SPARK ASSIGNMENT

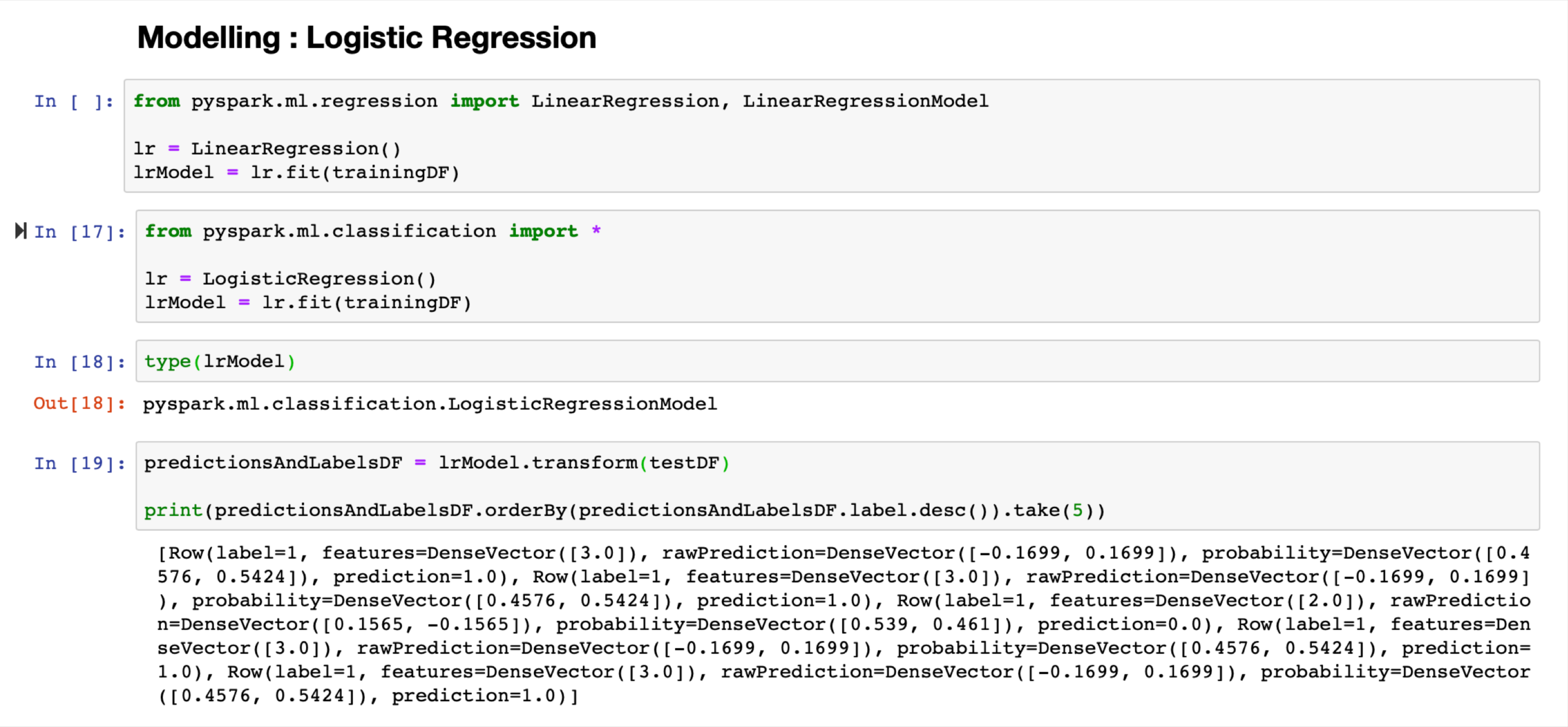
RAKESH RAVI K U

COMPUTING ID – rk9cx

Motivation

- The goal of the project is to check if there the responses to the survey questions from the Somerville happiness survey has a relationship with whether or not the respondents are actually happy.
- The respondents had to answer six questions on different elements ranging from the quality of school education in Somerville to the community and social events. In the end they were asked to answer if they felt happy or not.

Code Snippet and Explanation



Modelling : Logistic Regression

```
In [ ]:  from pyspark.ml.regression import LinearRegression, LinearRegressionModel

         lr = LinearRegression()
         lrModel = lr.fit(trainingDF)
```

```
In [17]: from pyspark.ml.classification import *

         lr = LogisticRegression()
         lrModel = lr.fit(trainingDF)
```

```
In [18]: type(lrModel)

Out[18]: pyspark.ml.classification.LogisticRegressionModel
```

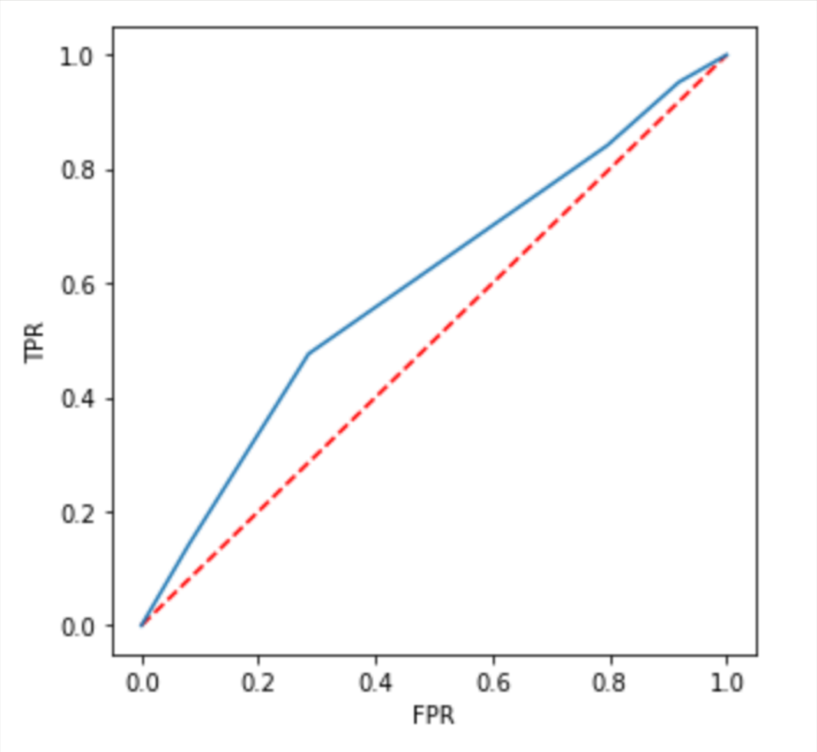
```
In [19]: predictionsAndLabelsDF = lrModel.transform(testDF)

         print(predictionsAndLabelsDF.orderBy(predictionsAndLabelsDF.label.desc()).take(5))

  [Row(label=1, features=DenseVector([3.0]), rawPrediction=DenseVector([-0.1699, 0.1699]), probability=DenseVector([0.4
  576, 0.5424]), prediction=1.0), Row(label=1, features=DenseVector([3.0]), rawPrediction=DenseVector([-0.1699, 0.1699]
  ), probability=DenseVector([0.4576, 0.5424]), prediction=1.0), Row(label=1, features=DenseVector([2.0]), rawPredictio
  n=DenseVector([0.1565, -0.1565]), probability=DenseVector([0.539, 0.461]), prediction=0.0), Row(label=1, features=Den
  seVector([3.0]), rawPrediction=DenseVector([-0.1699, 0.1699]), probability=DenseVector([0.4576, 0.5424]), prediction=
  1.0), Row(label=1, features=DenseVector([3.0]), rawPrediction=DenseVector([-0.1699, 0.1699]), probability=DenseVector
  ([0.4576, 0.5424]), prediction=1.0)]
```

I incorporated a logistic regression binary classifier in order to predict whether someone who responded to the survey was happy or not. I used the Decision variable as label and the quality of school as a feature to train the model

Visualization (ROC Curve)



The model’s area under ROC curve came out to be 0.605 which is a tad better than half chance.