Where Clause Column Selection using Decision Tree

Zihua Liu

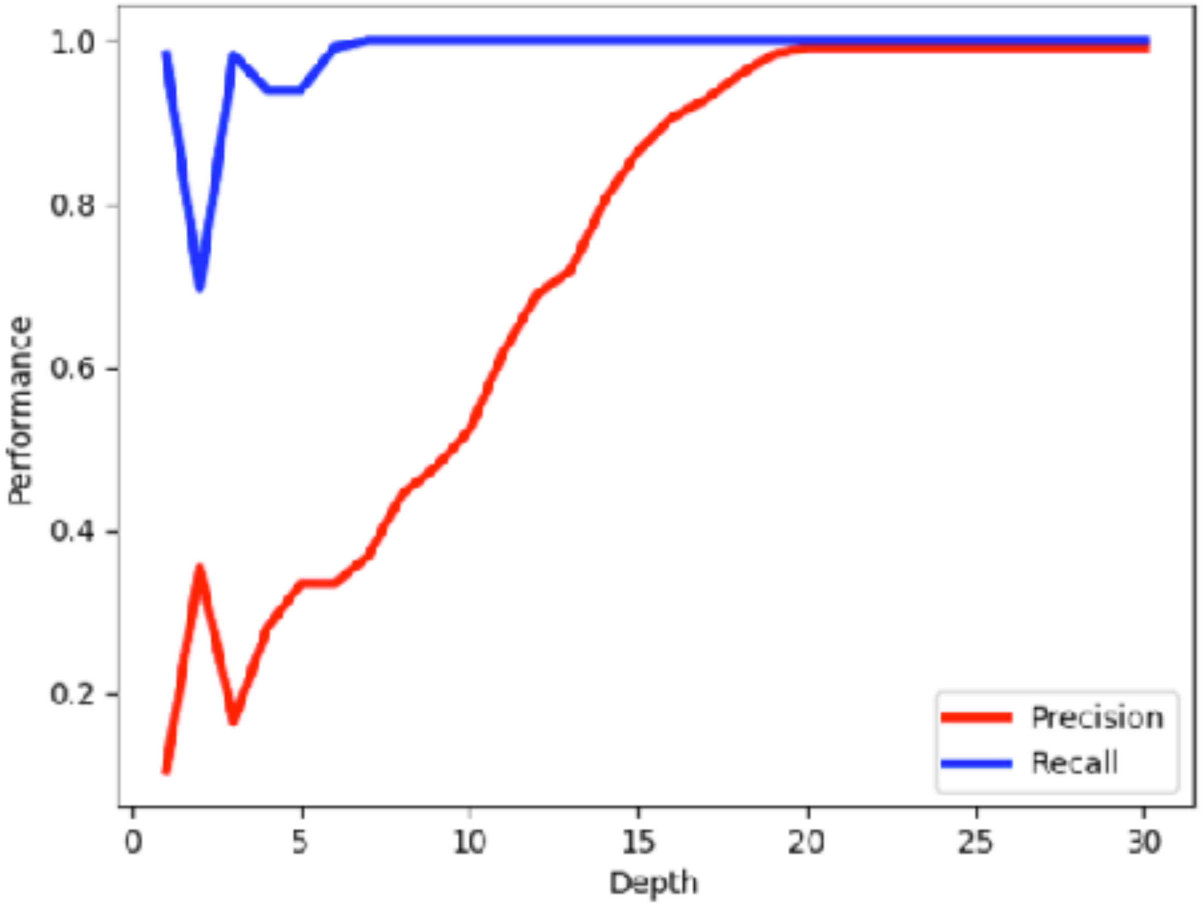
Dataset

- Training Set
    - Total 2046 columns
    - 115 Positive
    - 1931 Negative
- Test Set
    - Total 559 Columns
    - 30 Positive
    - 529 Negative

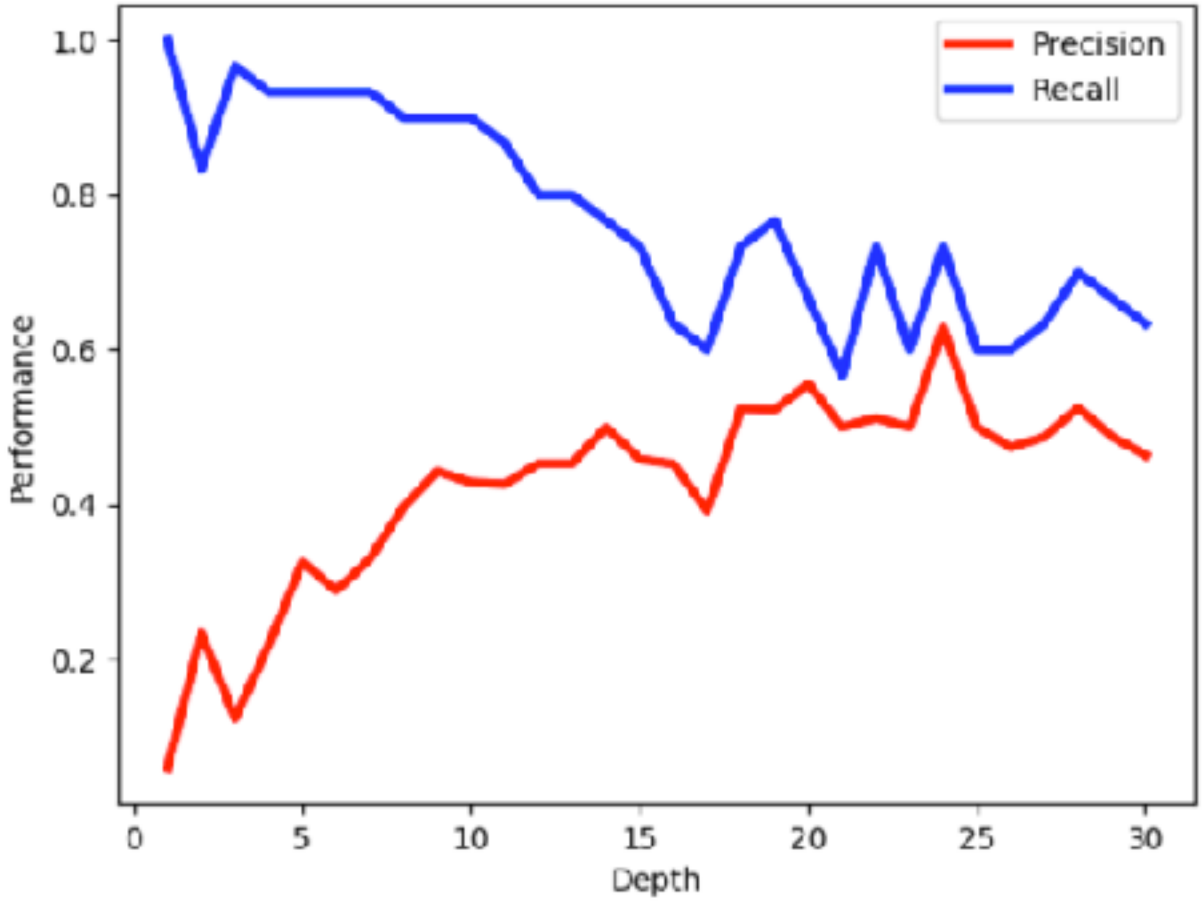
Features

- Number of Columns
- Word2Vec proximities
- Column Types Probabilities
- Question Types (One hot)
- Min Edit Distance between question and column content
- Average Cell Length
- Number of Rows
- Whether in Select Columns

Performance

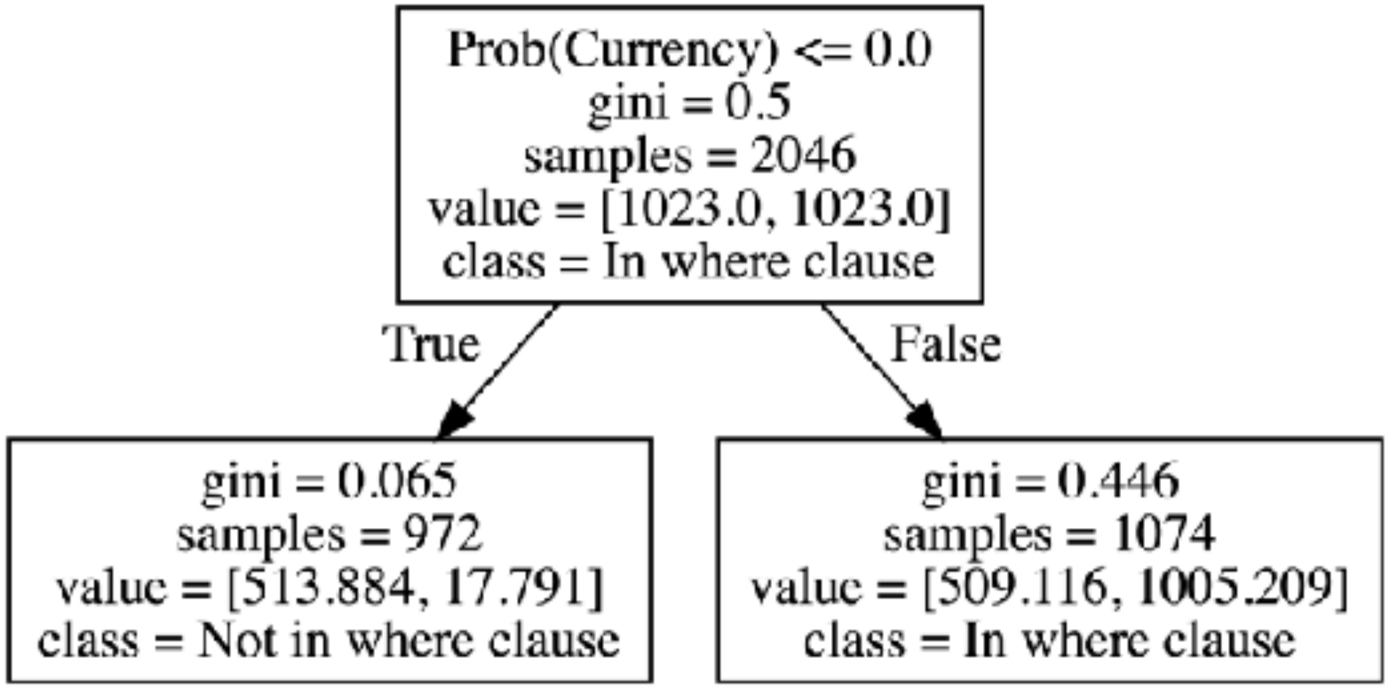


Training Set



Test Set

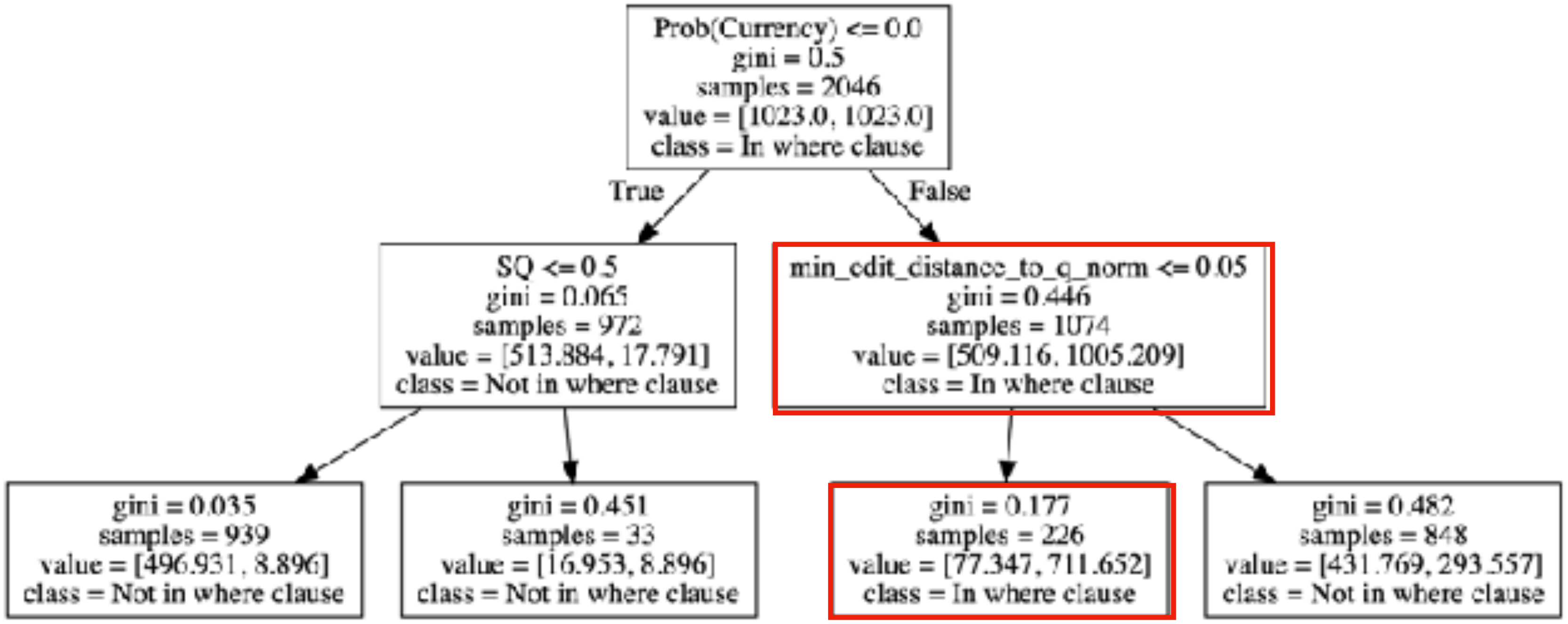
Depth = 1



	Label = 1	Label = 0
Pred = 1	113	961
Pred = 0	2	970

	Label = 1	Label = 0
Pred = 1	30	474
Pred = 0	0	55

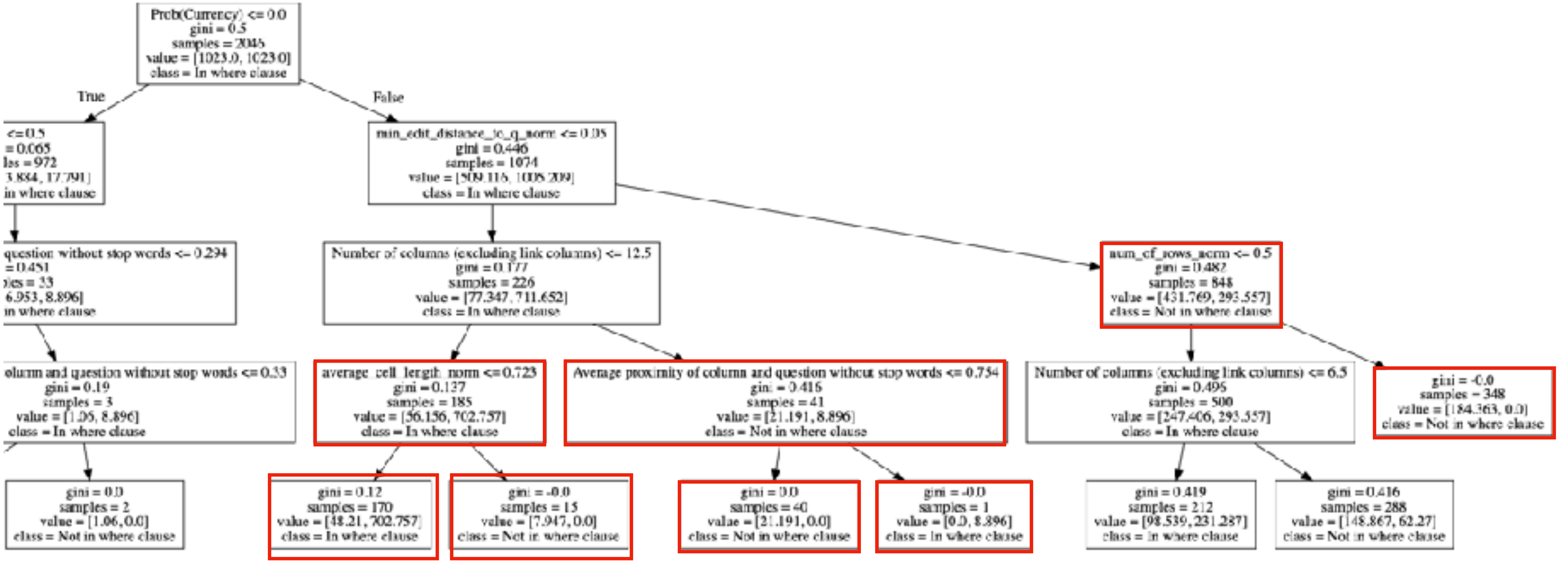
Depth = 2



	Label = 1	Label = 0
Pred = 1	80	146
Pred = 0	35	1785

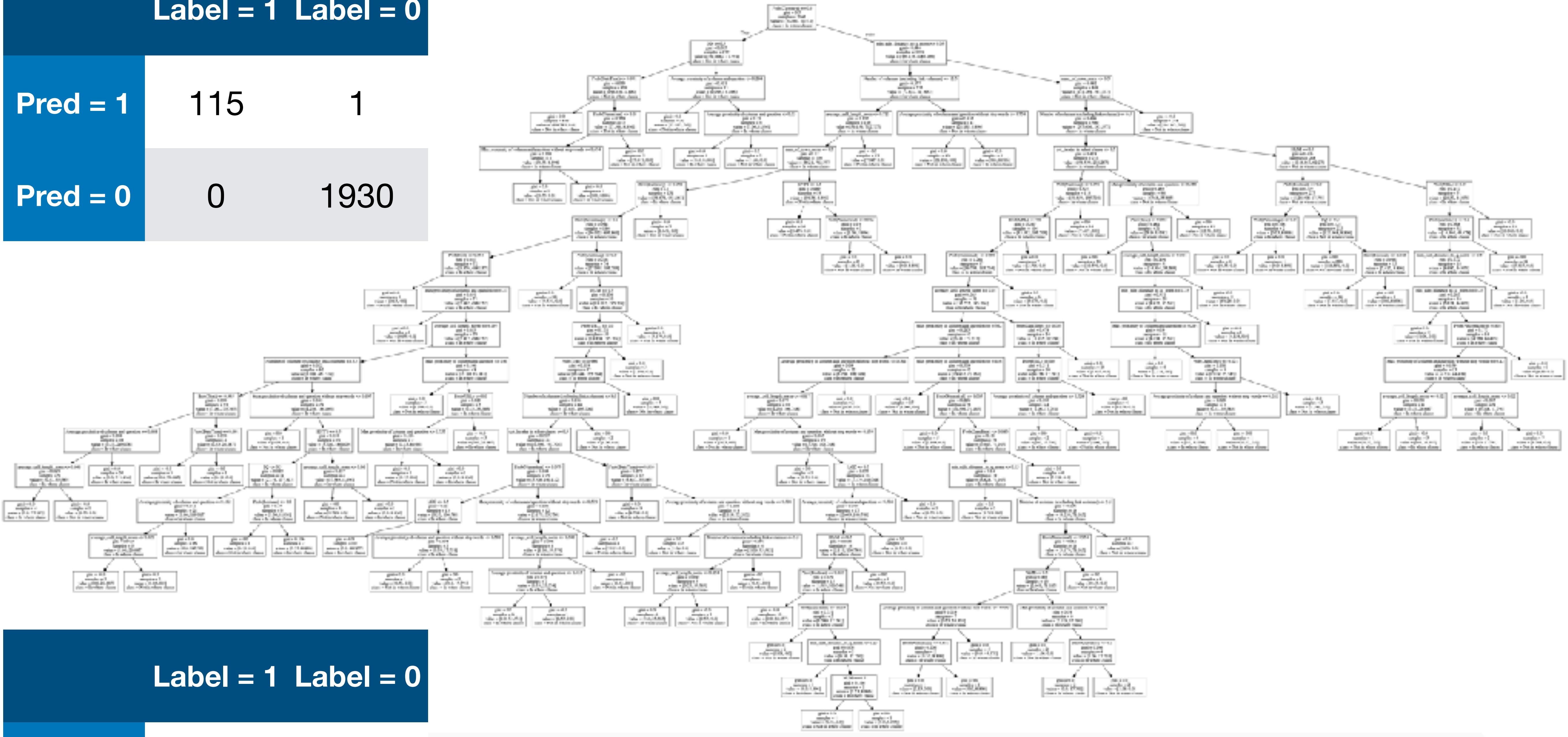
	Label = 1	Label = 0
Pred = 1	25	82
Pred = 0	5	447

Depth = 4


Prob(Currency) <= 0.0
gini = 0.5
samples = 2045
value = [1023.0, 1023.0]
class = In where clause
True
False
<= 0.5
= 0.065
les = 972
3.884, 17.791]
in where clause
min_edit_distance_to_q_norm <= 0.05
gini = 0.446
samples = 1074
value = [509.115, 1005.209]
class = In where clause
question without stop words <= 0.294
= 0.451
les = 33
6.953, 8.896]
in where clause
Number of columns (excluding link columns) <= 12.5
gini = 0.177
samples = 226
value = [77.347, 711.652]
class = In where clause
num_of_rows_norm <= 0.5
gini = 0.482
samples = 848
value = [431.769, 293.557]
class = Not in where clause
olumn and question without stop words <= 0.33
gini = 0.19
samples = 3
value = [1.05, 8.896]
class = In where clause
average_cell_length_norm <= 0.723
gini = 0.127
samples = 185
value = [56.156, 702.757]
class = In where clause
Average proximity of column and question without stop words <= 0.754
gini = 0.416
samples = 41
value = [21.191, 8.896]
class = Not in where clause
Number of columns (excluding link columns) <= 6.5
gini = 0.495
samples = 500
value = [247.406, 293.557]
class = In where clause
gini = -0.0
samples = 348
value = [184.363, 0.0]
class = Not in where clause
gini = 0.0
samples = 2
value = [1.06, 0.0]
class = Not in where clause
gini = 0.12
samples = 170
value = [48.21, 702.757]
class = In where clause
gini = -0.0
samples = 15
value = [7.947, 0.0]
class = Not in where clause
gini = 0.0
samples = 40
value = [21.191, 0.0]
class = Not in where clause
gini = -0.0
samples = 1
value = [0.0, 8.896]
class = In where clause
gini = 0.419
samples = 212
value = [98.539, 231.287]
class = In where clause
gini = 0.416
samples = 288
value = [148.867, 62.27]
class = Not in where clause

Depth = 24

	Label = 1	Label = 0
Pred = 1	115	1
Pred = 0	0	1930



	Label = 1	Label = 0
Pred = 1	22	13
Pred = 0	8	516

Thanks