18.70 394.12 5.21 28.70
8	0.08829 12.50 7.870 0 0.5240 6.0120 66.60 5.5605 5 311.0 15.20 395.60 12.43 22.90
9	0.14455 12.50 7.870 0 0.5240 6.1720 96.10 5.9505 5 311.0 15.20 396.90 19.15 27.10
10	0.21124 12.50 7.870 0 0.5240 5.6310 100.00 6.0821 5 311.0 15.20 386.63 29.93 16.50

در واقع ۱۴ ستون در یک ردیف نوشته شده اند و ما باید این مقادیر را به ستون‌های مورد نظر انتقال دهیم. برای جداسازی هر یک از ستون ها از فرمول‌های زیر استفاده می‌کنیم، (برای هر ستون، این فرمول‌ها را در یک سلول نوشته و برای بقیه سلول‌ها با دابل کلیک امتداد میدهیم):

جدول ۲-۲ دستورات اکسل برای آماده‌سازی داده‌ها

CRIM	=LEFT(A2,SEARCH(" ",A2,1)-1)
ZN	=RIGHT(LEFT(A2,SEARCH(" ",A2,2)+7),7)
INDUS	=RIGHT(LEFT(A2,SEARCH(" ",A2,11)+7),7)
CHAS	=RIGHT(LEFT(A2,SEARCH(" ",A2,27)-1),3)
NOX	=RIGHT(LEFT(A2,SEARCH(" ",A2,25)+7),7)
RM	=RIGHT(LEFT(A2,SEARCH(" ",A2,30)+7),7)
AGE	=RIGHT(LEFT(A2,SEARCH(" ",A2,40)+7),7)
DIS	=RIGHT(LEFT(A2,SEARCH(" ",A2,45)+7),7)
RAD	=RIGHT(LEFT(A2,SEARCH(" ",A2,55)+3),4)
TAX	=LEFT(RIGHT(A2,SEARCH(" ",A2,35)-2),5)
PTRATIO	=LEFT(RIGHT(A2,SEARCH(" ",A2,35)-9),5)
B	=LEFT(RIGHT(A2,SEARCH(" ",A2,35)-15),7)
LSTAT	=LEFT(RIGHT(A2,SEARCH(" ",A2,35)-22),7)
MEDV	=RIGHT(A2,SEARCH(" ",A2,5)-4)

سپس برای از بین بردن فرمول سلول‌ها، مقادیر تفکیک شده را در یک sheet دیگر اکسل Paste value می‌کنیم تا ثابت شوند. همچنین برای مشخص کردن رکورد ها، یک ستون ID اضافه می‌کنیم. ۱۰ رکورد اول آماده شده بصورت زیر است:

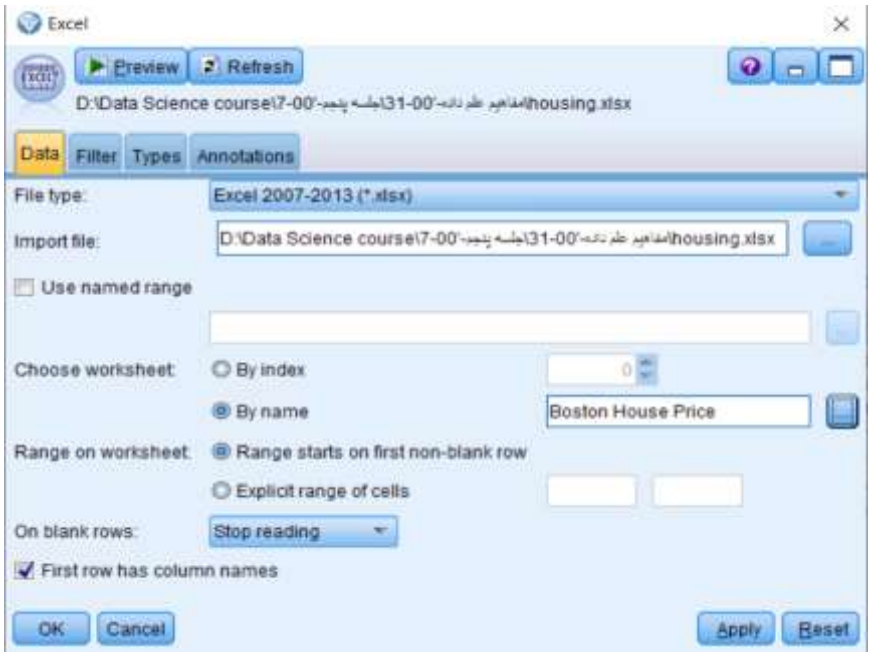
جدول ۲-۳ ده رکورد از فایل xslx آماده شده دیتاست

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	ID	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
2	1	0.00632	18.00	2.310	0	0.5380	6.5750	65.20	4.0900	1	296.0	15.30	396.90	4.98	24.00
3	2	0.02731	0.00	7.070	0	0.4690	6.4210	78.90	4.9671	2	242.0	17.80	396.90	9.14	21.60
4	3	0.02729	0.00	7.070	0	0.4690	7.1850	61.10	4.9671	2	242.0	17.80	392.83	4.03	34.70
5	4	0.03237	0.00	2.180	0	0.4580	6.9980	45.80	6.0622	3	222.0	18.70	394.63	2.94	33.40
6	5	0.06905	0.00	2.180	0	0.4580	7.1470	54.20	6.0622	3	222.0	18.70	396.90	5.33	36.20
7	6	0.02985	0.00	2.180	0	0.4580	6.4300	58.70	6.0622	3	222.0	18.70	394.12	5.21	28.70
8	7	0.08829	12.50	7.870	0	0.5240	6.0120	66.60	5.5605	5	311.0	15.20	395.60	12.43	22.90
9	8	0.14455	12.50	7.870	0	0.5240	6.1720	96.10	5.9505	5	311.0	15.20	396.90	19.15	27.10
10	9	0.21124	12.50	7.870	0	0.5240	5.6310	6.0821	6.0821	5	311.0	15.20	386.63	29.93	16.50

اکنون می‌توانیم وارد گام بعدی شویم.

۲-۲ وارد کردن داده

ابتدا فایل csv مرتب شده را بصورت اکسل ذخیره کرده و آن را می‌بندیم، سپس نرم افزار IBM را باز می‌کنیم. با استفاده از تب Sources، ابزار اکسل را به محیط کار اضافه کره و فایل اکسل مورد نظر را Import می‌کنیم.



شکل ۲-۱ Import کردن داده‌ها در IBM

۱۰ رکورد اول بصورت زیر نمایش داده می‌شوند:

Table Annotations

	ID	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	R...	TAX	PTRATIO	B	LST...	MEDV
1	1.0...	0.00632	18.00	2.310	0	0.5380	6.5750	65.20	4.0900	1	296.0	15.30	396.90	4.98	24.00
2	2.0...	0.02731	0.00	7.070	0	0.4690	6.4210	78.90	4.9671	2	242.0	17.80	396.90	9.14	21.60
3	3.0...	0.02729	0.00	7.070	0	0.4690	7.1850	61.10	4.9671	2	242.0	17.80	392.83	4.03	34.70
4	4.0...	0.03237	0.00	2.180	0	0.4580	6.9980	45.80	6.0622	3	222.0	18.70	394.63	2.94	33.40
5	5.0...	0.06905	0.00	2.180	0	0.4580	7.1470	54.20	6.0622	3	222.0	18.70	396.90	5.33	36.20
6	6.0...	0.02985	0.00	2.180	0	0.4580	6.4300	58.70	6.0622	3	222.0	18.70	394.12	5.21	28.70
7	7.0...	0.08829	12.50	7.870	0	0.5240	6.0120	66.60	5.5605	5	311.0	15.20	395.60	12.43	22.90
8	8.0...	0.14455	12.50	7.870	0	0.5240	6.1720	96.10	5.9505	5	311.0	15.20	396.90	19.15	27.10
9	9.0...	0.21124	12.50	7.870	0	0.5240	5.6310	6.0821	6.0821	5	311.0	15.20	386.63	29.93	16.50
10	10....	0.17004	12.50	7.870	0	0.5240	6.0040	85.90	6.5921	5	311.0	15.20	386.71	17.10	18.90

شکل ۲-۲ ده رکورد اول از فایل اکسل Import شده در IBM

اکنون زمان آن است که Data Type داده‌ها را مشخص کنیم، در بخش ۳-۱ فصل اول توضیحاتی را ارائه کردیم که ما را در تشخیص نوع داده‌ها کمک می‌کند، بر این اساس نوع داده‌ها به شرح زیر است:

جدول ۲-٤ نوع داده‌ها بر اساس تعریف آنها

AGE	RM	NOX	CHAS	INDUS	ZN	CRIM
Continuous	Continuous	Continuous	Flag	Continuous	Continuous	Continuous

MEDV	LSTAT	B	PTRATIO	TAX	DIS	RAD
Continuous	Continuous	Continuous	Continuous	Continuous	Continuous	Ordinal

اما نوع داده‌ای که به ما نشان می‌دهد، بصورت زیر است:

Field	Measurement	Values	Missing	Check	Role
ID	Continuous	[1,506.0]		None	Input
CRIM	Nominal	" 0.00632",...		None	Input
ZN	Nominal	" 0.00 "," 1...		None	Input
INDUS	Nominal	" 0.460"," ...		None	Input
CHAS	Flag	" 1"/" 0"		None	Input
NOX	Nominal	" 0 0.4"," 0...		None	Input
RM	Nominal	" 3.5610"," ...		None	Input
AGE	Nominal	" 2.90 "," ...		None	Input
DIS	Nominal	" 4 33"," ...		None	Input
RAD	Nominal	" 1"," 2"," ...		None	Input
TAX	Nominal	" 256.","18...		None	Input
PTRATIO	Nominal	" 15.1","12....		None	Input
B	Nominal	" 0.32 "," ...		None	Input
LSTAT	Nominal	" 3.95"," ...		None	Input
MEDV	Nominal	" 5.00"," 5....		None	Input

شکل ۲-۳ نوع داده‌های وارد شده در IBM

این موضوع بدین خاطر است که وقتی مقادیر را در sheet جدید بصورت Paste value انتقال دادیم، آنها را بصورت text ذخیره کرده است. برای رفع این مشکل از ابزار filler استفاده می‌کنیم:

شکل ۲-٤ تغییر نوع داده‌ها با استفاده از Filler

بنابراین با این دستور تمام مقادیر بصورت Continuous لحاظ می‌شود، سپس با ابزار Type، نوع داده ویژگی CHAS را به Flag تغییر می‌دهیم.

Field	Measurement	Values	Missing	Check	Role
ID	Continuous	[1.0,506.0]		None	Record ID
CRIM	Continuous	[0.00632,88.9762]		None	Input
ZN	Continuous	[0.0,100.0]		None	Input
INDUS	Continuous	[0.46,27.74]		None	Input
CHAS	Flag	1.0/0.0		None	Input
NOX	Continuous	[0.385,0.871]		None	Input
RM	Continuous	[3.561,8.78]		None	Input
AGE	Continuous	[1.137,99.3]		None	Input
DIS	Continuous	[1.1296,9.2229]		None	Input
RAD	Continuous	[1.0,24.0]		None	Input
TAX	Continuous	[187.0,711.0]		None	Input
PTRATIO	Continuous	[12.6,22.0]		None	Input
B	Continuous	[0.32,396.9]		None	Input
LSTAT	Continuous	[1.73,37.97]		None	Input
MEDV	Continuous	[5.0,50.0]		None	Input

شکل ۲-۵ نمایش و تغییر نوع داده‌ها با استفاده از Type

۲-۳ ساختاربندی و مجتمع کردن داده

در این دیتاست از آنجا که داده‌ها پراکنده نیست و فقط یک دیتاست داریم، بنابراین نیازی به انجام کاری برای مجتمع کردن داده‌ها نیست!

۲-٤ رفع خطا یا نویز

برای پیدا کردن خطا به مقادیرvalues در شکل ۲-۵ توجه می‌کنیم وآن را با جدول ۱-۱ مقایسه می‌کنیم. مشاهده می‌کنیم مغایرتی بین این دو وجود ندارد. لذا دیتاست ما خطایی ندارد.

۲-۵ شناسایی داده‌های پرت

۲-۵-۱ تشخیص توزیع داده‌ها

اولین گام اجرایی در شناسایی داده‌های پرت، مشخص نمودن توزیع داده‌هاست، بنابراین ابتدا با استفاده از Data Audit در تب Output بصورت چشمی توزیع داده‌ها رو مشاهده می‌کنیم.

شکل ۲-٦ نمایش توزیع داده‌ها با Data Audit

اینطور که از شکل‌ها مشخص است، بصورت چشمی ویژگی‌های RM، RAD، MEDV دارای توزیع نرمال به نظر می‌رسند، تشخیص توزیع سایر ویژگی‌ها کمی مبهم است، بدین منظور با استفاده از ابراز Sim Fit در

تب Output و با استفاده انجام آزمون کولموگروف–اسمیرنوف متوجه می‌شویم که توزیع هر یک از ویژگی‌ها بصورت زیر است:

Field	Storage	Status		Distribution
ID	Real	✓		Uniform
CRIM	Real	✓		Lognormal
ZN	Real	✓		Normal
INDUS	Real	✓		Exponential
NOX	Real	✓		Triangular
RM	Real	✓		Gamma
AGE	Real	✓		Normal
DIS	Real	✓		Lognormal
RAD	Real	✓		Exponential
TAX	Real	✓		Lognormal
PTRATIO	Real	✓		Triangular
B	Real	✓		Normal
LSTAT	Real	✓		Gamma
MEDV	Real	✓		Lognormal
CHAS	Real	✓		Categorical

شکل ۲–۷ نمایش توزیع داده‌ها با Sim Fit

۲–۵–۲ شناسایی تعداد داده‌های پرت

می‌توان با استفاده از Data Audit و از تب Quality با تنظیم روی ۳ تا ۵ سیگما یا ۱.۵ تا ۳ IQR، تعداد داده‌های پرت و خیلی پرت (Extreme) را برای هر مؤلفه مشاهده کرد. . از آنجا که ستون MEDV همان تارگت هست نقش آن را None می‌کنیم تا در خروجی Data Audit نمایش داده نشود.

Field	Measurement	Outliers	Extremes
ID	Continuous	0	0
CRIM	Continuous	4	4
ZN	Continuous	14	0
INDUS	Continuous	0	0
CHAS	Flag	--	--
NOX	Continuous	0	0
RM	Continuous	8	0
AGE	Continuous	0	0
DIS	Continuous	0	0
RAD	Continuous	0	0
TAX	Continuous	0	0
PTRATIO	Continuous	0	0
B	Continuous	25	0
LSTAT	Continuous	5	0

شکل ۲–۸ تعداد داده‌های پرت و خیلی پرت با تنظیم روی ۳ تا ۵ سیگما

از شکل ۲-۸ واضح است که لازم است برای مؤلفه‌های CRIM، ZN، RM، B و LSTAT داده‌های پرت و خیلی پرت را مدیریت کنیم.

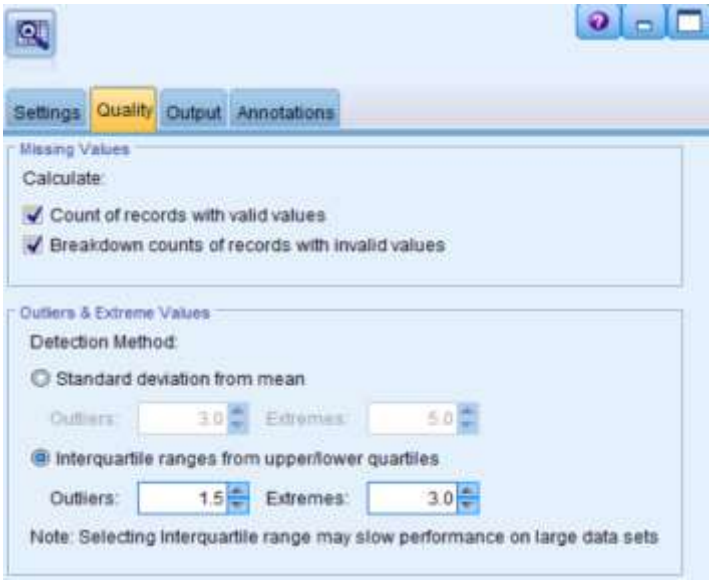
Field	Measurement	Outliers	Extremes
ID	Continuous	0	0
CRIM	Continuous	36	30
ZN	Continuous	23	45
INDUS	Continuous	0	0
CHAS	Flag	--	--
NOX	Continuous	0	0
RM	Continuous	29	1
AGE	Continuous	0	0
DIS	Continuous	0	0
RAD	Continuous	0	0
TAX	Continuous	0	0
PTRATIO	Continuous	15	0
B	Continuous	18	58
LSTAT	Continuous	6	0

شکل ۲-۹ تعداد داده‌های پرت و خیلی پرت با تنظیم روی ۱.۵ تا ۳ IQR

از شکل ۲-۹ واضح است که لازم است برای مؤلفه‌های CRIM، ZN، RM، PTRATIO، B و LSTAT داده‌های پرت و خیلی پرت را مدیریت کنیم. بنابراین تنها در مورد این مؤلفه‌ها در ادامه بحث می‌کنیم.

۲-۵-۳ استفاده از روش ۱.۵ تا ۳ IQR

برای مؤلفه‌های RM، PTRATIO و LSTAT با استفاده از Data Audit و با انجام تنظیمات زیر، داده‌ها پرت و خیلی پرت را از روش ۱.۵ تا ۳ IQR مدیریت می‌کنیم:



شکل ۲-۱۰ تنظیم ۱.۵ تا ۳ IQR در Data Audit

حال برای مؤلفه‌های فوق الذکر از طریق Action هایی که در شکل زیر نمایش داده شده است سوپر نود Outlier & Extreme را ایجاد می‌کنیم:

Field	Measurement	Outliers	Extremes	Action
ID	Continuous	0	0	None
CRIM	Continuous	36	30	None
ZN	Continuous	23	45	None
INDUS	Continuous	0	0	None
CHAS	Flag	--	--	--
NOX	Continuous	0	0	None
RM	Continuous	29	1	Coerce outliers / nullify extremes
AGE	Continuous	0	0	None
DIS	Continuous	0	0	None
RAD	Continuous	0	0	None
TAX	Continuous	0	0	None
PTRATIO	Continuous	15	0	Coerce
B	Continuous	18	58	None
LSTAT	Continuous	6	0	Coerce

شکل ۲-۱۱ مدیریت داده‌های پرت و خیلی پرت مؤلفه‌های غیر نرمال

۲-۵-٤ استفاده از روش ۲ تا ٤ IQR

در بازه ۲ تا ٤ IQR داده‌های پرت و خیلی پرت مؤلفه‌های غیر نرمال را مجدداً مدیریت می‌کنیم:

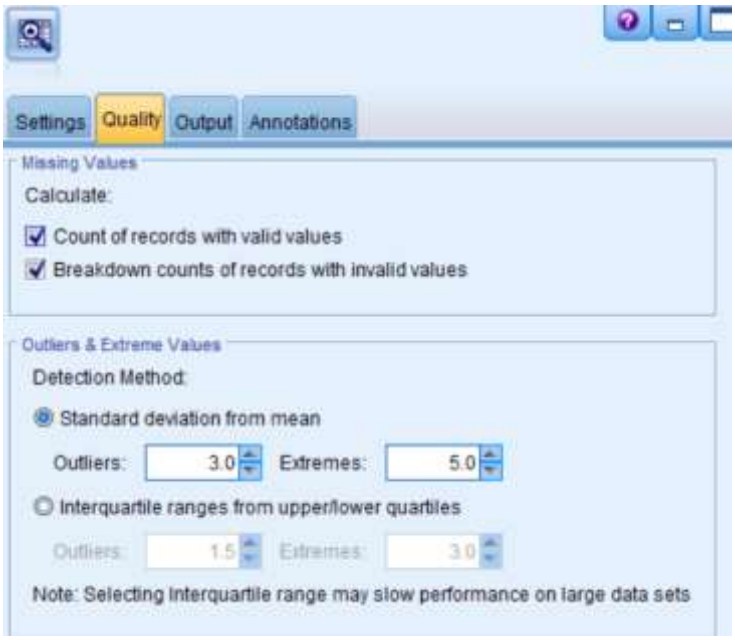
Field	Measurement	Outliers	Extremes	Action
ID	Continuous	0	0	None
CRIM	Continuous	28	22	None
ZN	Continuous	23	35	None
INDUS	Continuous	0	0	None
CHAS	Flag	--	--	--
NOX	Continuous	0	0	None
RM	Continuous	0	0	None
AGE	Continuous	0	0	None
DIS	Continuous	0	0	None
RAD	Continuous	0	0	None
TAX	Continuous	0	0	None
PTRATIO	Continuous	0	0	None
B	Continuous	17	52	None
LSTAT	Continuous	0	0	None

شکل ۲-۱۲ مدیریت مجدد داده‌های پرت و خیلی پرت مؤلفه‌های غیر نرمال

واضح است که برای سه مؤلفه‌ی RM، PTRATIO و LSTAT هیچ داده پرت یا خیلی پرتی نداریم.

۲-۵-۵ استفاده از روش ۳ تا ۵ سیگما

برای مؤلفه‌هایی که دارای توزیع نرمال و لاگ نرمال هستند از این روش برای کشف نقاط پرت و خیلی پرت استفاده می‌کنیم. بنابراین با استفاده از Data Audit تنظیمات زیر را انجام می‌دهیم:



شکل ۲-۱۳ تنظیم ۳ تا ۵ سیگما در Data Audit

حال برای مؤلفه‌های CRIM، ZN و B از طریق Action هایی که در شکل زیر نمایش داده شده است سوپر نود Outlier & Extreme را ایجاد می‌کنیم، بهتر است از قبل از ایجاد سوپر نود جدید، ابتدا با ابزار Type داده‌ها را Read values کنیم و سپس سوپر نود را ایجاد کنیم (دقت کنید در این مرحله فقط به مؤلفه‌های CRIM، ZN و B توجه می‌کنیم به سایر مؤلفه‌ها کاری نداریم).

Field	Measurement	Outliers	Extremes	Action
ID	Continuous	0	0	None
CRIM	Continuous	4	4	Coerce outliers / nullify extremes
ZN	Continuous	14	0	Coerce
INDUS	Continuous	0	0	None
CHAS	Flag	--	--	--
NOX	Continuous	0	0	None
RM	Continuous	0	0	None
AGE	Continuous	0	0	None
DIS	Continuous	0	0	None
RAD	Continuous	0	0	None
TAX	Continuous	0	0	None
PTRATIO	Continuous	0	0	None
B	Continuous	25	0	Coerce
LSTAT	Continuous	0	0	None

شکل ۲-۱٤ مدیریت داده‌های پرت و خیلی پرت مولفه‌های نرمال یا لاگ نرمال

۲-۵-٦ استفاده از روش ٤ تا ٦ سیگما

در این مرحله، مجدداً با استفاده از ابزار Data Audit، اما این بار در بازه ٤ تا ٦ سیگما، تعداد داده‌های پرت و خیلی پرت را بررسی می‌کنیم.

Field	Measurement	Outliers	Extremes	Action
ID	Continuous	0	0	None
CRIM	Continuous	6	0	Coerce
ZN	Continuous	0	0	None
INDUS	Continuous	0	0	None
CHAS	Flag	--	--	--
NOX	Continuous	0	0	None
RM	Continuous	0	0	None
AGE	Continuous	0	0	None
DIS	Continuous	0	0	None
RAD	Continuous	0	0	None
TAX	Continuous	0	0	None
PTRATIO	Continuous	0	0	None
B	Continuous	0	0	None
LSTAT	Continuous	0	0	None

شکل ۲-۱۵ مدیریت مجدد داده‌های پرت و خیلی پرت مؤلفه‌های نرمال و لاگ نرمال

مشاهده می‌شود مؤلفه‌ی CRIM همچنان دارای داده پرت است. بنابراین با استفاده از Coerce آن را مدیریت کرده و سوپر نود جدید را ایجاد می‌کنیم. اگر یکبار دیگرروی بازه بزرگتر مثلا ۵ تا ۷ سیگما با Data Audit خروجی بگیریم، ملاحظه می‌شود هیچ داده پرتی نداریم.

Field	Measurement	Outliers	Extremes	Action
ID	Continuous	0	0	None
CRIM	Continuous	0	0	None
ZN	Continuous	0	0	None
INDUS	Continuous	0	0	None
CHAS	Flag	--	--	--
NOX	Continuous	0	0	None
RM	Continuous	0	0	None
AGE	Continuous	0	0	None
DIS	Continuous	0	0	None
RAD	Continuous	0	0	None
TAX	Continuous	0	0	None
PTRATIO	Continuous	0	0	None
B	Continuous	0	0	None
LSTAT	Continuous	0	0	None

شکل ۲-۱۶ نمایش مدیریت داده‌های پرت کل مؤلفه‌ها

پس کار شناسایی داده‌های پرت به اتمام می‌رسد.

۲-۶ شناسایی داده‌های مفقوده

۲-۶-۱ تشخیص توزیع داده‌ها

قبل از ایجاد سوپر نود داده‌های مفقوده، از آنجا که عملیاتی برای شناسایی و مدیریت داده‌های پرت و خیلی پرت انجام دادیم، بهتر از Sim Fit بار دیگر استفاده کنیم تا توزیع داده ها را تشخیص دهیم.

Field	Storage	Status		Distribution
ID	Real			Uniform
CRIM	Real			Lognormal
ZN	Real			Normal
INDUS	Real			Exponential
NOX	Real			Triangular
RM	Real			Lognormal
AGE	Real			Normal
DIS	Real			Lognormal
RAD	Real			Exponential
TAX	Real			Lognormal
PTRATIO	Real			Lognormal
B	Real			Weibull
LSTAT	Real			Gamma
MEDV	Real			Lognormal
CHAS	Real			Categorical

شکل ۲-۱۷ نمایش مجدد توزیع داده‌ها با Sim Fit

۲-۶-۲ شناسایی درصد داده‌های مفقوده

ابتدا با استفاده از Data Audit و تنظیم ۵ تا ۷ سیگما، درصد داده‌های مفقوده هر فیلد را بررسی می‌کنیم.

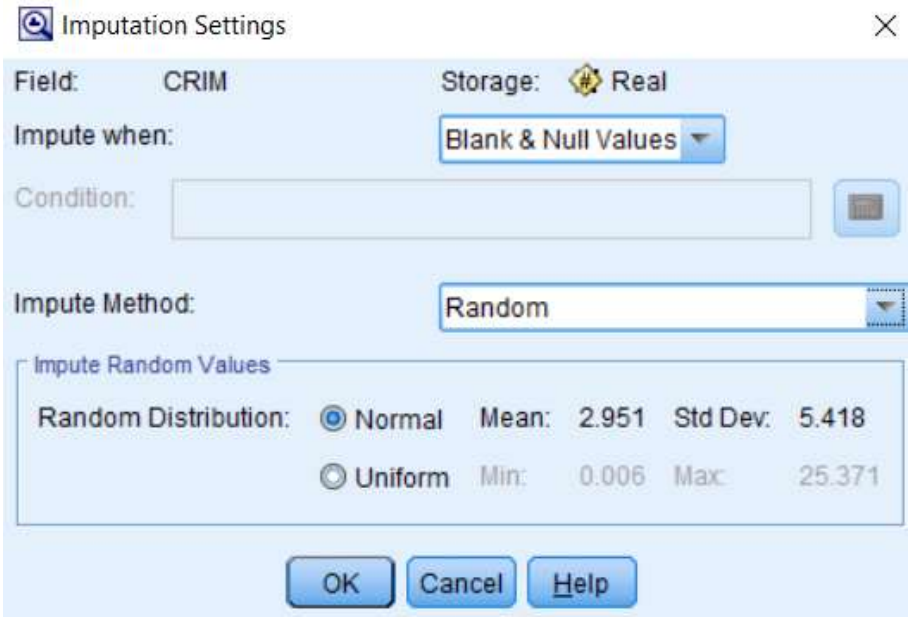
Field	Measurement	% Complete	Impute Missing	Method
ID	Continuous	100	Never	Fixed
CRIM	Continuous	99.209	Never	Fixed
ZN	Continuous	100	Never	Fixed
INDUS	Continuous	100	Never	Fixed
CHAS	Flag	100	Never	Fixed
NOX	Continuous	99.802	Never	Fixed
RM	Continuous	99.802	Never	Fixed
AGE	Continuous	100	Never	Fixed
DIS	Continuous	99.012	Never	Fixed
RAD	Continuous	100	Never	Fixed
TAX	Continuous	100	Never	Fixed
PTRATIO	Continuous	100	Never	Fixed
B	Continuous	100	Never	Fixed
LSTAT	Continuous	100	Never	Fixed

شکل ۲-۱۸ درصد داده‌های مفقوده مؤلفه‌ها

واضح است که برای مؤلفه‌های CRIM، NOX، RM و DIS داده مفقوده داریم. بنابراین تنها برای این مؤلفه‌ها مدیریت داده مفقوده انجام می‌دهیم.

۲-۶-۳ استفاده از روش Random-Normal

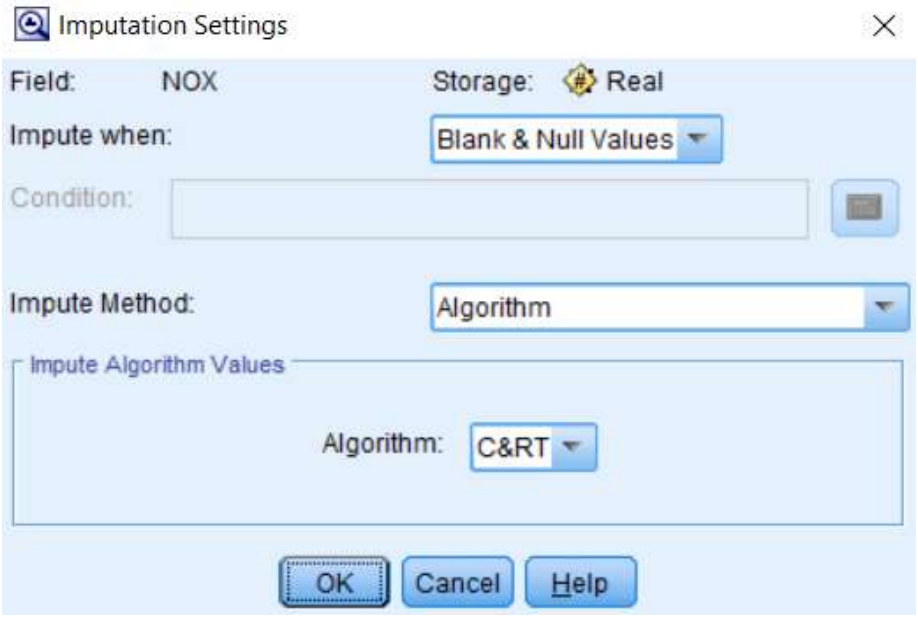
برای مؤلفه‌هایی که نرمال یا لاگ نرمال هستند یعنی CIRM، RM و DIS از این روش استفاده می‌کنیم، به عنوان مثال در شکل زیر برای مؤلفه CRIM چگونگی تنظیم Impute Missing نمایش داده شده است.



شکل ۲-۱۹ مدیریت داده‌های مفقوده با توزیع نرمال یا لاگ نرمال

۲-۶-٤ استفاده از روش Algorithm

برای مدیریت داده‌های مفقوده فیلد NOX از روش الگوریتم استفاده می‌کنیم. این روش مبتنی بر الگوریتم CART است.



شکل ۲-۲۰ مدیریت داده‌های مفقوده با توزیع غیر نرمال

در نهایت سوپر نود Missing value imputation را ایجاد کرده و با استفاده از Data Audit و تنظیم ۵ تا ۷ سیگما خروجی به صورت زیر است:

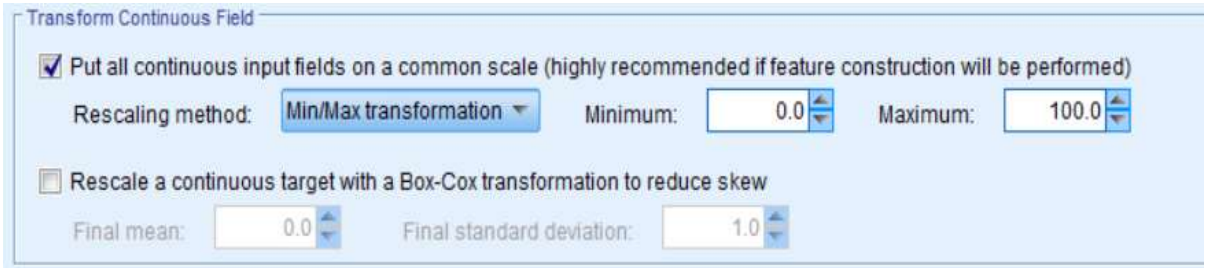
Field	Measurement	Outliers	Extremes	% Complete
ID	Continuous	0	0	100
CRIM	Continuous	0	0	100
ZN	Continuous	0	0	100
INDUS	Continuous	0	0	100
CHAS	Flag	--	--	100
NOX	Continuous	0	0	100
RM	Continuous	0	0	100
AGE	Continuous	0	0	100
DIS	Continuous	0	0	100
RAD	Continuous	0	0	100
TAX	Continuous	0	0	100
PTRATIO	Continuous	0	0	100
B	Continuous	0	0	100
LSTAT	Continuous	0	0	100

شکل ۲-۲۱ نمایش مدیریت داده‌های پرت و مفقوده کل مؤلفه‌ها

در اینجا کار مدیریت و شناسایی داده‌های مفقوده به پایان می‌رسد.

۲-۷ هم مقیاس سازی داده‌ها

با استفاده از Auto Data Prep در بخش Field Ops داده‌ها را هم مقیاس می‌کنیم. از روش Min/Max استفاده می‌کنیم و داده ها رو به مقیاس ۰ و ۱۰۰ میبریم. توجه کنید که فیلد ID را انتخاب نمی‌کنیم و فیلد MEDV را به عنوان تارگت تعیین می‌کنیم.



شکل ۲-۲۲ هم مقیاس سازی با روش Min/Max

سر آخر بر اساس ID داده ها رو Sort می‌کنیم. شکل زیر ۱۰ داده اول تمیز شده و مرتب شده را نشان می‌دهد.

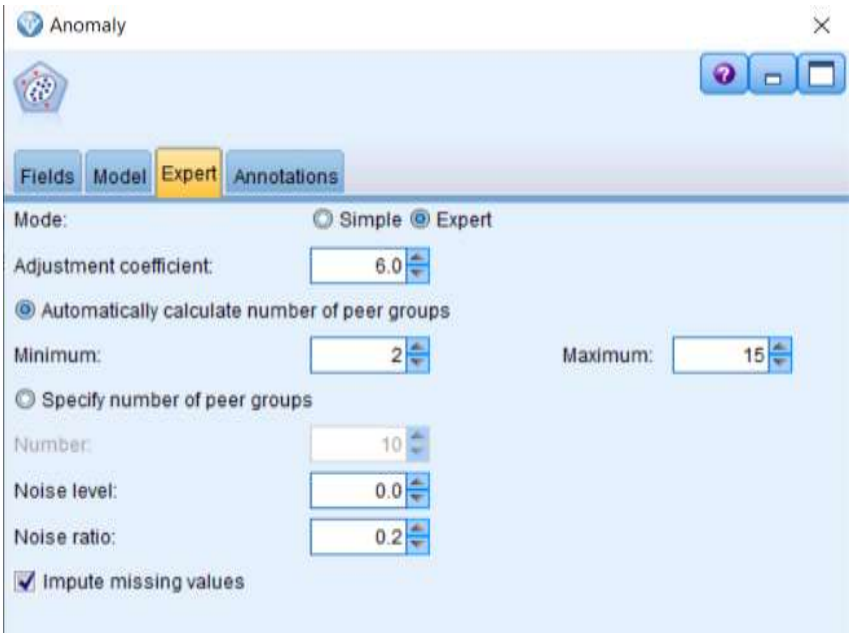
Table Annotations

	ID	CHAS	CRIM_t...	ZN_tran...	INDUS...	NOX_tr...	RM_tr...	AGE_tr...	DIS_tra...	RAD_tr...	TAX_tr...	PTRATI...	B_transf...	LSTAT_...	MEDV
1	1.000	0.000	9.987	22.132	6.782	31.481	60.811	65.262	44.937	0.000	20.802	23.864	100.000	10.735	24.000
2	2.000	0.000	10.062	0.000	24.230	17.284	55.608	79.218	54.346	4.348	10.496	52.273	100.000	24.476	21.600
3	3.000	0.000	10.061	0.000	24.230	17.284	81.419	61.085	54.346	4.348	10.496	52.273	98.704	7.597	34.700
4	4.000	0.000	10.080	0.000	6.305	15.021	75.101	45.499	66.094	8.696	6.679	62.500	99.277	3.997	33.400
5	5.000	0.000	10.210	0.000	6.305	15.021	80.135	54.056	66.094	8.696	6.679	62.500	100.000	11.891	36.200
6	6.000	0.000	10.071	0.000	6.305	15.021	55.912	58.640	66.094	8.696	6.679	62.500	99.115	11.495	28.700
7	7.000	0.000	10.278	15.369	27.163	28.601	41.791	66.688	60.712	17.391	23.664	22.727	99.586	35.343	22.900
8	8.000	0.000	10.478	15.369	27.163	28.601	47.196	96.740	64.896	17.391	23.664	22.727	100.000	57.539	27.100
9	9.000	0.000	10.714	15.369	27.163	28.601	28.919	5.038	66.307	17.391	23.664	22.727	96.730	93.146	16.500
10	10.000	0.000	10.568	15.369	27.163	28.601	41.520	86.349	71.778	17.391	23.664	22.727	96.756	50.768	18.900

شکل ۲-۲۳ نمایش ده رکورد اول داده‌های تمیز و مرتب شده

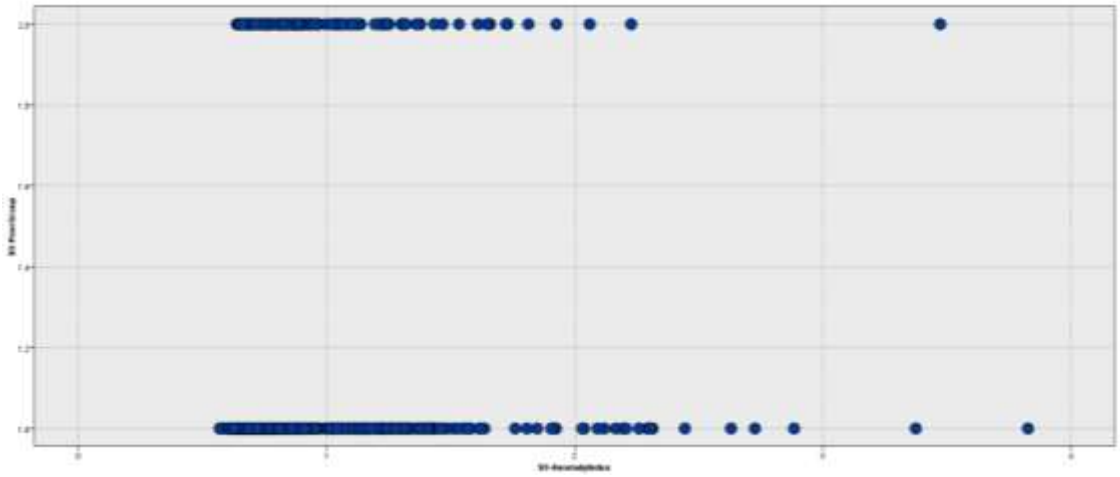
۲-۸ تشخیص ناهنجاری

برای تشخیص ناهنجاری کافیست ابزار Anomaly را از تب Modeling انتخاب کرده و بعد از اتصال Run کنیم، همچنین میتوان با تنظیمات Expert از ابزار بخواهیم نتایج ناهنجاری بین خوشه ۲ تا ۱۵ را نمایش دهد.



شکل ۲-۲٤ نمایش تنظیمات Anomaly

خروجی به ما یک الماس نشان می‌دهد. با استفاده از ابزار Plot موجود در تب Graph لازم است برای تشخیص ناهنجاری، بر اساس $O-AnomalyIndex (به عنوان x field) که نمایانگر فاصله بین هر رکورد و نقطه مرکزی گروه است و $O- PeerGroup (به عنوان y field) که نمایانگر این است که هر رکورد متعلق به کدام گروه است، نمودار دو بعدی رسم می‌کنیم.



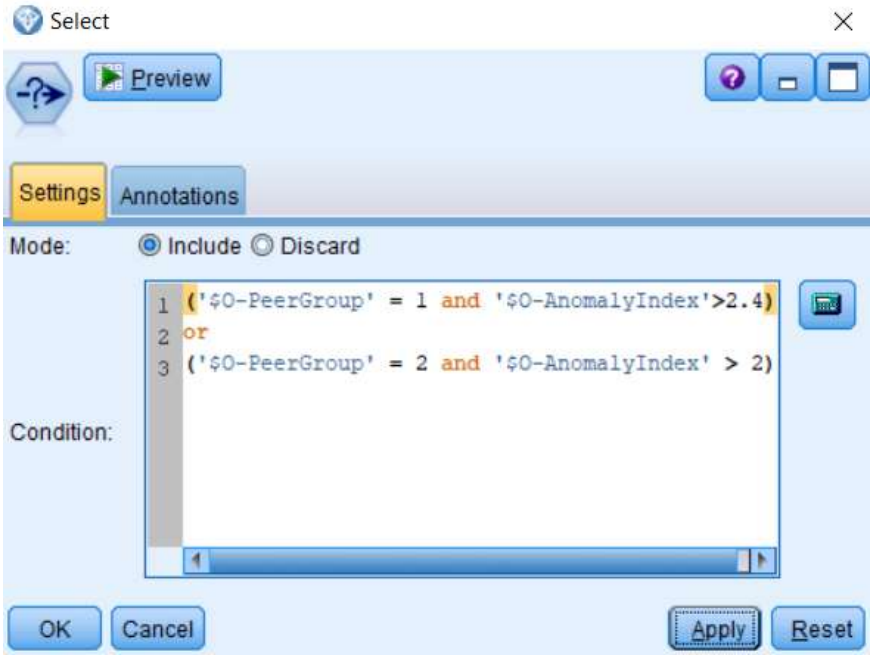
شکل ۲-۲۵ نمودار Anomaly

برای خوشه اول و دوم، ناهنجاری را بصورت زیر تعریف می‌کنیم:

خوشه اول: اگر $O-AnomalyIndex بیش از ۲.٤ باشد

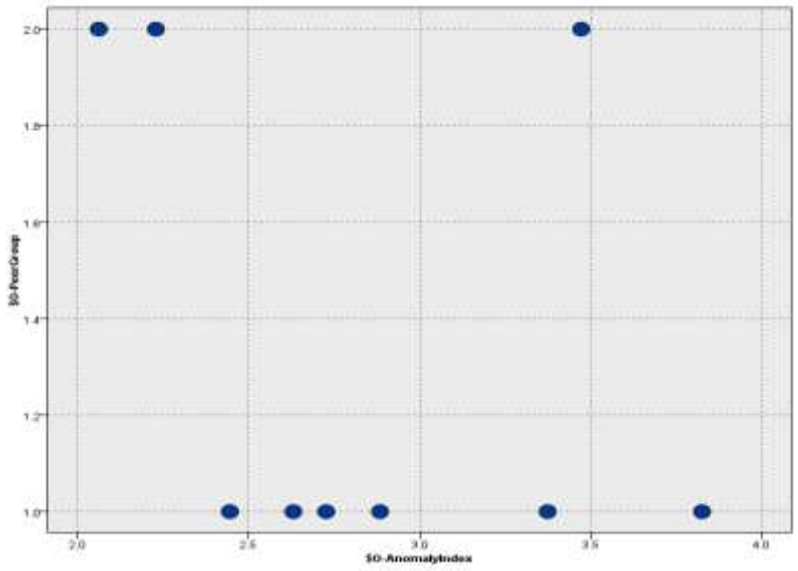
خوشه دوم: اگر $O-AnomalyIndex بیش از ۲ باشد

حال با استفاده از ابزار Select دستورات فوق را مطابق شکل زیر اجرا می‌کنیم:



شکل ۲-۲٦ حذف Anomaly با استفاده از Select

اگر از Select یک Plot بگیریم بر اساس $O-AnomalyIndex و $O- PeerGroup، داده‌های پرت بصورت زیر قابل ملاحظه هستند.



شکل۲-۲۷ تعداد Anomaly ها بعد از اجرای Select

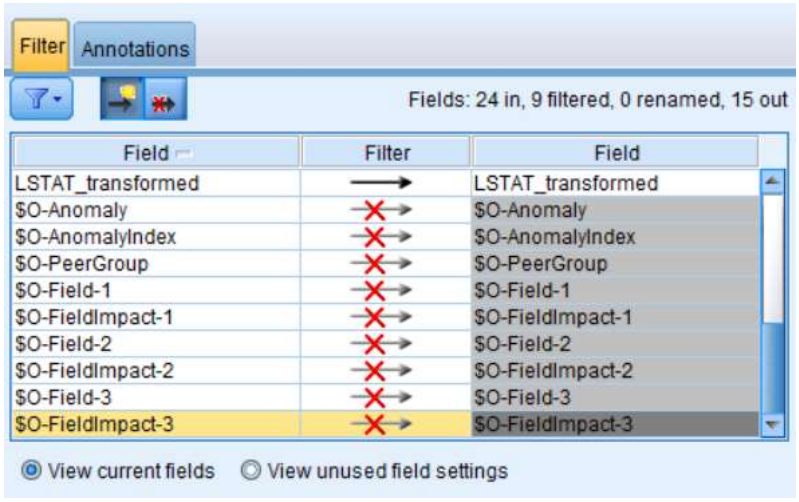
همچنین اگر از select با Table خروجی بگیریم داریم:

Table Annotations

	ID	$O-Anomaly	$O-AnomalyIndex	$O-PeerGroup	$O-Field-1	$O-Field-2	$O-Field-3
1	103....	F	2.445	1	B_transformed	PTRATIO_transform...	AGE_transformed
2	143....	T	2.727	1	CHAS	NOX_transformed	LSTAT_transfor...
3	146....	T	2.884	1	B_transformed	NOX_transformed	LSTAT_transfor...
4	147....	F	2.630	1	B_transformed	NOX_transformed	AGE_transformed
5	156....	T	3.827	1	B_transformed	CHAS	NOX_transformed
6	157....	T	3.375	1	B_transformed	NOX_transformed	RM_transformed
7	365....	F	2.228	2	CHAS	RM_transformed	LSTAT_transfor...
8	371....	F	2.061	2	CHAS	LSTAT_transformed	RM_transformed
9	491....	T	3.473	2	RAD_transformed	INDUS_transformed	CRIM_transform...

شکل ۲-۲۸ نمایش جزئیات رکوردهای ناهنجار

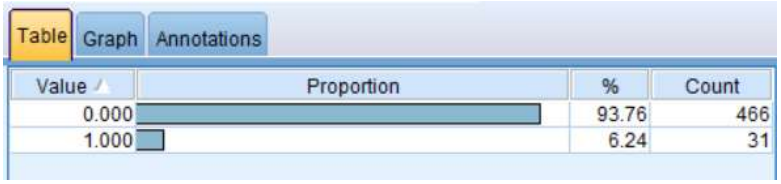
در این جدول واضح است که رکوردهای ۱۰۳، ۱۴۳، ۱۴۶، ۱۴۷، ۱۵۶، ۱۵۷، ۳۶۵، ۳۷۱ و ۴۹۱ از طریق دستور Select از سایر داده‌ها جدا شدند و در این بین تنها رکوردهای ۱۴۳، ۱۴۶، ۱۵۶، ۱۵۷ و ۴۹۱ به عنوان ناهنجاری یا Anomaly در سر ستون $O-Anomaly درست (True) تشخیص داده شده‌اند. همچنین این خروجی نشان می‌دهد، به عنوان مثال رکورد ۱۴۳، ابتدا بر اساس مؤلفه CHAS و سپس بر اساس مؤلفه NOX-transformed و در نهایت بر اساس مؤلفه LSTAT-transformed ناهنجار شناخته شده است. ما این داده ها را باید مدیریت کنیم. این کار دقت مدل را افزایش می‌دهد. بدین منظور با همان دستور Select و با تغییر گزینه از Include به Discard این امر امکان پذیر است. توجه کنید که برای ادامه مدلسازی باید یکسری فیلدهای اضافه را با استفاده از ابزار Filter حذف کرد:



شکل ۲-۲۹ حذف فیلدهای اضافی

۲-۹ مدیریت داده‌های نامتوازن

برای مدیریت داده‌های نامتوازن، از تب گراف، ابزار Distribution را انتخاب می‌کنیم و به Filter متصل می‌کنیم. بعد از انتخاب مؤلفه CHAS که از نوع Flag هست، ابزار را Run می‌کنیم و خروجی زیر نمایش داده می‌شود:



شکل ۲-۳۰ داده‌های نامتوازن

ملاحظه می‌شود که تعداد صفرها خیلی بیشتر از یک هست و این مؤلفه شدیداً نامتوازن هست. به همین منظور دو تا بالانس نود ایجاد می‌کنیم، یکی از نوع Boost و دیگری Reduce، با این نود ها کاری انجام نمی‌دهیم و تنها در انتها کار بوسیله حذف یا اضافه آنها دقت مدل را ارزیابی می‌کنیم.

۲-۱۰ نتیجه‌گیری

در این فصل به تمیز کردن داده‌ها و آماده سازی آنها قبل از انجام مدلسازی پرداختیم. ملاحظه کردیم که در این دیتاست، داده‌ها برای تمیز شدن باید از هفت مرحله اصلی یعنی ساختاربندی و مجتمع کردن داده‌ها، رفع خطا یا نویز، شناسایی داده‌های پرت، شناسایی داده‌های مفقوده، هم مقیاس سازی داده‌ها، تشخیص ناهنجاری و مدیریت داده‌های نامتوازن عبور کنند. در فصل بعدی به انجام مدلسازی می‌پردازیم.

فصل ۳: مدلسازی

۳-۱ مقدمه

در این فصل به منظور مدلسازی، ابتدا با استفاده از ابزار Partition که از تب Field Ops انتخاب کردیم، دادها را به دو بخش Train و Test با نسبت ۸۰ به ۲۰ تقسیم می‌کنیم، .

۳-۲ داده‌های آموزش و آزمایش

برای تقسیم بندی داده‌های آموزش و آزمایش از ابزار Partition استفاده می‌کنیم. نحوه تقسیم بندی با نسبت ۸۰ به ۲۰ است که با هر بار Run گرفتن برای هر الگوریتم بصورت تصادفی تغییر می‌کند. با کلیک راست روی Partition و با enable سازی گزینه Cache می‌توانیم داده‌ها را فیکس کنیم تا تغییری در تقسیم بندی ۸۰ به ۲۰ رخ ندهد. پارتیشن مورد استفاده در شکل زیر نمایش داده شده است. لازم به ذکر است که قبل از تقسیم بندی داده‌ها، تارگت را مشخص می‌کنیم تا در مراحل بعدی به مشکل برنخوریم. بنابراین نقش MEDV را از None به Target تغییر می‌دهیم.

شکل ۳-۱ تقسیم بندی داده‌های آموزش و آزمایش

۳-۳ انتخاب بهترین الگوریتم

هدف مسئله پیش‌بینی قیمت مسکن است و در مراحل قبل نیز MEDV را تارگت قرار دادیم. MEDV متوسط قیمت خانه‌های دارنده مالک (یعنی خانه‌هایی که مالک در آنجا زندگی می‌کند) بر اساس هزار دلار می‌باشد. از آنجا که MEDV یک ویژگی پیوسته است، برای انتخاب بهترین الگوریتم از ابزار Auto Numeric جهت مدلسازی استفاده می‌کنیم و مدل‌ها را بصورت Default می‌سازیم. با انجام تنظیمات زیر برای Auto Numeric، خروجی‌های زیر نمایش داده شده‌است.

شکل ۳-۲ ساخت مدل با استفاده از Auto Numeric

همانطور که مشخص است انواع مدل‌هایی که انتخاب کردیم عبارتند از Regression، Generalized Linear، SVM، Linear، CHAID، C&R Tree و Neural Net که بصورت Default مدل مرتبط را ایجاد کردیم، از تب مدل تعداد مدلها را برابر با ۱۰ قرار دادیم. خروجی برای داده‌های آموزش و داده‌های آزمایش در شکل‌های زیر نمایش داده شده است:

شکل ۳-۳ خروجی Auto Numeric برای داده‌های آموزش

شکل ۳-٤ خروجی Auto Numeric برای داده‌های آزمایش

اختلاف بین Correlation داده‌های آموزش و آزمایش برای هر یک از مدل‌های ایجاد شده در جدول زیر نوشته شده است.

جدول ۳-۱ اختلاف بین دقت داده‌های آموزش و آزمایش

Subtraction	Testing	Training	Model
0.021	0.932	0.953	Neural Net
0.014	0.929	0.943	C&R Tree
0.005	0.923	0.928	CHAID
-0.043	0.908	0.865	SVM
-0.033	0.887	0.854	Regression
-0.033	0.887	0.854	Generalized Linear
-0.029	0.869	0.84	Linear

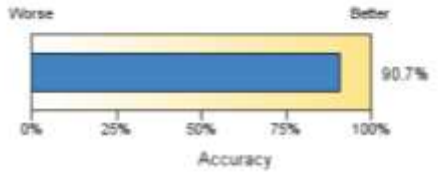
اینکه اختلاف منفی شده است به این علت است که وقتی داده‌های با پارتیشن جدا می‌شوند، ممکن است داده‌های آزمایش، داده‌های مشترک با آموزش داشته باشد و آنها را حفظ کرده باشد به این ترتیب دقت آنها می‌تواند بیشتر از داده‌های آموزش شود. بهتر است در پروژه‌های صنعتی از روش K-Fold بجای Hold out که همان تقسیم بندی ۸۰ به ۲۰ است استفاده شود. به هر حال خیلی این مسئله برای ما تفاوتی نمی‌کند و لذا قدرمطلق اختلاف را اگر در نظر بگیریم، کمترین میزان مربوط به مدل CHAID هست. در ادامه برای هر یک از مدلهای فوق الذکر نتایج مربوطه را تجزیه و تحلیل می‌کنیم.

۳-۳-۱ Neural Net

نتایج مدل Neural Net در شکل‌های زیر نمایش داده شده است.

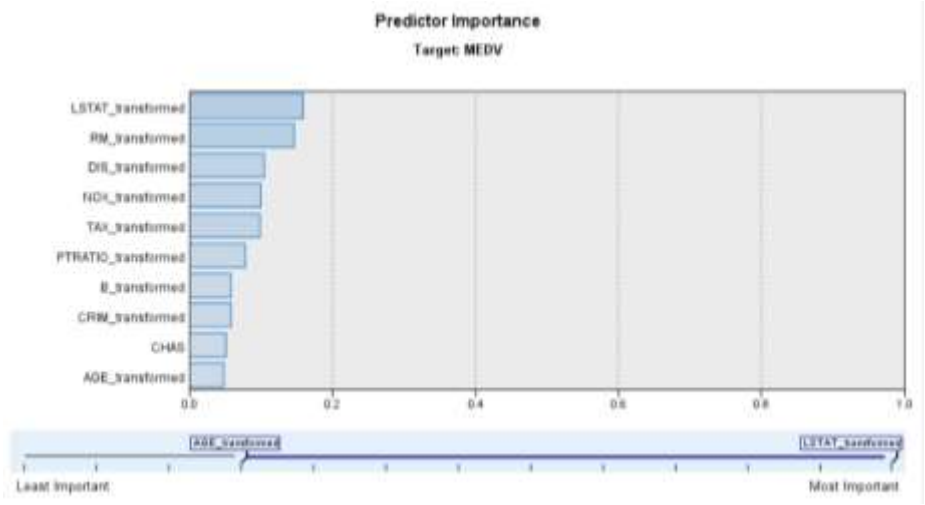
Model Summary

Target	MEDV
Model	Multilayer Perceptron
Stopping Rule Used	Error cannot be further decreased
Hidden Layer 1 Neurons	7



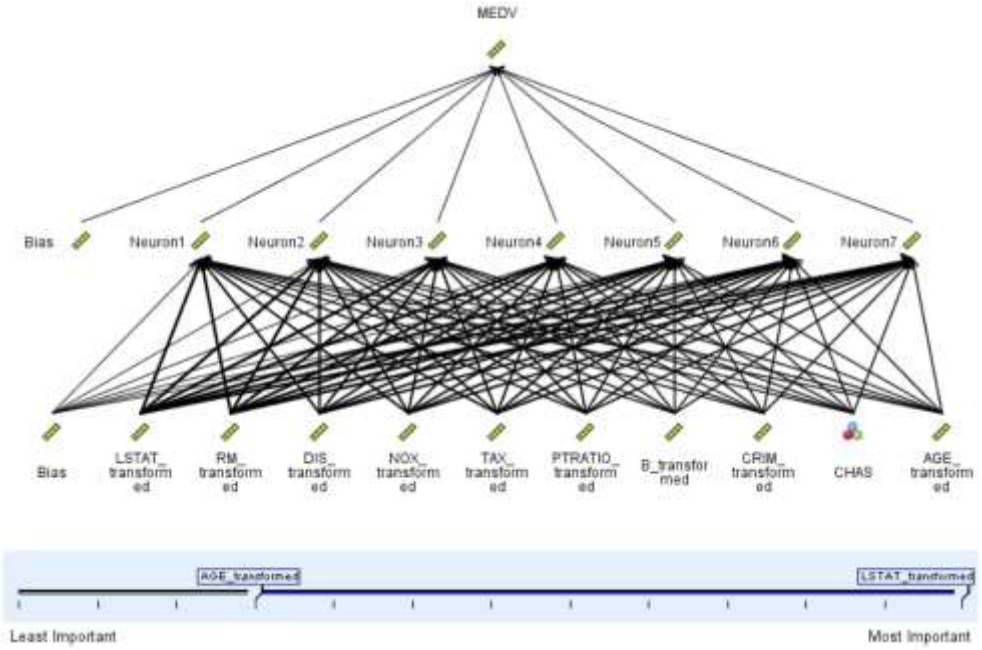
شکل ۳-۵ خلاصه مدل Neural Net

شکل زیر مهم ترین ویژگی‌هایی که در تخمین و پیش‌بینی MEDV در الگوریتم NN تأثیر می‌گذارد را به ترتیب اولویت نمایش می‌دهد.



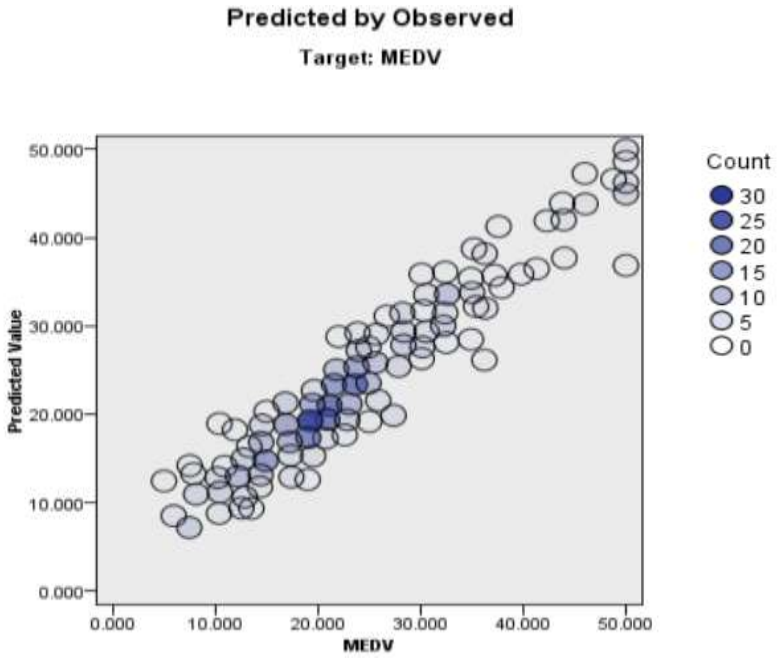
شکل ۳-٦ ترتیب اهمیت ویژگی‌ها برای تخمین MEDV

همانطور که از شکل ۳-٦ واضح است، ویژگی درصد قشر طبقه پایین جمعیت یا LSTAT نسبت به سایر ویژگی‌ها برای تخمین قیمت خانه بسیار حائز اهمیت تر است. بعد از آن متوسط تعداد اتاق‌های هر مسکن یا RM است که برای تخمین قیمت مسکن اهمیت پیدا می‌کند. ویژگی میانگین وزنی فاصله تا پنج مرکز شغلی شهر بوستون یا DIS در رده سوم این دسته بندی قرار دارد. به نظر این دسته بندی منطقی و عقلانی به نظر می‌رسد. همچنین کم اهمیت ترین ویژگی مربوط به تعداد واحدهای مسکونی که مالک در آنها زندگی می‌کند و قبل از سال ۱۹٤۰ ساخته شده‌اند می‌باشد. شکل زیر دسته بندی فوق را بر مبتنی بر روش Neutal Net با ۷ لایه نورون نمایش می‌دهد:



شکل ۳–۷ ساز و کار Neural Net برای اولویت بندی ویژگی‌ها

مقادیر پیش‌بینی شده برای MEDV از طریق مدل Neural Net در شکل ۳–۸ نمایش داده شده است، همچنین مقادیر عددی آن در جدول ۳–۲ گردآوری شده است:



شکل ۳–۸ ساز و کار Neural Net برای اولویت بندی ویژگی‌ها

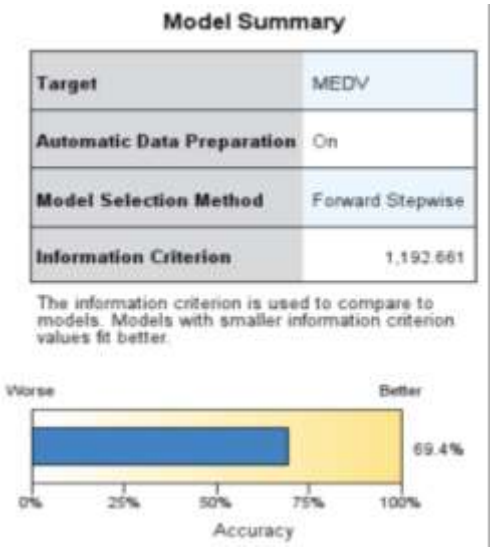
در شکل فوق به عنوان مثال برای MEDV برابر با 19.25 مقدار 19.2262 به تعداد ۲۶ بار تخمین صورت گرفته است.

جدول ۳-۲ مقدار و تعداد دفعات پیش‌بینی ویژگی MEDV

MEDV	Predicted Value	Count	MEDV	Predicted Value	Count
21.1105	20.9488	19	10.44	11.1199	5
19.45	21.111	10	8.125	10.8927	4
16.82	21.2398	5	13.5	9.2776	1
15	20.2842	1	12.5333	9.3166	3
23.0778	21.1118	9	10.35	8.7154	2
25.8667	21.5156	3	5.95	8.4377	2
20.9222	19.3538	18	7.46	7.1042	5
19.25	19.2262	26	21.5692	23.3478	13
16.8818	18.7867	11	19.6	22.6533	3
14.525	18.7222	4	23.6111	23.3301	18
11.9	18.1769	1	24.925	23.4925	8
10.4	18.8957	1	21.7857	25.0139	7
22.86	19.3185	5	23.7545	25.3133	11
25	19.1287	1	25.6	25.826	7
27.3667	19.8153	3	27.85	25.3747	4
20.7333	17.2199	3	30.1	26.2736	1
19.0625	17.2992	16	36.2	26.1101	1
17.2556	16.8196	9	23.98	27.1603	5
14.475	16.7787	8	24.9	27.5812	2
13.3	16.2966	1	28.32	27.8109	5
22.6	17.6038	1	30.175	27.591	4
19.5333	15.2149	3	32.5	28.0895	1
17.3	15.2741	3	34.9	28.448	1
14.8846	14.6757	13	22	28.7593	1
12.74	14.8731	5	23.85	29.1811	2
10.9	14.058	1	25.7	29.0454	2
7.5	14.1411	1	28.35	29.3403	4
19	12.5384	1	30.5667	29.3306	3
17.4	12.8206	3	32.2	29.9847	5
14.4	13.0688	6	26.7	31.0909	1
12.1875	12.9323	8	28.225	31.4441	4
10.225	12.7609	4	30.3	31.7314	1
7.85	13.2157	2	32.4	31.434	3
5	12.3905	1	35.4	32.1845	1
14.35	11.7214	2	36.3	31.991	2
12.8667	10.5398	3	30.55	33.5524	2
32.5714	33.5673	7	44	37.6866	1
35.025	33.7529	4	37.6	41.2627	1
37.9	34.32	1	42.3	41.8968	1
30.1	35.8496	1	43.95	41.9886	2
32.4	36.0748	1	43.8	43.9155	1
34.9	35.3614	1	46.05	43.812	2
37.2	35.7042	1	50	44.9602	4
39.8	35.8528	1	46	47.2847	1
41.3	36.4642	1	48.8	46.5982	1
50	36.8338	1	50	46.2501	3
35.2	38.7255	1	50	48.6107	1
36.25	38.1632	2	50	49.9512	3

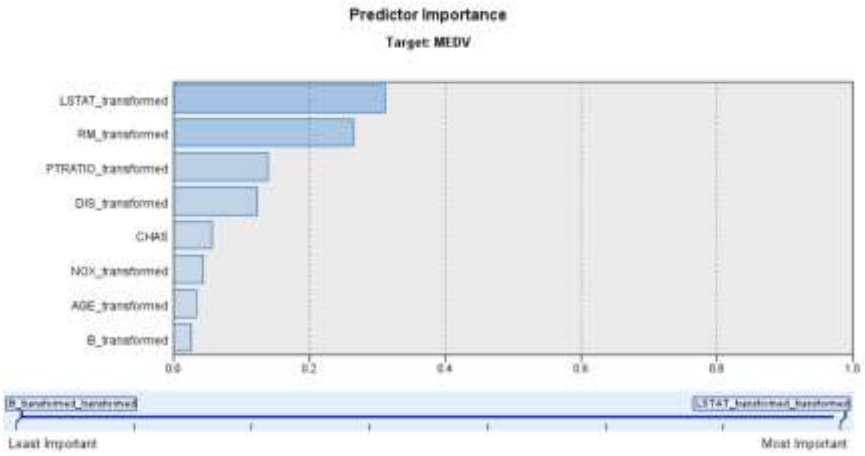
Linear ۲-۳-۳

نتایج مدل Linear در شکل‌های زیر نمایش داده شده است.



شکل ۳-۹ خلاصه مدل Linear

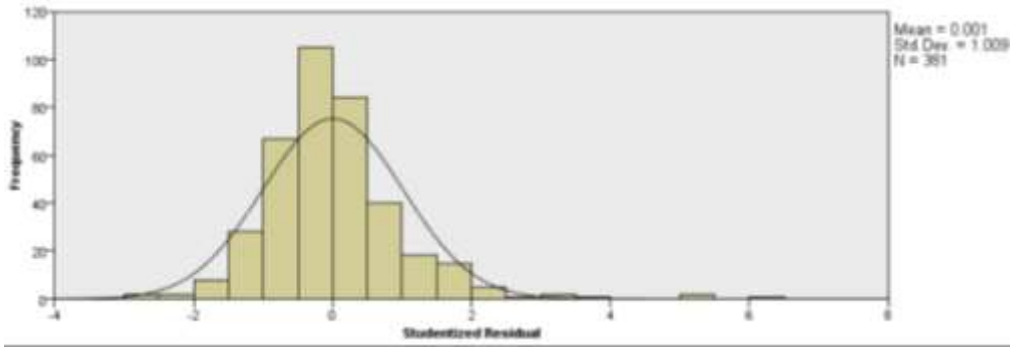
شکل زیر مهم ترین ویژگی‌هایی که در تخمین و پیش‌بینی MEDV در الگوریتم Linear تأثیر می‌گذارد را به ترتیب اولویت نمایش می‌دهد.



شکل ۳-۱۰ ترتیب اهمیت ویژگی‌ها برای تخمین MEDV

همانطور که از شکل ۳-۱۰ مشخص است، مانند الگوریتم NN دو ویژگی برجسته الگوریتم Linear عبارت است از LSTAT و RM ، اما بر خلاف NN که ویژگی سوم مهم DIS بود، در اینجا ویژگی سوم مهم نسبت دانش آموز به معلم بر اساس منطقه یا همان PTRATIO است. همچنین ویژگی B کم اهمیت‌ترین ویژگی عنوان شده است.

در علم آمار، Studentized residuals، خارج قسمت حاصل از تقسیم باقیمانده توسط برآورد انحراف معیار آن است. این یک شکل از هیستوگرام Studentized residuals که در آن توزیع بقایا را با توزیع عادی مقایسه می کند. خط صاف نشان دهنده توزیع عادی است. هرچه فرکانسهای باقیمانده به این خط نزدیکتر باشد، توزیع باقیمانده به توزیع نرمال نزدیکتر است.



شکل ۳-۱۱ مقایسه توزیع بقایا با توزیع عادی در مدل Linear

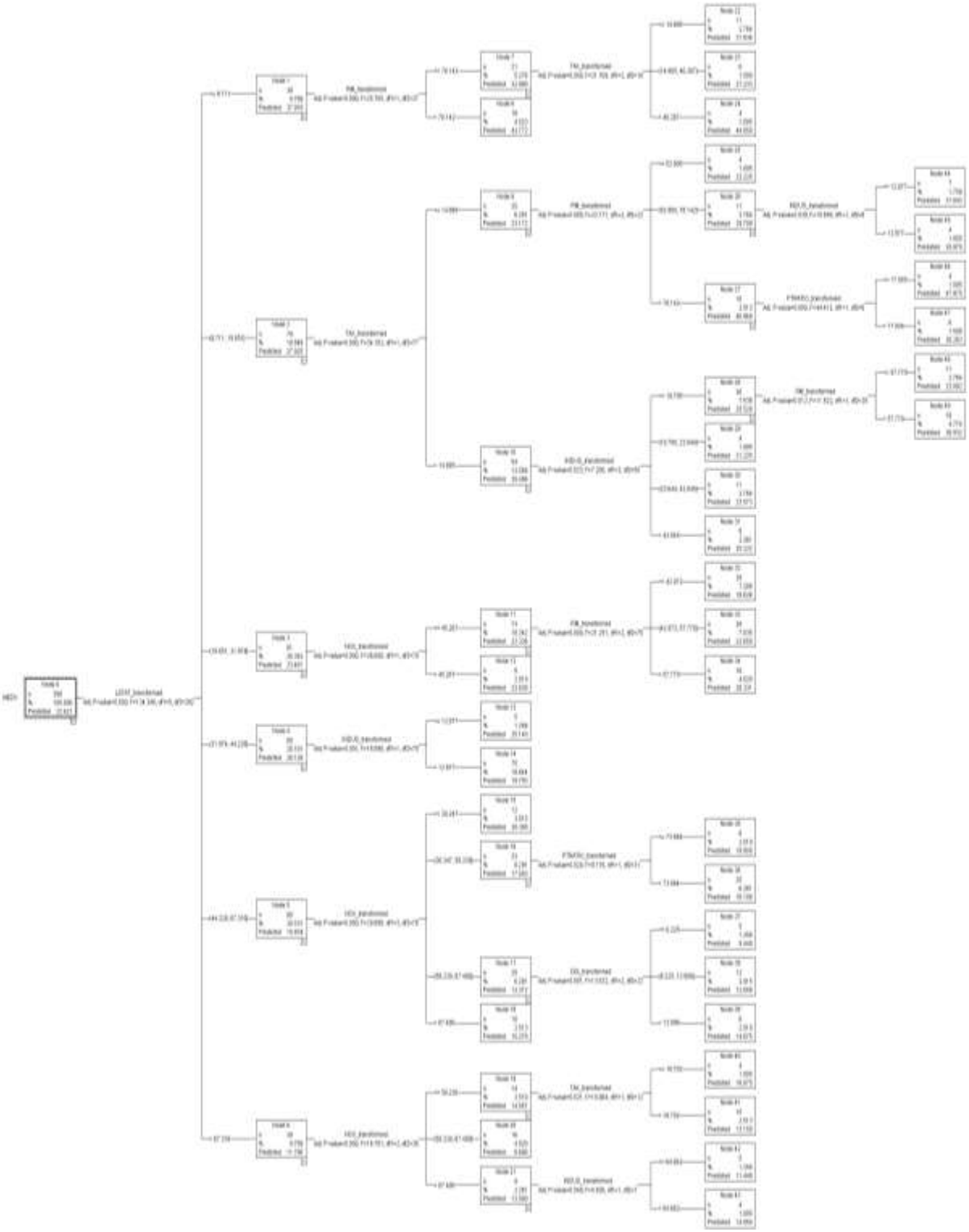
C&R Tree ۳-۳-۳

نتایج مدل CART در شکل زیر نمایش داده شده است. همانطور که مشخص هست خروجی یک درخت به عمق پنج هست.

شکل ۳-۱۲ خروجی مدل CART

CHAID ٤-٣-٣

نتایج مدل CHAID در شکل زیر نمایش داده شده است. همانطور که مشخص هست خروجی یک درخت به عمق چهار هست.



شکل ۳-۱۳ خروجی مدل CHAID

Regression ۵-۳-۳

نتایج مدل Regression در شکل زیر نمایش داده شده است.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.854[a]	.730	.721	4.737156

a. Predictors: (Constant), LSTAT_transformed, CHAS, AGE_transformed, B_transformed, PTRATIO_transformed, ZN_transformed, RAD_transformed, RM_transformed, DIS_transformed, INDUS_transformed, CRIM_transformed, NOX_transformed, TAX_transformed

شکل ۳-۱۴ خلاصه مدل Regression

نتایج حاصل از آنالیز واریانس یا ANOVA در شکل زیر نمایش داده شده است.

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	23292.875	13	1791.760	79.844	.000[b]
	Residual	8617.208	384	22.441		
	Total	31910.083	397			

b. Predictors: (Constant), LSTAT_transformed, CHAS, AGE_transformed, B_transformed, PTRATIO_transformed, ZN_transformed, RAD_transformed, RM_transformed, DIS_transformed, INDUS_transformed, CRIM_transformed, NOX_transformed, TAX_transformed

شکل ۳-۱۵ نتایج ANOVA

Generalized Linear ۶-۳-۳

نتایج مدل Generalized Linear در شکل‌های زیر نمایش داده شده است

Model Information

Dependent Variable	MEDV
Probability Distribution	Normal
Link Function	Identity

شکل ۳-۱۶ اطلاعات مدل Generalized Linear

Categorical Variable Information

			N	Percent
Factor	CHAS	0.0	369	92.7%
		1.0	29	7.3%
		Total	398	100.0%

Case Processing Summary

	N	Percent
Included	398	100.0%
Excluded	0	0.0%
Total	398	100.0%

Continuous Variable Information

		N	Minimum	Maximum	Mean	Std. Deviation
Dependent Variable	MEDV	398	5	50	22.42	8.965
Covariate	CRIM_transformed	398	13	100	26.52	17.954
	ZN_transformed	398	0	100	14.02	28.497
	INDUS_transformed	398	0	100	38.64	25.249
	NOX_transformed	398	0	100	34.68	23.656
	RM_transformed	398	0	100	50.20	20.500
	AGE_transformed	398	0	100	59.89	32.475
	DIS_transformed	398	0	100	32.71	25.054
	RAD_transformed	398	0	100	37.11	38.199
	TAX_transformed	398	0	100	41.90	32.586
	PTRATIO_transformed	398	0	100	59.97	24.072
	B_transformed	398	0	100	88.91	24.912
	LSTAT_transformed	398	1	100	35.61	22.694

شکل ۳–۱۷ اطلاعات داده‌های مورد استفاده در مدل Generalized Linear

نتایج حاصل از آزمون نیکویی برازش در شکل زیر آمده است.

Goodness of Fit

	Value	df	Value/df
Deviance	8617.208	384	22.441
Scaled Deviance	398.000	384	
Pearson Chi-Square	8617.208	384	22.441
Scaled Pearson Chi-Square	398.000	384	
Log Likelihood[b]	-1176.675		
Akaike's Information Criterion (AIC)	2383.351		
Finite Sample Corrected AIC (AICC)	2384.607		
Bayesian Information Criterion (BIC)	2443.147		
Consistent AIC (CAIC)	2458.147		

Dependent Variable: MEDV
Model: (Intercept), CHAS, CRIM_transformed, ZN_transformed, INDUS_transformed, NOX_transformed, RM_transformed, AGE_transformed, DIS_transformed, RAD_transformed, TAX_transformed, PTRATIO_transformed, B_transformed, LSTAT_transformed[a]

a. Information criteria are in smaller-is-better form.

b. The full log likelihood function is displayed and used in computing information criteria.

شکل ۳–۱۸ نتایج آزمون نیکویی برازش در مدل Generalized Linear

نتایج حاصل از Omnibus Test در شکل زیر آمده است.

Omnibus Test

Likelihood Ratio Chi-Square	df	Sig.
521.046	13	.000

Dependent Variable: MEDV
Model: (Intercept), CHAS, CRIM_transformed, ZN_transformed, INDUS_transformed, NOX_transformed, RM_transformed, AGE_transformed, DIS_transformed, RAD_transformed, TAX_transformed, PTRATIO_transformed, B_transformed, LSTAT_transformed[a]

a. Compares the fitted model against the intercept-only model.

شکل ۳-۱۹ نتایج آزمون Omnibus در مدل Generalized Linear

نتایج حاصل از آزمون اثرهای مدل در شکل زیر آمده است. این آزمون که به آزمون والد خی- دو معروف است یک آزمون فرض پارامتری با کاربردهای گوناگون برای آزمودن اینکه آیا پارامتر تخمین‌زده‌شده توسط نمونه برابر با پارامتر مورد نظر است یا خیر می‌باشد.

Tests of Model Effects

Source	Type III		
	Wald Chi-Square	df	Sig.
(Intercept)	181.540	1	.000
CHAS	10.202	1	.001
CRIM_transformed	.360	1	.549
ZN_transformed	1.001	1	.317
INDUS_transformed	.233	1	.629
NOX_transformed	8.147	1	.004
RM_transformed	74.485	1	.000
AGE_transformed	6.763	1	.009
DIS_transformed	37.829	1	.000
RAD_transformed	9.298	1	.002
TAX_transformed	9.249	1	.002
PTRATIO_transformed	43.041	1	.000
B_transformed	9.100	1	.003
LSTAT_transformed	73.644	1	.000

Dependent Variable: MEDV
Model: (Intercept), CHAS, CRIM_transformed, ZN_transformed, INDUS_transformed, NOX_transformed, RM_transformed, AGE_transformed, DIS_transformed, RAD_transformed, TAX_transformed, PTRATIO_transformed, B_transformed, LSTAT_transformed

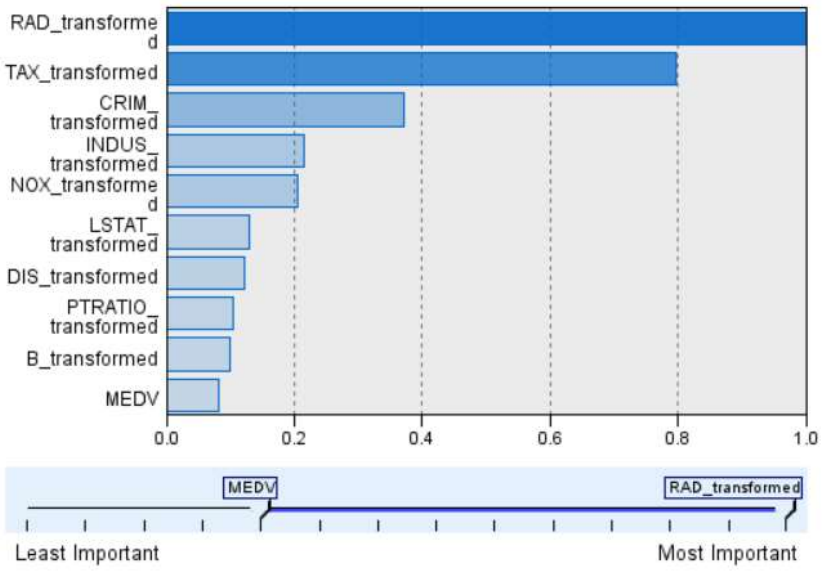
شکل ۳-۲۰ نتایج آزمون والد خی-دو در مدل Generalized Linear

۳-٤ خوشه بندی داده‌ها

می‌خواهیم در مجموعه داده‌ها، گروه و رکوردهای مشابه را بررسی کنیم. برای اینکار از خوشه بندی K-Means استفاده می‌کنیم. قبل از هر چیز باید تایپ تارگت را به Input تغییر دهیم. همچنین نیازی به استفاده از پارتیشن نیست. پس روی داده‌های تمیز ادامه کار را دنبال می‌کنیم. بدین منظور با استفاده از ابزار Auto cluster تعداد k را تنظیم می‌کنیم. مشاهده می‌کنیم که k=۲ از بین k های ۲، ۳، ٤، ٥، ٦، ۷، ۸، ۹، ۱۰، ۱۱، ۱۲، ۱۳، ۱٤، ۱٥، ۱٦، ۱۷، ۱۸، ۱۹، ۲۰، ۲٥، ۳۰، ۳٥، ٤۰ بهترین مقدار Silhouette را روی داده‌ها دارد.

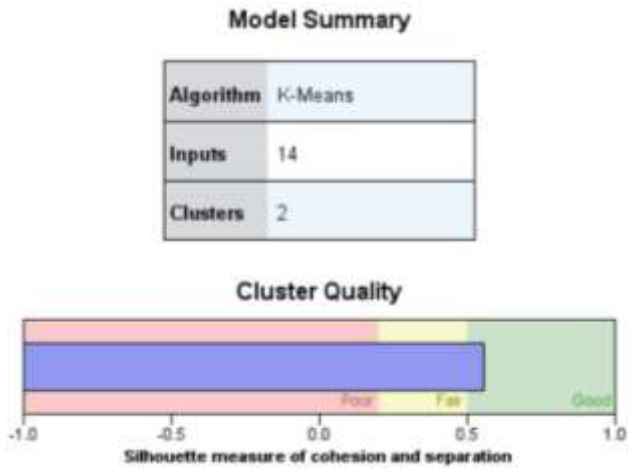
شکل ۳-۲۱ خوشه بندی K-Means

نتایج حاصل از مدل اول که نشان دهنده اهمیت ویژگی پیش‌بینی کننده می‌باشد، در شکل زیر نمایش داده شده است.



شکل ۳-۲۲ اهمیت ویژگی پیش‌بینی کننده در خوشه بندی K-Means

پر واضح است که ویژگی RAD یا شاخص دسترسی به بزرگراه‌های محوری سطح شهر بوستون و ویژگی TAX یا نرخ کل مالیات بر دارایی به ازای هر ۱۰۰۰۰ دلار، بیشترین اهمیت را دارند. در شکل زیر خلاصه مدل اول را مشاهده می‌کنید.



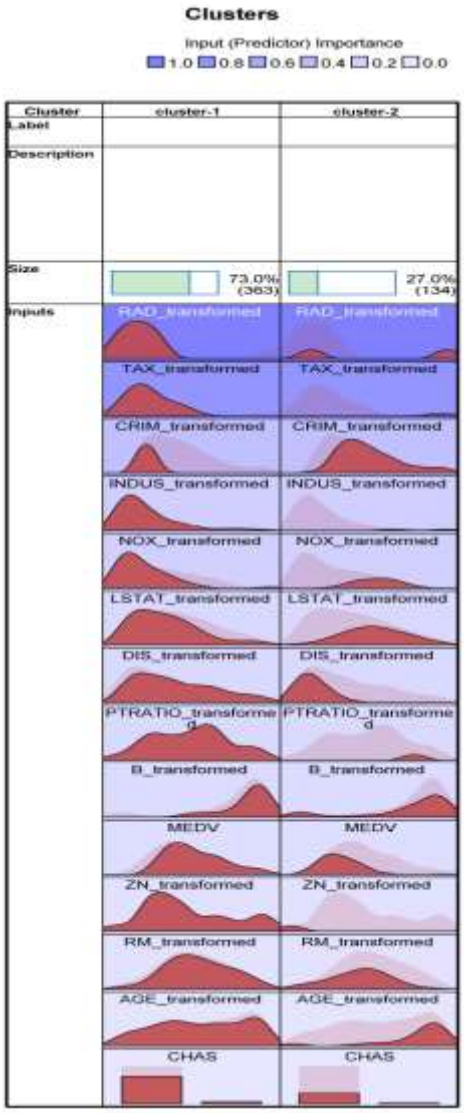
شکل ۳-۲۳ خلاصه مدل اول K-Means

شکل ۳-۲٤ دو خوشه‌بندی را نشان می‌دهد. کاملا مشخص هست که خوشه اول مربوط به ۳٦۳ مشاهده است که دسترسی کمی به بزرگراه‌های محوری سطح شهر بوستون دارند. همچنین این مشاهدات دارای نرخ کل مالیات بر دارایی کم است و از طرفی نرخ سرانه تبه کاری در این مشاهدات نسبت به خوشه دوم کمتر است. میزان زمین‌های مشاغل غیر خُرد در خوشه اول کم است. احتمالا به دلیل دسترسی کم به بزرگراه‌ها میزان غلظت اکسیدهای نیتریک در این مشاهدات کم است. همچنین درصد قشر طبقه پایین جمعیت در این مشاهدات نسبت به خوشه بندی دوم کمتر است. از طرفی میانگین فاصله تا پنج مرکز شغلی شهر بوستون نسبت به خوشه دوم نسبتا زیاد هست می‌باشد. نسبت دانش آموز به معلم در خوشه اول کم است. قیمت مسکن در خوشه اول زیاد است. تعداد اتاق خواب ها کمتر است و تعداد واحدهای مسکونی که مالک در آنها زندگی می‌کند و سال ساخت آنها قبل از ۱۹٤۰ است کم است. برای توضیح بهتر درصد متوسط برای هر ویژگی را برای هر یک از خوشه‌ها در جدول زیر گردآوری کردیم:

جدول ۳-۳ مقایسه بین دو خوشه

Brief explanation	Field	Cluster-1	Clister-2
دسترسی به بزرگراه‌های محوری	RAD	٪۱۷	٪۱۰۰
نرخ مالیات	TAX	٪۲۰	٪۹۰
نرخ تبه‌کاری	CRIM	٪۲۰	٪۳۰
زمین مشاغل غیر خُرد	INDUS	٪۲۰	٪۷۰
غلظت اکسیدهای نیتریک	NOX	٪۲۰	٪٦۵
قشر طبقه پایین جمعیت	LSTAT	٪۲۰	٪۵۳
فاصله تا پنج مرکز شغلی	DIS	٪۲۵	٪۱۰
نسبت دانش آموز به معلم	PTRATIO	٪۵۳	٪۸۰
نسبت سیاه پوستان	B	٪۹۹	٪۹۹

متوسط قیمت مسکن	MEDV	٪۲۲	٪۱۳
زمین مسکونی بالای ۲۵۰۰۰ ف.م	ZN	٪۰	٪۰
تعداد اتاق‌های مسکن	RM	٪٤۲	٪۵۳
مسکن دارای مالک و قبل ۱۹٤۰	AGE	٪۷۱	٪۹۷
محدود شدن راه توسط رودخانه	CHAS	٪۰	٪۰



شکل ۳-۲٤ اطلاعات حاصل از خوشه بندی مدل اول K-Means

اینطور که از دو خوشه بندی می‌توان استدلال کردن، قیمت مسکن بیش از هر چیز به نرخ مالیات مرتبط می‌باشد. هر چه نرخ مالیات بیشتر باشد قیمت مسکن کمتر است و بالعکس. همچنین در رکوردهایی که نرخ تبه‌کاری بیشتر است قیمت مسکن کمتر است. همچنین هرچی غلظت اکسیدهای نیتریک بیشتر باشد قیمت مسکن کمتر است. هرچه قشر طبقه پایین جمعیت زیاد باشد قیمت مسکن کمتر است. همچنین اگر مسکن دارای مالک باشد و سال ساخت آن قبل از ۱۹٤۰ باشد قیمت مسکن کمتر است.

نکته دیگر آن است که اگر مسکن دارای مالک و قبل از ۱۹٤۰ ساخته شده باشد تعداد اتاق‌های آن بیشتر است. همچنین هرچقدر دسترسی به بزرگراه‌های محوری بیشتر باشد، نسبت دانش‌آموزها به معلم‌ها بیشتر است، از طرفی میانگین وزنی فاصله تا پنج مرکز شغلی شهر بوستون کمتر است. همچنین هرچه زمین مشاغل غیر خُرد بیشتر باشند، نرخ مالیات بیشتر است. ویژگی‌های B، ZN و CHAS در این نوع دسته بندی دانشی به ما اضافه نمی‌کنند. حال در ادامه می‌خواهیم با استفاده از قوانین انجمنی دانشی که در مورد رابطه بین فیلدها بدست آوردیم را آزمایش کنیم.

۳-۵ قوانین انجمنی

یک نوع مدلسازی Unsupervised هستند و به ما سبک IF-Then Rules را برمی‌گردانند. در سامانه‌های توصیه‌گر کاربرد دارند. در این روش مدلسازی می‌خواهیم قوانینی را بدست آوریم و سپس ارزیابی کنیم. برای اینکار ما از ابزار Apriori، Carma و Sequence استفاده می‌کنیم. در ابتدا هر فیلدی را که برای استخراج قوانین انجمن به آن نیاز نداریم فیلتر می‌کنیم، در مدل طبقه بندی دیدیم روابطی وجود دارد بین فیلدها و حالا میخواهیم بصورت جداگانه برای آنها قوانینی را بیابیم پس فیلدها را بر اساس جدول ۳-٤ فیلتر می‌کنیم.

جدول ۳-٤ دسته بندی فیلدهای مورد نیاز برای استخراج قوانین انجمنی

شماره دسته	فیلدهای مورد نیاز
دسته اول	قیمت مسکن، نرخ مالیات، نرخ تبه‌کاری، غلظت اکسیدهای نیتریک، قشر طبقه پایین جمعیت، مسکن دارای مالک باشد و سال ساخت آن قبل از ۱۹٤۰
دسته دوم	دسترسی به بزرگراه‌های محوری، نسبت دانش‌آموزها به معلم‌ها، فاصله تا پنج مرکز شغلی
دسته سوم	مسکن دارای مالک و قبل از ۱۹٤۰ و تعداد اتاق‌ها
دسته چهارم	زمین مشاغل غیر خُرد و نرخ مالیات

ابتدا با استفاده از Filler باید فیلدها را بصورت Flag تعریف کنیم که بر اساس جدول ۳-۵ و شکل ۳-۲۵ این کار انجام می‌شود. برای Flag تعریف کردن فیلدهای دسته اول از دستورات زیر که بر اساس جدول ۳-۳ نوشته شده است استفاده می‌کنیم:

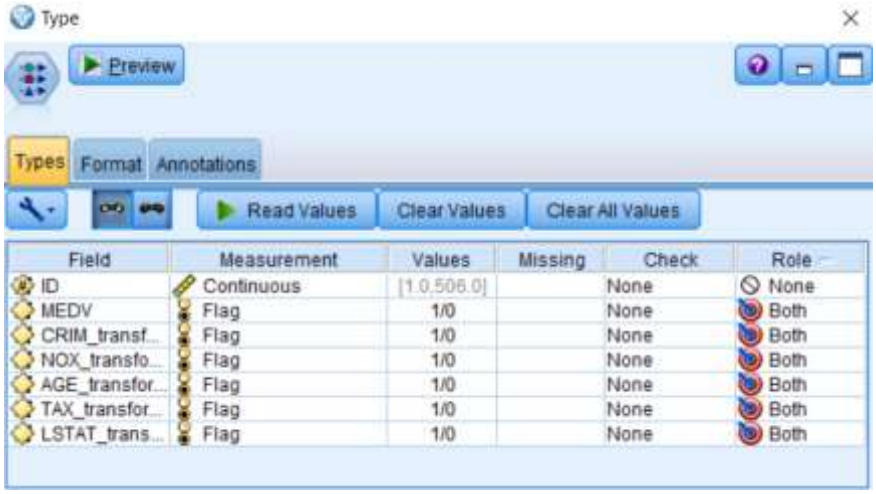
جدول ۳-۵ دستور Flag تعریف کردن فیلدهای دسته اول

نام فیلد	دستور
LSTAT ،TAX ،NOX ،CRIM	if @FIELD>20 then 1 else 0 endif
MEDV	if @FIELD>22 then 0 else 1 endif
AGE	if @FIELD>71 then 1 else 0 endif

برای ما رابطه بین نرخ مالیات بالا، نرخ تبه‌کاری بالا، میزان غلظت نیتریک‌های اسید زیاد، درصد قشر طبقه پایین جمعیت بالا و خانه‌های دارای مالک و قبل از ۱۹٤۰ زیاد با قیمت مسکن پایین مهم است، به همین خاطر دستورات بصورت فوق تعریف شده‌اند. به عنوان مثال برای سطر اول جدول فوق، فیلر را به شکل زیر تنظیم می‌کنیم:

شکل ۳-۲۵ تنظیم فیلر برای Flag کردن ویژگی‌ها

بعد از آن باید از طریق Type نقش فیلدها را از Input به Both تغییر دهیم (شکل ۳-۲٦)، به این معنی که هر فیلد می‌تواند ورودی یا خروجی باشد. همچنین نقش ID را به None تغییر می‌دهیم زیرا آن را در مدل در نظر نمی‌گیریم. همچنین مقیاس را باید Flag کرده و Read Values کنیم تا مقادیر درست اعمال شود.



شکل ۳-۲٦ تغییر Type ویژگی‌ها برای استخراج قوانین انجمنی

حال قوانین انجمنی حاصل از روش Apriori در شکل زیر نمایش داده شده است.

Consequent	Antecedent	Support %	Confidence %
LSTAT_transformed	MEDV	55.533	98.188
NOX_transformed	TAX_transformed LSTAT_transformed	54.326	87.778
LSTAT_transformed	TAX_transformed NOX_transformed	54.527	87.454
MEDV	TAX_transformed LSTAT_transformed	54.326	84.444
TAX_transformed	MEDV LSTAT_transformed	54.527	84.133
TAX_transformed	MEDV	55.533	83.333
LSTAT_transformed	NOX_transformed	68.813	83.333
TAX_transformed	NOX_transformed LSTAT_transformed	57.344	83.158
NOX_transformed	MEDV LSTAT_transformed	54.527	83.026
NOX_transformed	MEDV	55.533	82.609
NOX_transformed	LSTAT_transformed	69.819	82.133
TAX_transformed	NOX_transformed	68.813	79.24
NOX_transformed	TAX_transformed	69.014	79.009
MEDV	NOX_transformed LSTAT_transformed	57.344	78.947
LSTAT_transformed	TAX_transformed	69.014	78.717
MEDV	LSTAT_transformed	69.819	78.098
TAX_transformed	LSTAT_transformed	69.819	77.81
MEDV	TAX_transformed NOX_transformed	54.527	74.17
AGE_transformed	TAX_transformed NOX_transformed	54.527	70.849
MEDV	TAX_transformed	69.014	67.055
AGE_transformed	NOX_transformed LSTAT_transformed	57.344	67.018
MEDV	NOX_transformed	68.813	66.667
AGE_transformed	NOX_transformed	68.813	64.912
AGE_transformed	TAX_transformed LSTAT_transformed	54.326	64.074
AGE_transformed	MEDV LSTAT_transformed	54.527	62.731
AGE_transformed	MEDV	55.533	61.957
CRIM_transformed	TAX_transformed NOX_transformed	54.527	60.148
AGE_transformed	LSTAT_transformed	69.819	57.349
AGE_transformed	TAX_transformed	69.014	56.851
CRIM_transformed	TAX_transformed LSTAT_transformed	54.326	55.185
CRIM_transformed	NOX_transformed LSTAT_transformed	57.344	52.281

شکل ۳-۲۶ قوانین انجمنی با استفاده از روش Apriori

ملاحظه می‌شود که ۳۱ قانون تولید شده است. قانون اول با Confidence ٪۹۸ و Support ٪۵۵ نشان دهنده آن هست که اگر قشر طبقه پایین جمعیت بالا باشد آنگاه قیمت مسکن کم هست. سایر قوانین را به این شکل می‌توان تعبیر کرد. به عنوان مثال قانون ۲۶ نشان می‌دهد با Confidence٪۶۱ و Support ٪۵۵ اگر خانه دارای مالک و قبل از سال ۱۹۴۰ ساخته شده باشد آنگاه قیمت مسکن کم است. برای دسته‌های دیگر با همین ترفند می‌توان قانون بدست آورد.

۳-۶ نتیجه‌گیری

در این فصل به مدلسازی ریاضی پرداختیم. آموختیم که چطور می‌توان با استفاده از الگوریتم‌ها و روش‌های کمی برای پیش‌بینی و تخمین از جمله Regression، Generalized Linear، SVM، Linear، CHAID، C&R Tree و Neural Net وغیره به مدلسازی داده‌ها بپردازیم. همچنین نحوه خوشه بندی و ایجاد قوانین انجمنی برای داده‌های دیتاست پروژه را یادگرفتیم. در فصل بعدی به ارزیابی مدل‌های ساخته شده می‌پردازیم.

فصل ٤: ارزیابی مدل

٤-١ مقدمه

در این بخش قصد داریم تا مدل‌های ساخته شده را ارزیابی کنیم. در سه حالت می‌توان مدل‌ها را ارزیابی کرد. حالت اول مقایسه نتایج حاصل از مدل‌ها با در نظر گرفتن گام پاک‌سازی داده‌ها و بدون در نظر گرفتن آن است. حالت دوم مقایسه نتایج در حالتی که داده‌های نامتوازن مدیریت شوند با زمانی که مدیریت نشوند یعنی مقایسه نتایج بدست آمده از مدل‌ها با و بدون نودهای Boost و Reduce که پیش تر ایجاد کردیم. حالت سوم مقایسه میزان خطای متوسط بهترین مدل با خطای متوسط ۹.۲۱ هزار دلار می‌باشد.

٤-٢ ارزیابی مدل‌ها

برای ارزیابی مدل‌ها باید سه حالت متفاوت را بررسی کنیم. همانطور که پیش تر گفتیم حالت اول مقایسه نتایج مدل فصل سوم یعنی در نظر گرفتن گام‌های پاکسازی داده (بجز گام آخر) با حالتی که پاکسازی انجام نشود. حالت دوم در نظر گرفتن گام آخر پاکسازی داده (یعنی مدیریت داده‌های نامتوازن) ومقایسه با نتایج حالت قبل و حالت سوم مقایسه میزان خطای متوسط بهترین مدل با خطای متوسط ۹.۲۱ هزار دلار می‌باشد.

٤-٢-١ حالت اول

در این حالت نتایج حاصل از مدل‌ها را با در نظر گرفتن گام پاک‌سازی داده‌ها و بدون در نظر گرفتن آن بررسی می‌کنیم. نتایج حاصل از مدل‌ها برای داده‌های آموزش با و بدون در نظر گرفتن گام پاک‌سازی داده‌ها به ترتیب در شکل ٤-٢ و شکل ٤-١ آمده است و نتایج حاصل از مدل‌ها برای داده‌های آزمایش با و بدون در نظر گرفتن گام پاک‌سازی داده‌ها به ترتیب در شکل ٤-٤ و شکل ٤-٣ آمده است.

Sort by: Use | Ascending | Descending | Delete Unused Models | View: Training set

Use?	Graph	Model	Build Time (mins)	Correlation	No. Fields Used	Relative Error
☑		Neural Net 1	< 1	0.928	13	0.139
☑		CHAID 1	< 1	0.918	8	0.158
☑		Linear 1	< 1	0.869	12	0.245
☑		Regression 1	< 1	0.867	13	0.248
☑		Generalized Linear 1	< 1	0.867	13	0.248
☑		SVM 1	< 1	0.867	13	0.278
☑		C&R Tree 1	< 1	0.862	11	0.257

شکل ٤-١ نتایج مدل‌ها برای داده‌های آموزش بدون در نظر گرفتن گام پاک‌سازی داده‌ها

شکل ٤-٢ نتایج مدل‌ها برای داده‌های آموزش با در نظر گرفتن گام پاک‌سازی داده‌ها

نتایج حاصل از مقایسه مدل‌ها برای داده‌های آموزش در دو شکل فوق در جدول زیر خلاصه شده است:

جدول ٤-١ نتایج مقایسه مدل‌ها برای داده‌های آموزش با و بدون گام پاک‌سازی داده‌ها

Subtraction (1-2)	Without cleaning (2)	With cleaning (1)	Model
0.025	0.928	0.953	Neural Net
0.081	0.862	0.943	C&R Tree
0.01	0.918	0.928	CHAID
-0.002	0.867	0.865	SVM
-0.013	0.867	0.854	Regression
-0.013	0.867	0.854	Generalized Linear
-0.029	0.869	0.84	Linear

واضح هست که برای مدل‌های Neural Net، C&R Tree و CHAID پاکسازی داده‌ها تاثیر مثبت دارد و دقت مدل را بالا می‌برد. برای مدل‌های SVM، Regression، Generalized Linear و Linear پاکسازی داده‌ها تاثیر چندانی بر نتیجه نهایی نداشته است یا حتی تاثیر معکوس داشته است، این بدان علت هست که تمامی گام‌های پاک‌سازی داده برای این الگوریتم‌ها واجب هست حال آنکه ما در پاک‌سازی داده‌ها گام آخر یعنی مدیریت داده‌های متوازن را انجام ندادیم. در حالت دوم این مسئله بطور مفصل تری مورد بررسی قرار خواهد گرفت.

در ادامه نتایج حاصل از مدل‌ها برای داده‌های آزمایش با و بدون در نظر گرفتن گام پاک‌سازی داده‌ها (بجز گام آخر) مورد بررسی قرار گرفته شده است. به شکل‌های ٤-٤ و ٤-٣ توجه کنید.

Sort by: Use | Ascending | Descending | Delete Unused Models | View: Testing set

Use?	Graph	Model	Build Time (mins)	Correlation	No. Fields Used	Relative Error
✓		Neural Net 1	< 1	0.938	13	0.120
✓		CHAID 1	< 1	0.877	8	0.236
✓		Regression 1	< 1	0.851	13	0.277
✓		Generalized Linear 1	< 1	0.851	13	0.277
✓		Linear 1	< 1	0.851	12	0.278
✓		C&R Tree 1	< 1	0.838	11	0.3
✓		SVM 1	< 1	0.83	13	0.346

شکل ٤-٣ نتایج مدل‌ها برای داده‌های آزمایش بدون گام پاک‌سازی داده‌ها

Sort by: Use | Ascending | Descending | Delete Unused Models | View: Testing set

Use?	Graph	Model	Build Time (mins)	Correlation	No. Fields Used	Relative Error
✓		Neural Net 1	< 1	0.932	13	0.135
✓		C&R Tree 1	< 1	0.929	12	0.138
✓		CHAID 1	< 1	0.923	7	0.151
✓		SVM 1	< 1	0.908	13	0.23
✓		Regression 1	< 1	0.887	13	0.217
✓		Generalized Linear 1	< 1	0.887	13	0.217
✓		Linear 1	< 1	0.869	8	0.247

شکل ٤-٤ نتایج مدل‌ها برای داده‌های آزمایش با گام پاک‌سازی داده‌ها

نتایج حاصل از مقایسه مدل‌ها برای داده‌های آزمایش در دو شکل فوق در جدول زیر خلاصه شده است:

جدول ٤-٢ نتایج مقایسه مدل‌ها برای داده‌های آزمایش با و بدون گام پاک‌سازی داده‌ها

Subtraction (1-2)	Without cleaning (2)	With cleaning (1)	Model
-0.006	0.938	0.932	Neural Net
0.091	0.838	0.929	C&R Tree
0.046	0.877	0.923	CHAID
0.078	0.83	0.908	SVM
0.036	0.851	0.887	Regression
0.036	0.851	0.887	Generalized Linear
0.018	0.851	0.869	Linear

با توجه به نتایج حاصل از جدول فوق متوجه می‌شویم بجز در الگوریتم NN در سایر الگوریتم‌ها دقت داده‌های آزمایش بعد از انجام پاکسازی داده به مراتب بیشتر است. حال می‌خواهیم میزان خطای یا دقت مدل‌های حاصل از داده‌های آموزش و آزمایش را در دو حالت در نظر گرفتن پاک‌سازی و در نظر نگرفتن آن با یکدیگر مقایسه کنیم.

جدول ٤-٣ نتایج مقایسه دقت مدل‌های آموزش و آزمایش با و بدون گام پاک‌سازی داده‌ها

Max{Abs(1), Abs(2)}	Error Without cleaning (2)	Error With cleaning (1)	Model
0.021	-0.01	0.021	Neural Net
0.024	0.024	0.014	C&R Tree
0.041	0.041	0.005	CHAID
0.043	0.037	-0.043	SVM
0.033	0.016	-0.033	Regression
0.033	0.016	-0.033	Generalized Linear
0.029	0.018	-0.029	Linear

از جدول فوق ملاحظه می‌شود در الگوریتم‌های NN، SVM، Regression، Generalized Linear و Linear زمانیکه گام‌های پاک‌سازی داده (بجز گام آخر) طی شود، قدر مطلق تفاوت دقت بین داده‌های آموزی و آزمایش به مراتب بیشتر از حالتی است که داده‌ها پاک‌سازی نشوند. این شاید به این علت هست که برای این الگوریتم‌ها نیاز است تا **تمام** گام‌های پاکسازی طی شوند تا به دقت مطلوب برسند (برخلاف CHAID و CART که پاک‌سازی اهمیت چندانی ندارد). در بخش بعدی این مورد را به طور مفصل‌تری بررسی می‌کنیم و با نتایج این بخش مقایسه می‌کنیم.

٤-٢-٢ حالت دوم

در این حالت قصد داریم نتایج حاصل از مدل‌ها را با فرض انجام گام آخر پاکسازی داده یعنی مدیریت داده‌های نامتوازن در دو شکل Boost و Reduce مورد بررسی قرار دهیم. اگر داده‌های نامتوازن را به حالت Boost مدیریت کنیم (شکل ٤-٥) نتایج برای داده‌های آموزش در جدول زیر نوشته شده است:

جدول ٤-٤ نتایج حالت اول و مقایسه آن با پاک‌سازی و نود Boost برای داده‌های آموزش

Subtraction (1-3)	Subtraction (1-2)	Without cleaning (3)	With cleaning (2)	With cleaning +Boost (1)	Model
0.052	0.027	0.928	0.953	0.98	Neural Net
0.072	-0.009	0.862	0.943	0.934	C&R Tree
0.047	0.037	0.918	0.928	0.965	CHAID
0.006	0.008	0.867	0.865	0.873	SVM
-0.006	0.007	0.867	0.854	0.861	Regression
-0.006	0.007	0.867	0.854	0.861	Generalized Linear
-0.007	0.022	0.869	0.84	0.862	Linear

شکل ٤-٥ نتایج مدل برای داده‌های آموزش با پاک‌سازی و نود Boost

جدول ٤-٤ نشان می‌دهد زمانیکه تمام گام‌های پاک‌سازی داده را انجام می‌دهیم (گام آخر را بصورت Boost مدیریت می‌کنیم)، تمامی الگوریتم‌ها بجز CART برای داده‌های آموزش نسبت به حالتی که گام آخر پاک‌سازی یعنی مدیریت داده‌های نامتوازن انجام نمی‌شود، دقت بهتری داشتند. و در مقایسه با حالتی که هیچ پاک‌سازی انجام نمی‌شود برای همه الگوریتم‌ها بجز Regression، Generalized Linear و Linear دقت بهتری دارد. حال برای داده‌های تست نتایج در شکل ٤-٦ و نتایج مقایسه در جدول ٤-٥ نوشته شده است:

جدول ٤-٥ نتایج حالت اول و مقایسه آن با پاک‌سازی و نود Boost برای داده‌های آزمایش

Subtraction (1-3)	Subtraction (1-2)	Without cleaning (3)	With cleaning (2)	With cleaning +Boost (1)	Model
0.037	0.043	0.938	0.932	0.975	Neural Net
0.097	0.006	0.838	0.929	0.935	C&R Tree
0.084	0.038	0.877	0.923	0.961	CHAID
0.043	-0.035	0.83	0.908	0.873	SVM
0.004	-0.032	0.851	0.887	0.855	Regression
0.004	-0.032	0.851	0.887	0.855	Generalized Linear
0.004	-0.014	0.851	0.869	0.855	Linear

در جدول زیر **قدر مطلق** دقت مدل‌ها (تفاوت همبستگی داده‌های آموزش و آزمایش) را بدست آوردیم:

جدول ٤-٦ نتایج حاصل از قدر مطلق دقت مدل‌ها

Max{1, 2, 3}	Abs without cleaning (3)	Abs with cleaning (2)	Abs with cleaning +Boost (1)	Model
0.021	0.01	0.021	0.005	Neural Net
0.024	0.024	0.024	0.001	C&R Tree
0.041	0.041	0.041	0.004	CHAID
0.043	0.037	0.043	0	SVM
0.033	0.016	0.033	0.006	Regression
0.033	0.016	0.033	0.006	Generalized Linear
0.029	0.018	0.029	0.007	Linear

ملاحظه می‌شود در حالیکه تمام گام‌های پاک‌سازی طی نشود بیشترین Over fit اتفاق افتاده است، یعنی مدل اصلا خوب یاد نگرفته و صرفا حفظ کرده، برای همین اختلاف همبستگی داده‌های آموزش و آزمایش در این حالت زیاد هست. بعد از آن زمانیکه هیچ پاک‌سازی انجام ندهیم بیشترین Over fit را درمدل‌ها داریم. ملاحظه می‌شود وقتی تمام گام‌ها پاکسازی را طی کنیم، میزان Over fit بسیار ناچیز و نزدیک به صفر هست. بنابراین این خیلی مهم هست که تمام گام‌های پاکسازی انجام شود.

Use?	Graph	Model	Build Time (mins)	Correlation	No. Fields Used	Relative Error
☑		Neural Net 1	< 1	0.975	13	0.051
☑		CHAID 1	< 1	0.961	12	0.078
☑		C&R Tree 1	< 1	0.935	13	0.127
☑		SVM 1	< 1	0.873	13	0.264
☑		Linear 1	< 1	0.855	10	0.273
☑		Regression 1	< 1	0.855	13	0.273
☑		Generalized Linear 1	< 1	0.855	13	0.273

شکل ٤-٦ نتایج مدل برای داده‌های آزمایش با پاک‌سازی و نود Boost

حال اگر داده‌های نامتوازن را به حالت Reduce مدیریت کنیم (شکل ٤-٧) نتایج برای داده‌های آموزش در جدول زیر نوشته شده است:

جدول ٤-٧ نتایج حالات قبل و مقایسه آن با پاک‌سازی و نود Reduce برای داده‌های آموزش

Sub (1-4)	Sub (1-3)	Sub (1-2)	Without cleaning (4)	cleaning (3)	cleaning +Boost (2)	cleaning +Reduce (1)	Model
0.017	-0.008	-0.035	0.928	0.953	0.98	0.945	Neural Net
0.122	0.041	0.05	0.862	0.943	0.934	0.984	C&R Tree
0.073	0.063	0.026	0.918	0.928	0.965	0.991	CHAID
-0.088	-0.086	-0.094	0.867	0.865	0.873	0.779	SVM
0.026	0.039	0.032	0.867	0.854	0.861	0.893	Regression
0.026	0.039	0.032	0.867	0.854	0.861	0.893	Generalized Linear
0.009	0.038	0.016	0.869	0.84	0.862	0.878	Linear

جدول ٤-٧ نشان می‌دهد زمانیکه تمام گام‌های پاک‌سازی داده را به همراه نود Reduce طی می‌کنیم، تمامی الگوریتم‌ها برای داده‌های آموزش بجز SVM نسبت به حالتی که هیچ پاک‌سازی انجام نمی‌شود، دقت بهتری دارند. و در مقایسه با حالتی که گام آخر پاک‌سازی یعنی مدیریت داده‌های نامتوازن انجام نمی‌شود و یا با استفاده از نود Boost انجام می‌شود برای همه الگوریتم‌ها بجز SVM و NN دقت بهتری دارد.

شکل ٤-٧ نتایج مدل برای داده‌های آموزش با پاک‌سازی و نود Reduce

حال برای داده‌های تست نتایج در شکل ٤-٩ و نتایج مقایسه در جدول ٤-٨ نوشته شده است:

جدول ٤-٨ نتایج حالات قبل و مقایسه آن با پاک‌سازی و نود Reduce برای داده‌های آزمایش

Sub (1-4)	Sub (1-3)	Sub (1-2)	Without cleaning (4)	cleaning (3)	cleaning +Boost (2)	cleaning +Reduce (1)	Model
0.003	-0.022	-0.049	0.928	0.953	0.98	0.931	Neural Net
0.052	-0.029	-0.02	0.862	0.943	0.934	0.914	C&R Tree
-0.095	-0.105	-0.142	0.918	0.928	0.965	0.823	CHAID
0.094	0.096	0.088	0.867	0.865	0.873	0.961	SVM
0.02	0.033	0.026	0.867	0.854	0.861	0.887	Regression
0.02	0.033	0.026	0.867	0.854	0.861	0.887	Generalized Linear
0.028	0.057	0.035	0.869	0.84	0.862	0.897	Linear

در جدول زیر **قدر مطلق** دقت مدل‌ها (تفاوت همبستگی داده‌های آموزش و آزمایش) را بدست آوردیم:

جدول ٤-٩ نتایج حاصل از قدر مطلق دقت تمامی مدل‌ها

Max{1, 2, 3, 4}	Abs without cleaning (3)	Abs cleaning (2)	Abs cleaning +Boost (2)	Abs cleaning +Reduce (1)	Model
0.021	0.01	0.021	0.005	0.014	Neural Net
0.024	0.024	0.024	0.001	0.07	C&R Tree
0.168	0.041	0.041	0.004	0.168	CHAID
0.182	0.037	0.043	0	0.182	SVM
0.033	0.016	0.033	0.006	0.006	Regression
0.033	0.016	0.033	0.006	0.006	Generalized Linear
0.029	0.018	0.029	0.007	0.019	Linear

ملاحظه می‌شود بیشترین Over fit مربوط به حالت‌هایی است که یا تمام گام‌های پاک‌سازی طی نشده است یا مدیریت داده‌های نامتوارن با استفاده از نود Reduce انجام شده است، بنابراین مدل اصلا خوب یاد نگرفته

و صرفا حفظ کرده، برای همین اختلاف همبستگی داده‌های آموزش و آزمایش در این حالات زیاد هست. همچنین ملاحظه می‌شود وقتی تمام گام‌ها پاکسازی را با نود Boost طی کنیم، میزان Over fit بسیار ناچیز و نزدیک به صفر هست. بنابراین این خیلی مهم هست که تمام گام‌های پاکسازی با بهترین شیوه مدیریت داده که در اینجا روش Boost است انجام شود.

Sort by: Use Ascending Descending Delete Unused Models View: Testing set

Use?	Graph	Model	Build Time (mins)	Correlation	No. Fields Used	Relative ...
☑		Neural Net 1	< 1	0.931	13	0.455
☑		CHAID 1	< 1	0.823	6	0.763
☑		Regression 1	< 1	0.887	13	0.825
☑		Linear 1	< 1	0.897	7	0.91
☑		Generalized Linear 1	< 1	0.887	13	0.825
☑		SVM 1	< 1	0.961	13	0.108
☑		C&R Tree 1	< 1	0.914	13	0.611

شکل ٤-٨ نتایج مدل برای داده‌های آزمایش با پاکسازی و نود Reduce

٤-٢-٣ حالت سوم

در این حالت قصد داریم میزان خطای بهترین مدل را محاسبه کرده و با خطای متوسط ۹.۲۱ هزار دلار مقایسه کنیم. خطای متوسط از فرمول زیر محاسبه می‌شود:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(Predicted_i - Actual_i)^2}{N}}$$

با توجه به اینکه دقت مدل زمانیکه تمام گام‌های پاکسازی داده با نود Boost طی شود بیشتر است لذا خروجی الگوریتم‌ها درحالتی که از نود Boost استفاده کردیم را مورد بررسی قرار می‌دهیم. بنابراین با استفاده از Table از نود Cleaning with Boost خروجی میگیریم. میزان Actual و Predicted در جدول زیر گردآوری شده است:

جدول ٤-١٠ نتایج پیش بینی MEDV حاصل از بهترین مدل

Actual	Predicted	ID	Actual	Predicted	ID
34.9	35.614	254	24	28.324	1
37	32.295	255	21.6	24.392	2
30.5	29.904	256	34.7	33.116	3

36.4	37.943	**257**	33.4	33.009	**4**
31.1	32.138	**258**	36.2	30.858	**5**
29.1	29.632	**259**	28.7	24.916	**6**
50	44.141	**260**	22.9	20.519	**7**
33.3	36.174	**261**	27.1	18.464	**8**
30.3	32.544	**262**	16.5	12.497	**9**
34.6	32.696	**263**	18.9	17.758	**10**
34.9	31.617	**264**	15	17.473	**11**
32.9	32.253	**265**	18.9	19.414	**12**
24.1	26.009	**266**	21.7	19.065	**13**
42.3	42.444	**267**	20.4	20.026	**14**
48.5	42.439	**268**	18.2	18.548	**15**
50	44.82	**269**	19.9	19.776	**16**
22.6	21.607	**270**	23.1	20.823	**17**
24.4	22.571	**271**	17.5	17.058	**18**
22.5	17.873	**272**	20.2	15.907	**19**
24.4	22.869	**273**	18.2	17.657	**20**
24.4	22.869	**274**	13.6	11.979	**21**
24.4	22.869	**275**	19.6	16.966	**22**
24.4	22.869	**276**	15.2	14.907	**23**
24.4	22.869	**277**	14.5	15.138	**24**
24.4	22.869	**278**	15.6	14.565	**25**
24.4	22.869	**279**	13.9	12.759	**26**
24.4	22.869	**280**	16.6	15.42	**27**
24.4	22.869	**281**	14.8	13.996	**28**
24.4	22.869	**282**	18.4	17.493	**29**
24.4	22.869	**283**	21	19.364	**30**
24.4	22.869	**284**	12.7	11.1	**31**
24.4	22.869	**285**	14.5	18.556	**32**
24.4	22.869	**286**	13.2	9.625	**33**
24.4	22.869	**287**	13.1	13.509	**34**
20	17.951	**288**	13.5	13.32	**35**
20	17.951	**289**	18.9	21.794	**36**
20	17.951	**290**	20	21.022	**37**
20	17.951	**291**	21	21.539	**38**
20	17.951	**292**	24.7	21.617	**39**
20	17.951	**293**	30.8	29.377	**40**
20	17.951	**294**	34.9	36.661	**41**
20	17.951	**295**	26.6	29.208	**42**
20	17.951	**296**	25.3	24.597	**43**
20	17.951	**297**	24.7	24.274	**44**
20	17.951	**298**	21.2	22.045	**45**
20	17.951	**299**	19.3	21.269	**46**
20	17.951	**300**	20	20.101	**47**
20	17.951	**301**	16.6	17.376	**48**
20	17.951	**302**	14.4	11.196	**49**
21.7	21.282	**303**	19.4	17.207	**50**
21.7	21.282	**304**	19.7	19.917	**51**
21.7	21.282	**305**	20.5	22.958	**52**
21.7	21.282	**306**	25	26.696	**53**
21.7	21.282	**307**	23.4	22.468	**54**
21.7	21.282	**308**	18.9	16.333	**55**
21.7	21.282	**309**	35.4	32.09	**56**
21.7	21.282	**310**	24.7	23.75	**57**
21.7	21.282	**311**	31.6	32.985	**58**
21.7	21.282	**312**	23.3	21.802	**59**
21.7	21.282	**313**	19.6	19.605	**60**
21.7	21.282	**314**	18.7	17.257	**61**
21.7	21.282	**315**	16	15.957	**62**
21.7	21.282	**316**	22.2	22.996	**63**

21.7	21.282	**317**	25	23.424	**64**
19.3	16.926	**318**	33	25.667	**65**
19.3	16.926	**319**	23.5	27.673	**66**
19.3	16.926	**320**	19.4	22.648	**67**
19.3	16.926	**321**	22	21.343	**68**
19.3	16.926	**322**	17.4	18.393	**69**
19.3	16.926	**323**	20.9	21.106	**70**
19.3	16.926	**324**	24.2	25.783	**71**
19.3	16.926	**325**	21.7	21.975	**72**
19.3	16.926	**326**	22.8	24.314	**73**
19.3	16.926	**327**	23.4	24.962	**74**
19.3	16.926	**328**	24.1	25.881	**75**
19.3	16.926	**329**	21.4	24.583	**76**
19.3	16.926	**330**	20	23.052	**77**
19.3	16.926	**331**	20.8	23.382	**78**
19.3	16.926	**332**	21.2	21.7	**79**
22.4	21.919	**333**	20.3	22.305	**80**
22.4	21.919	**334**	28	28.634	**81**
22.4	21.919	**335**	23.9	25.513	**82**
22.4	21.919	**336**	24.8	24.379	**83**
22.4	21.919	**337**	22.9	23.311	**84**
22.4	21.919	**338**	23.9	24.237	**85**
22.4	21.919	**339**	26.6	27.036	**86**
22.4	21.919	**340**	22.5	21.164	**87**
22.4	21.919	**341**	22.2	23.644	**88**
22.4	21.919	**342**	23.6	31.462	**89**
22.4	21.919	**343**	28.7	32.01	**90**
22.4	21.919	**344**	22.6	25.704	**91**
22.4	21.919	**345**	22	25.147	**92**
22.4	21.919	**346**	22.9	26.397	**93**
22.4	21.919	**347**	25	26.646	**94**
28.1	25.45	**348**	20.6	23.761	**95**
23.7	13.544	**349**	28.4	27.726	**96**
25	24.015	**350**	21.4	22.575	**97**
23.3	24.156	**351**	38.7	36.955	**98**
23.3	24.156	**352**	43.8	39.842	**99**
23.3	24.156	**353**	33.2	33.489	**100**
23.3	24.156	**354**	27.5	24.87	**101**
23.3	24.156	**355**	26.5	26.312	**102**
23.3	24.156	**356**	19.3	19.83	**103**
23.3	24.156	**357**	20.1	20.688	**104**
23.3	24.156	**358**	19.5	16.752	**105**
23.3	24.156	**359**	19.5	16.751	**106**
23.3	24.156	**360**	20.4	19.137	**107**
23.3	24.156	**361**	19.8	20.425	**108**
23.3	24.156	**362**	19.4	19.554	**109**
23.3	24.156	**363**	21.7	20.456	**110**
23.3	24.156	**364**	22.8	24.829	**111**
23.3	24.156	**365**	18.8	19.267	**112**
28.7	26.497	**366**	18.7	19.177	**113**
21.5	22.21	**367**	18.5	22.625	**114**
21.5	22.21	**368**	18.3	19.479	**115**
21.5	22.21	**369**	21.2	22.321	**116**
21.5	22.21	**370**	19.2	22.522	**117**
21.5	22.21	**371**	20.4	18.924	**118**
21.5	22.21	**372**	19.3	20.169	**119**
21.5	22.21	**373**	22	20.542	**120**
21.5	22.21	**374**	20.3	20.616	**121**
21.5	22.21	**375**	20.5	18.763	**122**
21.5	22.21	**376**	17.3	15.474	**123**

21.5	22.21	**377**	18.8	18.676	**124**
21.5	22.21	**378**	21.4	20.13	**125**
21.5	22.21	**379**	15.7	13.348	**126**
21.5	22.21	**380**	16.2	15.517	**127**
21.5	22.21	**381**	18	18.993	**128**
23	25.89	**382**	14.3	14.16	**129**
23	25.89	**383**	19.2	19.735	**130**
23	25.89	**384**	19.6	19.555	**131**
23	25.89	**385**	23	19.684	**132**
23	25.89	**386**	18.4	16.189	**133**
23	25.89	**387**	15.6	13.41	**134**
23	25.89	**388**	18.1	17.764	**135**
23	25.89	**389**	17.4	16.265	**136**
23	25.89	**390**	17.1	19.297	**137**
23	25.89	**391**	13.3	13.882	**138**
23	25.89	**392**	17.8	16.302	**139**
23	25.89	**393**	14	13.948	**140**
23	25.89	**394**	14.4	9.296	**141**
23	25.89	**395**	15.6	12.154	**142**
23	25.89	**396**	11.8	9.305	**143**
26.7	30.078	**397**	14.6	9.16	**144**
26.7	30.078	**398**	17.8	9.718	**145**
26.7	30.078	**399**	15.4	13.101	**146**
26.7	30.078	**400**	21.5	17.353	**147**
26.7	30.078	**401**	19.6	17.526	**148**
26.7	30.078	**402**	15.3	16.042	**149**
26.7	30.078	**403**	15.3	16.042	**150**
26.7	30.078	**404**	15.3	16.042	**151**
26.7	30.078	**405**	15.3	16.042	**152**
26.7	30.078	**406**	15.3	16.042	**153**
26.7	30.078	**407**	15.3	16.042	**154**
26.7	30.078	**408**	15.3	16.042	**155**
26.7	30.078	**409**	15.3	16.042	**156**
26.7	30.078	**410**	15.3	16.042	**157**
26.7	30.078	**411**	15.3	16.042	**158**
21.7	21.191	**412**	15.3	16.042	**159**
21.7	21.191	**413**	15.3	16.042	**160**
21.7	21.191	**414**	15.3	16.042	**161**
21.7	21.191	**415**	15.3	16.042	**162**
21.7	21.191	**416**	15.3	16.042	**163**
21.7	21.191	**417**	19.4	15.121	**164**
21.7	21.191	**418**	17	17.793	**165**
21.7	21.191	**419**	17	17.793	**166**
21.7	21.191	**420**	17	17.793	**167**
21.7	21.191	**421**	17	17.793	**168**
21.7	21.191	**422**	17	17.793	**169**
21.7	21.191	**423**	17	17.793	**170**
21.7	21.191	**424**	17	17.793	**171**
21.7	21.191	**425**	17	17.793	**172**
21.7	21.191	**426**	17	17.793	**173**
27.5	29.392	**427**	17	17.793	**174**
27.5	29.392	**428**	17	17.793	**175**
27.5	29.392	**429**	17	17.793	**176**
27.5	29.392	**430**	17	17.793	**177**
27.5	29.392	**431**	17	17.793	**178**
27.5	29.392	**432**	17	17.793	**179**
27.5	29.392	**433**	41.3	37.452	**180**
27.5	29.392	**434**	24.3	27.669	**181**
27.5	29.392	**435**	23.3	24.717	**182**
27.5	29.392	**436**	27	30.374	**183**

27.5	29.392	**437**	27	30.374	**184**
27.5	29.392	**438**	27	30.374	**185**
27.5	29.392	**439**	27	30.374	**186**
27.5	29.392	**440**	27	30.374	**187**
27.5	29.392	**441**	27	30.374	**188**
30.1	28.255	**442**	27	30.374	**189**
44.8	38.23	**443**	27	30.374	**190**
50	39.347	**444**	27	30.374	**191**
37.6	39.52	**445**	27	30.374	**192**
31.6	32.08	**446**	27	30.374	**193**
46.7	40.293	**447**	27	30.374	**194**
31.5	34.972	**448**	27	30.374	**195**
24.3	22.132	**449**	27	30.374	**196**
31.7	34.589	**450**	27	30.374	**197**
41.7	39.526	**451**	50	44.358	**198**
48.3	39.423	**452**	50	46.187	**199**
29	29.945	**453**	50	46.187	**200**
29	29.945	**454**	50	46.187	**201**
29	29.945	**455**	50	46.187	**202**
29	29.945	**456**	50	46.187	**203**
29	29.945	**457**	50	46.187	**204**
29	29.945	**458**	50	46.187	**205**
29	29.945	**459**	50	46.187	**206**
29	29.945	**460**	50	46.187	**207**
29	29.945	**461**	50	46.187	**208**
29	29.945	**462**	50	46.187	**209**
29	29.945	**463**	50	46.187	**210**
29	29.945	**464**	50	46.187	**211**
29	29.945	**465**	50	46.187	**212**
29	29.945	**466**	50	46.187	**213**
29	29.945	**467**	50	45.516	**214**
24	22.734	**468**	50	45.516	**215**
25.1	28.009	**469**	50	45.516	**216**
25.1	28.009	**470**	50	45.516	**217**
25.1	28.009	**471**	50	45.516	**218**
25.1	28.009	**472**	50	45.516	**219**
25.1	28.009	**473**	50	45.516	**220**
25.1	28.009	**474**	50	45.516	**221**
25.1	28.009	**475**	50	45.516	**222**
25.1	28.009	**476**	50	45.516	**223**
25.1	28.009	**477**	50	45.516	**224**
25.1	28.009	**478**	50	45.516	**225**
25.1	28.009	**479**	50	45.516	**226**
25.1	28.009	**480**	50	45.516	**227**
25.1	28.009	**481**	50	45.516	**228**
25.1	28.009	**482**	22.7	23.552	**229**
25.1	28.009	**483**	25	22.645	**230**
31.5	34.364	**484**	50	43.41	**231**
23.7	26.45	**485**	23.8	20.669	**232**
23.3	27.027	**486**	23.8	22.983	**233**
22	24.749	**487**	22.3	23.432	**234**
20.1	21.582	**488**	17.4	20.42	**235**
22.2	21.909	**489**	19.1	22.326	**236**
23.7	26.197	**490**	23.1	20.167	**237**
17.6	16.848	**491**	23.6	27.083	**238**
18.5	14.503	**492**	22.6	23.416	**239**
24.3	21.934	**493**	29.4	31.157	**240**
20.5	19.425	**494**	23.2	23.191	**241**
24.5	21.835	**495**	24.6	27.01	**242**
26.2	26.154	**496**	29.9	30.883	**243**

24.4	24.552	**497**	37.2	34.167	**244**
24.8	28.821	**498**	39.8	34.533	**245**
29.6	30.944	**499**	36.2	25.886	**246**
42.8	32.489	**500**	37.9	34.639	**247**
21.9	22.979	**501**	32.5	28.845	**248**
20.9	21.356	**502**	26.4	20.668	**249**
44	40.618	**503**	29.6	22.663	**250**
50	41.577	**504**	50	37.421	**251**
36	35.992	**505**	32	31.794	**252**
30.1	34.209	**506**	29.8	30.737	**253**

با استفاده از اکسل میزان RMSE به میزان ٤.٧٠٢٠٣٣٣٢٢ هزار دلار برآورد شده است که از مقدار خطای متوسط ۹.۲۱ هزار دلار کمتر است.

٤-٣ نتیجه‌گیری

در این فصل حالات مختلفی که می‌توانستیم در ساخت مدل فصل سوم لحاظ کنیم را ارزیابی کردیم. ملاحظه کردیم زمانیکه گام‌های پاکسازی داده تماماً طی نشود نتیجه مدل قابل اتکا نیست. همچنین این مهم است که گام آخر پاک‌سازی داده یعنی مدیریت داده‌های نامتوازن با استفاده از چه نودی انجام می‌شود. در این پروژه استفاده از نود Reduce بخوبی نود Boost عمل نمی‌کند و حتی می‌تواند نتیجه را از حالتی که هیچ پاک‌سازی داده‌ای انجام نشده است، برای برخی الگوریتم‌ها همچون CHAID و SVM بد تر کند. ما توانستیم با طی کردن گام به گام مراحل پاک‌سازی داده و انتخاب بهترین شیوه پاک‌سازی در هر مرحله، خطای متوسط ۹.۲۱ هزار دلار را به میزان ٤.٥٠٧٩٦٦٦٧٨ هزار دلار کاهش دهیم.