Medical Appointment No Shows

Why do 30% of patients miss their scheduled appointments?

Jing Sun (js6mj)

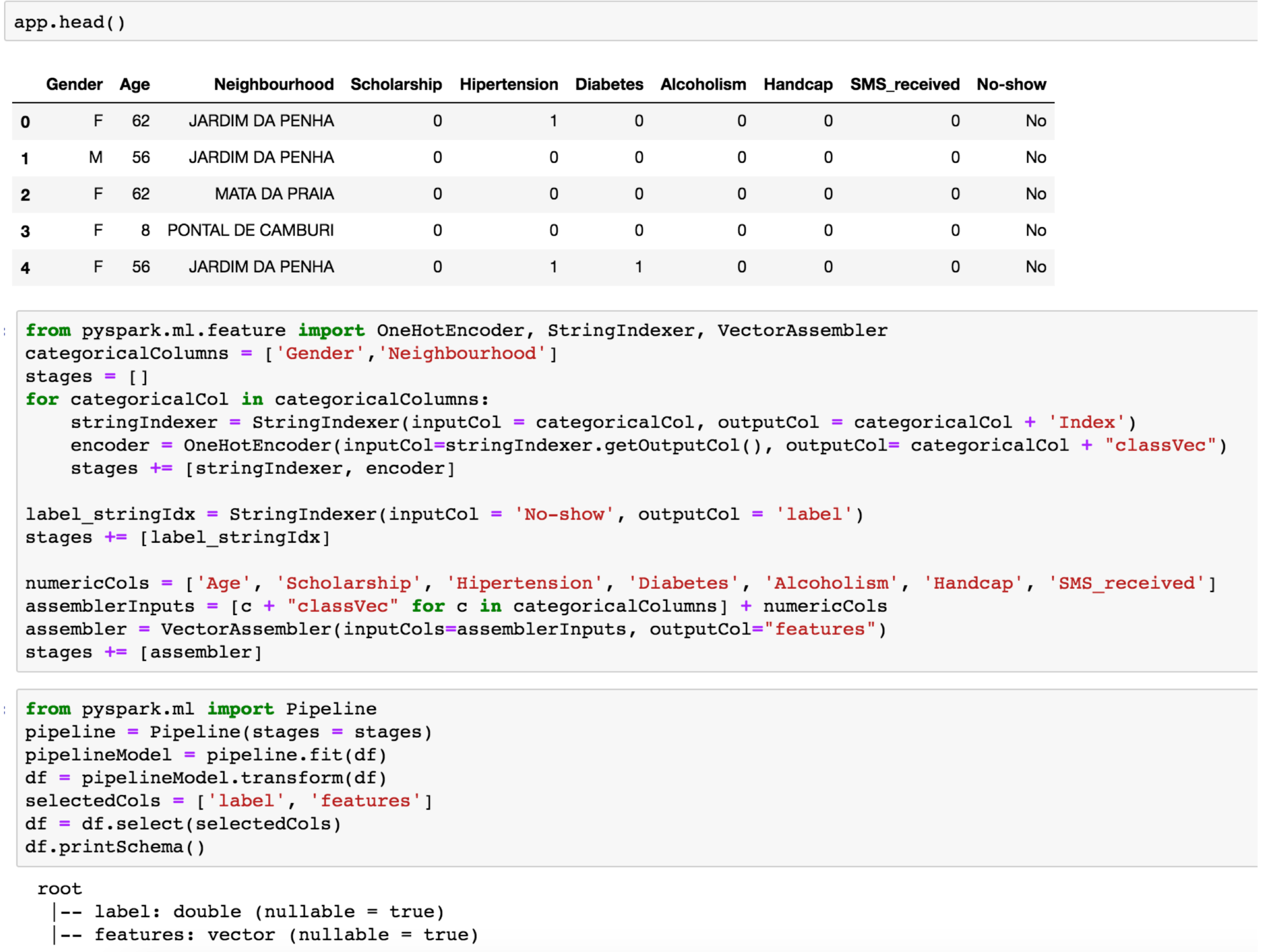
Motivation

Sometimes a person makes a doctor appointment, receives all the instructions and no-show. It’s not only a waste of time for the medical staff, but unethical since other people might have really needed this time slot. It would be interesting to see if we could develop an algorithm to predict patient no-shows.

https://www.kaggle.com/joniarroba/noshowappointments/home

Code Snippet

The code snippet on the right shows the process of constructing a pipeline to convert the original data frame into a MLlib-compatible format for analysis



```
app.head()
```

	Gender	Age	Neighbourhood	Scholarship	Hipertension	Diabetes	Alcoholism	Handcap	SMS_received	No-show
0	F	62	JARDIM DA PENHA	0	1	0	0	0	0	No
1	M	56	JARDIM DA PENHA	0	0	0	0	0	0	No
2	F	62	MATA DA PRAIA	0	0	0	0	0	0	No
3	F	8	PONTAL DE CAMBURI	0	0	0	0	0	0	No
4	F	56	JARDIM DA PENHA	0	1	1	0	0	0	No

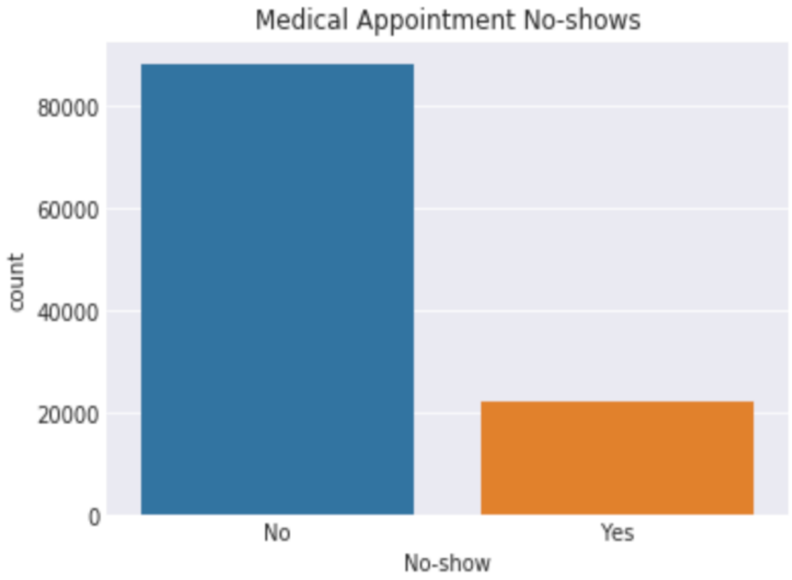
```
from pyspark.ml.feature import OneHotEncoder, StringIndexer, VectorAssembler
categoricalColumns = ['Gender','Neighbourhood']
stages = []
for categoricalCol in categoricalColumns:
    stringIndexer = StringIndexer(inputCol = categoricalCol, outputCol = categoricalCol + 'Index')
    encoder = OneHotEncoder(inputCol=stringIndexer.getOutputCol(), outputCol= categoricalCol + "classVec")
    stages += [stringIndexer, encoder]

label_stringIdx = StringIndexer(inputCol = 'No-show', outputCol = 'label')
stages += [label_stringIdx]

numericCols = ['Age', 'Scholarship', 'Hipertension', 'Diabetes', 'Alcoholism', 'Handcap', 'SMS_received']
assemblerInputs = [c + "classVec" for c in categoricalColumns] + numericCols
assembler = VectorAssembler(inputCols=assemblerInputs, outputCol="features")
stages += [assembler]
```

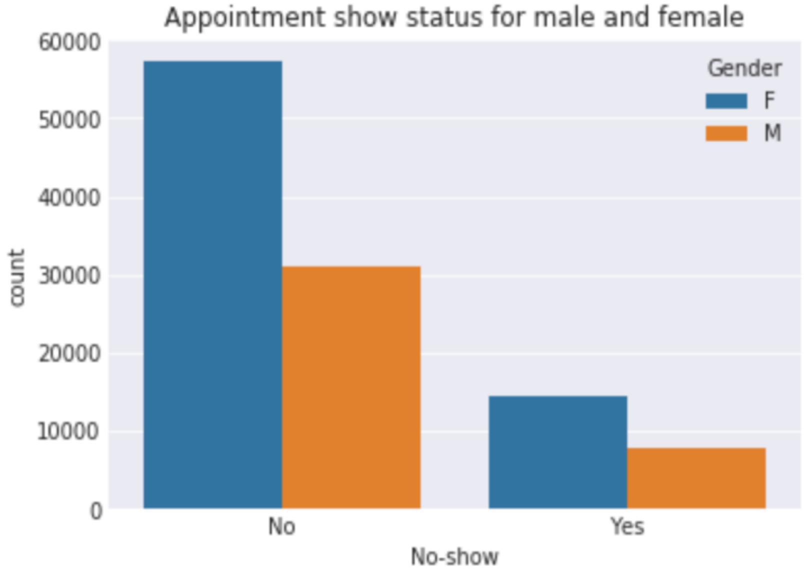
```
from pyspark.ml import Pipeline
pipeline = Pipeline(stages = stages)
pipelineModel = pipeline.fit(df)
df = pipelineModel.transform(df)
selectedCols = ['label', 'features']
df = df.select(selectedCols)
df.printSchema()
```

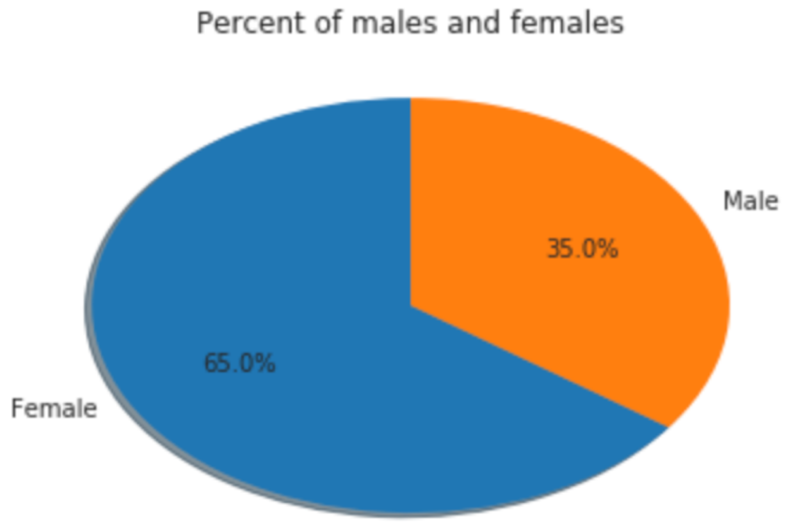
```
root
 |-- label: double (nullable = true)
 |-- features: vector (nullable = true)
```

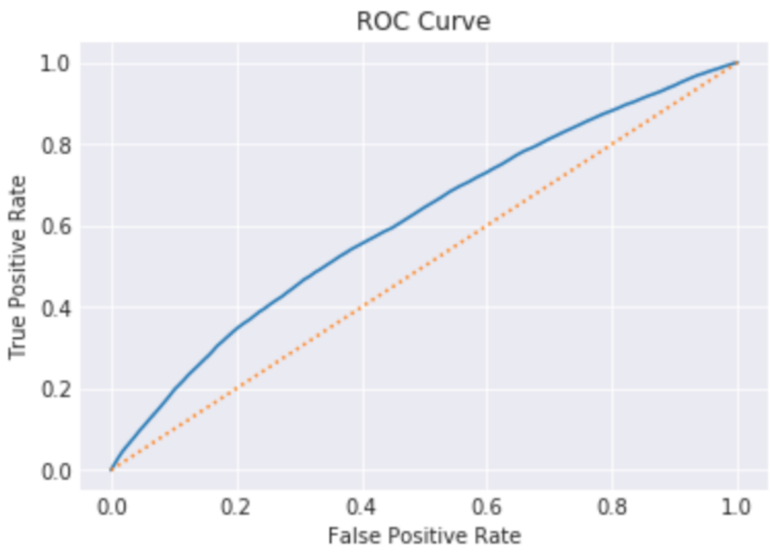
Visualization

The visuals on the right show that the dataset contains a lot more cases where patient did show up compared to no-shows. There are also a lot females in the dataset as well. The ROC curve on the bottom right suggests that we achieved 0.608 area under the curve on the training set using logistic regression.


Medical Appointment No-shows


Appointment show status for male and female


Percent of males and females


ROC Curve

```
Training set areaUnderROC: 0.608
```