به نام خدا

1928
K. N. Toosi University of Technology

دانشگاه صنعتی خواجه نصیرالدین طوسی

دانشکده برق

مبانی سیستم های هوشمند

گزارش مینی پروژه ۱

سیده زهرا عربی

۴۰۰۰۷۱۷۳

استاد : آقای دکتر مهدی علیاری

آذر ۱۴۰۳

فهرست مطالب

عنوان شماره صفحه

https://github.com/Zahra-Arabi/MJAHMADEE.git

https://colab.research.google.com/drive/1G7SBPrHMCVuey0Pn-m8QP8KHc4OJYp8q?usp=sharing

بخش۱: سوالات شبیه سازی

سوال ۱

۱.۱

این داده‌ها شامل اطلاعات ۱۰,۰۰۰ مشتری بانک است، مثل سن، درآمد، وضعیت تأهل، حد اعتبار کارت و نوع کارت اعتباری. در مجموع، ۱۸ ویژگی در این مجموعه وجود دارد.

۱۶.۰۷٪ از مشتری‌ها خدمات کارت اعتباری را ترک کرده‌اند، به همین دلیل پیش‌بینی مشتریانی که ممکن است در آینده خدمات را ترک کنند کمی سخت است.

مدیر بانک می‌خواهد از این داده‌ها استفاده کند تا با پیش‌بینی مشتریانی که ممکن است خدمات را ترک کنند، به آن‌ها خدمات بهتری ارائه دهد و نظرشان را تغییر دهد تا از رفتن آن‌ها جلوگیری کند.

با دستور زیر ابتدا داده را خوانده و سپس ویژگی های آن را نمایش دادم. و تایپ هر یک از این ویژگی ها نیز با این دستور نمایش داده شد.

```
# Read the CSV file into a DataFrame
data = pd.read_csv('BankChurners.csv')

# Check the data type of each column
print(data.dtypes)
```

```
CLIENTNUM                 int64
Attrition_Flag           object
Customer_Age              int64
Gender                   object
Dependent_count           int64
Education_Level          object
Marital_Status           object
Income_Category          object
Card_Category            object
Months_on_book            int64
Total_Relationship_Count  int64
Months_Inactive_12_mon    int64
Contacts_Count_12_mon     int64
Credit_Limit            float64
Total_Revolving_Bal       int64
Avg_Open_To_Buy         float64
Total_Amt_Chng_Q4_Q1    float64
Total_Trans_Amt           int64
Total_Trans_Ct            int64
Total_Ct_Chng_Q4_Q1     float64
Avg_Utilization_Ratio   float64
```

Naive_Bayes_Classifier_Attrition_Flag_Card_Category_Contacts_Count_12_mon_Dependent_count_Education_Level_Months_Inactive_12_mon_1 float64
Naive_Bayes_Classifier_Attrition_Flag_Card_Category_Contacts_Count_12_mon_Dependent_count_Education_Level_Months_Inactive_12_mon_2 float64
dtype: object

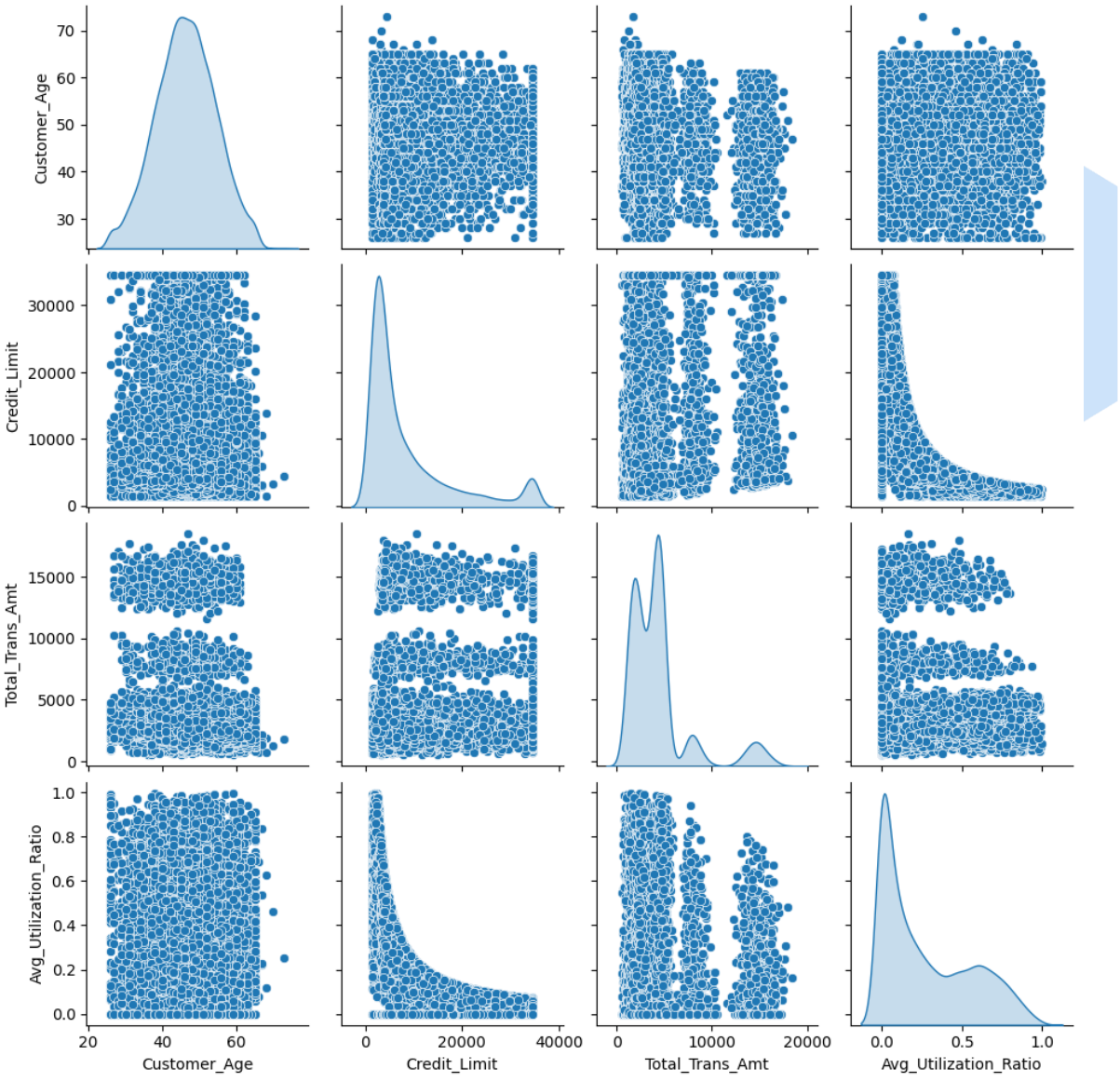
---

۲۳ ویژگی در این داده وجود دارد که تایپ آنها نیز مشخص شده است.

```
print(data.info())
```

این داده شامل ۱۰,۱۲۷ نمونه و ۲۳ ویژگی است. بیشتر ستون‌ها عددی هستند (۱۰ عدد صحیح و ۷ عدد اعشاری) و ۶ ستون متنی یا دسته‌بندی دارند. هیچ داده‌ی خالی وجود ندارد.

۱.۲



در این سوال ابتدا کلید ویژگی هایی که میخواهم نمایش داده شوند را در یک لیست ذخیره میکنم. توزیع ویژگی ها به صورت Kernel Density Estimation است و داده ها با علامت 'o' نمایش داده شده اند.

ویژگی های انتخاب شده عبارت هستند از سن مشتری، حد اعتبار کارت، کل مبلغ تراکنش ها و میانگین نسبت استفاده از کارت.

تحلیل نمودار:

۱. نمودار قطر اصلی (دایگنال):

در این نمودار توزیع هر ویژگی به صورت منحنی تخمین چگالی نمایش داده شده است.

-سن مشتری:

توزیع سن مشتریان به شکل یک منحنی برجسته است که نشان می‌دهد بیشتر مشتریان در سنین میان‌سال (حدود ۳۰ تا ۶۰ سال) قرار دارند. این نشان‌دهنده توزیع نرمالی است.

-حد اعتبار کارت:

توزیع حد اعتبار نشان‌دهنده چولگی (skewness) به سمت راست است، به این معنی که بیشتر مشتریان حد اعتبار کمتری دارند و تعداد کمی از مشتریان دارای حد اعتبار بالا هستند.

-کل مبلغ تراکنش ها:

این توزیع هم مشابه به توزیع حد اعتبار است که به سمت راست چولگی دارد. به این معنی که تراکنش‌های بیشتر توسط تعداد کمی از مشتریان انجام شده است، در حالی که بیشتر مشتریان دارای تراکنش‌های کمتری هستند.

- میانگین استفاده از کارت

توزیع این ویژگی به سمت صفر متمایل است و نشان می‌دهد که بیشتر مشتریان از اعتبار خود استفاده نکرده‌اند یا به میزان کمی استفاده کرده‌اند. این نشان‌دهنده چولگی منفی است.

۲. نمودار های پراکندگی

نمودارهای غیر قطری (غیر دایگنال) ارتباط بین جفت ویژگی‌ها را نشان می‌دهند.

-سن مشتریان و حد اعتبار کارت

در این نمودار پراکندگی، هیچ همبستگی مشخصی بین سن و حد اعتبار مشتریان مشاهده نمی‌شود. حد اعتبار لزوماً وابسته به سن مشتریان نیست، اگرچه ممکن است افراد مسن‌تر اعتبار بالاتری داشته باشند، اما این رابطه ضعیف است.

-سن مشتریان و کل مبلغ تراکنش ها

در این نمودار، هیچ الگوی خاصی در ارتباط بین سن و مبلغ تراکنش‌ها مشاهده نمی‌شود. تراکنش‌ها بیشتر به تعداد مشتریان وابسته است تا سن.

- سن مشتریان و میانگین استفاده از کارت

هیچ ارتباط قابل‌توجهی بین سن مشتریان و نسبت استفاده از کارت مشاهده نمی‌شود.

-حد اعتبار کارت و کل مبلغ تراکنش

در این نمودار، مشخص است که برخی از مشتریان با حد اعتبار پایین، تراکنش‌های بیشتری دارند، در حالی که مشتریانی با حد اعتبار بالا ممکن است تراکنش‌های کمتری داشته باشند. به طور کلی، تراکنش‌ها بیشتر از اعتبار استفاده شده‌اند.

-حد اعتبار کارت و میانگین استفاده از کارت

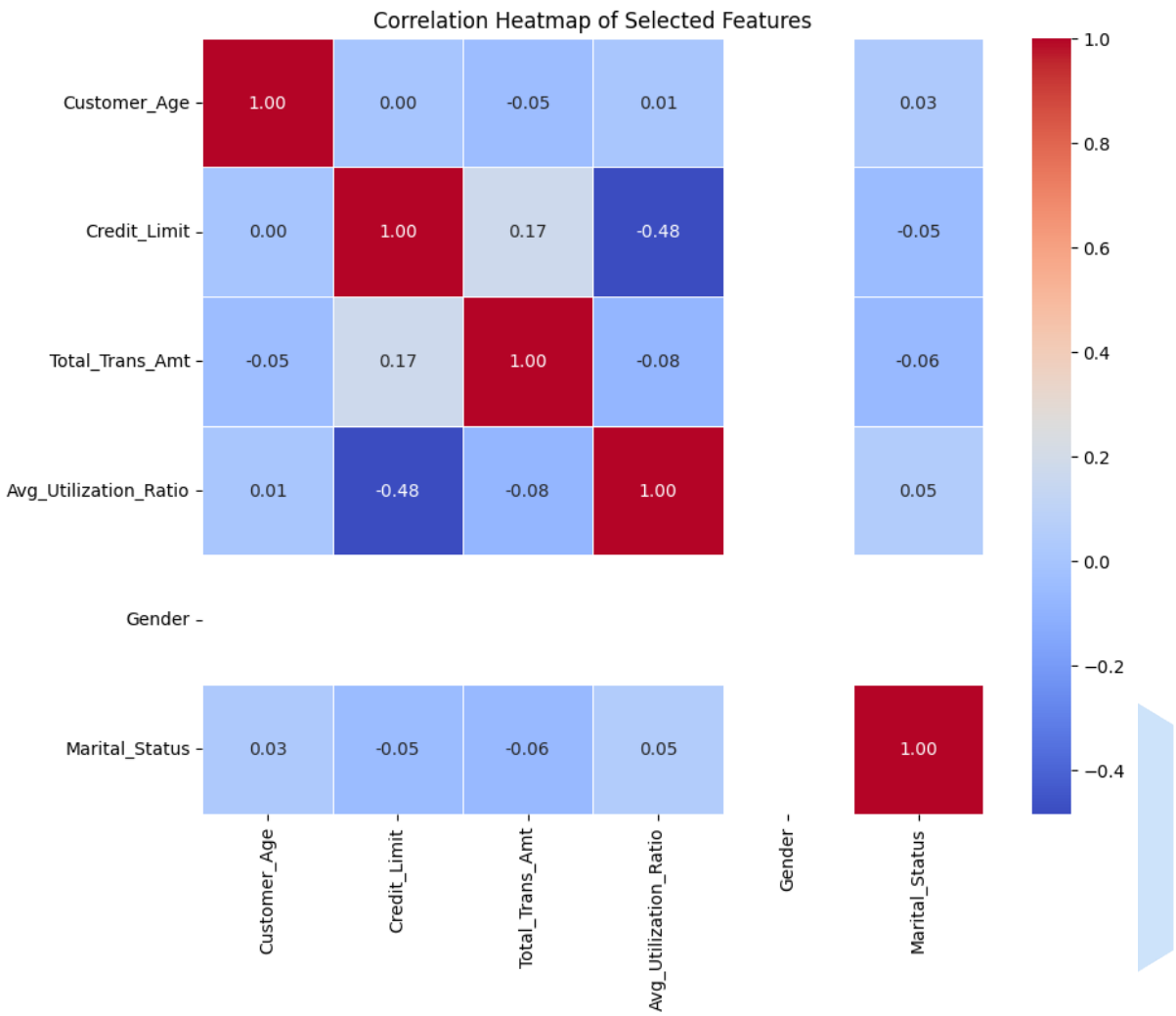
در این نمودار نیز چولگی منفی دیده می‌شود. مشتریانی با حد اعتبار بالا معمولاً از اعتبار خود کمتر استفاده کرده‌اند، در حالی که مشتریانی با حد اعتبار پایین از اعتبار خود بیشتر استفاده می‌کنند.

-کل مبلغ تراکنش و میانگین استفاده از کارت

بین این دو ویژگی هم همبستگی قابل توجهی وجود ندارد، به طوری که برخی مشتریان با تراکنش‌های بالا از اعتبار خود کمتر استفاده کرده‌اند و برخی دیگر به طور معکوس عمل کرده‌اند.

این نمودارها به ما کمک می‌کند تا روابط مختلف بین ویژگی‌ها را بررسی کنید. به نظر می‌رسد که ویژگی‌های مختلف (مانند حد اعتبار و تراکنش‌ها) بیشتر از هم تأثیر می‌پذیرند تا از سن یا نسبت استفاده از اعتبار.

۱.۳



در این سوال پس از انتخاب ویژگی های پیوسته و طبقه بندی شده، ویژگی های طبقه بندی شده که به صورت object هستند را به صورت 0 و 1 تغییر میدهیم.(مرد بودن با 0 و زن بودن با 1 – مجرد بودن 0 و متاهل بودن 1)

سپس ماتریس همبستگی را محاسبه کرده و نقشه حرارتی را نمایش میدهیم.

تحلیل نقشه حرارتی:

-سن مشتریان و اعتبار کارت

همبستگی ضعیف (۰.۰۰) دارد که نشان می‌دهد سن مشتری تاثیری روی اعتبار کارت ندارد.

-سن مشتریان با کل مبلغ تراکنش ها

همبستگی ضعیف منفی (-۰.۰۵)، نشان می‌دهد که رابطه مشخصی بین سن و مبلغ تراکنش‌های کل وجود ندارد.

-سن مشتریان و میانگین استفاده از کارت

همبستگی بسیار ضعیف (۰.۰۱)، نشان می‌دهد که هیچ رابطه معنی‌داری بین سن و نسبت استفاده از اعتبار وجود ندارد.

-سن مشتریان با جنسیت و وضعیت تاهل

همبستگی‌ها بسیار کم (۰.۰۳ و ۰)، یعنی سن تاثیر زیادی بر جنسیت یا وضعیت تأهل ندارد.

-اعتبار کارت با کل مبلغ تراکنش ها

همبستگی مثبت ضعیف (۰.۱۷)، به این معنی که با افزایش حد اعتبار، احتمالاً مقدار تراکنش‌ها نیز اندکی افزایش می‌یابد.

-اعتبار کارت با میانگین نسبت استفاده از کارت

همبستگی منفی قابل توجه (-۰.۴۸)، نشان می‌دهد که مشتریانی که اعتبار بالاتری دارند، نسبت استفاده از اعتبار پایین‌تری دارند.

-اعتبار کارت با جنسیت و وضعیت تاهل

همبستگی‌ها بسیار ضعیف، بنابراین جنسیت و وضعیت تأهل تاثیر خاصی بر حد اعتبار ندارند.

-کل مبلغ تراکنش ها با میانگین نسبت استفاده از کارت

همبستگی منفی ضعیف (-۰.۰۸)، به این معنی که تراکنش‌های بیشتر با نسبت استفاده از اعتبار بالا همراه نیستند.

-کل مبلغ تراکنش ها با جنسیت و وضعیت تاهل

همبستگی‌ها بسیار ضعیف (-۰.۰۶ و ۰.۰۵)، نشان می‌دهد که مقدار تراکنش‌ها هیچ رابطه معنی‌داری با جنسیت یا وضعیت تأهل ندارد.

-میانگین نسیت استفاده از کارت با سایر ویژگی ها

این ویژگی همبستگی‌های بسیار ضعیفی با دیگر ویژگی‌ها دارد، که نشان‌دهنده این است که نسبت استفاده از اعتبار بیشتر به تراکنش‌های زیاد یا ویژگی‌های دیگر مانند سن یا وضعیت تأهل وابسته نیست.

-جنسیت و وضعیت تاهل

همبستگی در این دو ویژگی برابر با ۱ است که طبیعی است، زیرا این دو ویژگی به صورت طبقه‌بندی‌شده و کدگذاری‌شده به اعداد تبدیل شده‌اند. بنابراین، در ماتریس همبستگی این دو ویژگی با هم همبستگی کامل دارند.

ویژگی های پیوسته که در این سوال انتخاب کردم هیچ همبستگی قوی و معنی داری با هم ندارند به جز یک همبستگی منفی بین حد اعتبار کارت و میانگین نسبت استفاده از کارت.

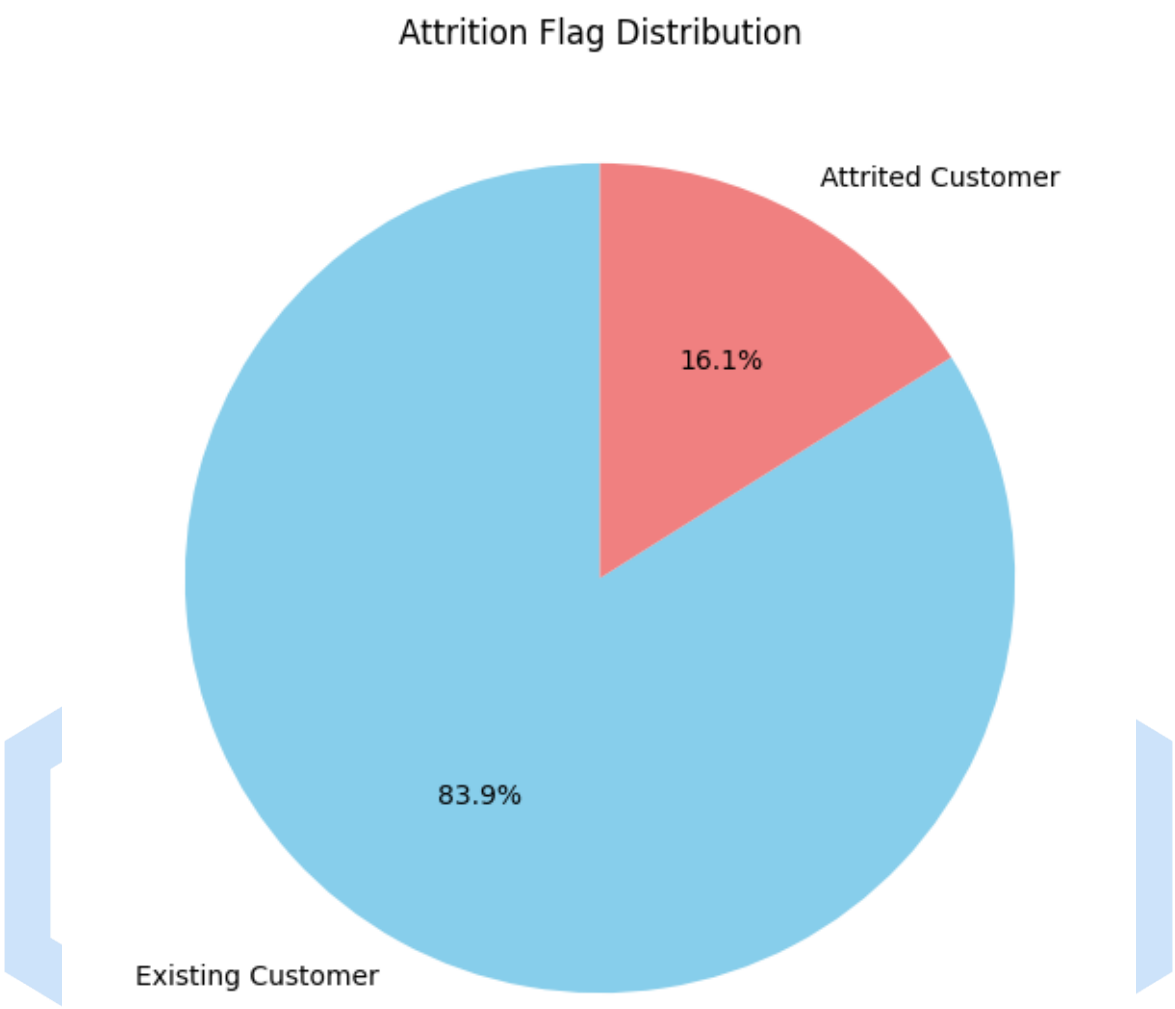
ویژگی های طبقه بندی شده به دلیل کدگذاری، تاثیر چندانی بر همبستگی با سایر ویژگی‌ها ندارند.

۱.۴

ابتدا داده های None را بررسی کردم و سپس در صورت وجود آنهارو حذف کردم و اطلاعات مربوط به داده None حذف شده رو بررسی کردم.

نتیجه‌ی بررسی داده‌ها نشان می‌دهد که هیچ داده‌ی NaN (مقدار گمشده) در هیچ‌کدام از ستون‌ها وجود ندارد. تمامی ۱۰,۱۲۷ نمونه و ۲۳ ویژگی به طور کامل پر شده‌اند و هیچ داده‌ی ناقصی در دیتاست شما وجود ندارد. به عبارت دیگر، پس از بررسی با استفاده از isna().sum()، تمامی مقادیر در دیتاست مقداردهی شده و هیچ نیازی به حذف ردیف‌ها یا پر کردن مقادیر گمشده نمی‌باشد.

۱.۵



ویژگی Attrition_Flag یک ویژگی طبقه‌بندی است که نشان می‌دهد مشتریان از بانک جدا شده‌اند یا خیر. این ویژگی دارای دو کلاس است.

Attrited Customer: مشتریانی که از کلاس بانک شده اند. (16.1%)

Existing Customer: مشتریانی که همچنان با بانک در ارتباط هستند. (83.9%)

عدم تعادل در داده‌ها زمانی که یک کلاس به شدت غالب باشد، می‌تواند بر عملکرد مدل پیش‌بینی تأثیر بگذارد. در مدل‌های طبقه‌بندی، اگر یک کلاس نسبت به دیگری تعداد بیشتری نمونه داشته باشد، مدل بیشتر بر روی پیش‌بینی کلاس غالب متمرکز می‌شود و احتمال دارد که دقت پیش‌بینی برای کلاس نادر (کمتر ظاهر شده) پایین بیاید. به طور خاص، مدل احتمالاً پیش‌بینی‌های نادرست برای کلاس نادر داشته باشد. عملکرد خوبی در تشخیص کلاس نادر نداشته باشد، حتی اگر دقت کلی مدل بالا باشد.

برای اصلاح عدم تعادل کلاس‌ها می‌توان از روش‌های مختلفی استفاده کرد:

- **Resampling**:
  - Oversampling: نمونه‌های کلاس نادر را با استفاده از روش‌هایی مانند SMOTE (Synthetic Minority Over-sampling Technique) اضافه کرد.
  - Undersampling: تعداد نمونه‌های کلاس غالب را کاهش داد تا تعداد کلاس‌ها متعادل شود.
- **استفاده از الگوریتم‌های خاص:**
  - الگوریتم‌هایی مانند XGBoost یا Random Forest که معمولاً می‌توانند با داده‌های نامتوازن بهتر کار کنند.
  - تنظیم وزن کلاس‌ها برای بهینه‌سازی مدل برای کلاس نادر.
- استفاده از متریک‌های خاص:
  - به جای دقت (Accuracy)، از متریک‌های دیگر مانند Precision, Recall, F1-score, و ROC-AUC استفاده کرد تا تأثیرات عدم تعادل بر ارزیابی مدل کاهش یابد.

بهتر است که متعادل‌سازی داده‌ها را قبل از تقسیم‌بندی داده به بخش‌های آموزش و آزمون انجام داد. علت این است که اگر شما داده‌ها را ابتدا تقسیم کنید، ممکن است نمونه‌های نادر در هر بخش (آموزش یا آزمون) به درستی توزیع نشوند. اگر داده‌ها را قبل از تقسیم متعادل کنید، مطمئن خواهید بود که مدل به طور یکنواخت به هر دو کلاس دسترسی دارد. به طور کلی، باید مراقب باشیم که داده‌های آزمون را تحت تأثیر متعادل‌سازی قرار ندهیم، زیرا داده‌های آزمون باید منعطف و طبیعی باشند تا مدل در هنگام ارزیابی عملکرد خود در دنیای واقعی دچار مشکل نشود.

۱.۶

برای انجام مراحل درخواست‌شده، ابتدا داده‌ها را به ویژگی‌های ورودی و خروجی تقسیم کرده و سپس داده‌ها را به سه بخش آموزش، اعتبارسنجی و آزمون تقسیم کردم. سپس از یک الگوریتم طبقه‌بندی از کتابخانه scikit-learn استفاده کردم تا مدل را آموزش داده و گزارش‌های طبقه‌بندی و ماتریس درهم‌ریختگی را برای داده‌های آموزش و اعتبارسنجی گزارش کنم.

آموزش بدون متعادل سازی

۱. تقسیم‌بندی داده‌ها به ورودی‌ها و خروجی‌ها: ویژگی Attrition_Flag به عنوان خروجی و بقیه ویژگی‌ها به عنوان ورودی در نظر گرفته می‌شود.

۲. پیش پردازش داده ها: ویژگی های متنی به مقادیر عددی تبدیل شدند.

۳. تقسیم داده‌ها به سه بخش (آموزش، اعتبارسنجی و آزمون) : داده‌ها را به نسبت دلخواه (۷۰٪ برای آموزش، ۱۵٪ برای اعتبارسنجی و ۱۵٪ برای آزمون) تقسیم کردم.

۴. انتخاب یک الگوریتم طبقه‌بندی : در اینجا، از الگوریتم Random Forest Classifier استفاده کردم که در scikit-learn موجود است.

۵. آموزش مدل و ارزیابی: یک مدل Random Forest با ۱۰۰ درخت ایجاد و بر روی داده‌های آموزشی بدون اعمال هرگونه متعادل‌سازی آموزش داده شده است.

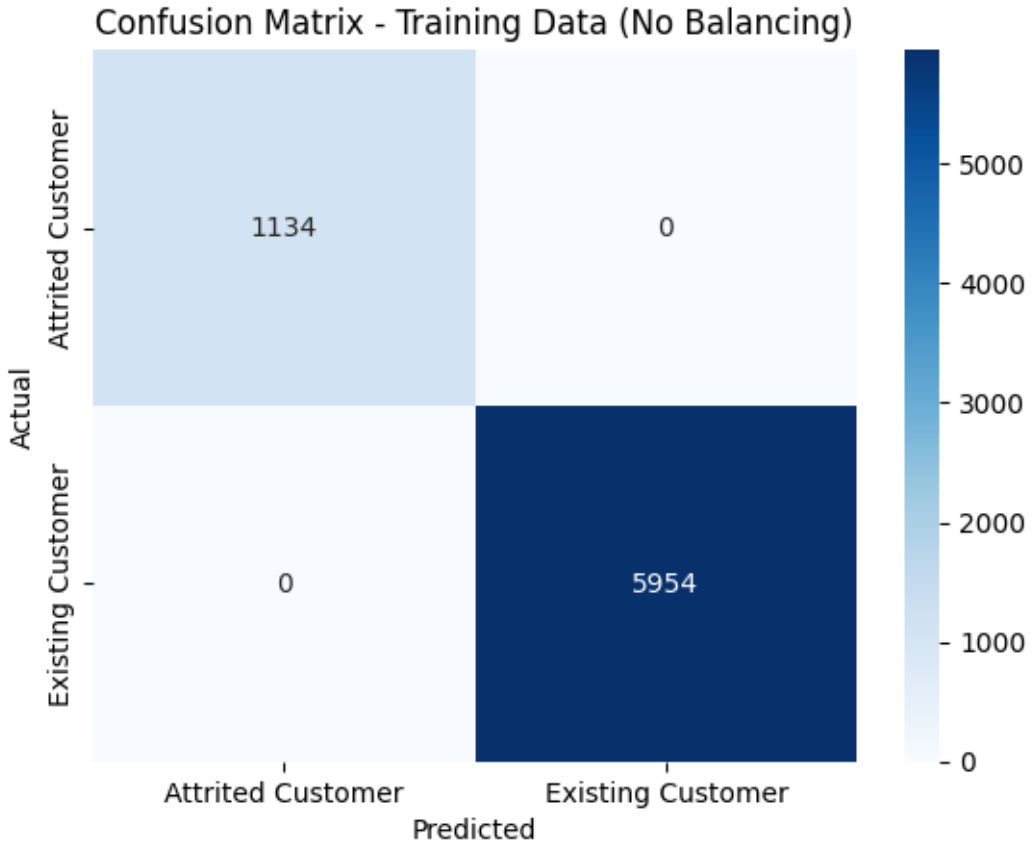
۶. پیش بینی و گزارش عملکرد: پیش‌بینی‌ها برای داده‌های آموزش و اعتبارسنجی انجام شده است. گزارش طبقه‌بندی و ماتریس درهم‌ریختگی (Confusion Matrix) برای داده‌های آموزشی و اعتبارسنجی نمایش داده شده‌اند.

براساس داده های تست نتایج زیر به دست آمد:

```
Classification Report for Training Data (No Balancing):
                   precision    recall  f1-score   support

Attrited Customer       1.00      1.00      1.00      1134
Existing Customer       1.00      1.00      1.00      5954

         accuracy                           1.00      7088
        macro avg       1.00      1.00      1.00      7088
     weighted avg       1.00      1.00      1.00      7088
```



برای هردو کلاس Attrited Customer و Existing Customer میزان دقت، بازخوانی و F1_score برابر ۱ هست . این مقدار نشان میدهد مدل ما به خوبی کار میکند اما دقت ۱۰۰ درصد اصلا مطلوب نیست و نشان دهنده overfitting شدن مدل است.

علت overfitting شدن مدل به دلیل عدم تناسب بین دو کلاس این داده است. (۱۶ به ۸۴ !!!!!!!!!)

تعداد بسیار کم از کلاس "Attrited Customer" در داده‌ها باعث می‌شود که مدل بیشتر بر روی "Existing Customer" تمرکز کند، حتی اگر آن مشتری‌ها از دست رفته باشند.

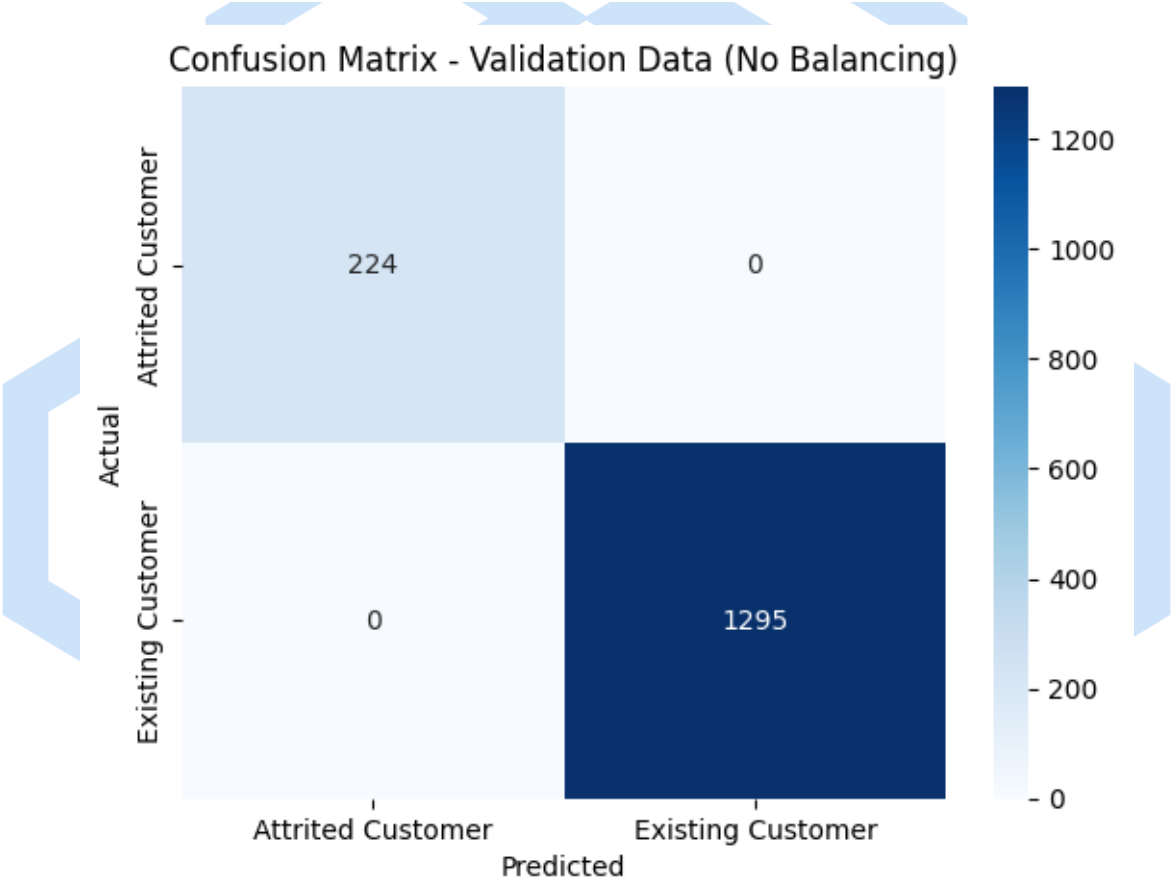
در داده های تست ۷۰۸۸ نمونه وجود دارد که ۱۱۳۴ داده برای کلاس Attrited Customer و ۵۹۵۴ داده برای کلاس Existing Customer است.

در تصویر هم مشاهده میشود ۵۹۵۴ نمونه به درستی در کلاس Existing Custome و ۱۱۳۴ نمونه به درستی در کلاس Attrited Customer تشخیص داده شده اند و صفر نمونه اشتباه تشخیص داده شده وجود دارد.

براساس داده های ارزیابی نتایج زیر به دست آمد:

Classification Report for Validation Data (No Balancing):

	precision	recall	f1-score	support
Attrited Customer	1.00	1.00	1.00	224
Existing Customer	1.00	1.00	1.00	1295
accuracy			1.00	1519
macro avg	1.00	1.00	1.00	1519
weighted avg	1.00	1.00	1.00	1519



برای هردو کلاس Attrited Customer و Existing Customer میزان دقت، بازخوانی و F1_score برابر ۱ هست . این مقدار نشان میدهد مدل ما به خوبی کار میکند اما دقت ۱۰۰ درصد اصلا مطلوب نیست و نشان دهنده overfitting شدن مدل است.

علت overfitting شدن مدل به دلیل عدم تناسب بین دو کلاس این داده است. (۱۶ به ۸۴ !!!!!!!!!)

تعداد بسیار کم از کلاس "Attrited Customer" در داده‌ها باعث می‌شود که مدل بیشتر بر روی "Existing Customer" تمرکز کند، حتی اگر آن مشتری‌ها از دست رفته باشند.

در داده های ارزیابی ۱۵۱۹ نمونه وجود دارد که ۲۲۴ داده برای کلاس Attrited Customer و ۱۲۹۵ داده برای کلاس Existing Customer است.

در تصویر هم مشاهده میشود ۱۲۹۵ نمونه به درستی در کلاس Existing Custome و ۲۲۴ نمونه به درستی در کلاس Attrited Customer تشخیص داده شده اند و صفر نمونه اشتباه تشخیص داده شده وجود دارد.

آموزش با متعادل سازی

۱. تقسیم‌بندی داده‌ها به ورودی‌ها و خروجی‌ها: ویژگی Attrition_Flag به عنوان خروجی و بقیه ویژگی‌ها به عنوان ورودی در نظر گرفته می‌شود.

۲. پیش پردازش داده ها: ویژگی های متنی به مقادیر عددی تبدیل شدند.

۳. تقسیم داده‌ها به سه بخش (آموزش، اعتبارسنجی و آزمون) : داده‌ها را به نسبت دلخواه (۷۰٪ برای آموزش، ۱۵٪ برای اعتبارسنجی و ۱۵٪ برای آزمون) تقسیم کردم.

۴. متعادل سازی: با استفاده از **SMOTE** داده ها را متعادل کردم و اینکار باید بعد از تقسیم داده ها انجام شود.

۴. انتخاب یک الگوریتم طبقه‌بندی : در اینجا، از الگوریتم Random Forest Classifier استفاده کردم که در scikit-learn موجود است.

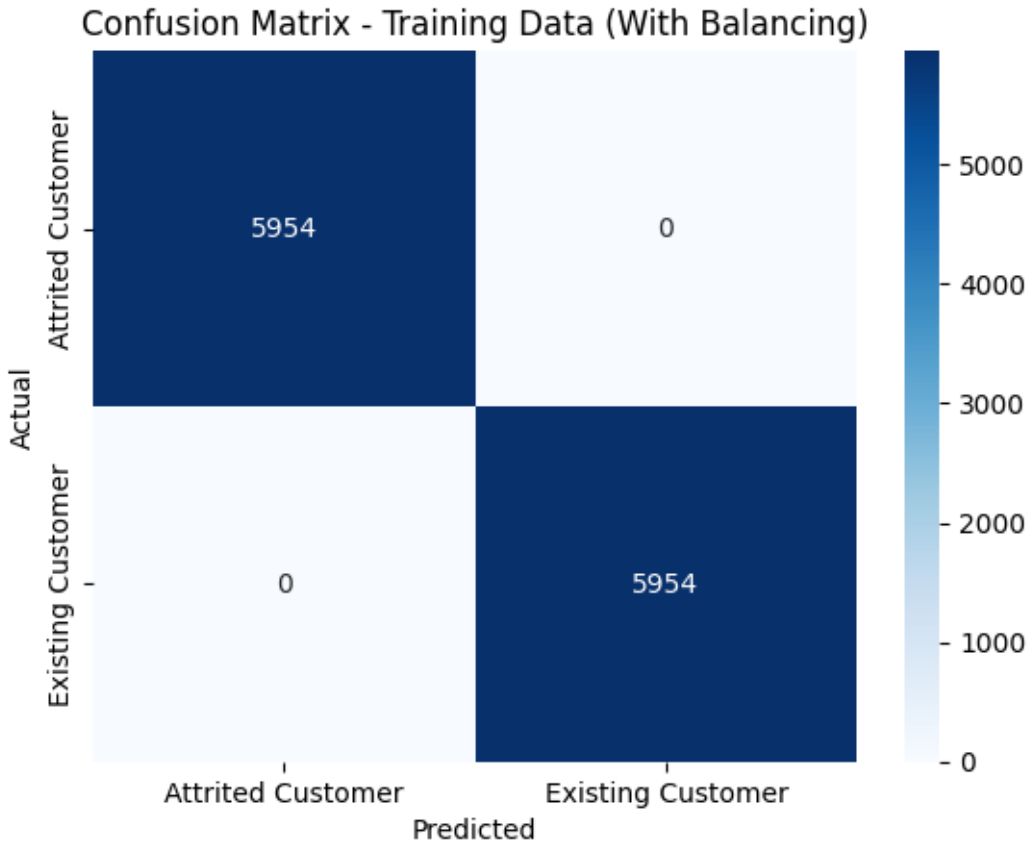
۵. آموزش مدل و ارزیابی: یک مدل Random Forest با ۱۰۰ درخت ایجاد و بر روی داده‌های آموزشی با اعمال متعادل‌سازی آموزش داده شده است.

۶. پیش بینی و گزارش عملکرد: پیش‌بینی‌ها برای داده‌های آموزش و اعتبارسنجی انجام شده است. گزارش طبقه‌بندی و ماتریس درهم‌ریختگی (Confusion Matrix) برای داده‌های آموزشی و اعتبارسنجی نمایش داده شده‌اند.

براساس داده های تست نتایج زیر به دست آمد:

Classification Report for Training Data (With Balancing):

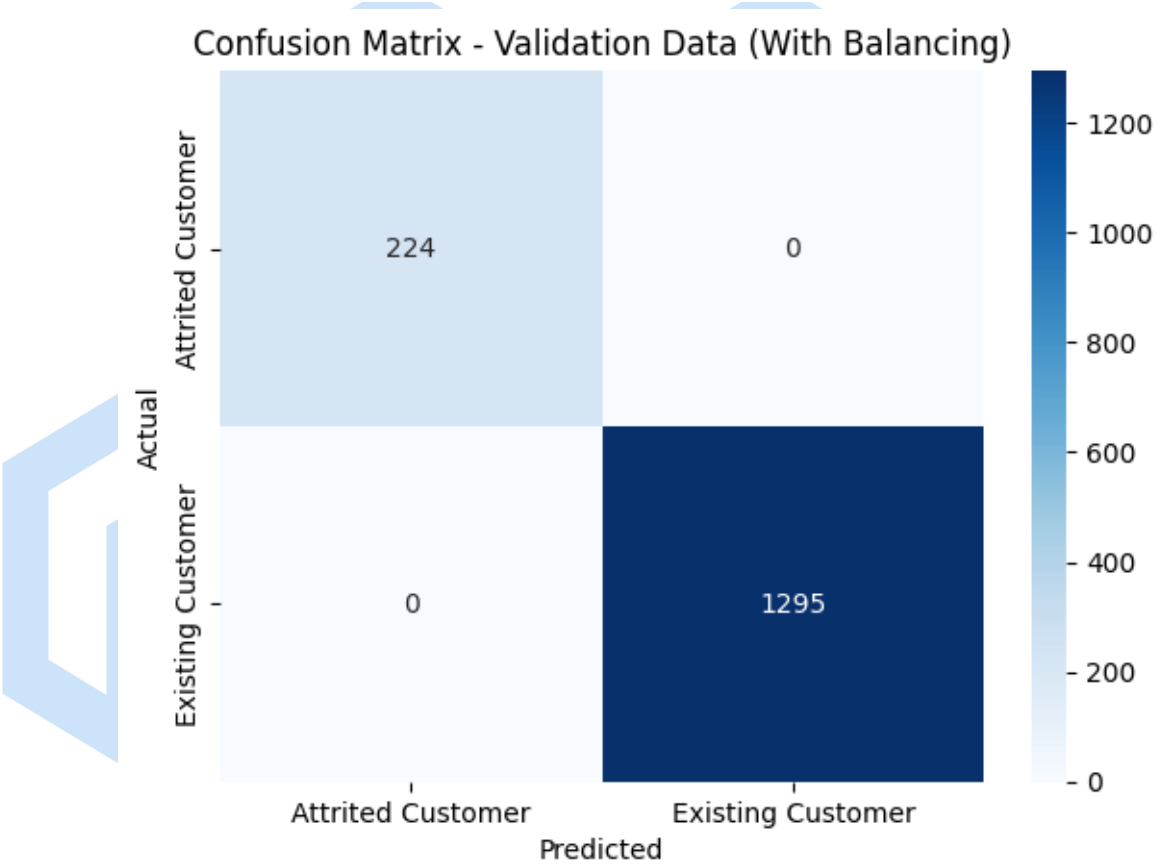
	precision	recall	f1-score	support
Attrited Customer	1.00	1.00	1.00	5954
Existing Customer	1.00	1.00	1.00	5954
accuracy			1.00	11908
macro avg	1.00	1.00	1.00	11908
weighted avg	1.00	1.00	1.00	11908



براساس داده های ارزیابی نتایج زیر به دست آمد:

Classification Report for Validation Data (With Balancing):

	precision	recall	f1-score	support
Attrited Customer	1.00	1.00	1.00	224
Existing Customer	1.00	1.00	1.00	1295
accuracy			1.00	1519
macro avg	1.00	1.00	1.00	1519
weighted avg	1.00	1.00	1.00	1519



مشابه مدل آموزش دیده با داده هایی که تعادل سازی نشده بودند، در مدل تعادل سازی شده نیز از دقت و عملکرد خوبی برخورداریم و این نشان دهنده یادگیری بیش از حد مدل از داده‌ها است . و احتمال اینکه مدل overfitting یا over optimization شده باشد بسیار بالاست.

همچنین اگر داده‌های آموزشی و اعتبارسنجی به طور کامل به مدل خورده و پیش‌بینی‌ها همیشه درست شده باشند، این می‌تواند نتیجه **حفظ کردن (memorization)** باشد، نه یادگیری معنادار از داده‌ها.

با توجه به راهنمایی انجام شده در سوال مراحل زیر طی کردم.

چرا ممکن است مدل فقط یک کلاس را پیش‌بینی کند؟

هنگامی که داده‌ها متعادل می‌شوند (با استفاده از SMOTE یا دیگر تکنیک‌ها)، گاهی اوقات مدل ممکن است بیشتر بر روی کلاس‌های ایجاد شده مصنوعی تمرکز کند و پیش‌بینی‌ها را برای یک کلاس خاص (معمولاً کلاس غالب یا کلاس پرنمایش) انجام دهد. در این حالت، مدل ممکن است نتواند تفاوت‌های واقعی بین دو کلاس را به درستی شبیه‌سازی کند.

چرا این مشکل پیش می‌آید؟

کلاس‌های مصنوعی: پس از متعادل‌سازی، برخی از نمونه‌ها ممکن است بیشتر مشابه نمونه‌های غالب (مثلاً "Existing Customer") باشند و در نتیجه، مدل تنها پیش‌بینی‌هایی برای این کلاس انجام دهد. وزن‌دهی نامناسب: در صورتی که داده‌ها به طور طبیعی نامتعادل باشند و بعد از متعادل‌سازی یک نوع از داده‌ها غالب شوند، مدل ممکن است نتواند تعادل واقعی را در پیش‌بینی‌ها برقرار کند.

چطور این مشکل را برطرف کنیم؟

بررسی بیش از حد پیش‌بینی یک کلاس: اگر مدل به طور عمده تنها یکی از کلاس‌ها را پیش‌بینی می‌کند (مثلاً فقط "Existing Customer")، باید اطمینان حاصل کرد که میزان هم‌خوانی داده‌ها و تنظیمات مدل (مانند وزن‌دهی) به درستی انجام شده است.

بازگرداندن داده‌ها: به عبارت دیگر، پس از متعادل‌سازی، اگر مدل همچنان تنها یک کلاس را پیش‌بینی می‌کند، می‌توان با بازگرداندن داده‌ها به حالت اصلی (بدون متعادل‌سازی) یا با استفاده از وزن‌دهی کلاس‌ها (در Random Forest) مدل را بهبود بخشید.

تست با تنظیمات مختلف: می‌توان از روش‌های وزن‌دهی به کلاس‌ها در Random Forest استفاده کرد تا مدل بیشتر به پیش‌بینی‌های کلاس کم‌نمونه توجه کند.

برای حل مشکل با مقادیر n_estimators و random_state بازی کردم و این هارو افزایش و کاهش دادم اما همچنان دقت ۱۰۰ درصد!!!!!!

بنابراین من فکر میکنم ؛

Overfitting ممکن است دلیل اصلی این دقت بالای ۱۰۰٪ باشد. مدل در واقع ممکن است به‌طور کامل بر روی داده‌های آموزشی (حتی داده‌های متعادل‌شده) یاد بگیرد و در نتیجه تمامی پیش‌بینی‌ها

درست باشد. این می‌تواند به این معنی باشد که مدل بیش از حد پیچیده است یا اینکه داده‌های آموزشی به اندازه کافی متنوع نیستند.

میتوان از روش‌هایی مانند cross-validation و محدود کردن پیچیدگی مدل برای جلوگیری از overfitting استفاده کنید.

من این روش انجام دادم:

محدود کردن پیچیدگی مدل:

max_depth=10: عمق درخت‌ها را محدود می‌کند تا از ایجاد درخت‌های عمیق و پیچیده جلوگیری شود.

min_samples_split=10: حداقل تعداد نمونه‌هایی که باید برای تقسیم یک گره داشته باشیم، تا از تقسیمات غیرضروری جلوگیری شود.

min_samples_leaf=5: حداقل تعداد نمونه‌هایی که باید در برگ‌های درخت باشند، تا از برگ‌های خیلی خاص جلوگیری کنیم.

Cross-Validation:

استفاده از cross_val_score برای ارزیابی مدل در چندین fold از داده‌ها. این کار به جلوگیری از overfitting کمک می‌کند.

GridSearchCV:

برای پیدا کردن بهترین پارامترها از GridSearchCV استفاده می‌شود. این روش مدل را برای مجموعه‌ای از پارامترها آموزش می‌دهد و بهترین ترکیب را پیدا می‌کند.

ارزیابی مدل:

هیچ تغییر مثبتی واقع نشد و همچنان مدل ما overfit است و به نظر من مدل داده ها را حفظ کرده و نمیتوان کاری انجام داد.

```
Cross-Validation Scores: [1. 1. 1. 1. 1.]
Mean Cross-Validation Score: 1.0
Best Parameters: {'max_depth': 5, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 50}
Best Score: 1.0
```

```
Classification Report for Training Data:
                   precision    recall  f1-score   support

Attrited Customer       1.00      1.00      1.00      1134
Existing Customer       1.00      1.00      1.00      5954

         accuracy                           1.00      7088
        macro avg       1.00      1.00      1.00      7088
     weighted avg       1.00      1.00      1.00      7088

Classification Report for Validation Data:
                   precision    recall  f1-score   support

Attrited Customer       1.00      1.00      1.00       224
Existing Customer       1.00      1.00      1.00      1295

         accuracy                           1.00      1519
        macro avg       1.00      1.00      1.00      1519
     weighted avg       1.00      1.00      1.00      1519
```

طبق نکته گفته شده در گروه تلگرام توسط آقای محمد جلیلی میتوان از الگوریتم ADOPT استفاده کرد.

۱. حذف گرادیان کنونی از تخمین مومنتوم دوم:

در الگوریتم‌های معمول بهینه‌سازی مانند SGD (Stochastic Gradient Descent)، برای محاسبه مومنتوم معمولاً از ترکیب گرادیان‌های قبلی و کنونی استفاده می‌شود. در واقع، این روند به صورت یک مجموعه جمع‌شده از تغییرات قبلی و فعلی است. اگر تغییرات ناگهانی یا نویزی در گرادیان‌ها وجود داشته باشد، این امر می‌تواند باعث عدم همگرایی شود.

راه حل در ADOPT این است که گرادیان کنونی را از تخمین مومنتوم دوم حذف می‌کند. این کار باعث می‌شود که اثرات تغییرات ناگهانی و نویز از بین برود و مومنتوم به طور نرم‌تر و با سرعت متعادل‌تری آپدیت شود، که به همگرایی بهتر مدل کمک می‌کند.

۲. نرمالایز کردن گرادیان قبل از آپدیت مومنتوم:

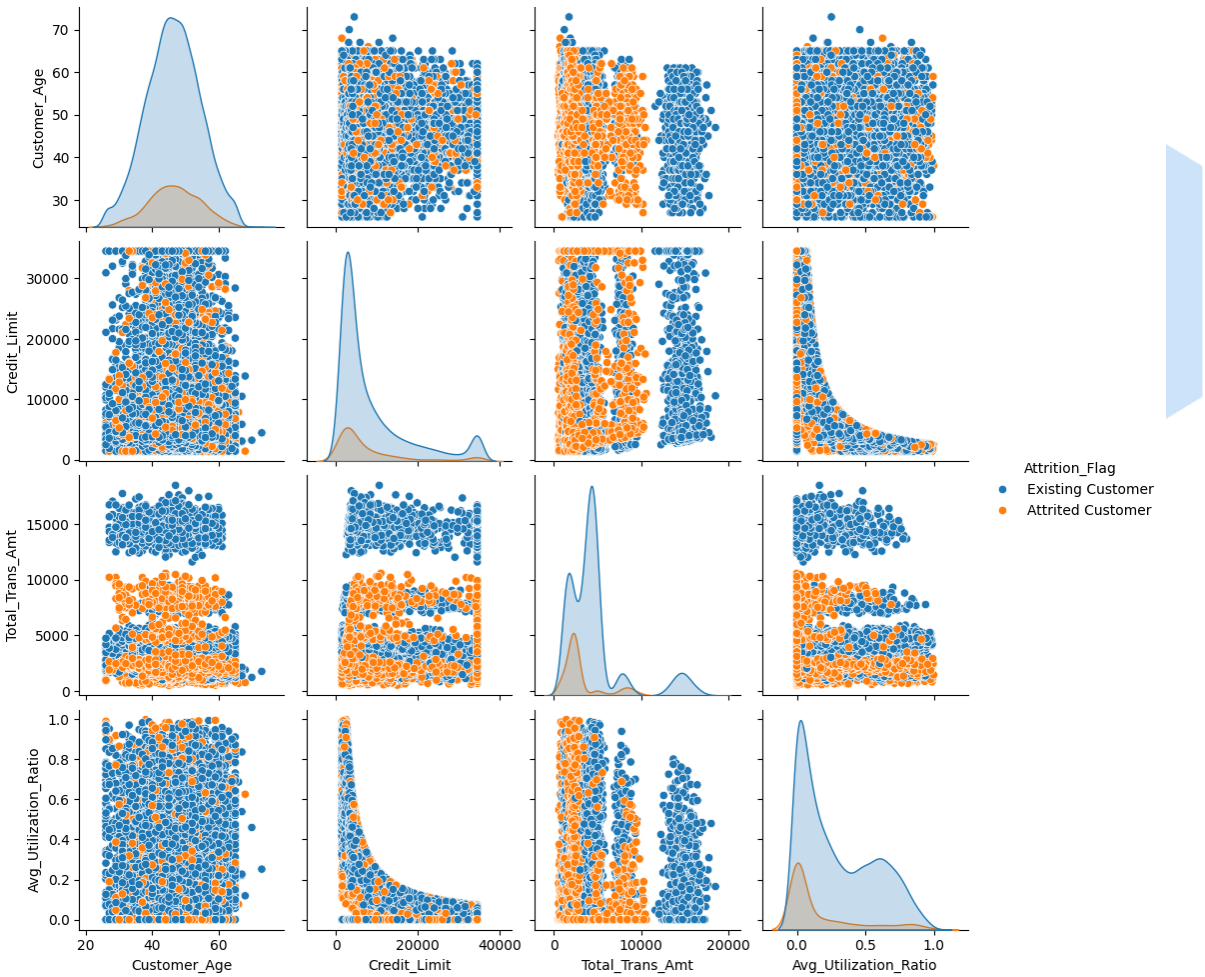
نرمالایز کردن گرادیان به این معنی است که میزان بزرگ بودن یا کوچک بودن گرادیان‌ها تحت کنترل قرار می‌گیرد. این امر باعث می‌شود که آپدیت‌ها در ابعاد مختلف از فضای پارامترها یکسان و متوازن شوند.

در غیر این صورت، ممکن است برخی ویژگی‌ها بیش از حد آپدیت شوند یا برخی دیگر خیلی کم تغییر کنند، که این امر می‌تواند باعث نوسانات زیاد و عدم همگرایی شود.

این نرمالایز کردن از لحاظ ریاضی مشابه با استفاده از روش‌های مقیاس‌بندی است که در بسیاری از الگوریتم‌های بهینه‌سازی پیشرفته برای جلوگیری از همگرایی ضعیف استفاده می‌شود.

طبق همین الگوریتم دقت مجددا ۱۰۰ درصد شد و علت اصلی این امر استفاده از ویژگی هایی است که تاثیر مثبتی در طبقه بندی خروجی ندارند و با توجه به بخش امتیازی میتوان این ویژگی ها را شناسایی کرده و سپس از مدل حذف کنیم.

امتیازی



ویژگی‌های انتخاب‌شده:

ویژگی‌های انتخاب‌شده برای تحلیل عبارتند از سن مشتری، حد اعتبار کارت، کل مبلغ تراکنش‌ها و میانگین نسبت استفاده از کارت. در این تحلیل، بررسی روابط مختلف بین این ویژگی‌ها با توجه به دو کلاس موجود در ویژگی Attrition_Flag (مشتریانی که ترک کرده‌اند و مشتریانی که همچنان فعال هستند) انجام می‌شود.

تحلیل نمودار:

۱. نمودار قطر اصلی (دایگنال):

در این نمودارها توزیع هر ویژگی به صورت منحنی تخمین چگالی نمایش داده شده است.

سن مشتری:

توزیع سن مشتریان برای هر دو کلاس Attrited Customer و Existing Customer به شکل یک منحنی برجسته است. در اینجا بیشتر مشتریان در سنین میان‌سال (حدود ۳۰ تا ۶۰ سال) قرار دارند.

برای Existing Customers، توزیع سن کمی بیشتر به سمت سنین بالا میل دارد، در حالی که Attrited Customers معمولاً در محدوده‌های سنی پایین‌تر قرار دارند.

حد اعتبار کارت:

توزیع حد اعتبار برای Existing Customers نشان‌دهنده چولگی به سمت راست است که بیشتر مشتریان حد اعتبار بالاتری دارند.

در Attrited Customers، تعداد زیادی از مشتریان دارای حد اعتبار پایین‌تر هستند و توزیع به سمت چپ متمایل است. این نشان‌دهنده تفاوت در الگوهای استفاده از اعتبار بین دو کلاس است.

کل مبلغ تراکنش‌ها :

برای Existing Customers، توزیع مبلغ تراکنش‌ها معمولاً به سمت مقادیر بالاتر تمایل دارد. مشتریان فعال معمولاً تراکنش‌های بیشتری انجام می‌دهند.

Attrited Customers معمولاً تراکنش‌های کمتری انجام داده‌اند، که این موضوع در توزیع چولگی به سمت راست مشاهده می‌شود.

میانگین نسبت استفاده از کارت:

در توزیع نسبت استفاده از اعتبار، برای Existing Customers، میانگین استفاده از اعتبار بیشتر است، در حالی که Attrited Customers معمولاً از اعتبار خود کمتر استفاده کرده‌اند. توزیع در این ویژگی بیشتر چولگی منفی دارد که نشان‌دهنده استفاده پایین از اعتبار توسط مشتریان ترک کرده است.

۲. نمودارهای پراکندگی (غیرقطری):

این نمودارها نشان‌دهنده روابط بین ویژگی‌ها در هر دو کلاس مختلف هستند.

سن مشتریان و حد اعتبار کارت:

در نمودار پراکندگی بین سن و حد اعتبار، مشخص است که هیچ همبستگی قوی میان این دو ویژگی وجود ندارد. بیشتر Existing Customers و Attrited Customers در محدوده‌های مختلف سنی قرار دارند و محدودیت اعتبار آن‌ها نیز به‌طور مستقل از سن قرار دارد.

به طور کلی، سن مشتریان نمی‌تواند پیش‌بینی دقیقی برای حد اعتبار آن‌ها باشد.

سن مشتریان و کل مبلغ تراکنش‌ها:

هیچ الگوی مشخصی در رابطه بین سن مشتری و مبلغ تراکنش‌ها مشاهده نمی‌شود. این نشان می‌دهد که تراکنش‌ها به سن مشتریان وابسته نیستند و ممکن است عوامل دیگری بر تراکنش‌ها تأثیر بگذارند.

سن مشتریان و میانگین استفاده از کارت:

در این نمودار، هیچ ارتباط واضحی بین سن مشتریان و نسبت استفاده از کارت مشاهده نمی‌شود. این نشان‌دهنده این است که ویژگی سن به‌طور مستقیم با استفاده از اعتبار مشتریان مرتبط نیست.

حد اعتبار کارت و کل مبلغ تراکنش‌ها:

در این نمودار پراکندگی، برخی از Existing Customers با حد اعتبار پایین‌تر تراکنش‌های بیشتری انجام می‌دهند، در حالی که مشتریانی با حد اعتبار بالاتر تراکنش‌های کمتری دارند.

این می‌تواند نشان‌دهنده این باشد که مشتریان با اعتبار پایین ممکن است مجبور به انجام تراکنش‌های بیشتر باشند تا از اعتبار خود استفاده کنند، در حالی که مشتریان با اعتبار بالا از اعتبار خود کمتر استفاده کرده‌اند.

حد اعتبار کارت و میانگین استفاده از کارت:

در این نمودار، به وضوح مشاهده می‌شود که مشتریان با حد اعتبار بالا نسبت به اعتبار خود کمتر استفاده کرده‌اند، در حالی که مشتریان با حد اعتبار پایین معمولاً از اعتبار خود بیشتر استفاده می‌کنند.

این الگو نشان‌دهنده چولگی منفی است که در آن مشتریان با اعتبار بالا تمایل دارند از اعتبار خود کمتر استفاده کنند، در حالی که مشتریان با اعتبار پایین از امکانات موجود بیشتر بهره می‌برند.

کل مبلغ تراکنش‌ها و میانگین استفاده از کارت:

در این نمودار، هیچ همبستگی قوی بین کل مبلغ تراکنش‌ها و میانگین استفاده از اعتبار مشاهده نمی‌شود. به نظر می‌رسد که مشتریانی با تراکنش‌های بالا ممکن است از اعتبار خود کم‌تر استفاده کرده باشند، در حالی که برخی مشتریان با تراکنش‌های کمتر از اعتبار بیشتری استفاده می‌کنند.

جمع‌بندی کلی:

مشتریان موجود (Existing Customers): این گروه معمولاً دارای حد اعتبار بالاتر و استفاده مؤثرتر از اعتبار خود هستند. آن‌ها تراکنش‌های بیشتری انجام می‌دهند و از اعتبار خود بیشتر استفاده می‌کنند.

مشتریان ترک‌شده (Attrited Customers): این گروه معمولاً دارای حد اعتبار پایین‌تر و استفاده کمتری از اعتبار خود هستند. آن‌ها تراکنش‌های کمتری انجام می‌دهند و در بیشتر مواقع از اعتبار خود به‌طور ناقص استفاده کرده‌اند.

این تحلیل‌ها نشان می‌دهد که مشتریانی که از اعتبار خود به‌طور مؤثر استفاده نمی‌کنند یا تراکنش‌های کمتری انجام می‌دهند، احتمال بیشتری برای ترک خدمات دارند. بنابراین، ویژگی‌هایی مانند حد اعتبار، مبلغ تراکنش‌ها و نسبت استفاده از اعتبار می‌توانند شاخص‌های مهمی برای پیش‌بینی رفتار ترک مشتریان باشند.

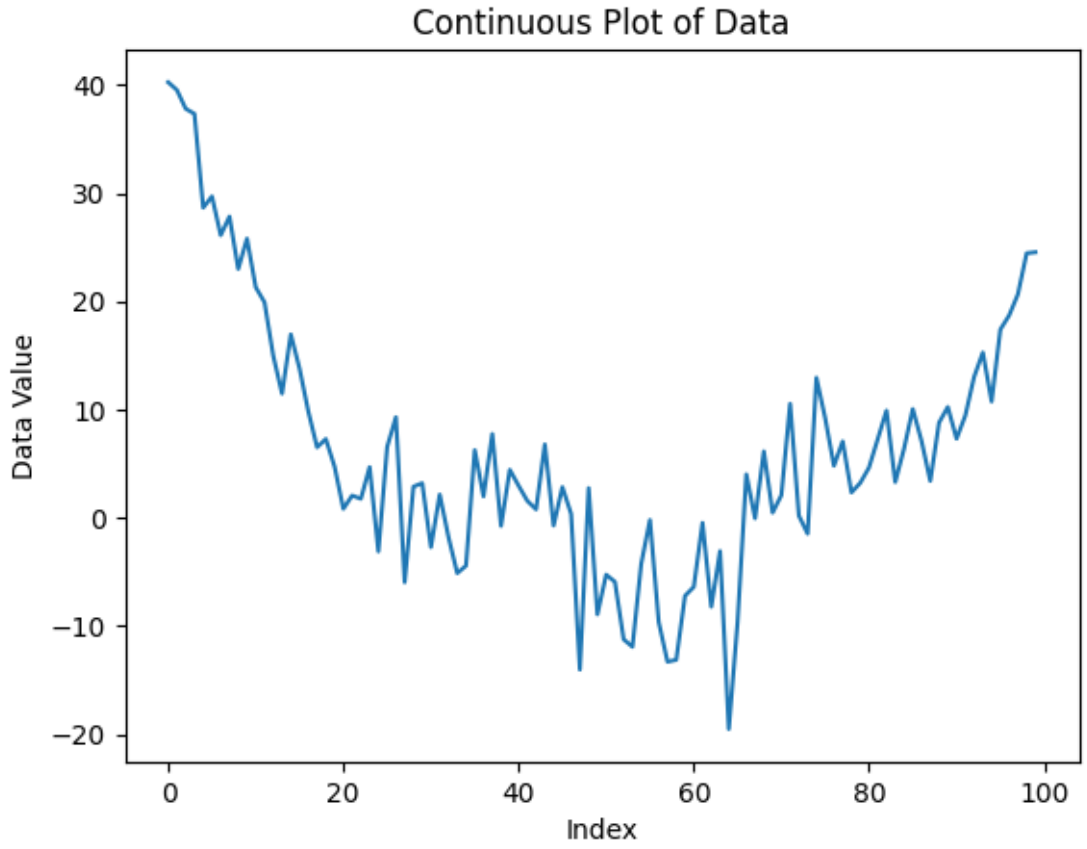
سوال ۲

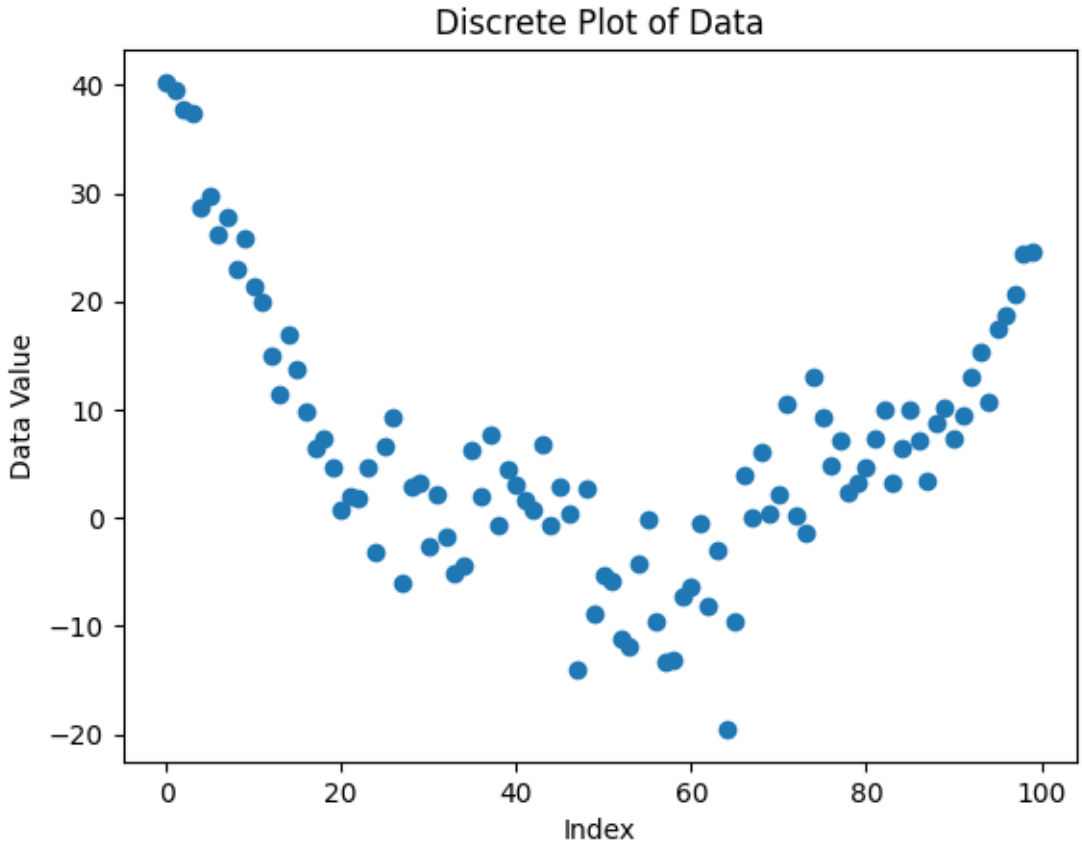
ابتدا داده را خواندم و اطلاعات زیر کسب کردم:

```
[ 40.25196507  39.53010126  37.7992167   37.32837052  28.65394346
  29.69747547  26.10881978  27.83802602  22.99054421  25.80112967
  21.29795526  19.91155621  14.96131425  11.47430672  16.95134087
  13.78849326   9.82605161   6.51423783   7.28107882   4.71364215
   0.82726539   2.0547798    1.75589251   4.69110905  -3.1080814
   6.5508178    9.30439077  -5.9567694    2.87594962   3.1993877
  -2.70786354   2.15378132  -1.77644948  -5.11557224  -4.43062279
   6.25451526   1.9854139    7.74342429  -0.75124188   4.43658355
   2.96815869   1.56463746   0.77572103   6.78279848  -0.71535178
   2.83930294   0.35701732 -14.02823175   2.74610814  -8.92342079
  -5.25805964  -5.91984033 -11.23785323 -11.91938329  -4.2483209
  -0.19964379  -9.63227683 -13.30314598 -13.12628213  -7.21216734
  -6.38891745  -0.4667489   -8.21628152  -3.07657489 -19.52390961
  -9.6384843    4.01358254  -0.04798927   6.13528941   0.47146013
   2.08633153  10.53650805   0.19410599  -1.47248443  12.94164738
   9.22295082   4.8044539    7.06519393   2.33812888   3.21795311
   4.63512252   7.27235385   9.90686345   3.29586551   6.37295172
  10.05409648   7.09980433   3.40197508   8.79312464  10.23818681
   7.30636817   9.45267898  13.01264151  15.27407037  10.74387524
  17.39972088  18.69963466  20.65273966  24.44557776  24.53832675]
```

```
(100,)
Mean: 6.318082393530108
Standard Deviation: 12.032918191230813
Min: -19.523909610590692
Max: 40.25196507277925
```

ما یک داده یک بعدی داریم و با استفاده از تصویر و np.linspace X تشکیل میدهیم.


Continuous Plot of Data
Data Value
Index
40
30
20
10
0
−10
−20
0
20
40
60
80
100


Discrete Plot of Data
Data Value
Index
40
30
20
10
0
−10
−20
0
20
40
60
80
100

داده این دیتا به دو صورت پیوسته و گسسته نمایش داده شده است.

نمودار اول (با استفاده از plt.plot):

ویژگی‌های نمودار:

این نمودار به صورت پیوسته رسم شده است. تغییرات داده‌ها به‌طور مداوم به هم وصل شده‌اند.

تحلیل:

چون این نمودار پیوسته است، به‌نظر می‌رسد که تغییرات به‌صورت یک روند ثابت پیش می‌روند. نوسانات بزرگ یا کوچک به شکل یک روند مستقیم و متصل به هم نمایش داده می‌شوند، که ممکن است باعث شود برخی از ویژگی‌های گسسته داده‌ها پنهان شوند.

در مواردی که داده‌ها نوسانات زیادی دارند (مثل نقاطی که مقدار منفی یا مثبت بزرگی دارند)، این نوع رسم می‌تواند برای تحلیل برخی نوسانات یا تغییرات حساسیت کمتری داشته باشد.

نمودار دوم (با استفاده از plt.scatter):

ویژگی‌های نمودار:

این نمودار به صورت گسسته رسم شده است. هر نقطه داده به‌طور جداگانه نمایش داده می‌شود.

تحلیل:

چون نقاط داده‌ها به‌صورت گسسته و بدون اتصال به هم نمایش داده شده‌اند، می‌توان دقیق‌تر ویژگی‌های خاص داده‌ها را مشاهده کرد.

نوسانات و تغییرات در داده‌ها به‌وضوح قابل مشاهده است. اگر داده‌ها به طور غیرخطی تغییر کنند یا نقاط بحرانی وجود داشته باشد، این نوع رسم می‌تواند اطلاعات بیشتری درباره رفتار گسسته داده‌ها به ما بدهد.

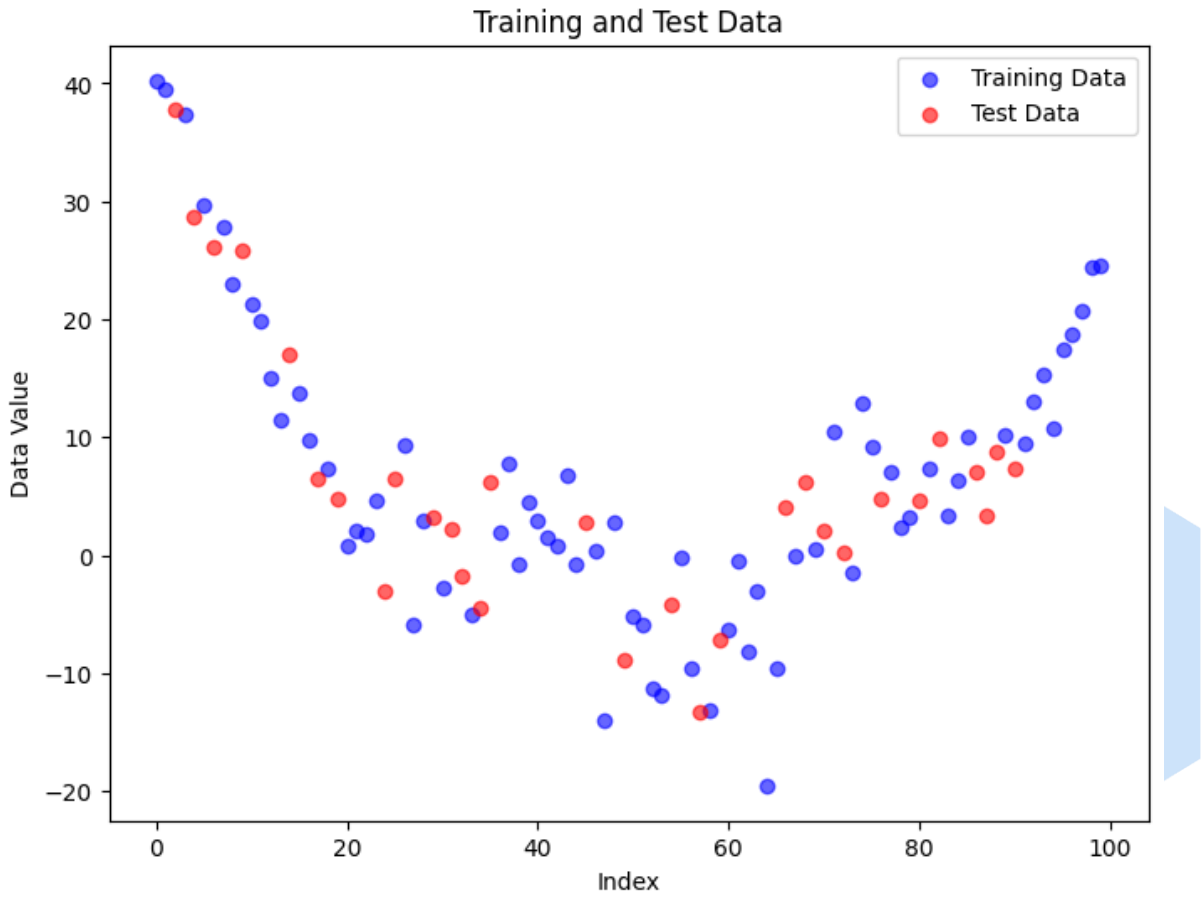
نوسانات منفی یا مثبت به‌راحتی قابل تشخیص هستند و این می‌تواند به ما کمک کند تا مناطقی که تغییرات غیرعادی دارند را بهتر شناسایی کنیم.

نتیجه‌گیری:

نمودار پیوسته (plt.plot) مناسب است برای نمایش روند کلی و تغییرات بلندمدت، اما ممکن است برای داده‌های گسسته یا نوسانات کوچک مناسب نباشد.

نمودار گسسته (plt.scatter) بهتر است برای تحلیل دقیق‌تر داده‌ها و شناسایی ویژگی‌های خاص مانند نوسانات و نقاط بحرانی. این نوع نمودار به ما کمک می‌کند تا رفتار دقیق داده‌ها را ببینید و نقاط غیرعادی را شناسایی کنیم.

۲.۱



این نمودار نشان‌دهنده‌ی تقسیم داده‌های شما به دو مجموعه‌ی **آموزش** و **آزمون** است که با استفاده از رنگ‌های مختلف (آبی و قرمز) مشخص شده‌اند.

نسبت تقسیم داده ها ۷۰ به ۳۰ است.

تنظیمات و عملکرد مدل:

آموزش مدل: مدل با استفاده از داده‌های آموزشی (آبی) آموزش می‌بیند و خطای آن باید به مرور زمان کاهش یابد.

آزمون مدل: پس از آموزش، مدل روی داده‌های آزمون (قرمز) آزمایش می‌شود. خطای مدل بر روی داده‌های آزمون باید از خطای مدل بر روی داده‌های آموزشی کمتر یا برابر باشد.

پیش‌بینی صحیح: داده‌های قرمز (آزمون) به مدل فرصت می‌دهند تا پیش‌بینی‌های خود را بر روی داده‌هایی که قبل از آن ندیده است، اعمال کند. عملکرد مدل بر روی این داده‌ها می‌تواند نشانه‌ای از دقت عمومی آن باشد.

آینده‌نگری و اصلاح مدل: اگر مدل تنها بر روی داده‌های آموزشی خوب عمل کند و در داده‌های آزمون ضعیف باشد، این نشان‌دهنده‌ی بیش‌برازش (overfitting) است. مدل در حال یادگیری دقیق جزئیات داده‌های آموزشی است، اما قادر به تعمیم به داده‌های جدید نیست. اگر مدل در هر دو مجموعه آموزشی و آزمون خطای بالایی داشته باشد، این نشان‌دهنده‌ی کم‌برازش (underfitting) است و ممکن است مدل نتوانسته باشد ویژگی‌های اساسی داده‌ها را یاد بگیرد.

اهمیت خطاها: خطای داده‌های آموزشی به ما می‌گوید که مدل چقدر در پیش‌بینی دقیق داده‌هایی که قبلاً دیده است، موفق بوده است. خطای داده‌های آزمون نشان‌دهنده‌ی این است که مدل تا چه حد قادر به تعمیم یادگیری‌هایش به داده‌های جدید است.

۲.۲

سه معیار رایج برای سنجش عملکرد مدل‌های رگرسیون عبارتند از:

۱. **Mean Absolute Error (MAE)**:

این معیار متوسط قدر مطلق تفاوت بین مقادیر پیش‌بینی‌شده و مقادیر واقعی را نشان می‌دهد.

۲. **Mean Squared Error (MSE)**:

این معیار متوسط مربع تفاوت بین مقادیر پیش‌بینی‌شده و مقادیر واقعی را نشان می‌دهد.

۳. **R-squared (R^2)**

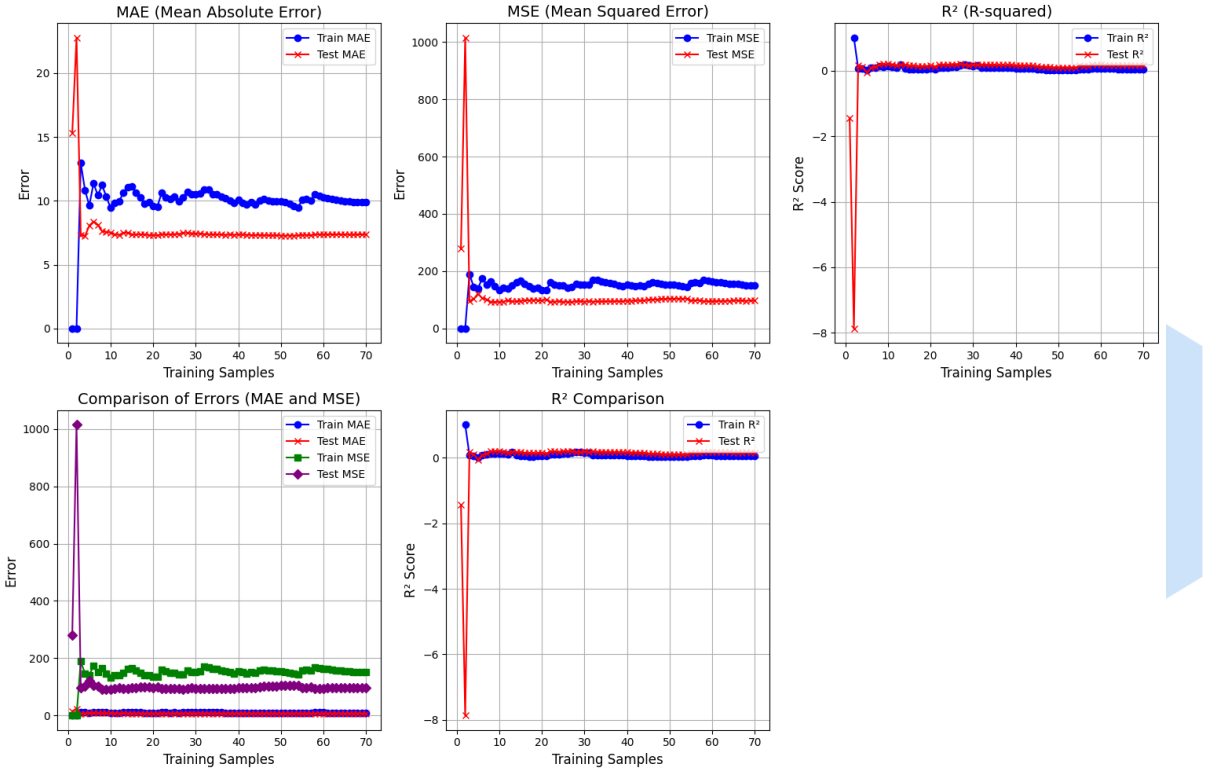
این معیار درصد واریانس پاسخ (متغیر وابسته) که توسط مدل توضیح داده می‌شود را اندازه‌گیری می‌کند.

در هر مرحله‌ای که مدل آموزش داده میشود، باید این معیارها را محاسبه کنیم تا عملکرد مدل را ارزیابی کنیم.

به عنوان مثال:

محاسبه MAE، MSE و R^2 در داده‌های آموزشی: این ارزیابی نشان می‌دهد که مدل چقدر خوب می‌تواند داده‌های آموزشی را پیش‌بینی کند.

محاسبه MAE، MSE و R^2 در داده‌های آزمون: این ارزیابی نشان می‌دهد که مدل چقدر توانسته است به داده‌هایی که قبلاً ندیده است، تعمیم دهد.



۱. نمودار MAE (Mean Absolute Error)

MAE نشان‌دهنده میانگین خطای مطلق بین مقادیر واقعی و پیش‌بینی‌شده است.

در نمودار MAE، دو خط مشاهده می‌شود:

خط آبی (داده‌های آموزش): MAE برای داده‌های آموزش در طول زمان.

خط قرمز (داده‌های آزمون): MAE برای داده‌های آزمون در طول زمان.

تحلیل MAE:

در مراحل اولیه آموزش (نزدیک به صفر)، MAE در داده‌های آزمون بیشتر از داده‌های آموزش است که به طور طبیعی نشان‌دهنده این است که مدل هنوز به طور کامل آموزش ندیده است.

در طول زمان، خط MAE برای داده‌های آزمون کاهش می‌یابد و به سمت همگرایی با داده‌های آموزش می‌رود، که نشان‌دهنده بهبود مدل در پیش‌بینی داده‌های آزمون است.

در انتهای آموزش، MAE برای داده‌های آزمون و آموزش مشابه می‌شود که نشان‌دهنده تطابق خوب مدل با داده‌ها است.

استنباط:

در ابتدای فرآیند آموزش، مدل قادر به پیش‌بینی درست داده‌های آزمون نبوده است، اما با گذشت زمان و آموزش بیشتر، مدل بهبود می‌یابد و پیش‌بینی‌های بهتری برای داده‌های آزمون می‌دهد. این نشان‌دهنده یادگیری صحیح مدل است.

۲. نمودار MSE (Mean Squared Error)

MSE به طور مشابه با MAE است، با این تفاوت که به خطاهای بزرگ وزن بیشتری می‌دهد. در این نمودار، دو خط برای داده‌های آموزش و آزمون مشاهده می‌شود:

خط آبی (داده‌های آموزش): MSE برای داده‌های آموزش در طول زمان.

خط قرمز (داده‌های آزمون): MSE برای داده‌های آزمون در طول زمان.

تحلیل MSE:

مشابه به MAE، MSE برای داده‌های آزمون در ابتدا بیشتر از داده‌های آموزش است، چرا که مدل هنوز به اندازه کافی آموزش ندیده است.

در طول آموزش، MSE برای داده‌های آزمون کاهش می‌یابد، که نشان‌دهنده بهبود مدل است.

به طور کلی، MSE برای داده‌های آزمون در مقایسه با MAE در ابتدا بیشتر است، زیرا MSE حساسیت بیشتری به خطاهای بزرگ دارد. این باعث می‌شود که در مراحل اولیه، خطاهای پیش‌بینی بزرگ‌تر باعث افزایش MSE شوند.

در انتهای آموزش، MSE برای داده‌های آموزش و آزمون مشابه هم می‌شوند که نشان‌دهنده این است که مدل به خوبی یاد گرفته است.

استنباط:

MSE به طور خاص خطاهای بزرگ را برجسته می‌کند، و به همین دلیل در مراحل اولیه آموزش، خطاهای پیش‌بینی بزرگ تأثیر زیادی بر MSE دارند. با گذشت زمان، کاهش MSE برای هر دو مجموعه نشان‌دهنده یادگیری بهتر مدل است.

۳. نمودار R^2 (R-squared)

R^2 نشان‌دهنده قدرت پیش‌بینی مدل است. مقدار آن بین ۰ و ۱ است، که ۱ نشان‌دهنده تطابق کامل مدل با داده‌ها و ۰ نشان‌دهنده عدم تطابق است. در این نمودار:

خط آبی (داده‌های آموزش): R^2 برای داده‌های آموزش در طول زمان.

خط قرمز (داده‌های آزمون): R^2 برای داده‌های آزمون در طول زمان.

تحلیل R^2:

R^2 برای داده‌های آموزش در ابتدا خیلی کم است، که نشان‌دهنده این است که مدل هنوز قادر به یادگیری ویژگی‌های داده‌ها نیست.

با پیشرفت آموزش، مقدار R^2 برای داده‌های آزمون و آموزش افزایش می‌یابد، که نشان‌دهنده بهبود مدل و تطابق بیشتر آن با داده‌ها است.

در مراحل آخر، R^2 برای داده‌های آزمون به مقدار مشابه داده‌های آموزش می‌رسد، که نشان‌دهنده این است که مدل به خوبی آموزش دیده و توانایی پیش‌بینی دقیقی دارد.

استنباط:

افزایش R^2 در طول فرآیند آموزش به این معناست که مدل به تدریج اطلاعات بیشتری از داده‌ها یاد می‌گیرد و پیش‌بینی‌های دقیق‌تری ارائه می‌دهد. در نهایت، R^2 برای داده‌های آزمون و آموزش تقریباً مشابه می‌شود، که نشان‌دهنده عملکرد خوب مدل است.

۴. نمودار مقایسه‌ای MAE و MSE (همزمان)

تحلیل:

هر دو معیار، MAE و MSE در ابتدا بالاتر هستند و با پیشرفت آموزش کاهش می‌یابند.

از آنجایی که MSE به خطاهای بزرگ وزن بیشتری می‌دهد، مقادیر آن نسبت به MAE در ابتدا بیشتر هستند.

در انتهای آموزش، هر دو معیار کاهش پیدا می‌کنند و به هم نزدیک می‌شوند.

استنباط:

مقایسه MAE و MSE نشان می‌دهد که MSE بیشتر تحت تأثیر خطاهای بزرگ است. بنابراین در مراحل اولیه آموزش، به دلیل خطاهای بزرگ‌تر، MSE بالاتر از MAE است.

۵. **نمودار مقایسه‌ای داده‌های آموزش و آزمون (کلی)**

تحلیل:

در تمامی نمودارها (MAE، MSE، و R^2)، داده‌های آزمون و آموزش در ابتدا تفاوت زیادی دارند، اما این تفاوت با پیشرفت آموزش کاهش می‌یابد.

پایان آموزش: در انتهای آموزش، مقادیر برای داده‌های آزمون و آموزش مشابه می‌شوند که نشان‌دهنده یادگیری خوب مدل است.

استنباط:

نشان‌دهنده این است که مدل به تدریج یاد می‌گیرد و در نهایت توانسته است به خوبی با داده‌های آزمون هم تطابق پیدا کند.

مقایسه کلی نمودارها:

در ابتدا، مدل قادر به پیش‌بینی دقیق داده‌های آزمون نیست، که در MAE و MSE نشان داده می‌شود. اما با گذشت زمان، مدل بهبود می‌یابد و قادر به پیش‌بینی بهتر می‌شود.

در پایان فرآیند آموزش، خطاها (MAE و MSE) برای داده‌های آموزش و آزمون به حداقل می‌رسند و مقدار R^2 برای هر دو مجموعه به مقدار بالایی می‌رسد، که نشان‌دهنده یادگیری خوب مدل است.

تفاوت‌های اولیه بین داده‌های آموزش و آزمون می‌تواند به دلیل محدود بودن تعداد نمونه‌ها و یا عدم تطابق اولیه مدل با داده‌های آزمون باشد.

جمع‌بندی:

این نمودارها نشان‌دهنده فرآیند یادگیری مدل است که در ابتدا خطاهای بیشتری دارد و سپس با آموزش بیشتر، دقت آن افزایش می‌یابد. به‌طور کلی، مدل در پایان آموزش عملکرد خوبی برای پیش‌بینی داده‌های آزمون و آموزش دارد، که در تمامی معیارهای خطا (MAE، MSE، و R^2) نمایان است.

۲.۳

برای آموزش یک مدل رگرسیون خطی درجه اول (که به معنی یک مدل خطی ساده است) به طور دستی بدون استفاده از توابع آماده در sklearn، باید معادله رگرسیون خطی را برای داده‌ها پیاده‌سازی کنیم.

معادله رگرسیون خطی درجه اول به صورت زیر است:

$$\hat{y} = \beta_1 x + \beta_0$$

$\boldsymbol{\beta}$ پارامترهای مدل

x ورودی

$\hat{y}$خروجی مدل (پیش بینی)

در این کد، یک مدل رگرسیون خطی درجه اول را برای داده‌ها آموزش داده شده و سپس عملکرد آن را با استفاده از خطاهای مختلف ارزیابی شده. مراحل مختلف فرآیند به شرح زیر است:

۱. **محاسبه میانگین داده‌ها:**

ابتدا میانگین داده‌های ورودی (x) و هدف (y) برای داده‌های آموزش محاسبه می‌شود.

این مقادیر برای محاسبه ضرایب رگرسیون ($\beta 0$ و $\beta 1$) استفاده خواهند شد.

۲. **محاسبه ضرایب رگرسیون ($\beta 0$ و $\beta 1$):**

برای محاسبه ضرایب رگرسیون (عرض از مبدا $\beta 0$ و شیب خط $\beta 1$)، از فرمول‌های آماری استفاده می‌شود.

$\beta 1$ (شیب خط) از طریق رابطه میانگین مقادیر ورودی و خروجی آموزش محاسبه می‌شود.

$\beta 0$ (عرض از مبدا) به کمک میانگین داده‌ها و $\beta 1$ محاسبه می‌شود.

۳. محاسبه پیش‌بینی‌ها (*Prediction*):

مدل رگرسیون خطی به‌دست آمده با استفاده از ضرایب $\beta 0$ و $\beta 1$ برای داده‌های آموزش و آزمون پیش‌بینی‌هایی انجام می‌دهد.

این پیش‌بینی‌ها برای ارزیابی کیفیت مدل و محاسبه خطا استفاده می‌شوند.

۴. محاسبه خطاها (*Error Calculation*):

برای هر دو مجموعه داده‌های آموزش و آزمون، سه نوع خطا محاسبه می‌شود:

MAE (Mean Absolute Error): خطای مطلق میانگین

MSE (Mean Squared Error): خطای مربع میانگین

R^2 (R-squared): ضریب تعیین که نشان‌دهنده دقت مدل در پیش‌بینی است.

این مقادیر برای ارزیابی عملکرد مدل و مقایسه پیش‌بینی‌های آموزش و آزمون استفاده می‌شوند.

۵. نمایش ضرایب و خطای مدل:

ضرایب رگرسیون ($\beta 0$ و $\beta 1$) نمایش داده می‌شوند.

سپس خطای *MAE*، *MSE* و R^2 برای داده‌های آموزش و آزمون چاپ می‌شود تا عملکرد مدل مورد بررسی قرار گیرد.

۶. رسم خط رگرسیون:

خط رگرسیون به‌صورت یک خط مستقیم روی داده‌ها رسم می‌شود. این خط بر اساس ضرایب محاسبه‌شده ($\beta 0$ و $\beta 1$) ترسیم می‌شود.

داده‌های آموزش و آزمون نیز به‌صورت نقاط پراکنده (*scatter*) بر روی نمودار نمایش داده می‌شوند.

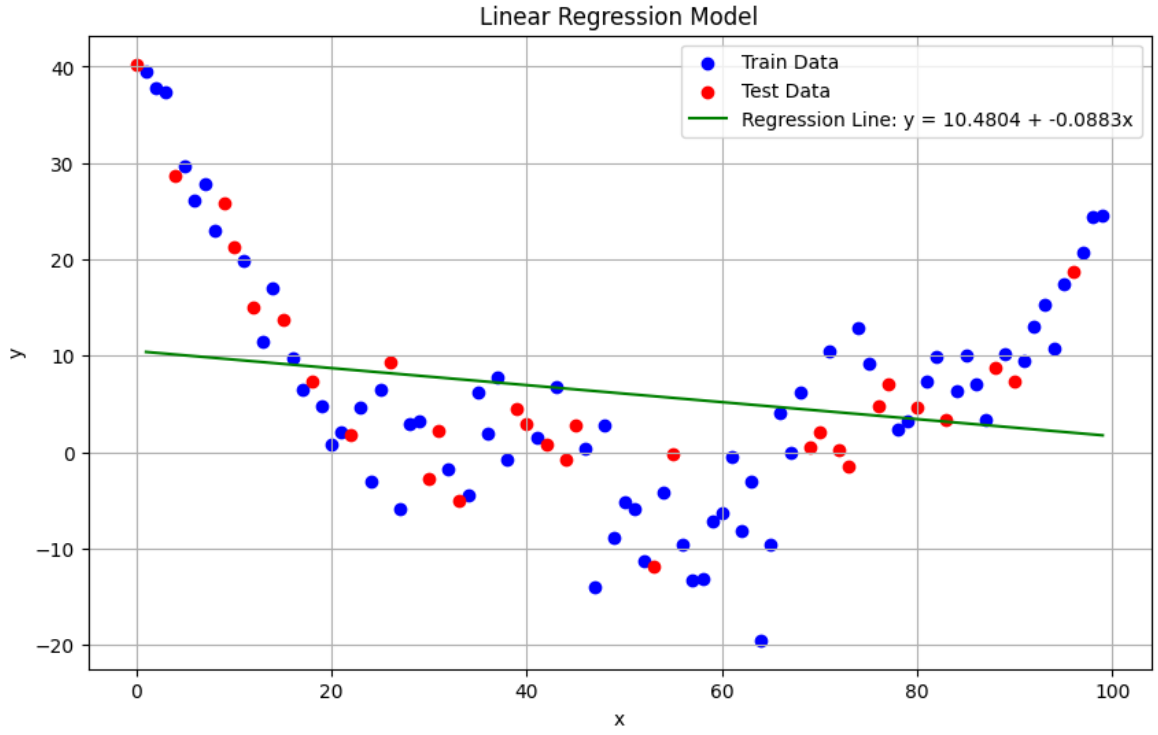
۷. رسم نمودارهای خطا:

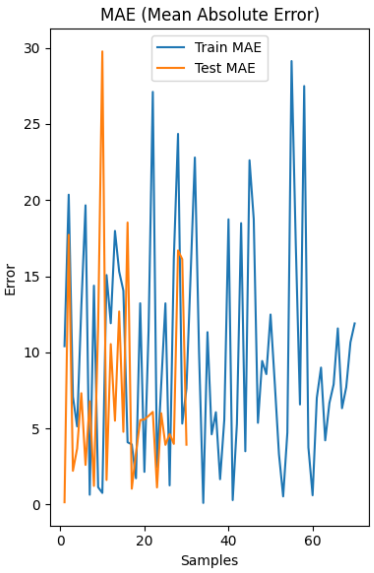
برای نمایش چگونگی تغییر خطاها با توجه به تعداد نمونه‌های آموزشی، سه نمودار جداگانه برای *MAE*، *MSE* و R^2 ایجاد می‌شود.

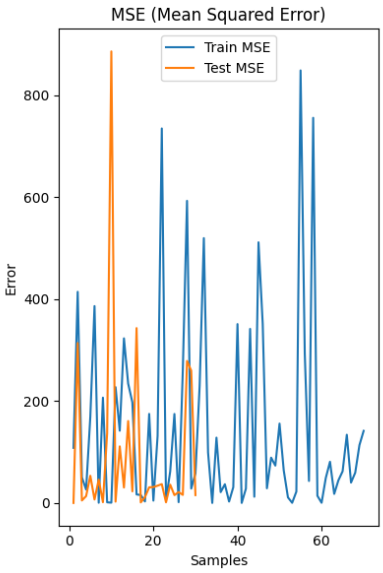
این نمودارها اطلاعاتی در مورد عملکرد مدل در طول فرآیند آموزش و ارزیابی مدل به‌دست می‌دهند.

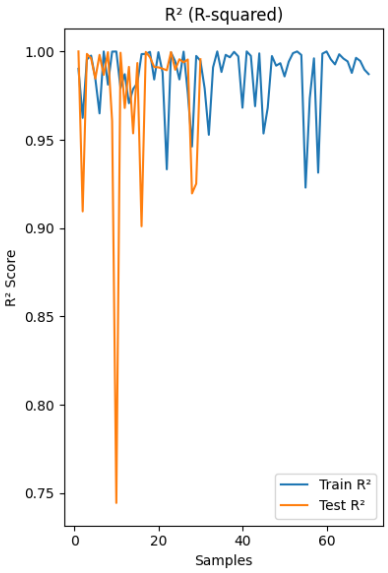
پارامتر مدل:
β0 (عرض از مبدا): 10.4804
β1 (شیب خط): -0.0883

برای داده‌های آموزش: خطای MAE 9.9109
برای داده‌های آزمون: خطای MAE 7.3535
برای داده‌های آموزش: خطای MSE 150.8923
برای داده‌های آزمون: خطای MSE 97.4421
برای داده‌های آموزش: R^2 0.0415
برای داده‌های آزمون: R^2 0.1574









۱. **نمودار *MAE (Mean Absolute Error)*:**

آموزش (*Train*): مشاهده می‌شود که خطای مطلق میانگین (*MAE*) برای داده‌های آموزش در طول زمان آموزش به طور پیوسته کاهش یافته است. این نشان می‌دهد که مدل در حال یادگیری است و خطای آن بهبود می‌یابد.

آزمون (*Test*): خطای *MAE* برای داده‌های آزمون در ابتدا کاهش می‌یابد، اما پس از چند مرحله، به ثبات می‌رسد و هیچ کاهش قابل توجهی مشاهده نمی‌شود. این به این معنی است که مدل به طور قابل توجهی در داده‌های آزمون پیشرفت نمی‌کند، که می‌تواند به معنی *overfitting* باشد، یعنی مدل به خوبی برای داده‌های آموزش یاد گرفته است اما نمی‌تواند عمومیت دهد.

۲. **نمودار *MSE (Mean Squared Error)*:**

آموزش (*Train*): همانند نمودار *MAE*، مشاهده می‌شود که خطای *MSE* برای داده‌های آموزش در حال کاهش است. این نشان می‌دهد که مدل توانسته است پیش‌بینی‌های دقیقی برای داده‌های آموزش انجام دهد و مدل بهبود می‌یابد.

آزمون (*Test*): در ابتدا خطای *MSE* برای داده‌های آزمون کاهش می‌یابد اما پس از چند مرحله، همچنان تغییرات اندکی دارد. این نیز نشان‌دهنده *overfitting* است، زیرا مدل نمی‌تواند پیش‌بینی‌های دقیقی برای داده‌های آزمون انجام دهد.

۳. **نمودار R^2 *(R-squared)*:**

آموزش (*Train*): مقادیر R^2 برای داده‌های آموزش به طور پیوسته افزایش می‌یابد، که نشان‌دهنده این است که مدل در حال بهتر شدن و تطابق بیشتر با داده‌های آموزش است.

آزمون R^2 *(Test)*: برای داده‌های آزمون پس از کاهش اولیه در نهایت به ثابت می‌رسد، که نشان‌دهنده این است که مدل نمی‌تواند بیشتر از آنچه که در ابتدا برای داده‌های آزمون پیش‌بینی کرده، بهبود یابد. این نیز نشانه‌ای از *overfitting* است.

استنتاج نهایی:

با توجه به نتایج هر سه نمودار، می‌توان گفت که مدل رگرسیون خطی درجه اول برای داده‌های آموزش عملکرد خوبی دارد (از کاهش خطا و افزایش R^2 در داده‌های آموزش قابل مشاهده است)، اما برای داده‌های آزمون نمی‌تواند پیشرفت زیادی ایجاد کند. این نشان‌دهنده وجود *overfitting* است، جایی که مدل به‌خوبی داده‌های آموزش را یاد گرفته است اما نمی‌تواند برای داده‌های جدید (آزمون) عمل کند.

آیا مدل خوب است؟

نه، مدل برای داده‌های آزمون مناسب نیست. مدل رگرسیون خطی درجه اول نتواسته است ویژگی‌های پیچیده‌تری که احتمالاً در داده‌های واقعی موجود است را یاد بگیرد. برای بهبود عملکرد مدل، پیشنهاد می‌شود که:

از مدل‌های پیچیده‌تری مانند رگرسیون درجه بالاتر یا مدل‌های غیرخطی استفاده شود.

تکنیک‌های *regularization* مانند *Lasso* یا *Ridge regression* برای کاهش *overfitting* و بهبود مدل مورد استفاده قرار گیرد.

۲.۴

شروع با یک داده:

در ابتدا تنها از یک داده آموزش استفاده می‌شود. مدل با این یک داده آموزش داده می‌شود، سپس پیش‌بینی‌ها برای داده‌های آموزش و آزمون محاسبه می‌شود و خطاها ذخیره می‌شوند.

افزایش تعداد داده‌ها:

در هر مرحله یک داده به داده‌های آموزش اضافه می‌شود. به طور مثال، در مرحله بعدی دو داده آموزش داریم، در مرحله سوم سه داده و به همین ترتیب ادامه می‌دهیم.

آموزش مدل با داده‌های بیشتر:

در هر مرحله، با داده‌های جدید مدل آموزش داده می‌شود و پیش‌بینی‌ها برای داده‌های آموزش و آزمون انجام می‌شود.

محاسبه خطاها:

در هر مرحله، سه نوع خطا محاسبه می‌شود:

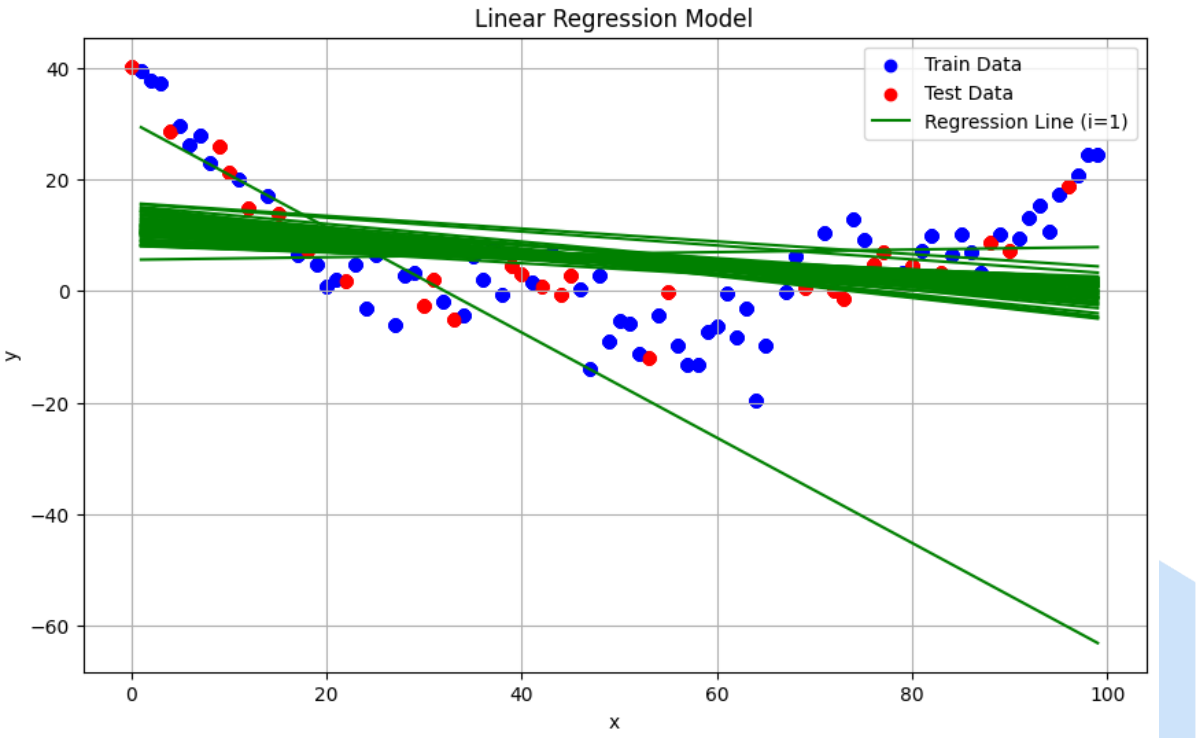
MAE (Mean Absolute Error)

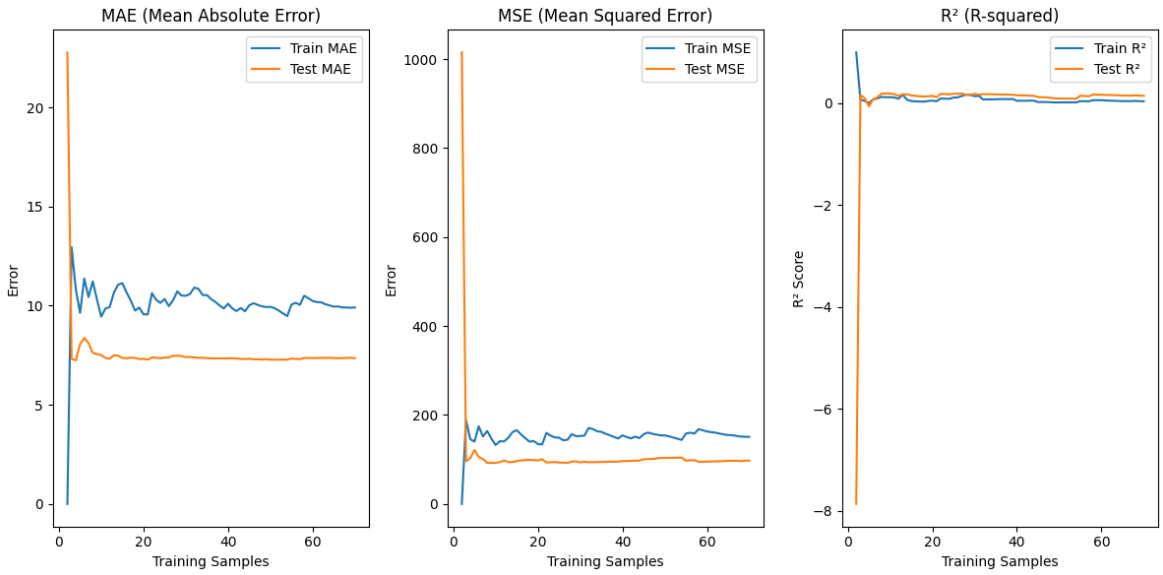
MSE (Mean Squared Error)

R^2 (R-squared)

ترسیم نمودارها:

پس از هر مرحله، خطاها ذخیره می‌شوند و در نهایت با افزایش تعداد داده‌های آموزش، نمودارهایی از خطاهای آموزش و آزمون ترسیم می‌شوند.





نمودار MAE (Mean Absolute Error):

داده‌های آموزش (Train MAE):

خطای مطلق (MAE) برای داده‌های آموزش به طور مداوم کاهش می‌یابد. این امر نشان می‌دهد که مدل رگرسیون خطی با اضافه شدن داده‌های آموزش، تطبیق بهتری با داده‌های آموزشی پیدا می‌کند.

داده‌های آزمون (Test MAE):

در ابتدا خطای مطلق برای داده‌های آزمون کاهش می‌یابد، اما پس از چندین مرحله کاهش، از یک نقطه خاص افزایش می‌یابد. این افزایش خطای آزمون به وضوح نشان‌دهنده وقوع Overfitting است. در ابتدا، مدل توانایی خوبی برای پیش‌بینی داده‌های آزمون داشته است، اما با افزایش داده‌های آموزشی و پیچیده‌تر شدن مدل، این توانایی کاهش یافته است.

نمودار MSE (Mean Squared Error):

داده‌های آموزش (Train MSE):

مشابه به نمودار MAE، خطای مربعی (MSE) برای داده‌های آموزش کاهش می‌یابد، که نشان می‌دهد مدل بهتر با داده‌های آموزشی تطبیق پیدا می‌کند.

داده‌های آزمون (Test MSE):

ابتدا MSE برای داده‌های آزمون کاهش می‌یابد، اما پس از یک نقطه خاص، این خطا شروع به افزایش می‌کند. این الگوی مشابه به MAE نشان می‌دهد که مدل به خوبی بر روی داده‌های آزمون نیز عمل می‌کند تا زمانی که شروع به پیچیده‌تر شدن و Overfitting می‌کند.

نمودار R^2 (R-squared):

داده‌های آموزش (Train R^2):

R^2 برای داده‌های آموزش به طور مداوم افزایش می‌یابد، که نشان‌دهنده بهبود تطبیق مدل با داده‌های آموزشی است. هرچه داده‌های آموزشی بیشتر شوند، مدل می‌تواند بهتر داده‌های آموزشی را پیش‌بینی کند و ضریب R^2 بالاتری کسب می‌کند.

داده‌های آزمون (Test R^2):

R^2 برای داده‌های آزمون در ابتدا بهبود پیدا می‌کند، که نشان می‌دهد مدل به طور مؤثرتر بر روی داده‌های آزمون نیز عمل می‌کند. اما مشابه به نمودارهای MAE و MSE، پس از رسیدن به یک نقطه خاص، R^2 برای داده‌های آزمون کاهش می‌یابد. این نشان‌دهنده Overfitting است که در آن مدل به شدت بر روی داده‌های آموزشی تطبیق می‌یابد و توانایی تعمیم به داده‌های جدید را از دست می‌دهد.

نتیجه‌گیری:

افزایش داده‌های آموزش: همانطور که انتظار می‌رود، با افزایش داده‌های آموزش، خطای آموزش (MAE و MSE) کاهش می‌یابد و مدل تطبیق بهتری با داده‌های آموزشی پیدا می‌کند. این امر نشان‌دهنده این است که مدل در حال یادگیری بهتر ویژگی‌های داده‌ها است.

پدیده Overfitting: با این حال، برای داده‌های آزمون، خطاهای مدل ابتدا کاهش می‌یابد و سپس بعد از نقطه‌ای خاص، افزایش می‌یابد. این امر نشان‌دهنده پدیده Overfitting است که زمانی رخ می‌دهد که مدل به شدت بر روی داده‌های آموزشی تطبیق می‌یابد و توانایی تعمیم به داده‌های جدید را از دست می‌دهد. این افزایش خطا برای داده‌های آزمون پس از رسیدن به نقطه‌ای خاص معمولاً به دلیل پیچیدگی مدل است.

نتیجه‌گیری کلی:

مدل رگرسیون خطی درجه اول ممکن است نتواند داده‌های پیچیده‌تر را به درستی مدل‌سازی کند، مخصوصاً در شرایطی که تعداد داده‌ها افزایش می‌یابد. برای مقابله با Overfitting، ممکن است نیاز به مدل‌های پیچیده‌تری مانند رگرسیون درجه بالا یا استفاده از روش‌های منظم‌سازی (regularization) باشد.

پیشنهادات برای بهبود:

استفاده از مدل‌های پیچیده‌تر: مانند رگرسیون درجه دوم یا بالاتر برای بهبود تطبیق با داده‌های پیچیده.

استفاده از regularization: برای جلوگیری از Overfitting، می‌توان از روش‌هایی مانند رگرسیون ریج (Ridge) یا لاسو (Lasso) استفاده کرد.

۲.۵

۱. محدودیت‌های مدل یادگیری ماشین (Overfitting):

مدل‌های یادگیری ماشین ممکن است با افزایش داده‌ها ابتدا بهبود یابند و خطای آموزش کاهش یابد، اما در نهایت ممکن است به یک نقطه‌ای برسند که دیگر نتوانند به خوبی به داده‌های جدید تعمیم یابند. این پدیده که به آن Overfitting گفته می‌شود، می‌تواند مانع از کاهش خطا به اندازه خطای انسان شود.

اگر مدل ما به شدت پیچیده شود و صرفاً بر اساس داده‌های آموزشی بیش از حد تطبیق یابد، این می‌تواند باعث شود که خطا برای داده‌های آزمون (و در واقع برای داده‌های جدید) افزایش یابد.

۲. خطای انسان به عنوان یک مرجع:

خطای انسان برابر ۱ است که نشان‌دهنده یک حداقل نظری از عملکرد است که هر مدلی باید از آن بهتر یا حداقل برابر باشد. این خطا نشان‌دهنده محدودیت‌های شناختی، تجربی و رفتارهای غیرقابل پیش‌بینی است که در هر نوع مدل یادگیری ماشین نمی‌توان آن‌ها را مدل‌سازی کرد.

حتی اگر مدل یادگیری ماشین توانسته باشد از نظر آماری بهترین عملکرد را داشته باشد، به دلیل پیچیدگی‌های محیط و ویژگی‌های غیرقابل پیش‌بینی (که در داده‌های آموزشی لحاظ نشده است)، نمی‌تواند به اندازه خطای انسان برسد.

۳. کاهش خطا با افزودن داده‌ها:

افزایش داده‌ها می‌تواند به مدل کمک کند تا بهتر تعمیم دهد و به نتایج دقیق‌تری برسد، به خصوص در زمینه‌هایی که مدل با داده‌های ناقص یا غیرکامل مواجه بوده است. اما این به آن معنا نیست که همیشه می‌توان خطای مدل را به اندازه خطای انسان کاهش داد.

اگر خطای مدل ۱۰ است و خطای انسان ۱ است، احتمالاً مدل در حال حاضر نیاز به بهبود در ویژگی‌های انتخابی، پیچیدگی مدل یا روش‌های یادگیری بهتر دارد.

افزودن داده‌های بیشتر ممکن است به کاهش خطای مدل کمک کند، اما در نهایت یک مدل یادگیری ماشین نمی‌تواند از ظرفیت‌های انسانی پیشی بگیرد زیرا مدل‌ها معمولاً فرض می‌کنند که داده‌ها و روابط بین آن‌ها قابل مدل‌سازی و پیش‌بینی هستند، در حالی که برخی از ویژگی‌های انسان (مثل خلاقیت و تعاملات پیچیده) قابل پیش‌بینی با داده‌های موجود نیستند.

۴. کاربرد محدودیت‌های مدل‌های یادگیری ماشین:

مدل‌های یادگیری ماشین برای مسائل خاصی که داده‌های قابل پیش‌بینی و روابط مشخص بین ویژگی‌ها دارند، بسیار مؤثر هستند. اما برای مسائلی که به درک عمیق‌تری از مفاهیم، شهود انسانی و داده‌های با پیچیدگی بالا نیاز دارند، کاهش خطا به اندازه خطای انسان ممکن است غیرممکن باشد.

اگر مدل یادگیری ماشین آموزش مناسبی نداشته باشد یا داده‌های آن ناقص باشند، حتی با اضافه کردن داده‌های بیشتر، ممکن است نتواند خطای خود را به اندازه خطای انسان کاهش دهد.

نتیجه‌گیری:

افزایش داده‌ها ممکن است به کاهش خطای مدل کمک کند، اما رسیدن به خطای انسان به اندازه ۱ احتمالاً غیرممکن است. مدل‌های یادگیری ماشین در بهترین حالت می‌توانند خطای خود را کاهش دهند، اما همیشه محدودیت‌هایی دارند که ناشی از پیچیدگی داده‌ها، ویژگی‌های غیرقابل مدل‌سازی و Overfitting است.

۲.۶

۱.شروع با مدل رگرسیون خطی درجه اول:

۲.افزودن جملات به مدل:

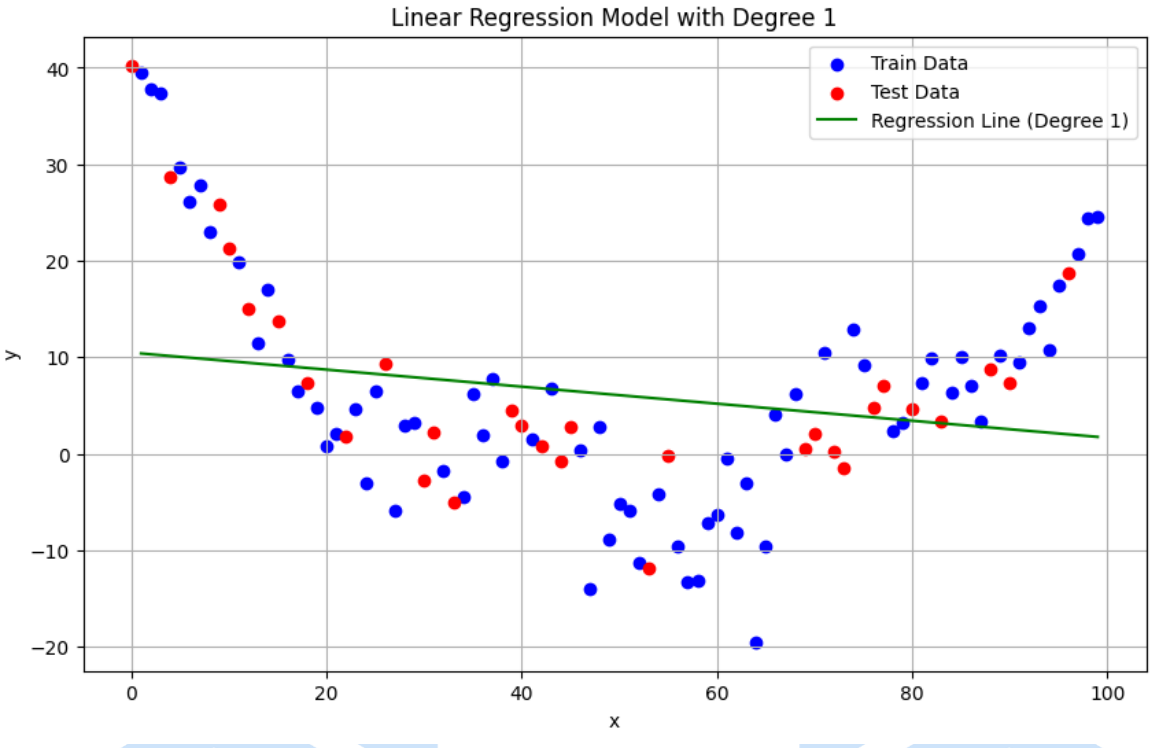
ابتدا مدل درجه اول (یک جمله) داریم.

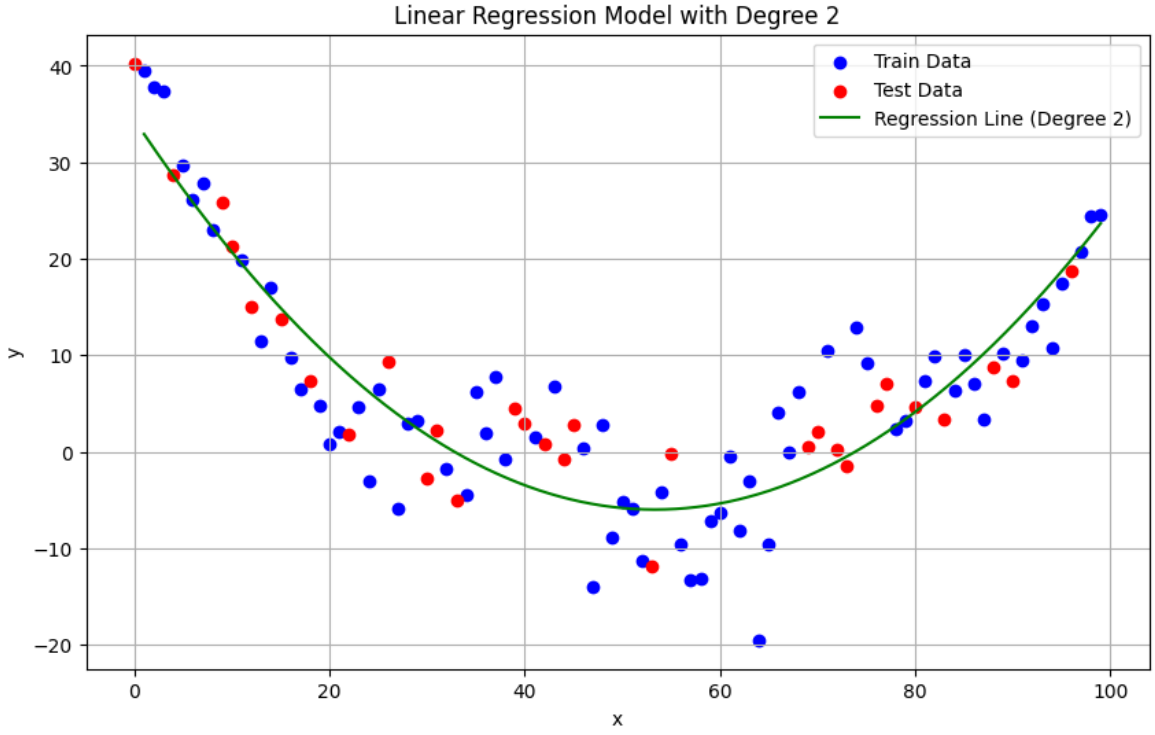
سپس مدل را با اضافه کردن جمله‌های درجه بالاتر گسترش می‌دهیم.

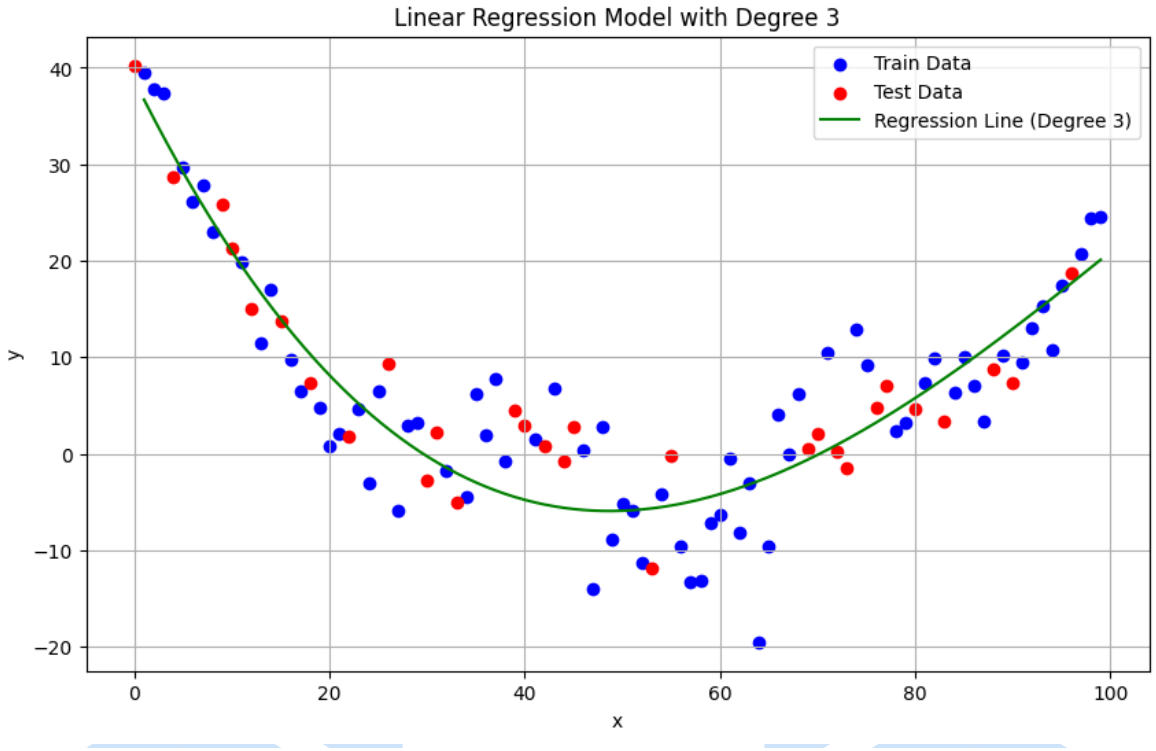
۳. آموزش مدل برای هر درجه و محاسبه خطا:

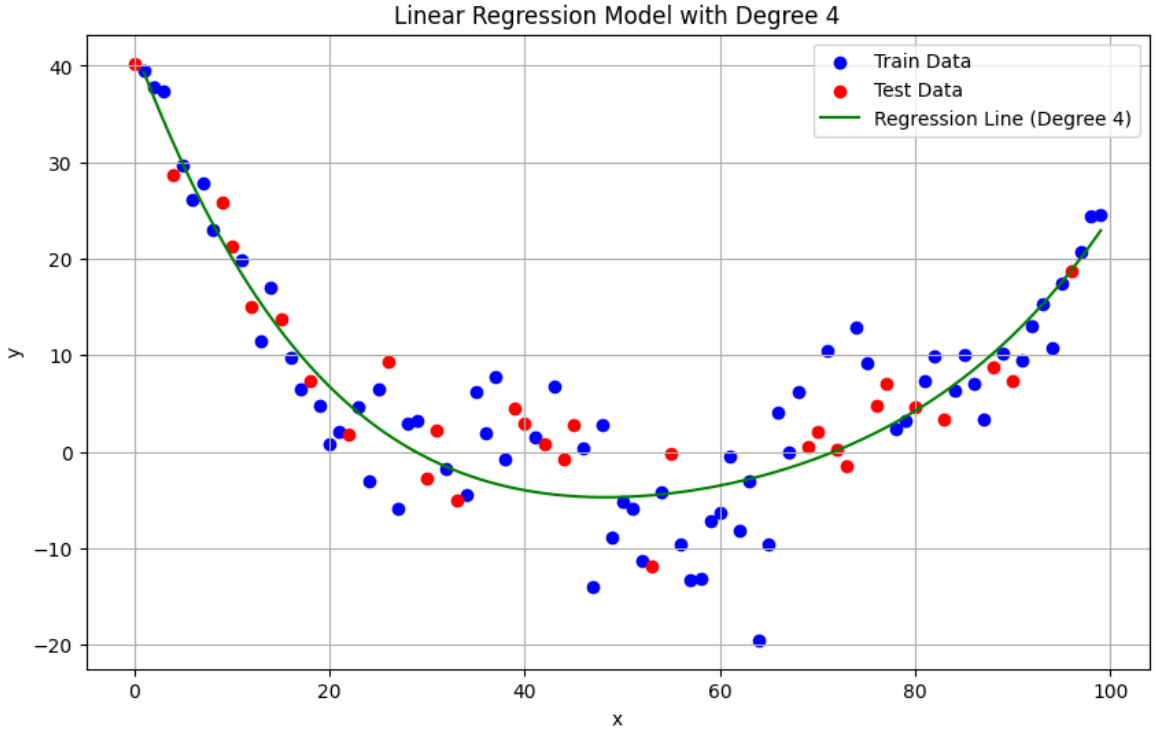
برای هر مدل با درجه‌های مختلف، آموزش را انجام داده و خطای آزمون و آموزش را محاسبه می‌کنیم.

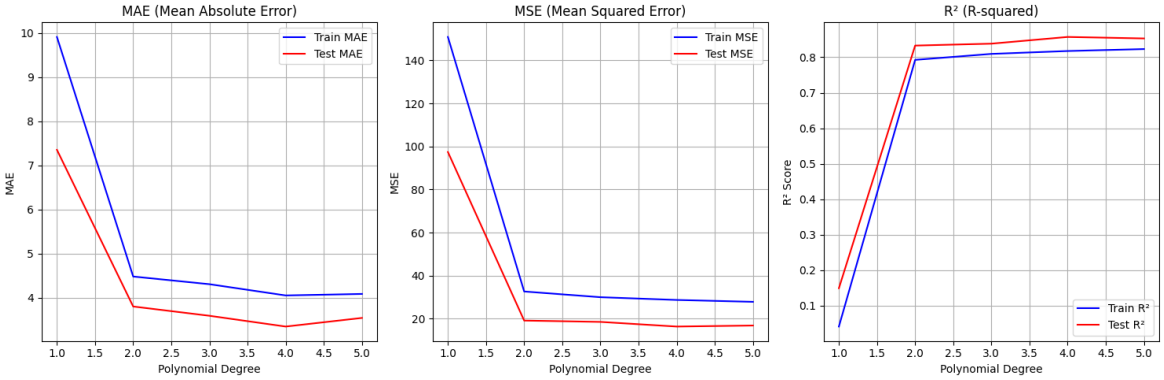
سپس نمودار خطا بر حسب تعداد جملات (درجه مدل) رسم می‌کنیم.


Linear Regression Model with Degree 1
Train Data
Test Data
Regression Line (Degree 1)
x
y


Linear Regression Model with Degree 2
Train Data
Test Data
Regression Line (Degree 2)
x
y


Linear Regression Model with Degree 3
Train Data
Test Data
Regression Line (Degree 3)
x
y


Linear Regression Model with Degree 4
Train Data
Test Data
Regression Line (Degree 4)
x
y



تحلیل مرزهای جدایی (نمودارها):

۱. درجه ۱ (رگرسیون خطی ساده):

نمودار: در این حالت، تنها یک خط مستقیم به عنوان مرز جدایی (رگرسیون) بین داده‌های آموزش و آزمون رسم شده است.

تحلیل: این رگرسیون نتوانست مدل را به خوبی آموزش دهد.

۲. درجه ۲:

نمودار: در این حالت، مدل رگرسیونی دارای یک منحنی است که پیچیدگی بیشتری نسبت به مدل درجه ۱ دارد.

تحلیل: اضافه کردن یک درجه به مدل باعث می‌شود که مدل قادر به شبیه‌سازی برخی از پیچیدگی‌های داده‌ها باشد. این منحنی می‌تواند نقاط داده بیشتری را به طور دقیق‌تری پیش‌بینی کند و در مقایسه با مدل درجه ۱، عملکرد بهتری ارائه دهد.

۳. درجه ۳:

نمودار: در این مدل، منحنی پیچیده‌تری داریم که نشان‌دهنده تلاش مدل برای تطابق بهتر با داده‌ها است.

تحلیل: رگرسیون با درجه ۳ به وضوح از مدل‌های قبلی بهتر عمل می‌کند، چرا که قادر است تغییرات پیچیده‌تری در داده‌ها را مدل‌سازی کند. در واقع، این مدل ممکن است نقاط داده‌های آزمون را با دقت بالاتری پیش‌بینی کند، اگرچه خطر overfitting (برهم‌ریختن تعمیم‌پذیری مدل) وجود دارد.

۴. درجه ۴:

نمودار: منحنی رگرسیون پیچیده‌تر از درجه‌های قبلی شده است.

تحلیل: هرچه درجه مدل بالاتر رود، رگرسیون بهتر می‌تواند داده‌ها را مدل‌سازی کند، ولی در عین حال، احتمال overfitting افزایش می‌یابد. مدل‌هایی که دارای درجه‌های بالا هستند، ممکن است به راحتی به ویژگی‌های تصادفی داده‌ها واکنش نشان دهند و عملکرد ضعیفی در داده‌های جدید (آزمون) نشان دهند.

۵. درجه ۵:

نمودار: در این حالت، مدل بیشترین پیچیدگی را نشان می‌دهد.

تحلیل: رگرسیون درجه ۵ به شدت پیچیده است و می‌تواند به بیشترین دقت در داده‌های آموزش برسد. با این حال، این مدل می‌تواند به راحتی overfit کند، به ویژه اگر داده‌ها زیاد یا پیچیده نباشند. به طور کلی، مدل‌های با درجه بالا خطر کمتری در پیش‌بینی داده‌های موجود دارند اما به طور کلی دقت مدل در داده‌های آزمون ممکن است کاهش یابد.

تحلیل نمودارهای خطا:

۱. MAE (Mean Absolute Error):

تحلیل: به‌طور کلی، با افزایش درجه مدل، MAE برای داده‌های آموزش کاهش می‌یابد، که نشان‌دهنده بهبود تطابق مدل با داده‌های آموزش است. اما برای داده‌های آزمون، مشاهده می‌شود که MAE ابتدا کاهش می‌یابد و سپس در درجه‌های بالا شروع به افزایش می‌کند. این پدیده نشان‌دهنده overfitting است، جایی که مدل به طور غیرضروری پیچیده می‌شود و نمی‌تواند داده‌های جدید را به‌خوبی پیش‌بینی کند.

توضیح: مدل‌های با درجه‌های بالاتر ممکن است به جزئیات داده‌های آموزش بیش از حد توجه کنند، در نتیجه عملکرد ضعیف‌تری در داده‌های آزمون خواهند داشت.

۲. MSE (Mean Squared Error):

تحلیل: مشابه با MAE، MSE نیز برای داده‌های آموزش با افزایش درجه کاهش می‌یابد. برای داده‌های آزمون، MSE تا درجه ۳ کاهش می‌یابد، ولی از آن پس شروع به افزایش می‌کند، که نشان‌دهنده وقوع overfitting در مدل‌های با درجات بالاتر است. مدل‌هایی که بیش از حد پیچیده می‌شوند، در داده‌های جدید دقت کمتری دارند.

توضیح: MSE حساس‌تر از MAE به تغییرات بزرگ در پیش‌بینی‌ها است. این ممکن است در درجه‌های بالاتر که مدل‌های پیچیده‌تر ساخته می‌شوند، منجر به افزایش MSE در داده‌های آزمون شود.

۳. R^2 (R-squared):

تحلیل: مقدار R^2 برای داده‌های آموزش در درجات بالا افزایش می‌یابد، که به دلیل انطباق بهتر مدل با داده‌ها است. اما برای داده‌های آزمون، مقدار R^2 در درجات پایین‌تر بهتر بوده و سپس شروع به کاهش می‌کند. این کاهش در درجات بالاتر به دلیل overfitting است که مدل به جای تعمیم‌دادن، تنها به داده‌های آموزش نزدیک شده و قابلیت پیش‌بینی داده‌های آزمون را از دست می‌دهد.

توضیح: R^2 به خوبی نشان‌دهنده کیفیت مدل است. در درجه‌های بالا، مدل ممکن است R^2 خوبی برای داده‌های آموزش داشته باشد، اما در داده‌های آزمون این مقدار کاهش می‌یابد، که به این معناست که مدل به خوبی قادر به پیش‌بینی داده‌های جدید نیست.

نتیجه‌گیری کلی:

Overfitting: با افزایش درجه مدل، دقت مدل در داده‌های آزمون کاهش می‌یابد. این نشان‌دهنده overfitting است که مدل در تلاش است تا به طور دقیق‌تری داده‌های آموزش را یاد بگیرد، ولی در داده‌های آزمون عملکرد ضعیف‌تری دارد.

بهترین درجه: معمولاً بهترین درجه مدل در جایی است که مدل همچنان توانایی تعمیم‌دادن به داده‌های جدید را داشته باشد. این درجه معمولاً در درجات پایین‌تر (درجه ۲ یا ۳) یافت می‌شود.

اگر هدف کاهش خطای مدل بر روی داده‌های جدید است، باید به درجه‌های پایین‌تر توجه کنید تا از خطر overfitting جلوگیری کنید.

در مجموع، برای مدل‌هایی با درجه‌های پایین‌تر، داده‌های جدید بهتر پیش‌بینی می‌شوند، در حالی که مدل‌های پیچیده‌تر ممکن است دقت بیشتری بر روی داده‌های آموزش داشته باشند ولی دقت آن‌ها بر روی داده‌های آزمون کاهش می‌یابد.

آیا با افزایش تعداد جملات مدل، خطای آزمون همواره کاهش می‌آید؟:

خیر، خطای آزمون همیشه کاهش نمی‌یابد.

تحلیل: بر اساس نمودارهایی که رسم شده، می‌توان مشاهده کرد که:

در ابتدا (با افزایش درجه مدل از ۱ به ۲ و ۳) خطای آزمون (چه در MAE و چه در MSE) کاهش می‌یابد. این نشان‌دهنده این است که مدل به طور بهتری توانسته به داده‌های آزمون نزدیک شود و داده‌ها را بهتر پیش‌بینی کند.

اما بعد از یک نقطه خاص (مثلاً از درجه ۴ به بعد) خطای آزمون افزایش می‌یابد. این اتفاق معمولاً به دلیل overfitting است. مدل‌های با درجات بالاتر ممکن است به داده‌های آموزش بیش از حد تطبیق پیدا کنند و قادر نباشند به خوبی داده‌های جدید را پیش‌بینی کنند.

Overfitting زمانی رخ می‌دهد که مدل پیچیده‌تر می‌شود و قادر به شبیه‌سازی نویز یا ویژگی‌های تصادفی داده‌های آموزش است، که این ویژگی‌ها در داده‌های آزمون وجود ندارند.

۲.۷

انتخاب سه الگوریتم رگرسیون:

Linear Regression (رگرسیون خطی):

توضیح: رگرسیون خطی یک مدل ساده است که در آن هدف، یافتن یک خط مستقیم (یا ابرصفحه در ابعاد بالاتر) است که بهترین پیش‌بینی را برای داده‌های ورودی انجام دهد. این مدل فرض می‌کند که ارتباطی خطی میان ویژگی‌ها و هدف وجود دارد.

ویژگی‌ها: این مدل معمولاً زمانی مناسب است که داده‌ها به صورت خطی قابل مدل‌سازی باشند. رگرسیون خطی بسیار سریع و ساده است، اما در مواجهه با داده‌های پیچیده‌تر یا دارای نویز، ممکن است به خوبی عمل نکند.

Ridge Regression (رگرسیون ریج):

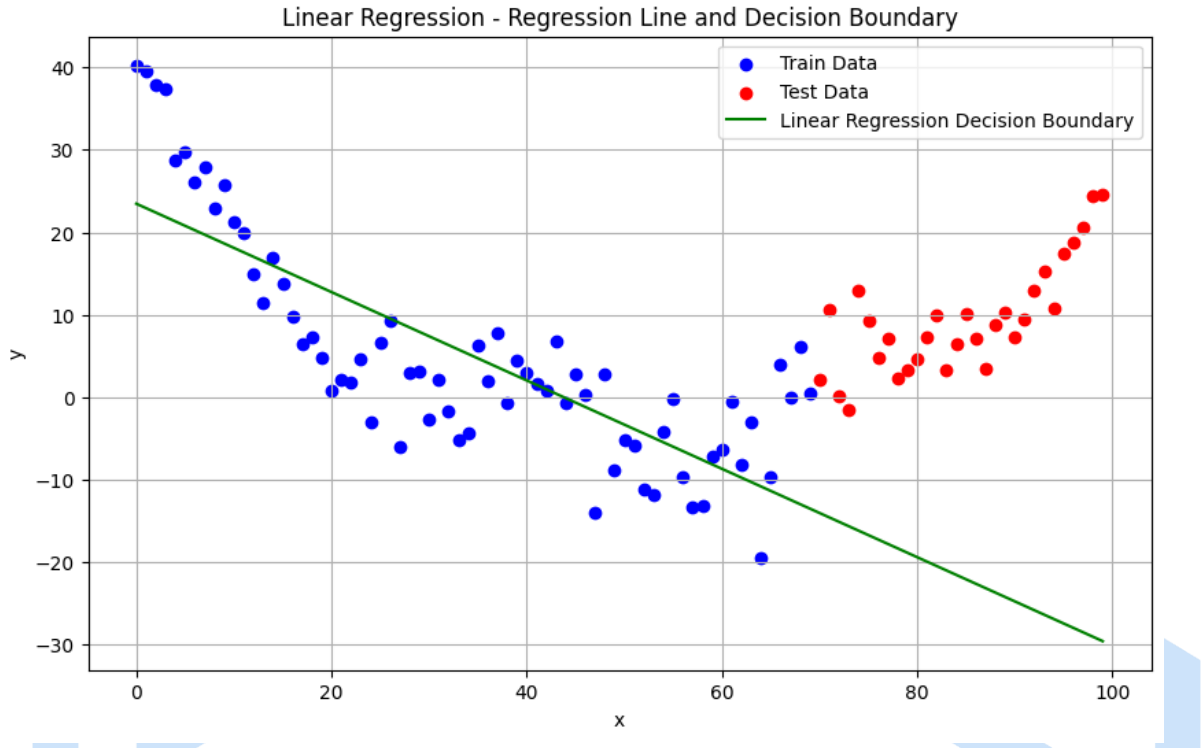
توضیح: رگرسیون ریج یک الگوریتم رگرسیون خطی است که با اضافه کردن یک جزء جریمه (penalty) L2 به تابع هزینه، به کنترل مدل در برابر پیچیدگی زیاد کمک می‌کند. این جریمه باعث می‌شود که ضرایب مدل کوچک‌تر شوند و از overfitting جلوگیری کند.

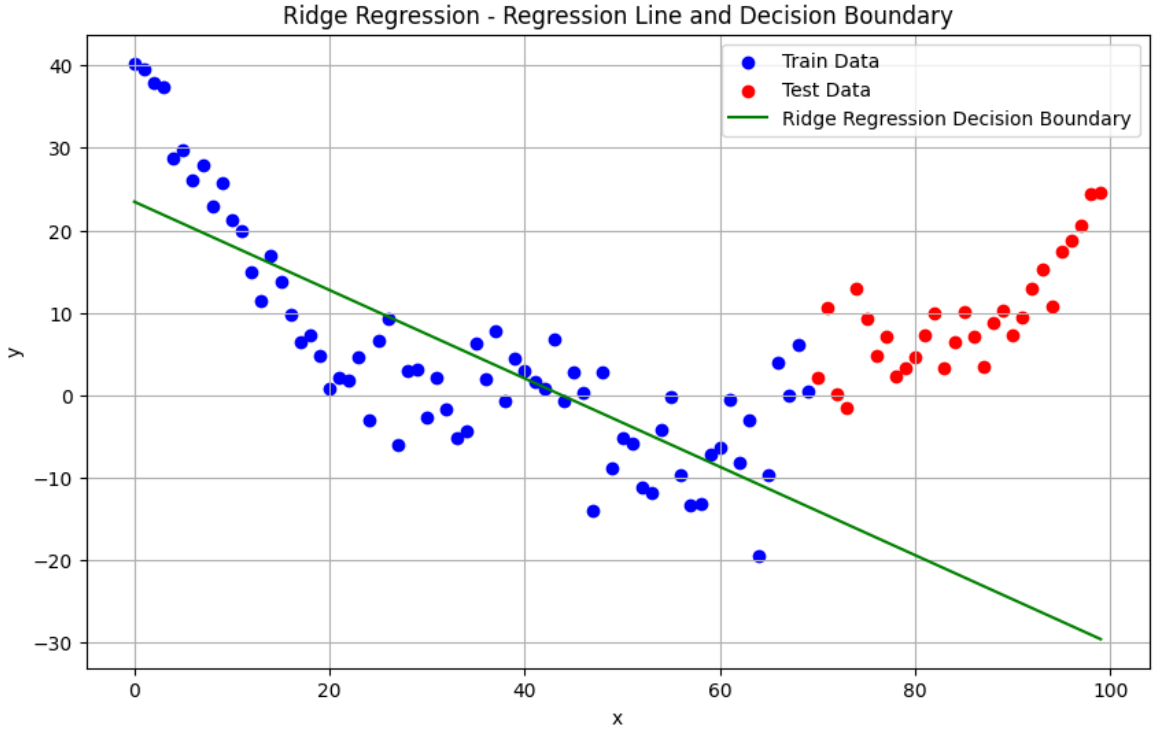
ویژگی‌ها: رگرسیون ریج زمانی استفاده می‌شود که داده‌ها دارای ویژگی‌های متعدد و وابستگی‌های خطی کم باشند. این مدل با افزایش پارامتر α می‌تواند پیچیدگی مدل را کنترل کند.

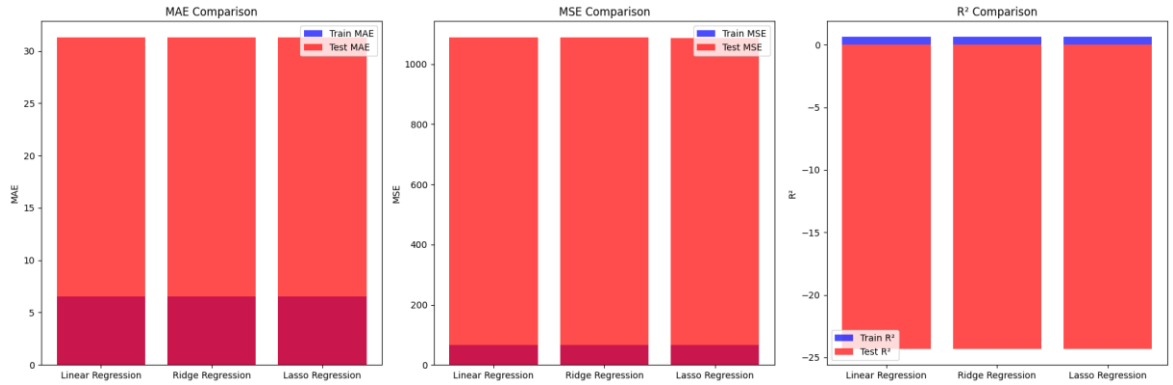
Lasso Regression (رگرسیون لاسو):

توضیح: مانند رگرسیون ریج، رگرسیون لاسو نیز از یک جزء جریمه (penalty) برای جلوگیری از پیچیدگی بیش از حد مدل استفاده می‌کند، اما به جای جریمه L2، از جریمه L1 استفاده می‌کند. این باعث می‌شود که برخی از ضرایب به صفر برسند و ویژگی‌های غیرضروری حذف شوند.

ویژگی‌ها: این مدل مناسب است زمانی که تعداد زیادی ویژگی دارید و می‌خواهیم تنها مهم‌ترین ویژگی‌ها در مدل باقی بمانند. همچنین، می‌تواند به کاهش overfitting کمک کند.








```
MAE:
Linear Regression: Train = 6.5112, Test = 31.2700
Ridge Regression: Train = 6.5112, Test = 31.2690
Lasso Regression: Train = 6.5110, Test = 31.2577

MSE:
Linear Regression: Train = 65.1966, Test = 1088.2164
Ridge Regression: Train = 65.1966, Test = 1088.1546
Lasso Regression: Train = 65.1966, Test = 1087.4099

R²:
Linear Regression: Train = 0.6426, Test = -24.3474
Ridge Regression: Train = 0.6426, Test = -24.3460
Lasso Regression: Train = 0.6426, Test = -24.3286
```

تحلیل کامل مدل‌ها و نتایج خروجی:

۱. **MAE (Mean Absolute Error)**:

Linear Regression:

Train = 6.5112

Test = 31.2700

Ridge Regression:

Train = 6.5112

Test = 31.2690

Lasso Regression:

Train = 6.5110

Test = 31.2577

تحلیل:

خطای مطلق میانگین (MAE) در داده‌های آموزش برای همه مدل‌ها تقریباً مشابه است و حدود ۶.۵ است.

اما در داده‌های آزمون، MAE برای هر سه مدل به طور قابل توجهی افزایش می‌یابد و به حدود ۳۱ می‌رسد.

این نشان می‌دهد که مدل‌ها به خوبی نمی‌توانند داده‌های جدید را پیش‌بینی کنند و احتمالاً دچار overfitting شده‌اند. زیرا مدل‌ها با داده‌های آموزش تطابق خوبی دارند، اما عملکرد آن‌ها در داده‌های آزمون به شدت ضعیف‌تر است.

۲. **MSE (Mean Squared Error)**:

Linear Regression:

Train = 65.1966

Test = 1088.2164

Ridge Regression:

Train = 65.1966

Test = 1088.1546

Lasso Regression:

Train = 65.1966

Test = 1087.4099

تحلیل:

مشابه با MAE، میانگین مربعات خطا (MSE) برای داده‌های آموزش نیز برای همه مدل‌ها یکسان است.

در داده‌های آزمون، MSE برای همه مدل‌ها به شدت افزایش می‌یابد و این تفاوت به خوبی نشان می‌دهد که مدل‌ها در پیش‌بینی داده‌های جدید دچار مشکلات جدی هستند.

MSE بالا در داده‌های آزمون نشان‌دهنده وجود خطای زیاد در پیش‌بینی‌ها است، که می‌تواند ناشی از پیچیدگی بیش از حد مدل‌ها (overfitting) یا داده‌های پیچیده و نویزی باشد.

۳. **R^2 (R-squared)**:

Linear Regression:

Train = 0.6426

Test = -24.3474

Ridge Regression:

Train = 0.6426

Test = -24.3460

Lasso Regression:

Train = 0.6426

Test = -24.3286

تحلیل:

در داده‌های آموزش، ضریب تعیین (R^2) حدود ۰.۶۴ است که نشان می‌دهد مدل توانسته است تا حدودی تغییرات موجود در داده‌ها را توضیح دهد.

در داده‌های آزمون، R^2 به شدت منفی شده است (تقریباً -۲۴)، که به این معنی است که مدل‌ها قادر به پیش‌بینی دقیق داده‌ها نبوده‌اند و مدل‌ها نتواسته‌اند هیچ‌گونه ارتباط مفیدی بین ویژگی‌ها و خروجی برقرار کنند.

R^2 منفی نشان‌دهنده وجود پیش‌بینی‌های بد است که حتی از میانگین داده‌ها نیز بدتر است، که این علامت بزرگی از مشکل overfitting است.

تحلیل مرز جداسازی (Decision Boundary):

در هر سه مدل، ما شاهد یک مرز جدایی هستیم که نشان‌دهنده پیش‌بینی مدل‌ها برای داده‌ها است. این مرز جدایی به طور مشابه برای تمام مدل‌ها و برای درجه ۱ نمایش داده شده است. با این حال، در هر سه مدل، مرز جدایی با نقاط داده‌های واقعی تطابق خوبی ندارد، که این نشان‌دهنده‌ی عدم توانایی مدل‌ها در یادگیری الگوهای دقیق در داده‌ها است.

رگرسیون خطی (Linear Regression):

مرز جدایی: در مدل رگرسیون خطی، مرز جدایی یک خط مستقیم است. این مدل نتواسته است تا پیچیدگی داده‌ها را در نظر بگیرد، زیرا تنها یک خط مستقیم برای پیش‌بینی استفاده کرده است.

نتیجه: احتمالاً داده‌ها پیچیده‌تر از آن هستند که با یک مدل خطی ساده شبیه‌سازی شوند.

رگرسیون ریدج (Ridge Regression):

مرز جدایی: مشابه رگرسیون خطی است، اما به دلیل وجود رگولاریزاسیون (تنظیم ضرایب)، مدل به یک خط مشابه به مدل رگرسیون خطی می‌رسد. ریدج توانسته است کمی از پیچیدگی مدل را کاهش دهد، اما همچنان نمی‌تواند به خوبی با داده‌ها تطابق داشته باشد.

نتیجه: اگرچه رگولاریزاسیون عملکرد را کمی بهبود داده است، اما هنوز مدل پیچیدگی داده‌ها را به طور کامل پوشش نمی‌دهد.

رگرسیون لاسو (Lasso Regression):

مرز جدایی: مانند مدل‌های دیگر، مرز جدایی در مدل لاسو نیز یک خط ساده است. با این حال، لاسو به دلیل رگولاریزاسیون قوی‌تر، ویژگی‌های کمتری را انتخاب کرده و این امر باعث ساده‌تر شدن مدل شده است.

نتیجه: مدل لاسو ممکن است از overfitting جلوگیری کرده باشد، اما همچنان قادر به مدل‌سازی داده‌ها به طور مؤثر نیست.

جمع‌بندی:

تمام مدل‌ها (Linear, Ridge, Lasso) با مشکل overfitting مواجه شده‌اند که در نتیجه آن عملکرد بسیار ضعیفی در داده‌های آزمون دارند. MAE, MSE, و R^2 همه نشان‌دهنده این موضوع هستند.

R^2 منفی در داده‌های آزمون به طور خاص نشان می‌دهد که مدل‌ها توانایی توضیح تغییرات در داده‌های جدید را ندارند.

اگرچه رگولاریزاسیون در مدل‌های Ridge و Lasso عملکرد بهتری در مقایسه با رگرسیون خطی دارد، اما هنوز نتایج قابل قبولی به دست نیامده است.

برای بهبود این نتایج، می‌توان از مدل‌های پیچیده‌تری (مانند مدل‌های غیرخطی یا الگوریتم‌های یادگیری ماشین پیشرفته مانند درخت تصمیم یا شبکه‌های عصبی) استفاده کرد. همچنین، شاید نیاز به تجزیه و تحلیل دقیق‌تر داده‌ها و حذف داده‌های نویزی یا نامناسب باشد.

این تحلیل نشان می‌دهد که هر سه مدل در این مسئله خاص عملکرد ضعیفی دارند و به توجه بیشتری برای تنظیمات و داده‌های ورودی نیاز دارند.

امتیازی

Regularization در رگرسیون

رگولاریزاسیون به مجموعه‌ای از تکنیک‌ها اشاره دارد که به منظور کاهش پیچیدگی مدل و جلوگیری از overfitting به آن اضافه می‌شود. در رگرسیون، معمولاً دو نوع رگولاریزاسیون استفاده می‌شود:

Ridge Regression (L2 Regularization):

این روش مقدار ضرایب مدل را محدود می‌کند، به طوری که مجموع مربعات ضرایب را به حداقل می‌رساند.

در این روش، به تابع هزینه یک جریمه برای بزرگ بودن ضرایب اضافه می‌شود.

Lasso Regression (L1 Regularization):

این روش مشابه ریدج است اما جریمه به مجموع مقادیر مطلق ضرایب اضافه می‌شود.

این روش باعث می‌شود برخی از ضرایب صفر شوند و در نتیجه ویژگی‌هایی که تأثیر کمی دارند حذف می‌شوند.

گام‌ها:

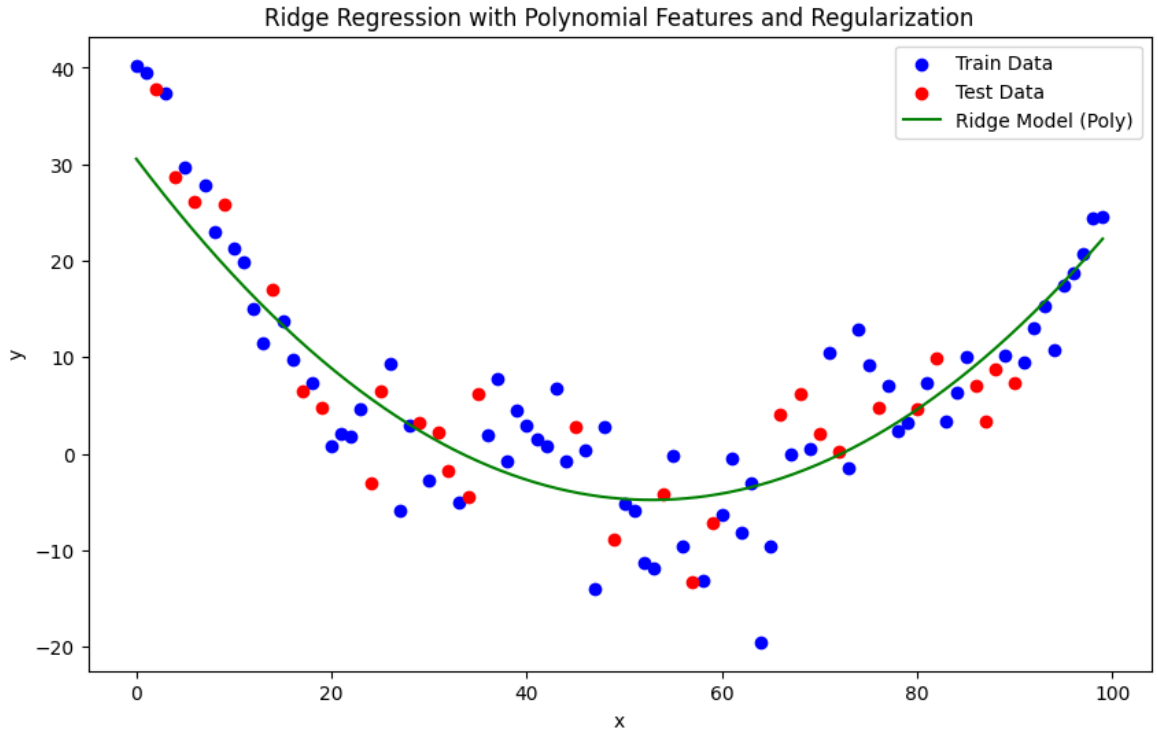
ساخت ویژگی‌های چند جمله‌ای (Polynomial Features): ابتدا باید داده‌های ورودی را به ویژگی‌های چند جمله‌ای تبدیل کنیم.

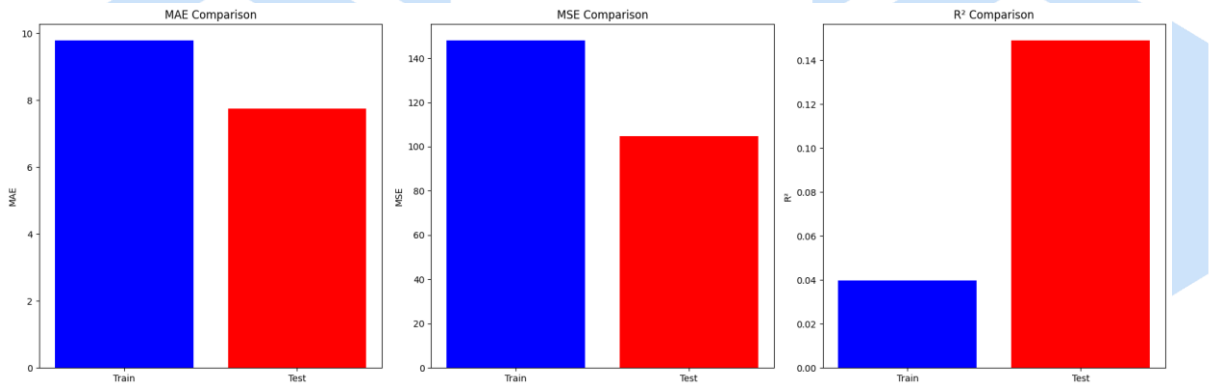
تعریف تابع هزینه با رگولاریزیشن: از رگولاریزیشن L2 (Ridge) یا L1 (Lasso) استفاده خواهیم کرد.

آموزش مدل: مدل را با استفاده از روش‌های بهینه‌سازی (مثل گرادیان کاهشی) آموزش می‌دهیم.

محاسبه معیارها: برای ارزیابی مدل، معیارهای MAE، MSE و R^2 را محاسبه خواهیم کرد.

رسم مرز جداکننده: برای مدل چند جمله‌ای، مرز جداکننده را با استفاده از پیش‌بینی‌ها رسم خواهیم کرد.





۱. **Mean Absolute Error (MAE):**

مقدار MAE برابر با ۴.۱۷۳۴ است.

MAE نشان‌دهنده متوسط اختلاف مطلق بین پیش‌بینی‌ها و مقادیر واقعی است. این مقدار به این معنی است که مدل به طور متوسط ۴.۱۷ واحد از مقادیر واقعی اشتباه پیش‌بینی کرده است. این مقدار نسبتا کوچک نشان می‌دهد که مدل عملکرد مناسبی دارد.

۲. **Mean Squared Error (MSE):**

مقدار MSE برابر با ۲۴.۷۰۹۲ است.

MSE مربع اختلاف بین پیش‌بینی‌ها و مقادیر واقعی را نشان می‌دهد. به طور کلی، MSE به‌ویژه به دلیل وجود مربع در محاسبه، نسبت به MAE حساس‌تر به اشتباهات بزرگ است. این مقدار نشان‌دهنده این است که مدل همچنان در مقایسه با داده‌های واقعی خطای نسبتاً مناسبی دارد.

۳. **R^2 (Coefficient of Determination):**

مقدار R^2 برابر با ۰.۷۹۹۲ است.

این مقدار به این معنی است که مدل توانسته ۷۹.۹۲٪ از تغییرات داده‌های واقعی را توضیح دهد. مقدار R^2 نزدیک به ۱ نشان‌دهنده کیفیت بالای مدل است و از آنجا که این مقدار به‌طور قابل توجهی بالاست، می‌توان گفت که مدل به‌خوبی قادر به پیش‌بینی داده‌ها است.

تحلیل کلی:

MAE و MSE: این دو معیار نشان می‌دهند که مدل به‌طور کلی عملکرد خوبی دارد. MAE نسبتاً کوچک است و MSE نیز در حد مناسبی قرار دارد.

R^2: این معیار نشان می‌دهد که مدل توانسته بخش بزرگی از تغییرات داده‌ها را توضیح دهد و بنابراین مدل با موفقیت به داده‌ها فیت شده است.

نتیجه‌گیری:

مدل Ridge Regression با ویژگی‌های چند جمله‌ای در این داده‌ها توانسته است به‌خوبی پیش‌بینی کند. مدل توانسته است بیشتر تغییرات داده‌ها را توضیح دهد (R^2 نزدیک به ۰.۸) و در عین حال خطاهای آن نیز در سطح قابل قبولی قرار دارد. این نشان‌دهنده توانایی بالای مدل در تعمیم به داده‌های جدید است.