Motivation

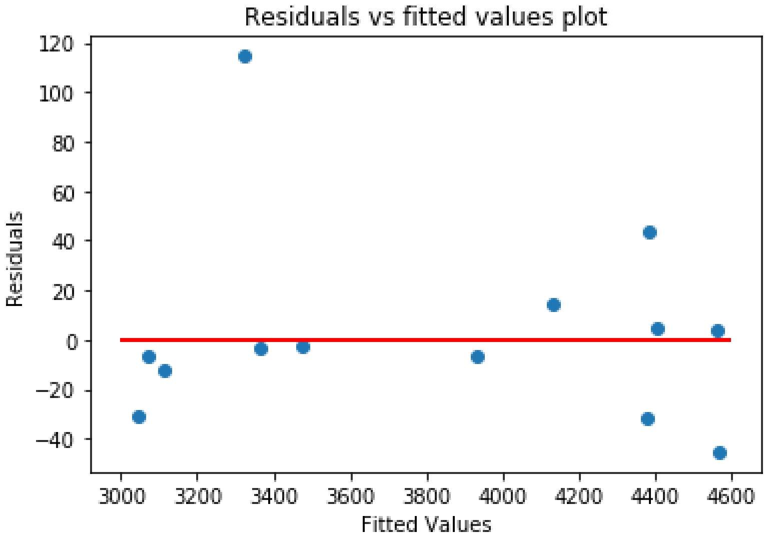
The motivation to use Spark is to have multiple instances run the code. This will allow parallel processing and hence it will be faster. Over here, in this assignment, I have used a dataset of thrust data. I have used a linear regression model to calculate thrust based on 6 variables.

Code snippet and explanation (show off)

The following code is for model building-

```
# Data pre-processing before building a model
from pyspark.ml import Pipeline
from pyspark.ml.feature import StringIndexer, VectorAssembler
assembler = VectorAssembler(inputCols=['x1','x2','x3','x4','x5','x6'], outputCol="features")
stages = [assembler]
transf_df = assembler.transform(df)
transf_df = transf_df.select(['features', 'y'])
# Linear regression
from pyspark.ml.regression import LinearRegression
#Using the linear regression function to model
lr = LinearRegression(featuresCol = 'features', labelCol='y', maxIter=10, regParam=0.3, elasticNetParam=0.8)
lr_model = lr.fit(train_df) #Model is fit over here
#Prediction
lr_predictions = lr_model.transform(test_df) #Making predictions on the test data
```

Visualization



After I fit the model on the data, this is the residuals vs fitted values plot.