PySpark assignment (khg8mh)

MOTIVATION:

- Find relationship between variables like GRE score, TOEFL score, Statement of Purpose and Letter of recommendations quality and see how that affects chances of admit to graduate course in the United States.
- Predict chance of admit of a student to any graduate course in the United States based on the student's GRE score, Undergrad GPA, TOEFL score, quality of his Statement of Purpose and Letters of recommendations.

CODE SNIPPET AND EXPLANATION:

MLlib based analysis

In [4]:
```
# Data pre-processing before building a model
from pyspark.ml import Pipeline
from pyspark.ml.feature import StringIndexer, VectorAssembler

assembler = VectorAssembler(inputCols=['GRE','TOEFL','UniversityRating','SOP','LOR','CGPA','Research'], outputCol="featu
stages = [assembler]

label_stringIdx = StringIndexer(inputCol="ChanceOfAdmit", outputCol="label")
stages += [label_stringIdx]

partialPipeline = Pipeline().setStages(stages)
pipelineModel = partialPipeline.fit(df)
preppedDataDF = pipelineModel.transform(df)

selectedcols = ["label", "features"] + df.columns
dataset = preppedDataDF.select(selectedcols)
```

In [5]:
```
# train test split
(trainingData, testData) = dataset.randomSplit([0.7, 0.3], seed=100)
```

In [21]:
```
# Logistic regression
from pyspark.ml.classification import LogisticRegression
lr = LogisticRegression(labelCol="label", featuresCol="features", maxIter=10)

# Train model with Training Data
lrModel = lr.fit(trainingData)

# Predict on testing data
predictions = lrModel.transform(testData)
```

- Pre-processed data in order to get it into spark dataframe
- Converted data to a format required by Logistic regression module in PySpark
- Performed Logistic regression and achieved an ROC score of 79%

VISUALISATION:

Distribution of probability of chance of an admit: