SPARK PRESENTATION

NAME: KANIKA DAWAR

MODEL: RANDOM FOREST CLASSIFIER

DATA: UCI WINE DATASET (MULTI CLASS CLASSIFICATION)

Motivation

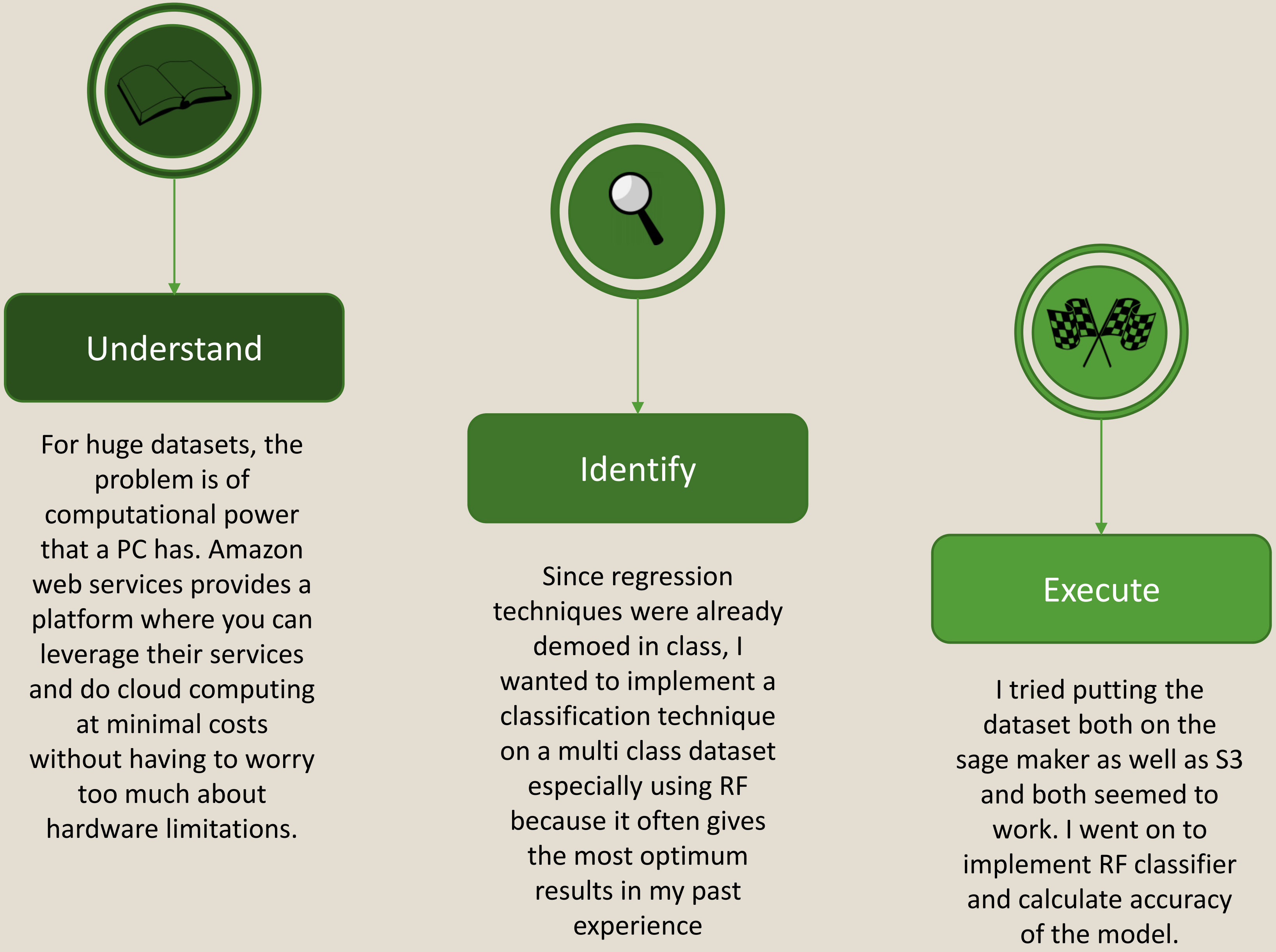
SECTION I

Motivation

Why Spark?

Why this problem?

How to?



Understand

For huge datasets, the problem is of computational power that a PC has. Amazon web services provides a platform where you can leverage their services and do cloud computing at minimal costs without having to worry too much about hardware limitations.

Identify

Since regression techniques were already demoed in class, I wanted to implement a classification technique on a multi class dataset especially using RF because it often gives the most optimum results in my past experience

Execute

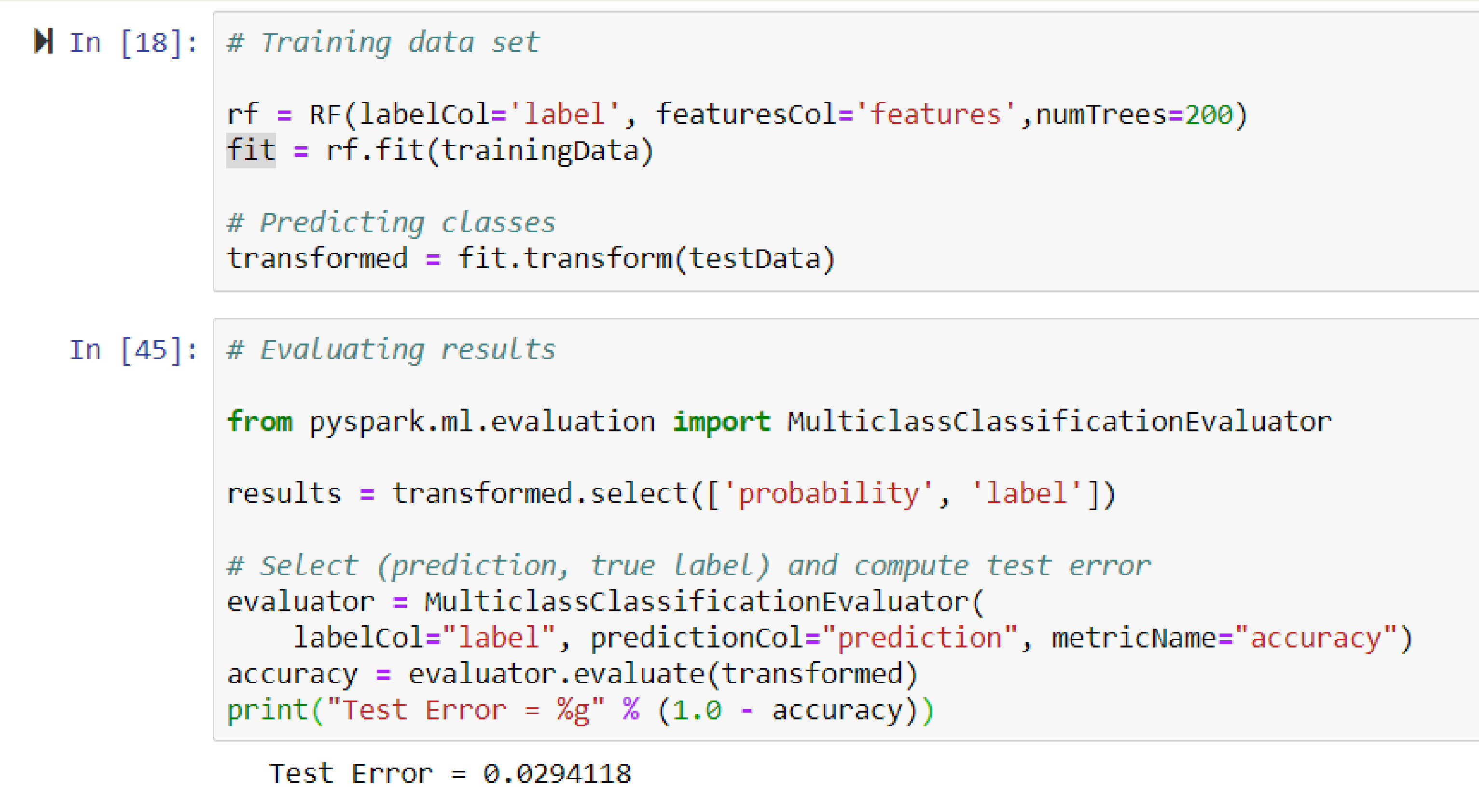
I tried putting the dataset both on the sage maker as well as S3 and both seemed to work. I went on to implement RF classifier and calculate accuracy of the model.

I went on to UCI dataset library that would help me achieve my aim of experimenting with Spark for classification techniques

The Code Snippet

SECTION II

ML lib – Random Forest



In [18]:

```
# Training data set

rf = RF(labelCol='label', featuresCol='features',numTrees=200)
fit = rf.fit(trainingData)

# Predicting classes
transformed = fit.transform(testData)
```

In [45]:

```
# Evaluating results

from pyspark.ml.evaluation import MulticlassClassificationEvaluator

results = transformed.select(['probability', 'label'])

# Select (prediction, true label) and compute test error
evaluator = MulticlassClassificationEvaluator(
    labelCol="label", predictionCol="prediction", metricName="accuracy")
accuracy = evaluator.evaluate(transformed)
print("Test Error = %g" % (1.0 - accuracy))
```

```
Test Error = 0.0294118
```

Here, the data has been trained to classify alcohol into the three labels (Wine classes) using 12 features and 200 trees.
Then prediction has been done on the test dataset using the random forest model.

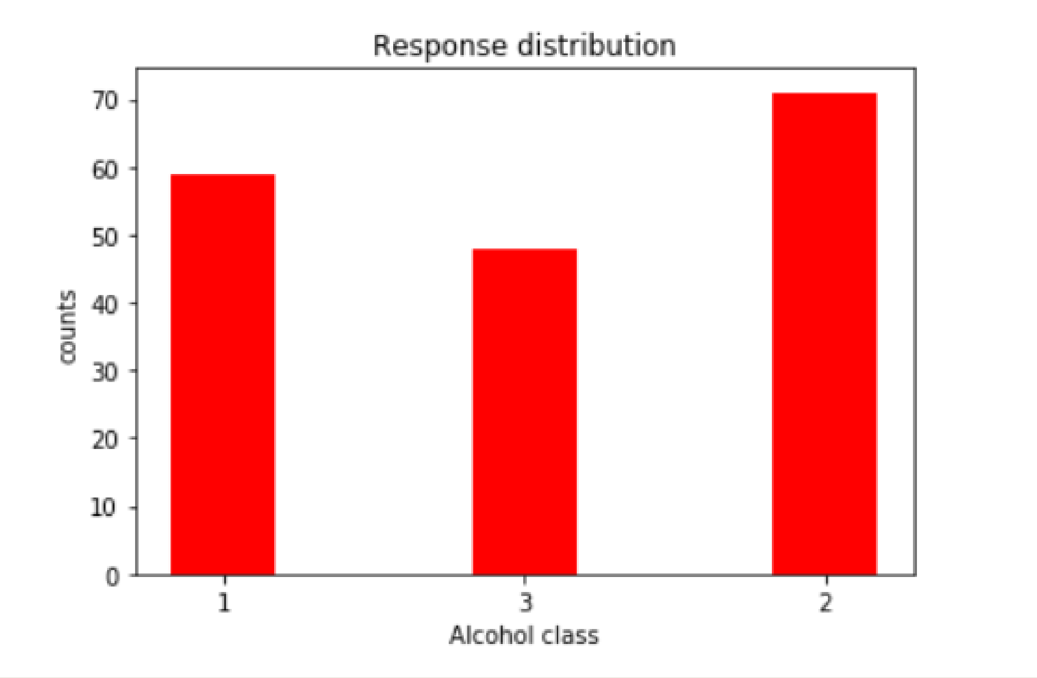
I have tried to use Accuracy of predictions (for a multi class classification) as the metric for model evaluation.

Visualizations

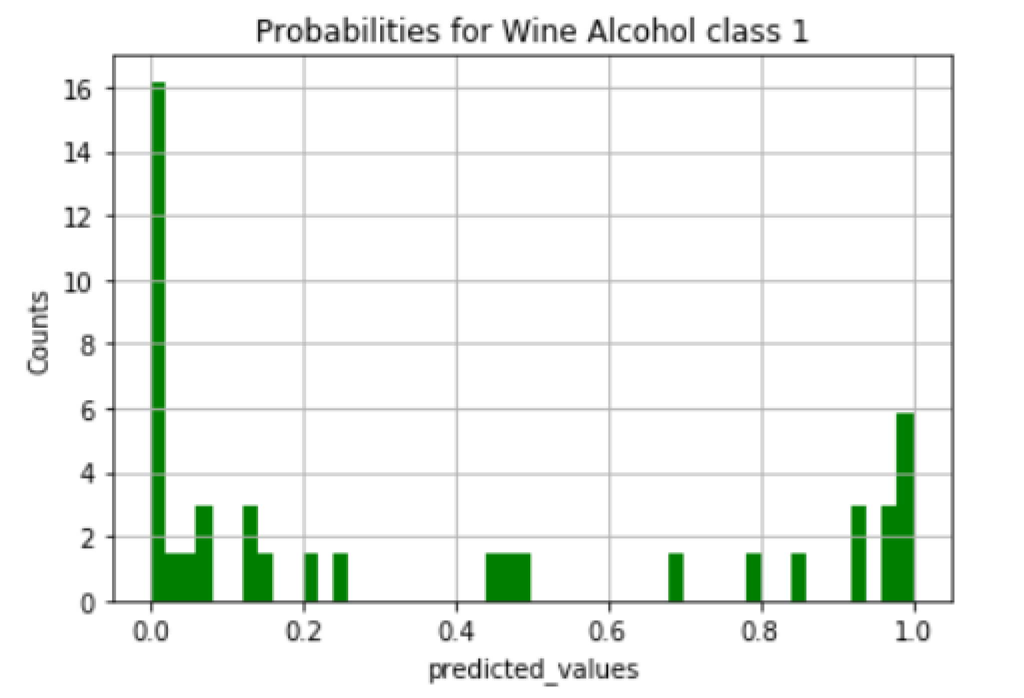
SECTION III

Result Visualization

Response distribution

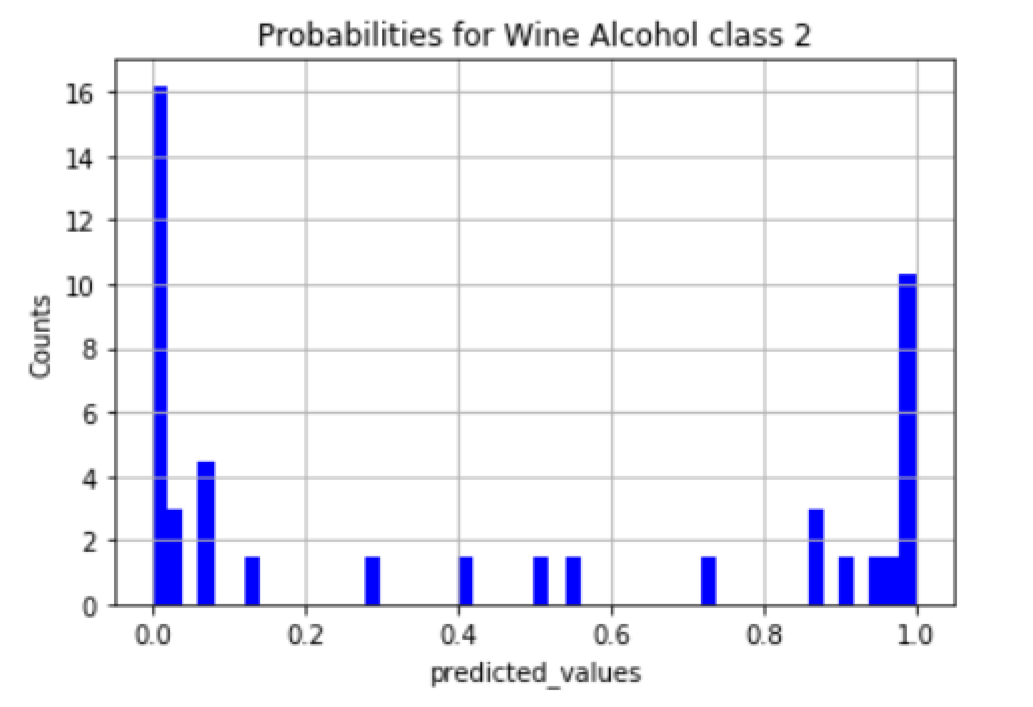


Alcohol Class 1 prob prediction

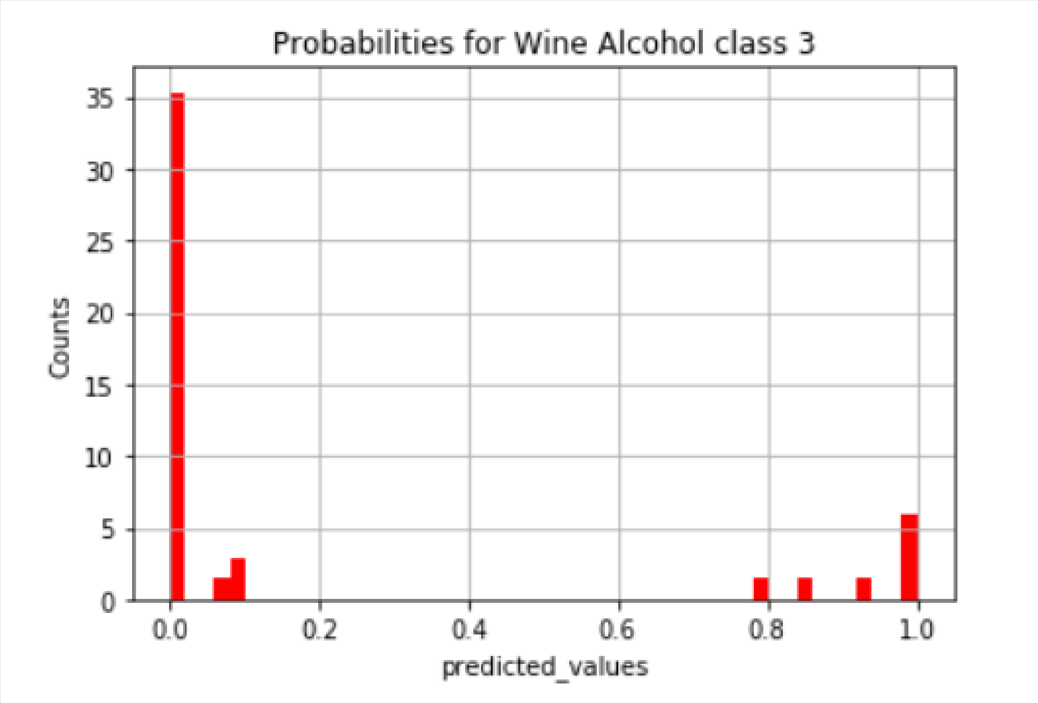


Result Visualization

Alcohol Class 2 prob prediction



Alcohol Class 3 prob prediction


Probabilities for Wine Alcohol class 3
Counts
predicted_values