به‌نام خداوند بخشنده و مهربان

**تمرین چهارم:** لاجستیک رگرسیون
**درس:** تحلیل سیستم داده‌های حجیم

**استاد:** محمدعلی نعمت‌بخش
دستیاران: فاطمه ابراهیمی، پریسا لطیفی، امیر سرتیپی

**نام و نام‌خانوادگی:** زلفا شفرئی
**آدرس گیت:** https://github.com/zolfaShefreie/Spark_ML/

- **لطفا پاسخ تمارین حتما در سامانه‌ی کوئرا ارسال شود.**
- **لطفا پاسخ‌های خود را در خود سند سوال نوشته و همراه نوتبوک تمرین در کوئرا ارسال کنید.**
- **نام سند ارسالی HW-{homework number}-{Name Family}-{student number}**
- **تمامی فایل‌های مورد نیاز این تمرین در این لینک قابل دسترس است.**
- **خروجی از هر مرحله‌ی تمرین را در سند خود بارگذاری کنید.**

در این تمرین هدف کار با کتاب‌خانه‌ی pyspark و همچنین کتاب‌خانه‌ی یادگیری ماشین آن است.

برای این منظور دیتاستی در اختیار شما قرار گرفته است. اطلاعات کاربران شرکتی در اختیار شما قرار داده شده است. این شرکت اطلاعات چند ماه از کاربرانش را برچسب گذاری کرده است. این برچسب به معنای این است که آیا مشتری شرکت را ترک کرده و دیگر از خدمات آن استفاده نمی‌کند یا خیر. انتظار می‌رود با بررسی دقیق مجموعه‌ی داده و تحلیل داده‌گان آن در نهایت مدل پیشبینی کننده‌ای برای این شرکت طراحی کنید.

هر یک از موارد زیر را به دقت بررسی کنید و نتایج آن را در قالب اسکرین شات و تحیل خود در سند ذکر کنید.

- **قدم اول:** دیتاست داده شده را پیش پردازش کنید. مقادیر NA را مقدار دهی کنید تحلیل داده اکتشافی (EDA) را به خوبی انجام دهید. این ستونها براساس ماهیت خود میتواند تولید کننده ویژگیهای بیشتری باشند که ممکن است دقت مدل شما را باالاتر ببرند. در این مرحله همبستگی و ارتباط بین تمام ویژگی هایی که میتوانید استخراج کنید را بررسی کنید. (نمودارهای لازم برای تحلیل دادگان ترسیم شود.)
- **قدم دوم:** عملیات feature engineering را به خوبی برای داده‌گان خود انجام دهید و دلیل انتخاب هریک از ستون‌ها یا عدم انتخاب آن‌ها را به صورت منطقی بیان کنید. (با نمودار و تحلیل آن، با کمک EDA انجام شده)
- **قدم سوم:** الگوریتم Logestic Regression را بر روی داده‌های خود اعمال کنید.

- **قدم چهارم:** دقت مدل خود را ارزیابی کنید. ( در این مرحله شما باید مراحل آزمایش، تعداد دادگان ترین و تست، احتمال صحیح بودن یک برچسب که مدل پیشبینی کرده است، را تعیین کنید)
- نتایج مدل قبل و بعد از پیش‌پردازش را مقایسه کنید.

**نکات مهم**

- برای پیش‌پردازش دادگان و یادگیری ماشین فقط از کتابخانه پای‌اسپارک استفاده شود.(pandas مجاز نیست).
- نمودارهای ترسیمی حتما همراه با تحلیل در سند آورده شوند.
- کپی نکنید! از قبل تمام کدهای نوشته شده در اینترنت جمع‌اوری شده است کپی کردن شما مشخص می‌شود.

---

در ابتدا مجموعه داده دانلود می‌شود و نصب و تنظیمات اسپارک صورت می‌گیرد سپس با دستور read.csv مجموعه داده در یک دیتافریم لود می‌شود. پارامترهای header و inferSchema به ترتیب برای شناسایی سطر اول به‌عنوان نام ستون‌ها و گرفتن تایپ‌ها براساس فایل می‌باشد.

```
import requests

customer_url = 'https://raw.githubusercontent.com/zolfaShefreie/Spark_ML/main/data.csv'
customer_file_path = "data.csv"

def download_file(url: str, file_path):
  """
  download file and save on file path
  """
  file_content = requests.get(url).text
  file = open(file_path, 'w')
  file.write(file_content)
  file.close()

download_file(customer_url, customer_file_path)
```

```
!pip install pyspark

Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Collecting pyspark
  Downloading pyspark-3.2.1.tar.gz (281.4 MB)
     |████████████████████████████████| 281.4 MB 34 kB/s
Collecting py4j==0.10.9.3
  Downloading py4j-0.10.9.3-py2.py3-none-any.whl (198 kB)
     |████████████████████████████████| 198 kB 56.4 MB/s
Building wheels for collected packages: pyspark
  Building wheel for pyspark (setup.py) ... done
  Created wheel for pyspark: filename=pyspark-3.2.1-py2.py3-none-any.whl size=281853642 sha256=f225
  Stored in directory: /root/.cache/pip/wheels/9f/f5/07/7cd8017084dce4e93e84e92efd1e1d5334db05f2e83
Successfully built pyspark
Installing collected packages: py4j, pyspark
Successfully installed py4j-0.10.9.3 pyspark-3.2.1

import pyspark
from pyspark.sql import SparkSession

spark = SparkSession.builder.config("spark.driver.memory", "8g").appName('spark_ml').getOrCreate()
sc = spark.sparkContext
```

```
df = spark.read.csv(customer_file_path, inferSchema=True, header=True)

df.printSchema()
```

## تحلیل داده

تنها نتایج اصلی در سند آورده شده است اما کدها و نتایج به‌صورت کامل در گیت‌هاب قابل مشاهده می‌باشد.

### تحلیل اولیه

مجموعه داده شامل ۲۲۹۹۹۰ سطر است و با دستور summery می‌توان اطلاعات اولیه‌ی داده‌ها را یافت که در شکل زیر قسمتی از آن قابل مشاهده است. نتایج زیر برای داده‌های عددی از این دستور بدست می‌آید:

- تعدادهای کمتر از تعداد سطرهای مجموعه نشان‌دهنده‌ی وجود نال در ستون‌ها است.
- seniorCitizen دارای مقدار ماکسیسم ۱۷ می‌باشد اما دیگر اطلاعات آن بیشتر به باینری بودن مقادیر اشاره دارد که می‌تواند نشان دهنده‌ی اوت لایر باشد.
- Tenure طبق توضیحات مجموعه داده به تعداد ماه‌هایی که شخص از سرویس‌های شرکت استفاده می‌کرده است اشاره دارد. مقدار مینیمم که عدد منفی است غیرقابل قبول می‌باشد.
- مقادیر اوت لایر در ستون MonthlyCharges قابل مشاهده است.

```
df.summary().show()
```

```
+-------+----------+------+------------------+-------+----------+------------------+------------+------------+---------------+--------------+----------+
|summary|customerID|gender|     SeniorCitizen|Partner|Dependents|            tenure|PhoneService|MultipleLines|InternetService|OnlineSecurity|OnlineB|
+-------+----------+------+------------------+-------+----------+------------------+------------+------------+---------------+--------------+-------+
|  count|    229737|229755|            229724| 229765|    229748|            229765|      229721|      229727|         229760|        229760|      2|
|   mean|      null|  null|0.21505371663387368|   null|      null|49.435018388353 32|        null|        null|           null|          null|       |
| stddev|      null|  null| 0.8971977539088777|   null|      null|36.632996142938 01|        null|        null|           null|          null|       |
|    min|0002-ORFBO|Female|               0.0|     No|        No|            -598.0|          No|          No|            DSL|            No|       |
|    25%|      null|  null|               0.0|   null|      null|              37.0|        null|        null|           null|          null|       |
|    50%|      null|  null|               0.0|   null|      null|              56.0|        null|        null|           null|          null|       |
|    75%|      null|  null|               0.0|   null|      null|              68.0|        null|        null|           null|          null|       |
|    max|9995-HOTOH|  Male|              17.0|    Yes|       Yes|              72.0|         Yes|         Yes|             No|           Yes|       |
+-------+----------+------+------------------+-------+----------+------------------+------------+------------+---------------+--------------+-------+
```

با توجه به احتمال وجود مقادیر نال تعداد آن را در هر ستون بدست می‌آوریم. که در تمامی ستون مقادیر نال موجود دارد.

```
df.select([count(when(isnan(c) | col(c).isNull(), c)).alias(c) for c in df.columns]).show()
```

```
+----------+------+-------------+-------+----------+------+------------+-------------+---------------+--------------+------------+----------------+--
|customerID|gender|SeniorCitizen|Partner|Dependents|tenure|PhoneService|MultipleLines|InternetService|OnlineSecurity|OnlineBackup|DeviceProtection|Te
+----------+------+-------------+-------+----------+------+------------+-------------+---------------+--------------+------------+----------------+--
|       253|   235|          266|    225|       242|   225|         269|          263|            230|           230|         243|             254|
+----------+------+-------------+-------+----------+------+------------+-------------+---------------+--------------+------------+----------------+--
```

وجود نال در ستون شناسه عجیب به نظر می‌رسد. داده‌های شناسه باید یونیک و غیر نال باشند زیرا هویت مشتری را نشان می‌دهند که بررسی‌ها دقیق‌تر در ادامه توضیح داده می‌شود. در اینجا امکان وجود سطر تکراری در مجموعه داده بررسی می‌شود. طبق نتایج بعضی از سطرها بارها تکرار شده است و با حذف مقادیر تکراری تعداد سطرهای مجموعه داده به ۷۸۳۸ می‌رسد که نشان‌دهنده‌ی حجم زیادی از سطرهای تکراری می‌باشد.

وجود سطر تکراری به‌شرطی که توازن بین داده‌ها تغییر نکند ضرری ندارد و در شرایطی برای افزایش یک دسته‌ی خاص از برچسب که تعداد سطرهای آن بسیار کم است، این کار انجام می‌شود تا مدل آن برچسب را بهتر یاد بگیرد.

طبق نتایج توازن بین برچسب‌ها تغییر کرده است به‌همین علت داده‌های تکراری حذف خواهند شد تا نتایج تحلیل‌ها به علت وجود تعداد زیادی از سطرهای تکراری تغییر نکند.

```
df.select(df.columns).groupBy(df.columns).count().filter(col('count')>1).agg({'count': 'sum'}).show()
```

```
+----------+
|sum(count)|
+----------+
|    228579|
+----------+
```

```
df.dropDuplicates().count()
```

```
7838
```

```
df.select('Label').groupBy('Label').count().show()
```

```
+-----+------+
|Label| count|
+-----+------+
| null|   208|
|   No|195878|
|  Yes| 33904|
+-----+------+
```

```
df.dropDuplicates().select('Label').groupBy('Label').count().show()
```

```
+-----+-----+
|Label|count|
+-----+-----+
| null|  208|
|   No| 5681|
|  Yes| 1949|
+-----+-----+
```

```
195878/5681, 33904/1949, 195878/33904, 5681/1949
```

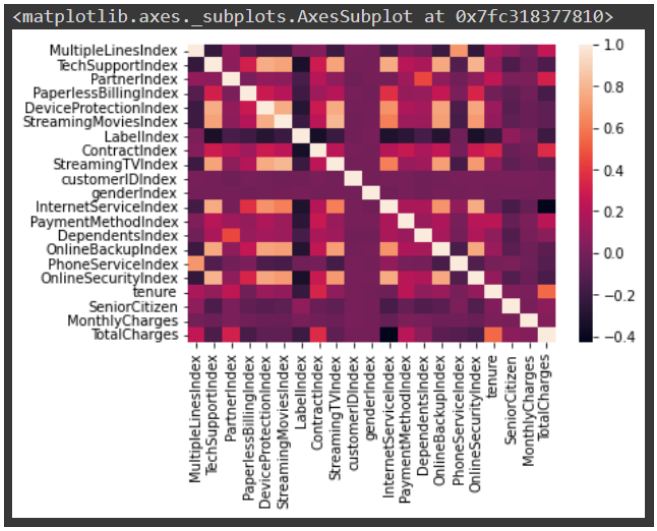
```
(34.47949304699877, 17.395587480759364, 5.777430391694195, 2.914828116983068)
```

برای بررسی اولیه‌ی همبستگی‌ها می‌توان از یک نمودار کمک گرفت. محاسبه‌ی همبستگی فقط برای داده‌های عددی امکان‌پذیر است پس با stringIndexer تمام داده‌ها را عددی می‌کنیم و سپس با استفاده از vectorAssembler آن‌ها را تبدیل به یک وکتور کرده و با استفاده از correlation.core() اسپارک یک ماتریس همبستگی بدست آورده می‌شود. این ماتریس با پکیج seaborn به‌راحتی قابل نمایش است.

نتایج همبستگی:

- مواردی که مربوط به داشتن اینترنت است یه رابطه‌ی مستقیمی باهم دارند که بهتر است بررسی شود.
- رابطه‌ی بین چند خط داشتن و سرویس موبایل نیز رابطه‌ی نسبتا مستقیمی دارند که باید بررسی شود.

- در نگاه اول برچسب رابطه‌ی مستقیمی با بیشتر از داده‌ها ندارد که بهتر است این وظیفه برعهده‌ی مدل گذاشته شود




در ادامه به‌بررسی روابط بین ستون‌ها پرداخته می‌شود که می‌تواند در پیش پردازش استفاده شود. تنها نتایج مهم به‌صورت موردی در سند ذکر می‌شوند.

- مقادیر ۱۷ و ۱۴ برای seniorCitizen مقادیر اوت لایر هستند که باید اصلاح شوند.
- کسانی که phoneService آن‌ها No است مقدار InternetService آن‌ها برابر با DSL می‌باشد. طبق نتایج می‌توان گفت کسانی که phoneService شان برابر با No است InternetService شان برابر با No نیست و کسانی که InternetService شان برابر با No است، phoneService برابر Yes می‌باشد.

| PhoneService | InternetService | count |
|---|---|---|
| No | DSL | 768 |
| null | null | 73 |
| Yes | No | 1642 |
| null | Fiber optic | 78 |
| Yes | DSL | 1780 |
| null | No | 58 |
| No | null | 44 |
| Yes | null | 113 |
| Yes | Fiber optic | 3222 |
| null | DSL | 60 |

- بیشتر افرادی که شرکت را ترک کردند (از سرویس‌های شرکت استفاده نمی‌کنند) مقدار InternetService برابر Fiber optic می‌باشد اما افرادی که شرکت را ترک نکردند این مقدار برابر با DSL می‌باشد.

| InternetService | Label | count |
|---|---|---|
| DSL | Yes | 477 |
| Fiber optic | Yes | 1345 |
| null | null | 66 |
| No | No | 1538 |
| Fiber optic | null | 44 |
| No | null | 49 |
| null | No | 150 |
| null | Yes | 14 |
| Fiber optic | No | 1911 |
| No | Yes | 113 |
| DSL | null | 49 |
| DSL | No | 2082 |

- هنگامی که phoneService برابر با No باشد MultipleLines باید برابر با No phone service باشد.
- بیشتر افرادی که شرکت را ترک کردند چندخطه بودند اما بیشتر افرادی که ترک نکردند چند خطه نبودند.

| MultipleLines | Label | count |
|---|---|---|
| No phone service | null | 32 |
| No phone service | No | 596 |
| Yes | Yes | 892 |
| null | null | 73 |
| No | No | 2702 |
| Yes | No | 2216 |
| No phone service | Yes | 184 |
| No | null | 56 |
| null | No | 167 |
| Yes | null | 47 |
| null | Yes | 23 |
| No | Yes | 850 |

- هنگامی که internetService برابر با No باشد ستون‌های OnlineSecurity، OnlineBackup، DeviceProtection، TechSupport، StreamingTV و StreamingMovies باید برابر با No internet service باشند.
- بیشتر افرادی که مقدار OnlineSecurity برابر با No بود شرکت را رها کردند و بیشتر افرادی که OnlineSecurity داشتند شرکت را رها نکردند که می‌توان برای پر کردن مقادیر برچسب از آن استفاده کرد.

| OnlineSecurity | Label | count |
|---|---|---|
| No internet service | No | 1542 |
| Yes | Yes | 310 |
| null | null | 65 |
| No | No | 2185 |
| Yes | No | 1811 |
| null | No | 143 |
| No | null | 60 |
| Yes | null | 35 |
| null | Yes | 22 |
| No | Yes | 1504 |
| No internet service | Yes | 113 |
| No internet service | null | 48 |

- بیشتر افرادی که شرکت را رها کردند OnlineBackup نداشتند و بیشتر افرادی که رها نکردند OnlineBackup داشتند که می‌توان از این رابطه برای پر کردن مقادیر OnlineBackup درصورت مشخص بودن برچسب استفاده کرد.

| OnlineBackup | Label | count |
|---|---|---|
| No internet service | No | 1527 |
| Yes | Yes | 534 |
| null | null | 57 |
| No | No | 1975 |
| Yes | No | 2021 |
| null | No | 158 |
| No | null | 60 |
| null | Yes | 28 |
| Yes | null | 32 |
| No | Yes | 1274 |
| No internet service | Yes | 113 |
| No internet service | null | 59 |

- بیشتر افرادی که TechSupport دارند DSL دارند و بیشتر افرادی که TechSupport ندارند Fiber optic دارند.

| InternetService | TechSupport | count |
|---|---|---|
| DSL | Yes | 1235 |
| Fiber optic | Yes | 900 |
| null | null | 91 |
| Fiber optic | null | 52 |
| No | No internet service | 1643 |
| No | null | 57 |
| null | No | 65 |
| null | Yes | 36 |
| Fiber optic | No | 2348 |
| DSL | null | 64 |
| DSL | No | 1309 |
| null | No internet service | 38 |

- بیشتر افرادی که InternetService آن‌ها برابر با DSL است StreamingTV ندارند و بیشتر افرادی مقدار برابر با Fiber optic است StreamingTV دارند.

| InternetService | StreamingTV | count |
|---|---|---|
| DSL | Yes | 1022 |
| Fiber optic | Yes | 1825 |
| null | null | 71 |
| Fiber optic | null | 66 |
| No | No internet service | 1653 |
| null | No | 55 |
| No | null | 47 |
| null | Yes | 57 |
| Fiber optic | No | 1409 |
| DSL | null | 65 |
| DSL | No | 1521 |
| null | No internet service | 47 |

- بیشتر افرادی که InternetService آن‌ها برابر با DSL است StreamingMovies ندارند و بیشتر افرادی مقدار برابر با Fiber optic است StreamingMovies دارند.

| InternetService | StreamingMovies | count |
|---|---|---|
| DSL | Yes | 1066 |
| Fiber optic | Yes | 1792 |
| null | null | 65 |
| Fiber optic | null | 55 |
| No | No internet service | 1649 |
| No | null | 51 |
| null | No | 62 |
| null | Yes | 50 |
| Fiber optic | No | 1453 |
| DSL | null | 48 |
| DSL | No | 1494 |
| null | No internet service | 53 |

- بیشتر کسانی که ترک کردند StreamingMovies نداشتند اما بیشتر کسانی که ترک نکردند StreamingMovies را داشتند.

| StreamingMovies | Label | count |
|---|---|---|
| No internet service | No | 1533 |
| Yes | Yes | 834 |
| null | null | 54 |
| No | No | 1975 |
| Yes | No | 2028 |
| No | null | 52 |
| null | No | 145 |
| null | Yes | 20 |
| Yes | null | 46 |
| No | Yes | 982 |
| No internet service | Yes | 113 |
| No internet service | null | 56 |

- بین StreamingMovies و StreamingTV رابطه‌ی مستقیم است اما یک پنجم از داده‌ها از این قاعده پیروی نمی‌کنند. پس نمی‌توان آن را به‌عنوان قانون صددرصدی بیان کرد تا یکی از ستون‌ها حذف شود.

| StreamingTV | StreamingMovies | count |
|---|---|---|
| Yes | Yes | 2042 |
| null | null | 72 |
| No | No | 2140 |
| Yes | No | 806 |
| No internet service | No internet service | 1655 |
| No | null | 46 |
| null | No | 63 |
| Yes | null | 56 |
| null | Yes | 67 |
| No | Yes | 799 |
| null | No internet service | 47 |
| No internet service | null | 45 |

- بیشتر افرادی که OnlineSecurity دارند OnlineBackup دارند و بیشتر کسانی که OnlineSecurity ندارند OnlineBackup ندارند.

| OnlineSecurity | OnlineBackup | count |
|---|---|---|
| Yes | Yes | 1160 |
| null | null | 61 |
| No | No | 2296 |
| Yes | No | 942 |
| No internet service | No internet service | 1649 |
| No | null | 74 |
| null | No | 71 |
| null | Yes | 48 |
| Yes | null | 54 |
| No | Yes | 1379 |
| null | No internet service | 50 |
| No internet service | null | 54 |

- بیشتر افرادی که OnlineSecurity دارند DeviceProtection دارند و بیشتر کسانی که OnlineSecurity ندارند DeviceProtection ندارند.

| OnlineSecurity | DeviceProtection | count |
|---|---|---|
| Yes | Yes | 1120 |
| null | null | 72 |
| No | No | 2298 |
| Yes | No | 988 |
| No internet service | No internet service | 1643 |
| No | null | 74 |
| null | No | 76 |
| null | Yes | 32 |
| Yes | null | 48 |
| No | Yes | 1377 |
| No internet service | null | 60 |
| null | No internet service | 50 |

- DeviceProtection و OnlineBackup رابطه‌ی مستقیم دارند اما دو پنجم از داده‌ها از این قاعده پیروی نمی‌کنند. پس نمی‌توان از این حالت برای یک قانون صد در صدی استفاده کرد و یکی از ستون‌ها را حذف کرد.

| DeviceProtection | OnlineBackup | count |
|---|---|---|
| Yes | Yes | 1355 |
| null | null | 73 |
| No | No | 2092 |
| Yes | No | 1141 |
| No internet service | No internet service | 1641 |
| null | No | 76 |
| No | null | 85 |
| null | Yes | 47 |
| Yes | null | 33 |
| No | Yes | 1185 |
| null | No internet service | 58 |
| No internet service | null | 52 |

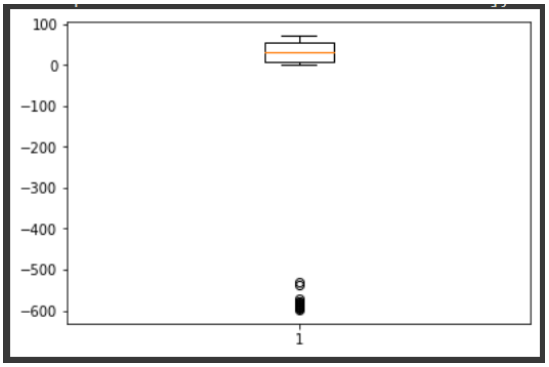
- بیشتر کسانی که پارتنر دارند خویشاوند هم دارند و بیشتر کسانی که خویشاوند ندارند پارتنر هم ندارند.

| Partner | Dependents | count |
|---|---|---|
| Yes | Yes | 1874 |
| null | null | 64 |
| No | No | 3489 |
| Yes | No | 1704 |
| No | null | 101 |
| null | No | 99 |
| null | Yes | 62 |
| Yes | null | 77 |
| No | Yes | 368 |

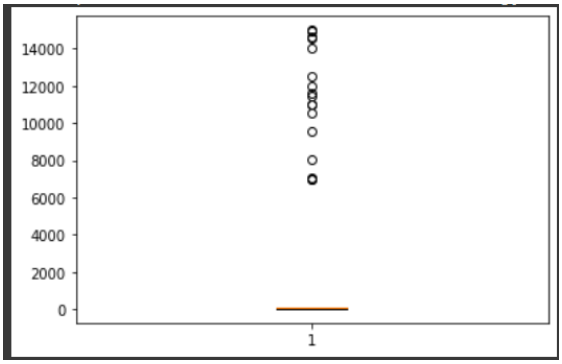
- بیشتر افرادی که شرکت را ترک کردند پارتنر ندارند اما بیشتر افرادی که شرکت را ترک نکردند پارتنر دارند.

| Partner | Label | count |
|---|---|---|
| Yes | Yes | 702 |
| null | null | 60 |
| No | No | 2658 |
| Yes | No | 2887 |
| No | null | 82 |
| null | No | 136 |
| Yes | null | 66 |
| null | Yes | 29 |
| No | Yes | 1218 |

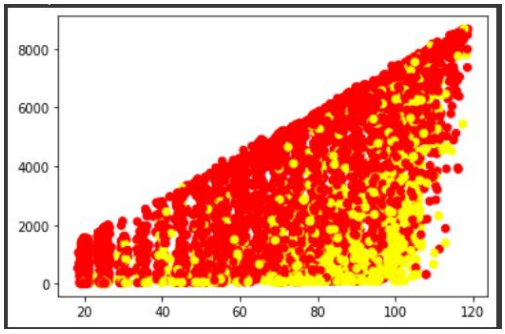
- مقادیر منفی برای tenure اوت لایر به حساب می‌آیند و با توجه به معنی این ستون غیر قابل قبول می‌باشند.




- برخی از مقادیر در ستون MonthlyCharges اوت لایر هستند زیرا از مقادیر TotalCharges بزرگتر هستند.



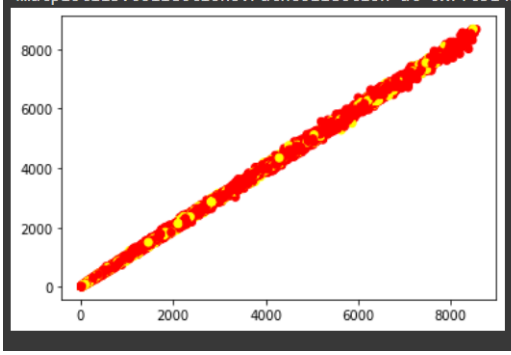

- MonthlyCharges و TotalCharges باهم رابطه‌ی مستقیم دارند.



- طبق شکل زیر قانون TotalCharges = MonthlyCharges * tenure رد نمی‌شود ولی این قانون به‌صورت دقیق صدق نمی‌کند که می‌تواند میزان تغییر در MonthlyCharges برای افراد را در ماه‌ها نشان دهد. از این قانون می‌توان برای پر کردن مقادیر نال استفاده کرد.

```
plt.scatter([x*y for x, y in zip(MonthlyCharges_list, tenure_list)], TotalCharges_list, c=[colors[x] for x in label_list])
```

```
<matplotlib.collections.PathCollection at 0x7fc3142e6d10>
```




- بیشتر افرادی که PaymentMethod آن‌ها برابر با Mailed check می‌باشد، مقدار PaperlessBilling آن‌ها برابر با No می‌باشد. در باقی حالات بیشتر برابر با Yes است.

| PaymentMethod | PaperlessBilling | count |
|---|---|---|
| Credit card (auto... | null | 31 |
| null | null | 84 |
| Credit card (auto... | No | 642 |
| Mailed check | null | 23 |
| Bank transfer (au... | No | 702 |
| Mailed check | Yes | 663 |
| Bank transfer (au... | null | 48 |
| Credit card (auto... | Yes | 960 |
| null | No | 40 |
| null | Yes | 122 |
| Electronic check | No | 625 |
| Electronic check | Yes | 1885 |
| Bank transfer (au... | Yes | 927 |
| Electronic check | null | 71 |
| Mailed check | No | 1015 |

- بیشتر هنگامی که PaymentMethod برابر با Mailed check و Electronic check است، Contract برابر با Month-to-month و هنگامی که دو حالت دیگر باشد بیشتر برابر با Two year می‌باشد.

همچنین هنگامی که Contract برابر با One year باشد بیشتر مقدار PaymentMethod برابر با Credit card و هنگامی که برابر با Two year باشد متد برابر با Bank transfer و هنگامی که برابر با Month-to-month بیشتر Electronic check است.

| PaymentMethod | Contract | count |
|---|---|---|
| Credit card (auto... | null | 32 |
| Mailed check | One year | 336 |
| null | Month-to-month | 59 |
| Mailed check | Two year | 439 |
| null | null | 70 |
| null | One year | 24 |
| Mailed check | null | 23 |
| Bank transfer (au... | One year | 393 |
| Bank transfer (au... | Month-to-month | 597 |
| null | Two year | 93 |
| Bank transfer (au... | null | 44 |
| Bank transfer (au... | Two year | 643 |
| Electronic check | Two year | 214 |
| Credit card (auto... | One year | 429 |
| Credit card (auto... | Two year | 602 |
| Electronic check | null | 61 |
| Mailed check | Month-to-month | 903 |
| Electronic check | Month-to-month | 1923 |
| Credit card (auto... | Month-to-month | 570 |
| Electronic check | One year | 383 |

- مقادیر customerID دارای موارد تکراری است که این مورد باید اصلاح شود. طبق نتیجه‌ی یکی از شناسه‌ها که دارای تکرار است می‌توان این سطرها باهم ادقام شوند تا مقادیر نال‌های آن‌ها از بین برود و سطر کامل‌تری در اختیار بگذارد. این تکرار برای موارد غیر نال در نظر گرفته می‌شود و نال‌ها به‌عنوان فرد مستقل در نظر گرفته می‌شود.

| customerID | gender | SeniorCitizen | Partner | Dependents | tenure | PhoneService | MultipleLines | InternetService | OnlineSecurity | OnlineBackup | DeviceProtection | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2928-HLDBAcsas | null | 0.0 | No | null | 6.0 | null | No | No | No internet service | No internet service | null | No |
| 2928-HLDBAcsas | Female | null | null | No | 6.0 | Yes | No | No | No internet service | No internet service | No internet service | No |
| 2928-HLDBAcsas | Female | null | No | null | 6.0 | null | No | No | No internet service | No internet service | No internet service | No |
| 2928-HLDBAcsas | null | 0.0 | null | No | 6.0 | Yes | No | No | No internet service | No internet service | null | |

## پیش‌پردازش داده

موارد زیر به ترتیب برای پیش پردازش انجام شده‌اند.

- ادغام سطرها با آیدی یکسان
- پر کردن مقادیر نال gender با استفاده از مد
- نال کردن مقادیر اوت لایر seniorCitizen و سپس پر کردن مقادیر نال با استفاده از مد
- پر کردن مقادیر نال Partner با استفاده از رابطه با ستون خویشاوند و لیبل (اگر باز دارای مقادیر نال بود با مد)
- پر کردن مقادیر نال Dependents با استفاده از مد
- پر کردن مقادیر نال phoneService با استفاده از رابطه بین چند خطه و internetService و در آخر با مد
- پر کردن مقادیر نال MultipleLines با استفاده از رابطه بین phoneService و برچسب و در آخر مد. نگاشت No phone Service با No به دلیل وجود این اطلاعات در ستون phoneService
- پر کردن مقادیر نال internetService با ستون‌های OnlineSecurity، OnlineBackup، DeviceProtection، TechSupport، StreamingTV و StreamingMovies و رابطه‌ی بین با phoneService و Label و در آخر با نال

- پر کردن مقادیر نال TechSupport با استفاده از internetService و مد و در آخر نگاشت مقادیر No internet service با No به دلیل وجود این اطلاعات در internetService
- پر کردن مقادیر نال OnlineSecurity با استفاده از internetService و مد و در آخر نگاشت مقادیر No internet service با No به دلیل وجود این اطلاعات در internetService
- پر کردن مقادیر نال OnlineBackup با استفاده از روابط برچسب و internetService و OnlineSecurity و در آخر نگاشت مقادیر No internet service با No به دلیل وجود این اطلاعات در internetService
- پر کردن مقادیر نال DeviceProtection با استفاده از روابط OnlineSecurity و internetService و در آخر نگاشت مقادیر No internet service با No به دلیل وجود این اطلاعات در internetService
- پر کردن مقادیر نال StreamingTV با استفاده از internetService و در آخر نگاشت مقادیر No internet service با No به دلیل وجود این اطلاعات در internetService
- پر کردن مقادیر نال StreamingMovies با استفاده از internetService و در آخر نگاشت مقادیر No internet service با No به دلیل وجود این اطلاعات در internetService
- پر کردن مقادیر نال Contract با استفاده از روابط PaymentMethod و در آخر با مد
- پر کردن مقادیر نال PaymentMethod با استفاده از Contract
- پر کردن مقادیر نال PaperlessBilling با استفاده از PaymentMethod و مد
- جایگزین کردن مقادیر اوت لایر tenure و MonthlyCharges با نال. پر کردن مقادیر tenure وMonthlyCharges و TotalCharges با استفاده از قانون بدست آمده (اگر دو تای آن‌ها نال باشد با استفاده از میانگین یکی از آنها پر می‌شود)
- حذف سطرهایی با مقادیر نال در Label (زیرا پر کردن دستی آن می‌تواند خطا در پیش‌بینی مدل را افزایش دهد)
- حذف ستون customerID زیرا این ستون دارای مقادیر یونیک می‌باشد و آموزش مدل را سخت‌تر می‌کند.
- تبدیل ستون Contract به تعداد ماه‌ها (نگاشت هر کدام از دسته‌ها به تعداد ماه‌های قرار داد. زیرا این دسته‌ها به صورت ترتیبی هستند و مقادیر آن‌ها بهتر است با تعداد ماه‌ها نگاشت شود تا مدل تفاوت بین آن‌ها را حس کند)
- تبدیل تمامی استرینگ‌های باقی مانده به عدد با استفاده از stringIndexer و تبدیل به one hot دو ستون PaymentMethod و internetservice با استفاده از oneHotEncoder (به‌دلیل اینکه به مدل برای هر لیبل عددی که استرینگ داده شده است ارزش گذاری نکند)
- نرمالیز کردن مقادیر عددی Contract، tenure، MonthlyCharges و TotalCharges با استفاده از vectorAssembler و MinMaxScaler
- تبدیل به وکتور تمامی ستون‌ها به جز Label با استفاده از vectorAssembler

## آموزش مدل

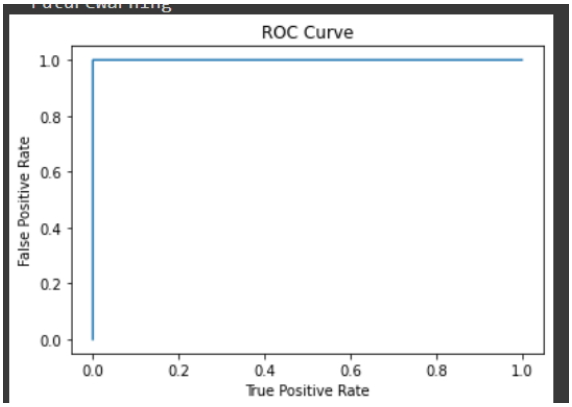
پس از پیش‌پردازش داده‌ها با نسبت ۲ و ۸ به داده‌های تست و داده‌های آموزش تقسیم می‌شوند.

```
train, test = normalized_df.randomSplit([0.8, 0.2], 42)
```

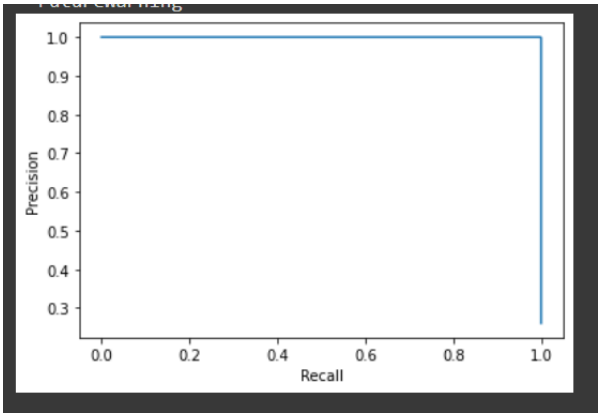
مدل لاجستیک رگرسیون با استفاده از داده‌های آموزش، آموزش می‌بینند.

```
logistic_regression = LogisticRegression(featuresCol='features', labelCol='LabelIndex', maxIter=30)
lr_trained_model = logistic_regression.fit(train)
```

نتایج به‌صورت زیر می‌باشد.



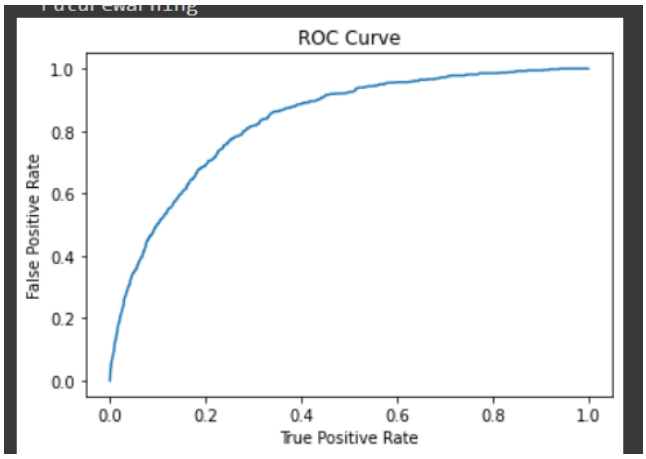

```
Training set areaUnderROC: 0.999999550110088
```



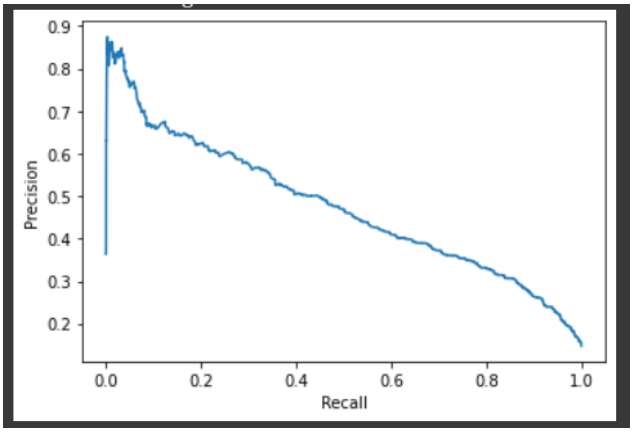

```
Test Area Under ROC 1.0
```

## بررسی اثر پیش‌پردازش داده

برای حالت بدون پیش پردازش داده‌ها تنها مقادیر نال‌ها حذف می‌شوند و تبدیل به عدد و در آخر وکتور صورت می‌گیرد تا مدل به ارور نخورد. نتایج بدون پیش پردازش به‌صورت زیر می‌باشد. که می‌توان اثر پیش پردازش را در نتایج نسبت به نتایج قبلی به وضوح مشاهده کرد.




```
Training set areaUnderROC: 0.8327346769189856
```




```
Test Area Under ROC 0.833292152387083
```