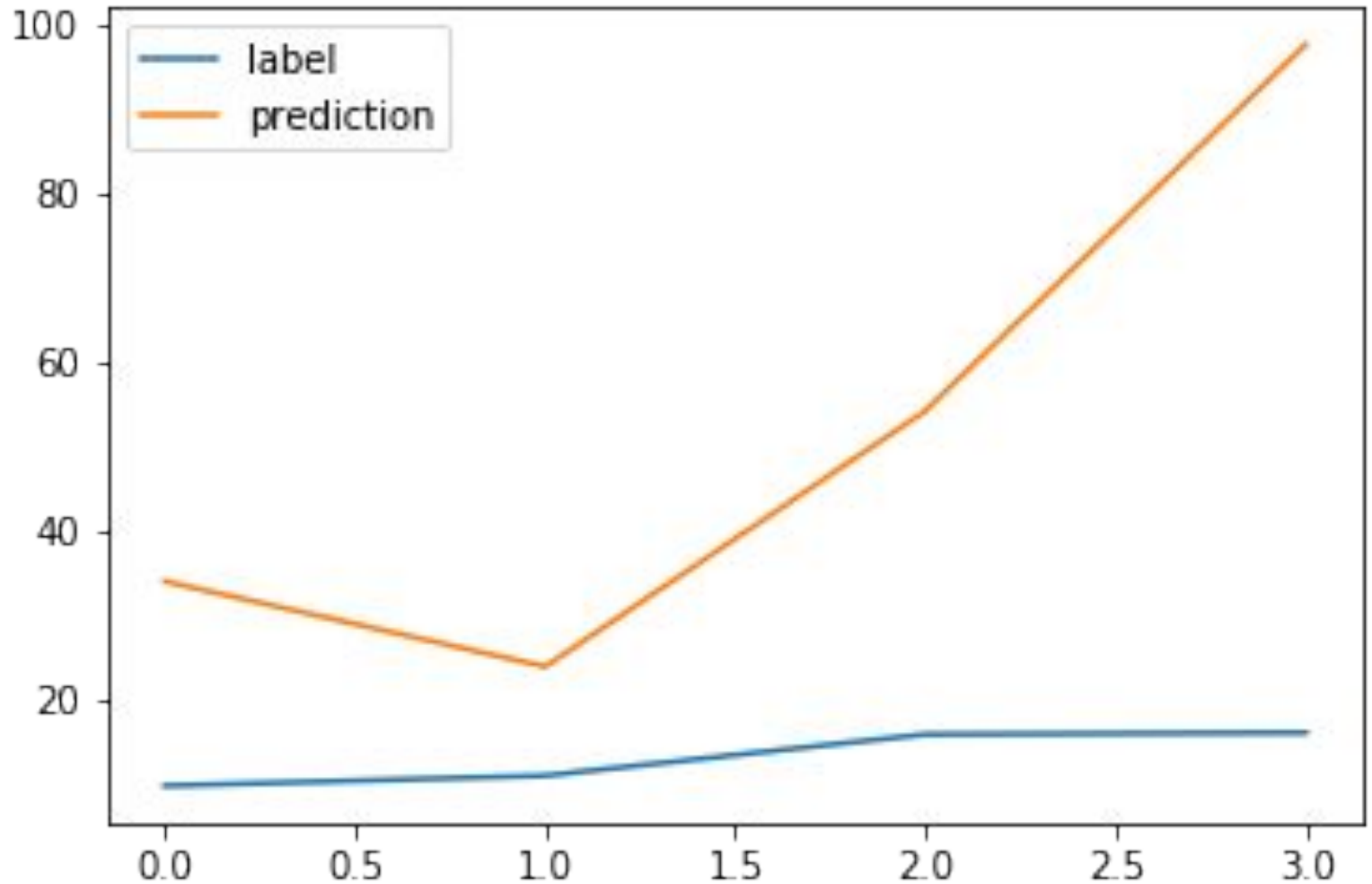
Motivation

- This is a sample data set of methanol's oxidation in supercritical water.
- The response variable is the percent of conversion; x5 is ratio of oxygen to methanol.

- I used the s3 and parquet to read the data.
- Vectorized the features to use the machine learning tools in spark.
- Applied linear regression model to the data selected.
- Generated a plot of predicted value and real label value.

- The result is not good because the dataset I chose is too small.
- I tried with the dataset of our capstone. Obviously some imputation is required as we have too much data missing. We also have some variable missing 100%, which should be removed in order for the spark machine learning to work.


label
prediction
100
80
60
40
20
0.0
0.5
1.0
1.5
2.0
2.5
3.0