The Problem:

Breast Cancer Diagnosis

The Problem:

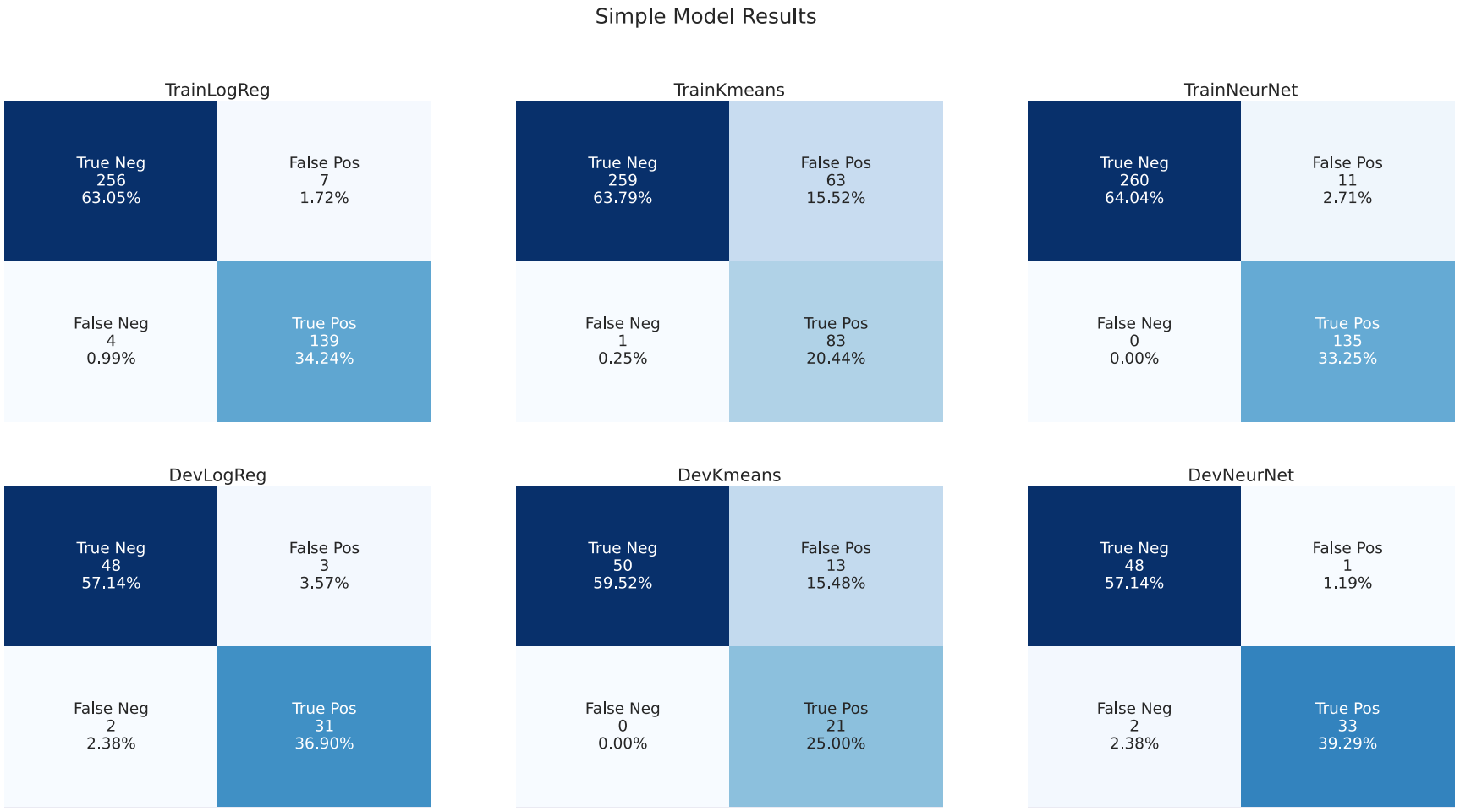
Breast Cancer Diagnosis:
False Negatives
(minimize)

The Data:

31 Variable: Diagnosis (“M” of “B”) and 30 Data Points

a) radius (mean of distances from center to points on the perimeter)
b) texture (standard deviation of gray-scale values)
c) perimeter
d) area
e) smoothness (local variation in radius lengths)
f) compactness (perimeter^2 / area - 1.0)
g) concavity (severity of concave portions of the contour)
h) concave points (number of concave portions of the contour)
i) symmetry
j) fractal dimension ("coastline approximation" - 1)

Simple Models



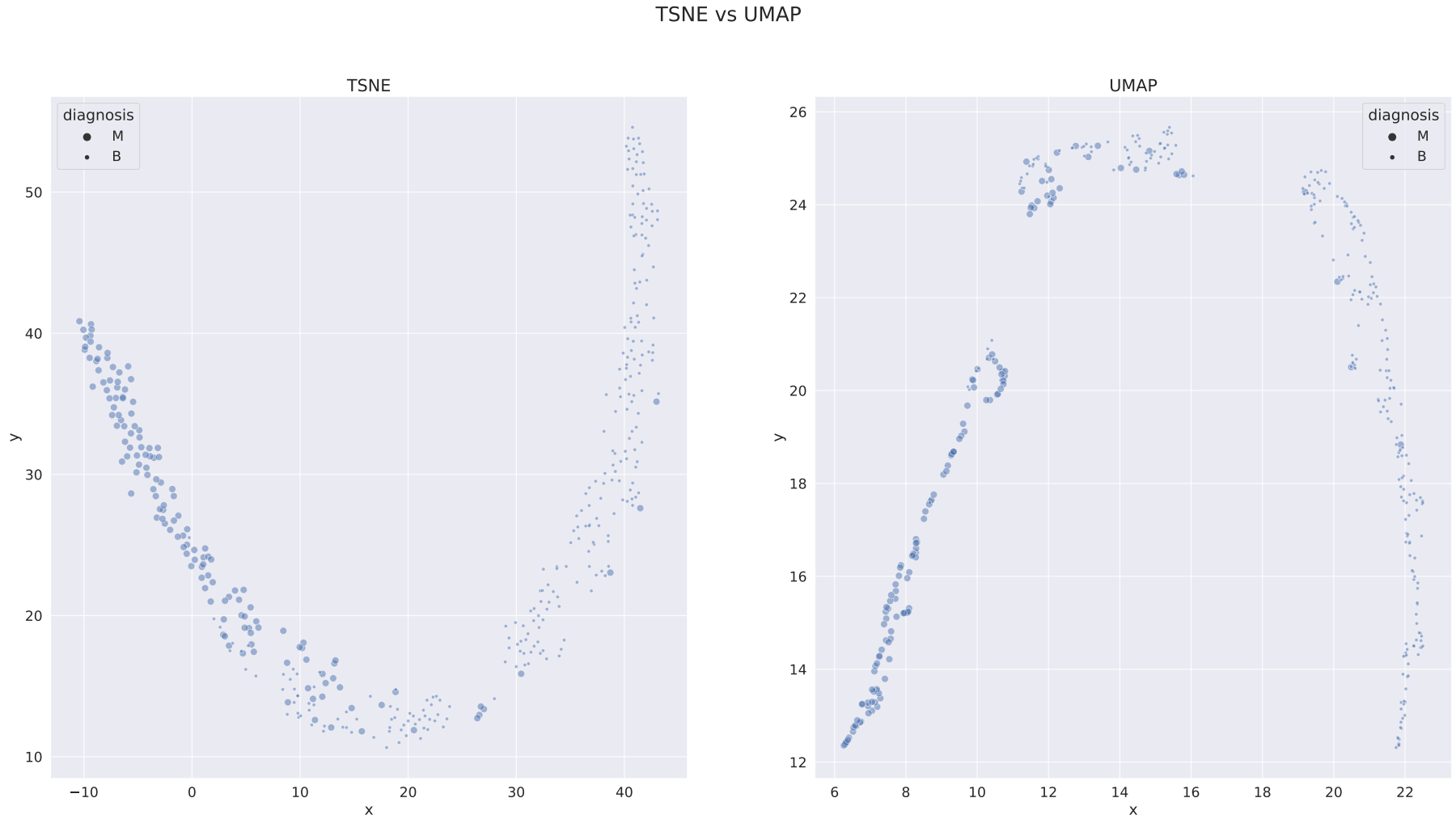
Ideal Case

(Whiteboard)

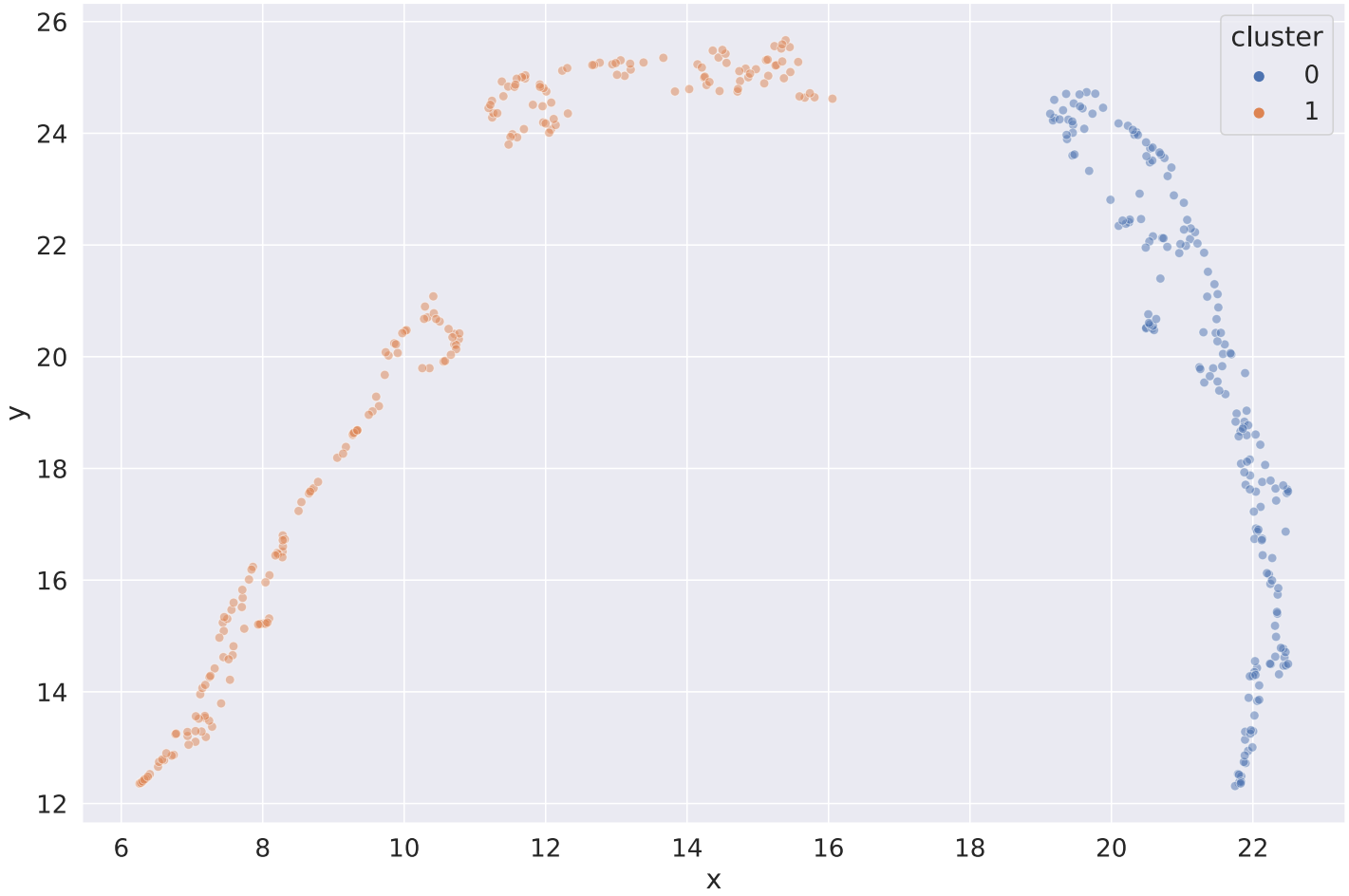
The Ensemble

(Whiteboard)

TSNE vs UMAP

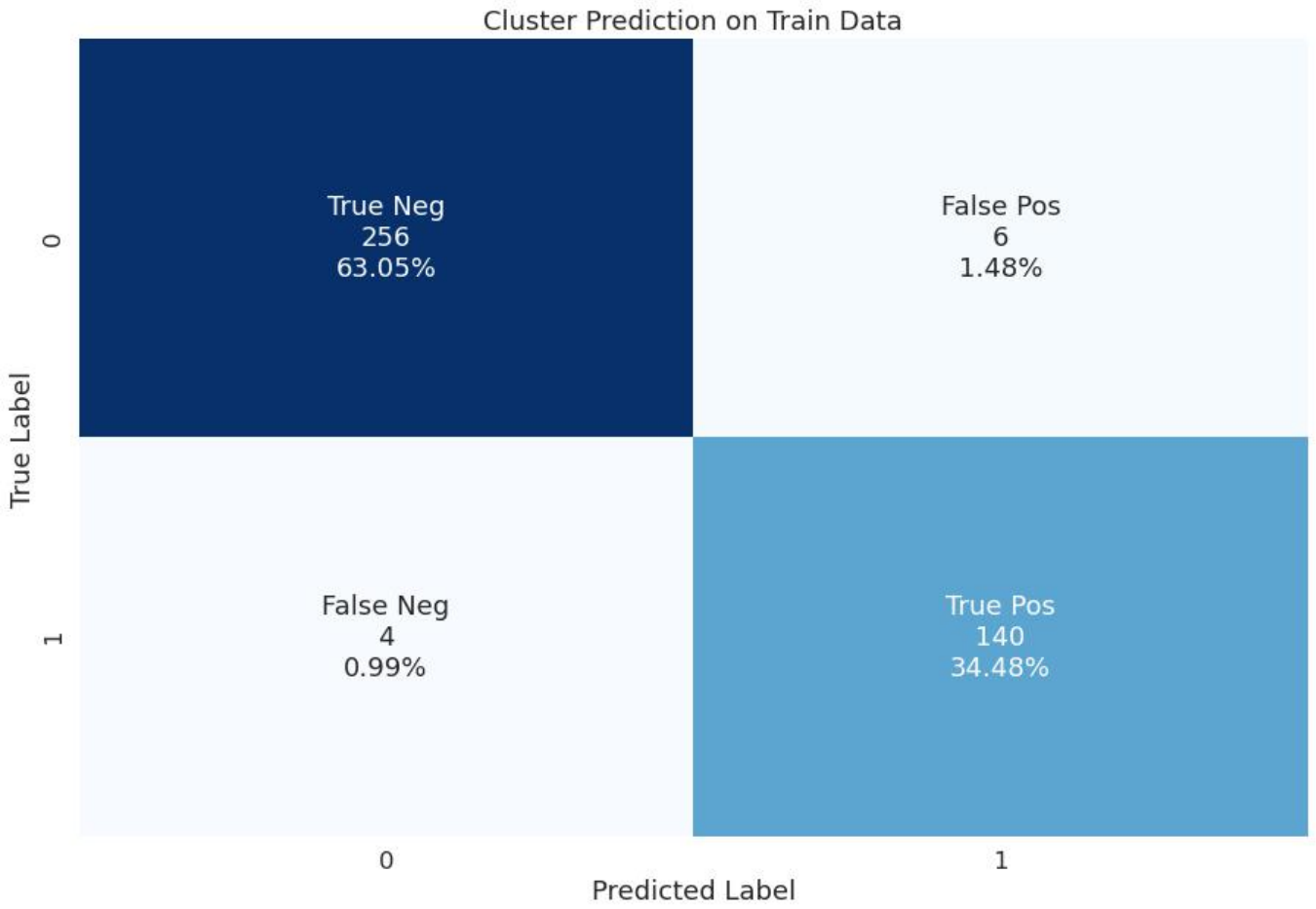

TSNE vs UMAP
TSNE
diagnosis
M
B
50
40
30
20
10
y
−10
0
10
20
30
40
x
UMAP
diagnosis
M
B
26
24
22
20
18
16
14
12
y
6
8
10
12
14
16
18
20
22
x

Train Kmeans



Train NN on each Cluster

To predict classification



Problems with UMAP

(Streamlit)

UMAP Predict Model

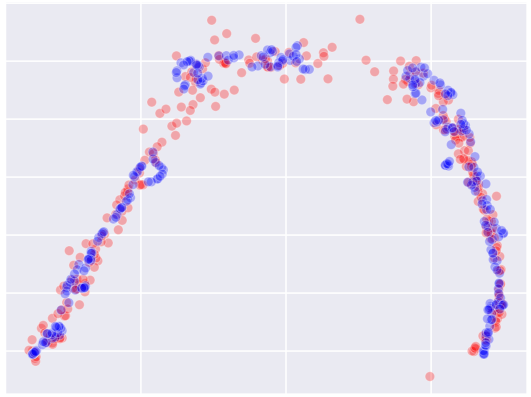
(Whiteboard)

Test Loss Functions:
MSLE
MAPE
Huber

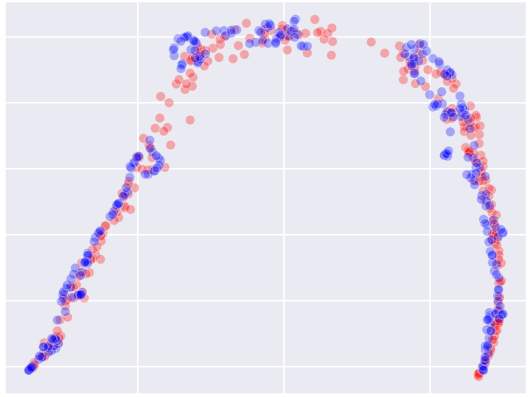
UMAP Parmeters

Neural Net on UMAP Results

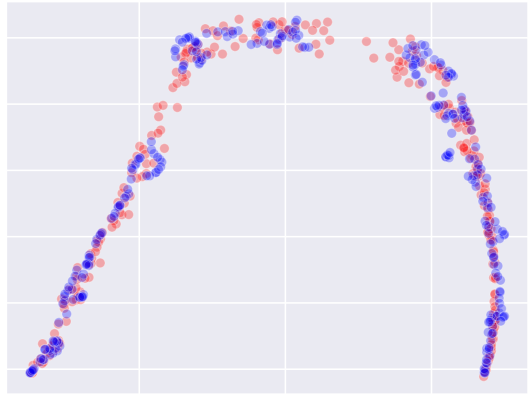
MSLE - Train



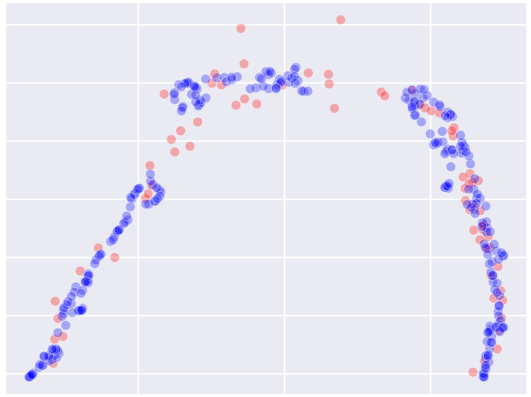
MAPE - Train



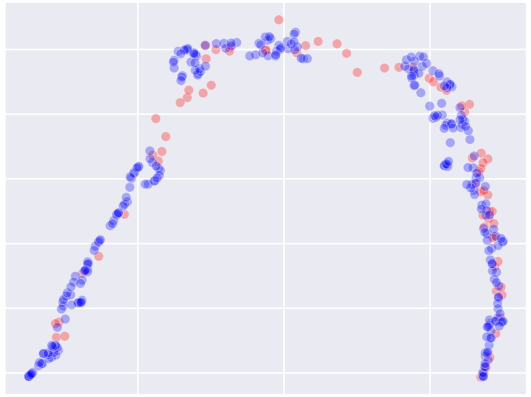
Huber - Train



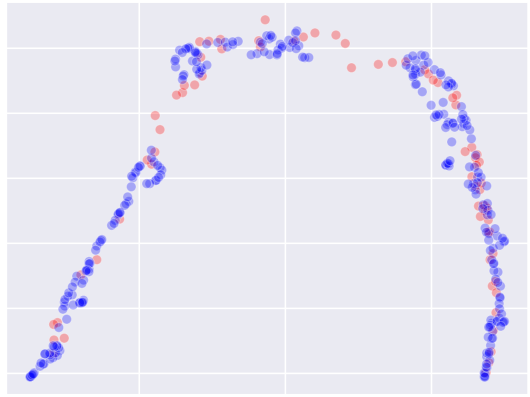
MSLE - Dev



MAPE - Dev



Huber - Dev

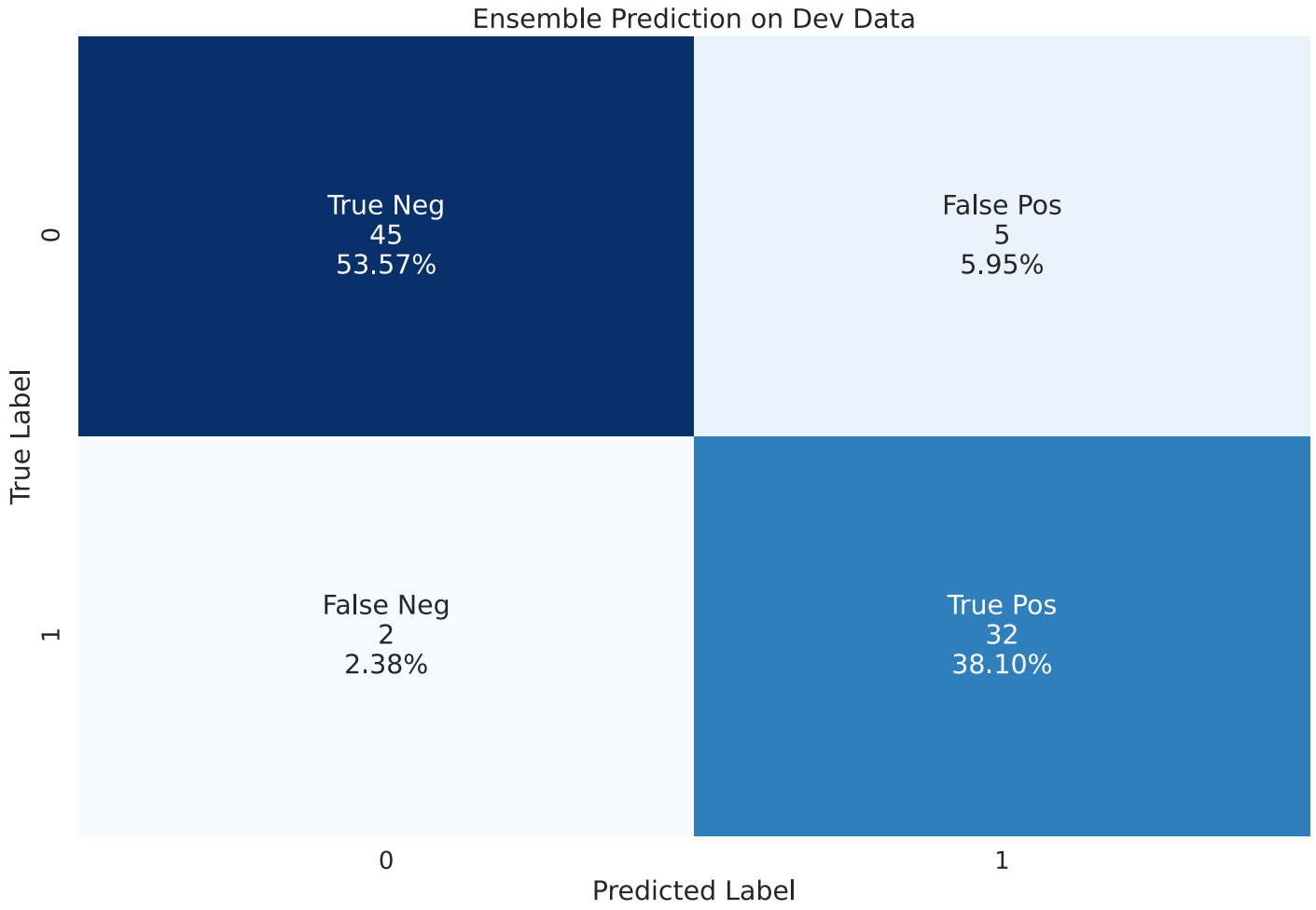


UMAP Parameters

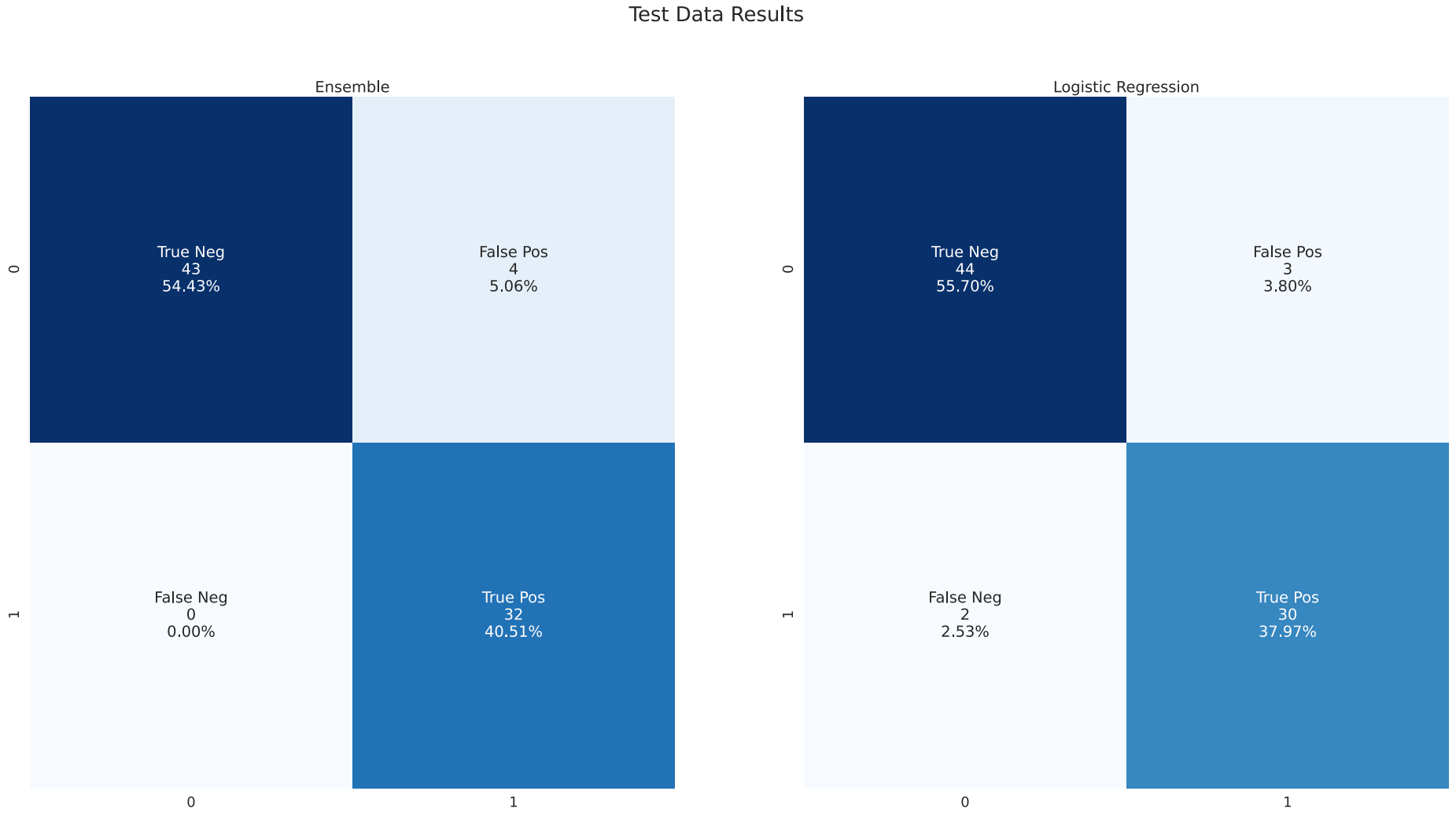
the KMeans model was able to predict 66 correct memberships with the MSLE loss function the Dev set
the KMeans model was able to predict 67 correct memberships with the MAPE loss function the Dev set
the KMeans model was able to predict 67 correct memberships with the Huber loss function the Dev set
Total observations in Dev set is: 70

the KMeans model was able to predict 54 correct memberships with the MSLE loss function the Test set
the KMeans model was able to predict 55 correct memberships with the MAPE loss function the Test set
the KMeans model was able to predict 56 correct memberships with the Huber loss function the Test set
Total observations in Test set is: 56

Ensemble Model Results



Ensemble Model Results



Conclusion

- I would not rely on this model.
- Does not give consistent results; any good results could be because of data split
- Does not significantly outperform a simple Logistic Regression Model.
- **UMAP** uses a **stochastic method** of reducing dimensionality. Because our **data was normalized**, we **assumed data** was **homogeneous**. If data is not homogeneous, any stochastic method would be inappropriate.
- Have less than **600 data points; need more data**.