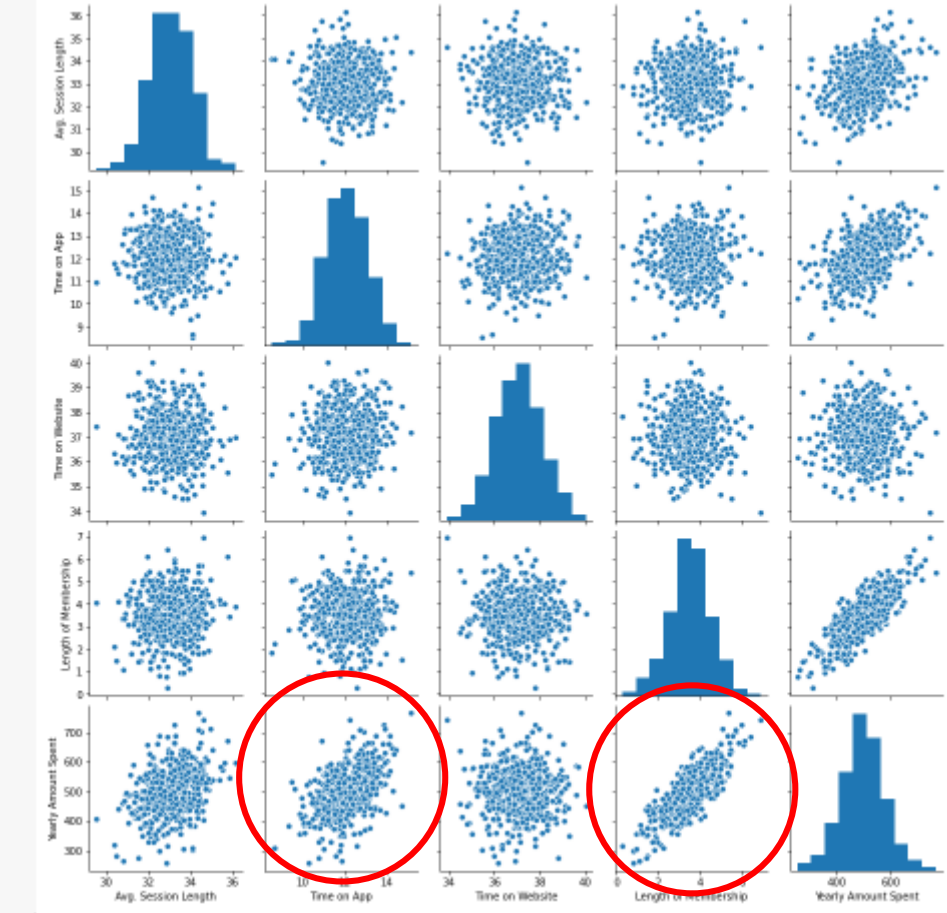
پیاده سازی و مراحل جواب همه در فایل Ecommerce Customers.ipynb

آ)

۵ ستون داده عددی داریم. آخرین ستون یعنی ستون Yearly Amount Spent را به عنوان هدف یعنی y تعین میکنم. بعد از نمودارهای بدست آمده :

اول ۴ سطر و ۴ ستون اول را که بررسی کنیم نشان می دهد که بین متغیرهای x_i رابطه خطی و همبستگی ای وجود ندارد. در مورد سطر آخر اما یعنی رابطه همبستگی متغیر هدف با هر یک از متغیرها به نظر میرسد تنها ارتباط خطی قابل توجه بین Yearly Amount Spent و ستون ها ی Time on App و Length of Membership وجود دارد که از بین این دو، دومی ارتباط معنادارتر و منسجم تری به نظر میرسد.



برای بررسی دقیق تر نمودار بعدی را داریم :

با توجه به شاخص راهنمای کنار، روی ۴ سطر و ستون اولی یعنی متغیرهای براوردیابی همبستگی دیده نمیشود (بین Time on App و Time on Website یک مقدار همبستگی ناچیز وجود دارد – در حد ۰٫۱ که قابل چشم پوشی است) اما در سطر و ستون آخر به غیر از Time on Website بقیه ۳ ستون دیگر درجات خوبی از همبستگی را با متغیر هدف دارند. این مقدار برای Time on App و Length of Membership به

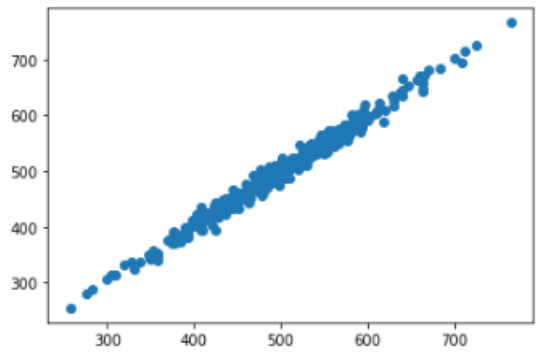
ترتیب حدود ۰٫۵ و ۰٫۸ است ودر مورد ستون Avg. Session Length به مقدار ۰٫٤ افت میکند. این نمودار به طور دقیق تری نتایج نمودار قبلی را تایید میکند و علاوه بر آن معیار عددی هم به دست میدهد. از روی این نایج اولیه مشخص میشود که مقدار Length of Membership بیشترین تاثیر را روی Yearly Amount Spent دارد و بعد از ان Time on App و Avg. Session Length به ترتیب و در نهایت Time on Website کمترین تاثیر را دارند.

ب)

اول ۲۵ درصد داده ها را برای تست جدا میکنم. (کل داده ها ۵۰۰ تا بوده) حالا از روی ۳۷۵ داده باقیمانده مدل رگرسیون خطی چندگانه را برازش میدهم. برای محاسبه خطای اموزش هم مقدار MSE و جذر ان را فقط روی همین مجموعه ۳۷۵ تا داده اموزشی بدست میاورم. مقدار خطاهای بدست امده به صورت زیر :

	Train set
MSE	*103.18965151113609*
RMSE	*10.158230727402096*

و اگر بخواهیم نتیجه همسانی مقادیر واقعی متغیرهدف داده های اموزش و مقدار براورد شده انها را ببینیم نمودار انحراف کمی نسبت به خط نیمساز دارد که نشان میدهد مدل قابل قبولی داریم.



داخل کادر سبز در سطر اول مقدار عرض از مبدا برابر 1049- و در سطرهای بعدی مقدار ضریب هر متغیر در مقابل ان امده. همانطور که از قسمت قبلی هم انتظار داشتیم، ضریب همبستگی با متغیر Time on Website یک مقدار بسیار ناچیز نزدیک به صفر است و بیشترین ضریب همبستگی را مقدار +٦٠ دارد که مربوط به Length of Membership است که بزرگی این هم از روی نتایج قسمت قبل قابل پیشبینی بود.

داخل کادر قرمز مقادیر پی نوشته شده که برای ارزیابی مدل استفاده میکنیم. تمام سطر ها مقدار پی بسیار کوچک و برای ٣ متغیر برابر صفر نشان داده اند و این یعنی میتوان فرض اچ صفر را با اطمینان زیادی رد کرد. یعنی رابطه خطی بین y و x های ما وجود دارد. تنها موردی که مشکل ساز است مقدار نظیر سطر چهارم برابر 0.173 است. اگر مطابق معمول مقدار الفا را برای ازممون فرض برابر 0.05 بگیریم فرض اچ صفر در مورد این یکی را نمیتوان رد کرد. اگر مقدار بازه اطمینان داده شده برای ضریب نظیر این متغیر را در دو ستون بعدی (کادر زرد) چک کنیم تنها ضریبی که صفر داخل بازه اطمینان قرار میگیرد مربوط به این مورد است. فرض اچ صفر در این مورد را نمی توان رد کرد ولی به غیر از این متغیر، بقیه نتایج قابل قبولی بدست داده اند.

آخرین موردی که برای ارزیابی مدل بررسی می کنم (با توجه به اینکه رگرسیون چندگانه است) مقدار نظیر فلش قرمز است که برابر Adj. R-squared است و هرچه این مقدار بزرگتز و نزدیک به یک باشد مدل وضعیت مطلوب تری دارد. این عدد 0.981 بدست امده و بیسار خوب است.

Dep. Variable:	Q("Yearly Amount Spent")	R-squared:	0.984
Model:	OLS	Adj. R-squared:	0.984
Method:	Least Squares	F-statistic:	5674.
Date:	Fri, 29 May 2020	Prob (F-statistic):	0.00
Time:	11:03:10	Log-Likelihood:	-1401.5
No. Observations:	375	AIC:	2813.
Df Residuals:	370	BIC:	2833.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1049.4457	27.160	-38.639	0.000	-1102.853	-996.038
Q("Avg. Session Length")	25.7315	0.516	49.900	0.000	24.717	26.745
Q("Time on App")	38.8872	0.540	71.975	0.000	37.825	39.950
Q("Time on Website")	0.3199	0.539	0.593	0.554	-0.741	1.381
Q("Length of Membership")	61.5267	0.531	115.868	0.000	60.483	62.571

Omnibus:	0.355	Durbin-Watson:	1.734
Prob(Omnibus):	0.837	Jarque-Bera (JB):	0.271
Skew:	0.064	Prob(JB):	0.873
Kurtosis:	3.028	Cond. No.	2.64e+03

ج)

از ۵ دسته استفاده کردم. روس همان ۳۷۵ تا داده اموزش تابع را اجرا کردم تا تقریبی از خطا بدست بیاید. مقدار MSE وRMSE که برای تست پیش بینی می شود برابر است با :

	Predicted
MSE	*104.86734425049633*
RMSE	*10.233088271425057*

د)

خطای بدست امده برای ۱۲۵ داده تست که کنار گذاشته بودیم برابر است با:

	Test set
MSE	*84.86573160592083*
RMSE	*9.212259853365016*

و از قسمت قبل برای داده های اموزش داشتیم :

	Train set
MSE	*103.18965151113609*
RMSE	*10.158230727402096*

خطای RMSE را که مقایسه کنیم اختلاف بسیار ناچیز است. یعنی خطای تست و اموزش مقداری بسیار نزدیک به هم دارند. از این جهت میتوان گفت که مدل overfit نیست. چون در آن صورت این اختلاف عدد بسیار بزرگی میشد و RMSE داده های تست مقدار خیلی بزرگتری از داده های اموزشی داشت. پس اینجا بیش برازش یا overfit نداریم و سطح پیچیدگی انتخاب شده (خطی) برای براورد هدف زیادتر از چیزی که نیاز داریم نیست.

از طرف دیگر مقدار خطای بدست آمده عدد بزرگی نیست اگر به عدد های بردار y نگاه کنیم اختلاف ۱۰ تا به ازای اعدادی در اندازه ی 500 (میانگین) مقدار خطای قابل توجهی نیست. اما اگر بخواهم دلیل قطعی تری مطرح کنم، به قسمت قبل ارجاع می دهم. معمولا خطای underfit را برای زمانی مطرح میکنیم که پیچیدگی مدل برای داده های استفاده شده و رابطه انها با متغیر هدف کم است. مثلا وقتی مدلی که ضریب همبستگی خطی خوبی ندارد را روی یک مدل خطی برازش میدهیم مدلی که بدست میاید خطای زیادی دارد چون به اندازه ای پیچیده نیست که روابط بین متغیرها را پوشش دهد. پس خطا در نهایت برای هم داده های اموزش و هم تست زیاد میشود. حالا به قسمت قبلی برگردم و مقدار پی هایی که بدست امد و نشان دهنده ی خطی بودن رابطه x و y ها بود. و مدل استفاده شده اینجا هم مدل خطی است. میتوان گفت که پیچیدگی مدل کمتر از داده ها نیست و اگر مدل غیرخطی پیچده تری انتخاب میکردیم، دچار بیش برازش میشدیم. پس نتیجه اینکه underfit هم نیست و مدل دقیقا پیچیدگی مطلوب و خطای قابل قبولی دارد. البته برای اینکه دقیقا تعین کنیم بهترین نقطه ی bias–variance tradeoff کجا اتفاق می افتد باید از محاسبات بیشتر که بلد نیستم استفاده کنم یا اینکه از روش cross-validation استفاده کنم که اینجا خواسته نشده. پس صرف اطلاعاتی که در این سطح از مدل و خطاها دارم به نظرم مدل در تعادل bias–variance هست و عملکرد مطلوبی دارد.

۵)

روی متغیری باید سرمایه گذاری کرد که افزایش آن افزایش بیشتری در مقدار هدف داشته باشد. یعنی آن که وابستگی بیشتری داشته باشد. اگر این را یک مسئله بهینه سازی در نظر بگیریم، آن متغیری باید مقدار بگیرد که ضریب بزرگتری در تابع هدف دارد. تابع هدف را برابر معادله رگرسیونی بدست آمده در قسمت ب فرض کنیم و اعداد را از کادر سبز جدول قسمت ب جاگذاری کنم:

$$y = 25.731489\, x_1 + 38.887210\, x_2 + 0.319876\, x_3 + 61.5267\, x_4 - 1049.44$$

حالا سوال بین ستون Time on Website و Time on App دنبال متغیر موثرتر میگردد. بین ضریب x_2 و x_3 باید بزرگترین را انتخاب کنیم. نتیجه اینکه از مدل بدست امده، تمرکز روی اپلیکیشن میتواند زمان استفاده از ان را افزایش دهد و این مقدار با ضریب حدود ۴۰ بر مقدار هدف تاثیرگذار است در حالی که مقدار استفاده از وبسایت تاثیر خطی چندانی ندارد!