به نام خدا

دانشگاه اصفهان

دانشکده مهندسی کامپیوتر

درس مدیریت پایگاه دانش

گزارش پروژه سری ..1...

نام و نام خانوادگی	ابوالفضل عابدینی
شماره دانشجویی	4023614026
تاریخ ارسال گزارش	1403/01/26

فهرست گزارش

مقدمه

ابتدا مراحل پیش پردازش از قبیل حذف سطر های تکراری و پر کردن مقادیر خالی , نرمالسازی و...

سپس در مرحله EDA کارهایی از قبیل بازبینی داده ها و نشان دادن نمودار ها و روابط ها و .... پرداختیم بعد از انجام مرحله پیش پردازش دیتا ست پیش پردازش شده را به مدل logistic regression داده ایم قبل مرحله پیش پردازش داده دقت مدل 79 درصد بود بعد از مرحله دقت مدل به 86 درصد افزایش پیدا کرد.

پیش پردازش

اصلی ترین کار در این پروژه پیش پردازش است زیرا باعث میشود بتوانیم دقت مدل خود را افزایش دهیم و همچنین زمان اجرا را کاهش دهیم. در این مرحله 5 کار انجام شده که به ترتیب انها را بیان میکنیم:

1.حذف سطر های تکراری

تعداد سطر های دیتا ست 229990 است ما در این مرحله میخواهیم سطرهای تکراری را حذف کنیم برای این منظور یک تابع به نام deduplication نوشته شده است بعد اجرا این تابع تعدادسطر ها به 7838 کاهش پیدا کرد.خروجی این تابع به صورت زیر است.

]:

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	...	DeviceProtection	Tech
0	5331-RGMTT	Male	1.0	Yes	No	54.0	Yes	Yes	Fiber optic	No	...	Yes	
1	5161-XEUVX	Male	0.0	Yes	No	37.0	Yes	Yes	Fiber optic	No	...	Yes	
2	0336-PIKEI	Male	1.0	Yes	No	72.0	Yes	No	DSL	Yes	...	Yes	
3	3345-PBBFH	Male	0.0	Yes	No	8.0	Yes	No	DSL	No	...	No	
4	5067-XJQFU	Male	1.0	Yes	Yes	66.0	Yes	Yes	Fiber optic	No	...	Yes	
...	...	...	...	...	...	...	...	...	...	...	...	...	
229256	8587-XYZSFcsas	NaN	0.0	No	No	67.0	Yes	No	DSL	No	...	NaN	
229329	2654-VBVPB	Female	0.0	No	No	1.0	Yes	No	No	No internet service	...	No internet service	N
229614	5043-TRZWM	Female	0.0	No	No	1.0	Yes	No	Fiber optic	No	...	Yes	
229792	5256-SKJGOcsas	NaN	NaN	Yes	Yes	64.0	NaN	NaN	NaN	No	...	NaN	
229917	2609-IAICY	Female	0.0	No	No	1.0	Yes	Yes	Fiber optic	No	...	No	

7838 rows × 21 columns

2. حذف سطر هایی که مقادیر خالی زیادی دارند

سطرهای بسیار زیادی وجود دارند که مقادیر خالی در ان وجود دارد تابع ایی به نام DeleteRowMoreMissingValue نوشته شده این تابع سطر هایی را که بیشتر از 7 مقدار خالی باشد یعنی چیزی حدود 35 درصد مقادیر خالی باشد را حذف میکند.بعد از انجام این مرحله تعداد سطر ها از 7838 به 7613 کاهش یافت.

3.پرکردن مقادیر خالی:

چون داده ها هم numerical وcategorical هستند پر کردن مقادیر خالی به دو صورت است است در داده های عددی تابعی به نام fillMisingValueWithMean نوشته شده که مقادیر خالی را با میانگین ان ستون پر میکند و همچنین داده های پرت و منفی را نیز را حذف و بجایش میانگین را میگذارد.در داده های categorical تابعی به نام fillMissingValueWithMode نوشته شده که مقادیر خالی را با مد mode ان ستون پر میکند. تصویر زیر دو تابع بالا را نشان میدهد.

In [6]:
```
def fillMisingValueWithMean(df):
    deep_copy = df.copy()
    deep_copy = deep_copy.drop(columns=['Label'])
    column_names = list(deep_copy.columns)
    for i in column_names:
        if deep_copy[i].dtypes=="float64" and  not(i == "SeniorCitizen"):
            print(i)
            deep_copy[i] = deep_copy[i].fillna(deep_copy[i].mean())
            data=list(deep_copy[i])
            for j in range(len(data)):
                if data[j]<0:
                    data[j]=df[i].mean()
    deep_copy[i]=data
    deep_copy.insert(20, "Label",df["Label"])
    return deep_copy

```

In [7]:
```
def fillMissingValueWithMode(df):
    deep_copy = df.copy(deep=True)
    column_names = list(df.columns)
    for i in column_names:
        if deep_copy[i].dtypes=="O"or i=="SeniorCitizen":
            most_frequent_category = deep_copy[i].mode()[0]
            deep_copy[i].fillna(most_frequent_category, inplace=True)
    return deep_copy
```

4.تبدیل داده های categorical به numerical

بعد از انکه داده های خالی در ستون های categorical پر کردیم برای بهتر شدن دقت مدل وهمچنین تحلیل و بررسی در EDA داده ها را به صورت عددی تبدیل میکنیم برای اینکار تابع ایی به نام modifyCatToNum نوشته شده است.

5.نرمالسازی

در نهایت داده های عددی خود را نرمالسازی میکنیم به طوری که اعداد در بازه [1,0] قرار بگیرند برای اینکار تابع ایی به نام normalization نوشته شده است.

در نهایت تمام این توابع در یک تابع مادر به نام preProcessing قرار میگیرند دیتا ست اولیه به عنوان ورودی به مرحله 1 داده میشود و خروجی ان ورودی مرحله بعدی است تا در نهایت دیتا پیش پردازش شده ایجاد شود خروجی دیتا ست پیش پردازش شده به صورت زیر است:

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	...	DeviceProtection	Te
0	5331-RGMTT	1	1.0	1	0	0.973134	1	2	1	0	...	2	
1	5161-XEUVX	1	0.0	1	0	0.947761	1	2	1	0	...	2	
2	0336-PIKEI	1	1.0	1	0	1.000000	1	0	0	2	...	2	
3	3345-PBBFH	1	0.0	1	0	0.904478	1	0	0	0	...	0	
4	5067-XJQFU	1	1.0	1	1	0.991045	1	2	1	0	...	2	
...	...	...	...	...	...	...	...	...	...	...	...	...	
229256	8587-XYZSFcsas	0	0.0	0	0	0.992537	1	0	0	0	...	0	
229329	2654-VBVPB	0	0.0	0	0	0.894030	1	0	2	1	...	1	
229614	5043-TRZWM	0	0.0	0	0	0.894030	1	0	1	0	...	2	
229792	5256-SKJGOcsas	0	0.0	1	1	0.988060	1	0	1	0	...	0	
229917	2609-IAICY	0	0.0	0	0	0.894030	1	2	1	0	...	0	

7613 rows × 21 columns

EDA

بعد از انجام پیش پردازش، میتوانیم به تحلیل EDA بپردازیم. در این مرحله، به طور مقدماتی داده ها را بررسی میکنیم و آمارها و الگوهای مختلفی را بررسی کنید، مانند میانگین، واریانس، توزیعها و نمودارهای مختلف در این بخش ثبت و انجام شدم و شامل مراحل زیر است.

1.بازبینی داده ها

در این مرحله سه کار اساسی انجام میشود نمایش تعداد ردیفها و ستونها در دیتاست و نمایش خالصه آماری از دادهها مانند میانگین، میانه، حداکثر و حداقل و در نهایت مایش نمونهای از ردیفهای دیتاست انجام و در خروجی نمایش داده میشود. تابعی به نام DataReview نوشته شده است.تصویر زیر خروجی تابع اطاعات اماری هر ستون از دیتا ست را نشان میدهد.

```
In [41]:  1 result=dataReview()
          2 result
```

```
Name: Partner, dtype: object
count     229748
unique         2
top           No
freq      147990
Name: Dependents, dtype: object
count    229765.000000
mean         49.435018
std          36.632996
min        -598.000000
25%          37.000000
50%          56.000000
75%          68.000000
max          72.000000
Name: tenure, dtype: float64
count     229721
unique         2
top          Yes
freq      207904
```

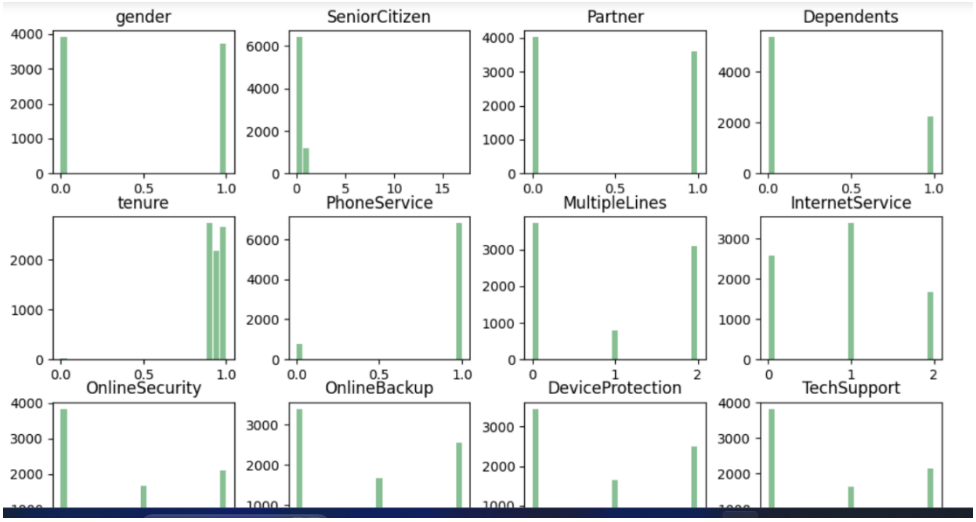
2. نمایش نمودارهای توصیفی

در این قسمت تحلیل داده از طریق نمودار میپردازیم نمودار های استفاده شده در این پروژه عبارت اند از histogeram,boxplot,scatterplot,pairplot هستند که به دقت به انها میپردازیم:

Histogeram.2-1

این روش یک روش راحت برای تجسم توزیع داده ها در دیتا ست است هیستوگرام برای هر مقدار منحصر به فرد در ستون مشخص شده ایجاد می کند و این امکان می دهد توزیع ها را در گروه های مختلف مقایسه کنیم.

تصویر زیر هیستوگرام بعضی از ستون ها را نشان میدهد.



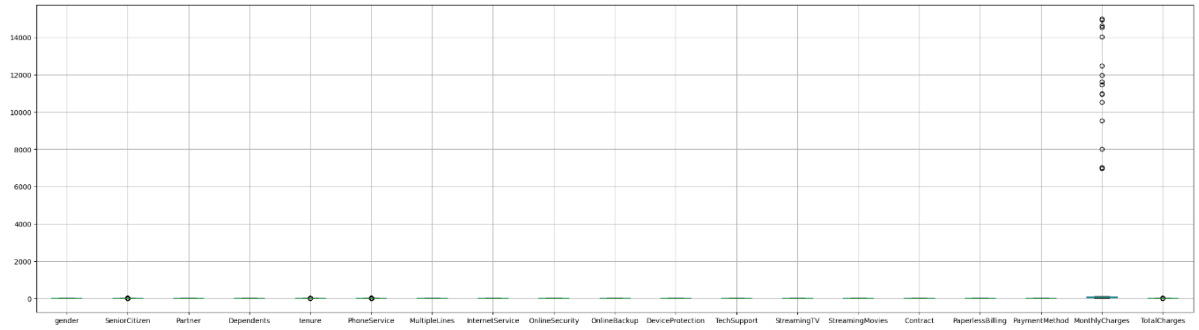
یکی از نتیجه هایی که از این نمودار میتوان گرفت این است که با توجه به نمودار هیستوگرام gender تعداد کاربران زن و مرد در این شرکت تقریبا با هم برابر است .

BoxPlot.2-2

(برای نمایش قدرت تمایل، پراکندگی و تشکیل از داده ها استفاده میشود . این روشی برای نمایش بصری توزیع داده ها، نشان دادن میانه، چارک ها و نقاط پرت بالقوه است . اما چون داده ها نرمال شده بودند روش خوبی برای تحلیل داده نبود.تصویر زیر به وضوح روشن است بعد پیش پردازش که داده های عددی نرمال شده اند و داده های کتگوریکال به عددی شده است. نمودار اطلاعات خوبی به ما نمیدهد.

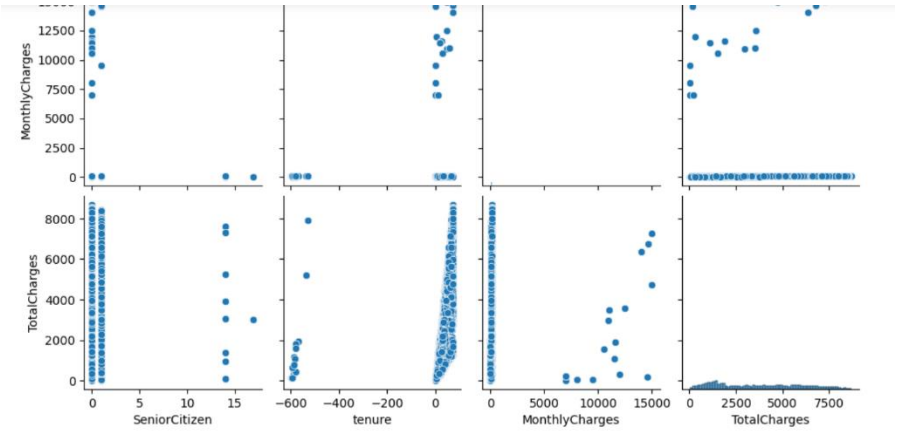
boxplot

```
plt.figure(figsize=(30, 8))
preProcessed_df.boxplot()
plt.show()
```



pairplot.2-3

نمودار زوجی که با استفاده از تابع Pairplot ایجاد شده است، ابزار قدرتمندی برای تجسم روابط بین متغیرهای متعدد در یک مجموعه داده است. نمودارهای زوجی به ویژه برای تجزیه و تحلیل داده های اکتشافی مفید هستند، زیرا راهی سریع برای تجسم روابط بین چندین متغیر در یک مجموعه داده ارائه می دهند.



3.تحلیل روابط

در این قسمت تحلیل داده از طریق روابط میپردازیم روابط های استفاده شده در این پروژه عبارت اند از correlation coefficient,boxplot,sensitivity analysis,pricipal component analysis هستند که به دقت به انها میپردازیم:

correlation coefficient-3-1

ضریب همبستگی یک معیار آماری است که قدرت و جهت رابطه خطی بین دو متغیر را کمی می کند.

رایج ترین نوع ضریب همبستگی، ضریب همبستگی پیرسون است که قدرت و جهت رابطه خطی بین دو متغیر را اندازه گیری می کند.در این پروژه نیز ضریب همبستگی پیرسون استفاده شده است.خروجی زیرتحلیل روابط ضریب همبستگی پیرسون است.

Correlation coefficient

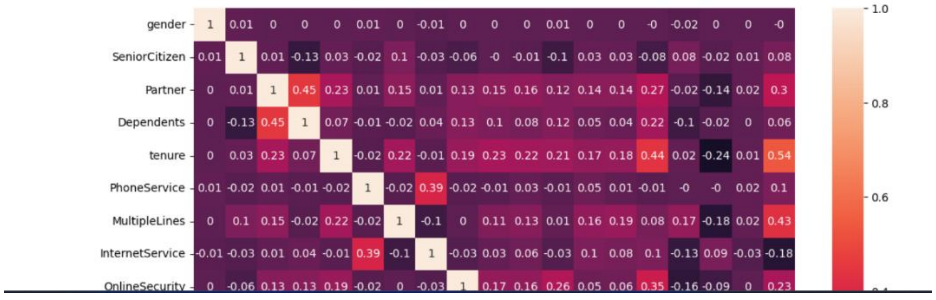
In [16]:
```
correlation_matrix = preProcessed_df.corr(method='pearson')
correlation_matrix
```

Out[16]:

	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	Device
gender	1.000000	0.008364	0.000886	0.000904	0.000788	0.007001	0.003252	-0.008915	0.002055	0.000267	
SeniorCitizen	0.008364	1.000000	0.011637	-0.130146	0.028128	-0.021352	0.098879	-0.035000	-0.061420	-0.000601	
Partner	0.000886	0.011637	1.000000	0.450618	0.227990	0.014038	0.146906	0.008730	0.131909	0.152987	
Dependents	0.000904	-0.130146	0.450618	1.000000	0.073794	-0.010872	-0.020522	0.039456	0.132428	0.095693	
tenure	0.000788	0.028128	0.227990	0.073794	1.000000	-0.017065	0.215918	-0.011631	0.194229	0.231343	
PhoneService	0.007001	-0.021352	0.014038	-0.010872	-0.017065	1.000000	-0.015300	0.387943	-0.018611	-0.005905	
MultipleLines	0.003252	0.098879	0.146906	-0.020522	0.215918	-0.015300	1.000000	-0.099690	0.004804	0.113425	
InternetService	-0.008915	-0.035000	0.008730	0.039456	-0.011631	0.387943	-0.099690	1.000000	-0.026417	0.028167	
OnlineSecurity	0.002055	-0.061420	0.131909	0.132428	0.194229	-0.018611	0.004804	-0.026417	1.000000	0.174346	
OnlineBackup	0.000267	-0.000601	0.152987	0.095693	0.231343	-0.005905	0.113425	0.028167	0.174346	1.000000	
DeviceProtection	0.003289	-0.014931	0.160654	0.078620	0.221264	0.030841	0.126978	0.055034	0.162496	0.184393	
TechSupport	0.011880	-0.099947	0.122880	0.124591	0.208656	-0.009164	0.007819	-0.026483	0.264516	0.185234	
StreamingTV	0.004946	0.025562	0.137757	0.054474	0.172694	0.045405	0.164908	0.099407	0.047779	0.156801	
StreamingMovies	0.004195	0.026069	0.138962	0.040199	0.182960	0.013988	0.185366	0.081932	0.062815	0.150880	

heatmap

In [17]:
```
plt.figure(figsize=(12,10))
plt=sns.heatmap(correlation_matrix.round(2),annot=True)
```



sensitivity analysis-3-2:

تحلیل حساسیت روشی است که برای تعیین اینکه چگونه مقادیر مختلف یک متغیر مستقل بر یک متغیر وابسته خاص تحت مجموعه ای از مفروضات تأثیر می گذارد، استفاده می شود. این روشی برای درک

تأثیر تغییرات متغیرهای ورودی بر خروجی یک مدل یا سیستم است. تجزیه و تحلیل حساسیت می تواند به شناسایی متغیرهایی که بیشترین تأثیر را بر نتیجه دارند کمک کند و می تواند برای ارزیابی استحکام یک مدل یا سیستم در برابر تغییرات در ورودی های آن استفاده شود.

sensitivity analysis

In [24]:
```
def my_model(x_1, x_2):
    """
    Represents the model function
    """
    return x_1 ** x_2
```

In [36]:
```
from sensitivity import SensitivityAnalyzer

dict_={
    'x_1':preProcessed_df["gender"],
    'x_2':preProcessed_df["SeniorCitizen"]
}

```

In [ ]:
```
sa = SensitivityAnalyzer(dict_, my_model)
```

: PCA-3-3

تجزیه و تحلیل مؤلفه اصلی (PCA) یک روش آماری است که این تبدیل به گونه ای تعریف می شود که اولین جزء اصلی بیشترین واریانس ممکن را داشته باشد و هر جزء بعدی به نوبه خود بیشترین واریانس ممکن را تحت محدودیت متعامد بودن آن نسبت به مؤلفه های قبلی داشته باشد.PCA را روی داده های استاندارد شده خود اعمال کردیم و داده های خود را به 5 جزء اصلی کاهش دادیم.

Principal Component Analysis (PCA)

In [51]:
```
from sklearn.decomposition import PCA
def PCA_Handler(preProcessed_df):
    preProcessed_df=preProcessed_df.drop(columns=['customerID','Label'])
    pca = PCA(n_components=5)
    pca.fit(preProcessed_df)
    principalComponents = pca.transform(preProcessed_df)
    principal_df = pd.DataFrame(data=principalComponents, columns=['PCA%i' % i for i in range(1, pca.n_components +
    return principal_df
```

In [52]:
```
PCA_Handler(preProcessed_df)
```

Out[52]:

	PCA1	PCA2	PCA3	PCA4	PCA5
0	9.379339	0.696852	1.690760	1.775618	-0.679566
1	4.629201	0.698361	-0.963381	0.932107	0.680085
2	-15.270555	2.355453	-0.618626	-1.934113	-0.321573
3	-40.120622	-1.043357	1.353263	-1.644132	-0.299191
4	18.779388	2.038672	-0.107805	1.578866	0.580128
...	...	...	...	...	...
7608	-39.120835	-0.999201	-0.901237	-1.421336	-0.070064

نتیجه گیری

بعد از مرحله پیش پردازش و EDA باید دیتا ست خود را مورد ارزیابی قرار دهیم در این پروژه مدلی که استفاده شد logestic regression بود یک الگوریتم یادگیری ماشینی نظارت شده که عمدتاً برای مسائل طبقه بندی باینری استفاده می شود. در این دیتاست نیز طبقه بندی به صورت باینری بوده و ستون Label

دارای دو مقدار yes,no بوده است قبل از مرحله پیش پردازش دقت مدل 79 درصد بوده است بعد از مرحله پیش پردازش دقت مدل به 86 درصد افزایش یافته است. تصویر زیر خروجی مدل را نشان میدهد.

```
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
  n_iter_i = _check_optimize_result(

Cross-validation scores: [0.8645468  0.86593814 0.86396313 0.86555742 0.86352832]
Test set accuracy: 0.8622560784722947

Results as a table:
```

Out[32]:

	Target Label	Cross-Validation Scores	Test Accuracy
0	Label	[0.8645467984579264, 0.8659381431345836, 0.863...	0.862256

منابع

https://www.geeksforgeeks.org/what-is-exploratory-data-analysis/

https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.plot.box.html

https://www.geeksforgeeks.org/data-visualization-with-pairplot-seaborn-and-pandas/

https://pandas.pydata.org/docs/reference/