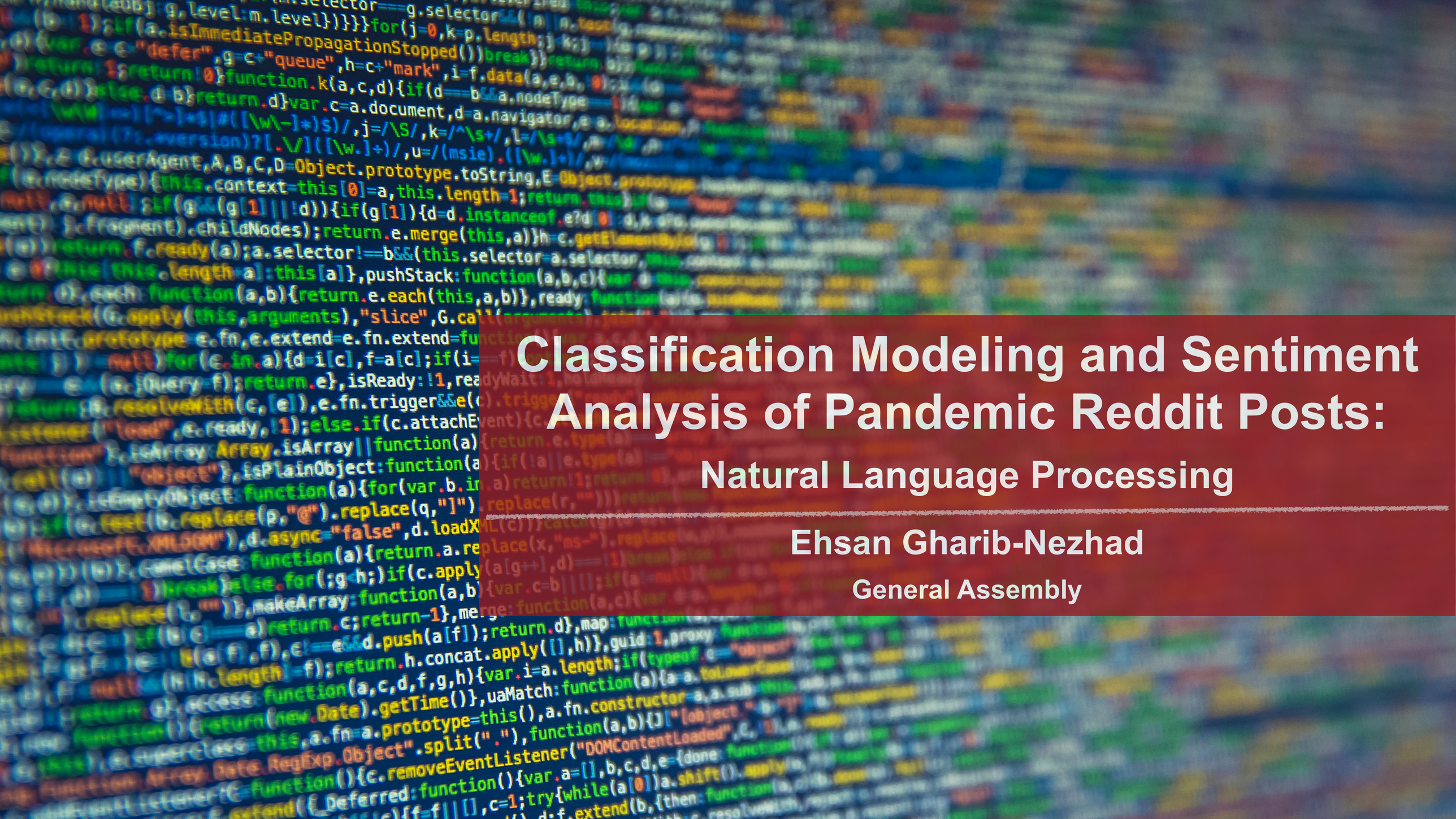


Classification Modeling and Sentiment Analysis of Pandemic Reddit Posts:

Natural Language Processing

Ehsan Gharib-Nezhad

General Assembly

Is this post from PandemicPreps or Covid19Positive subReddit?

For those that have tested positive I hope every single one of you feels better soon!

Is this post from PandemicPreps or Covid19Positive subReddit?

For those that have tested positive I hope every single one of you feels better soon!

Covid19Positive

Is this post from PandemicPreps or Covid19Positive subReddit?

For those that have tested positive I hope every single one of you feels better soon!

Covid19Positive

I see a lot of people on here buying around the same types of stuff (canned beans, rice, etc). Just remember to know how to cook each item you buy and

Is this post from PandemicPreps or Covid19Positive subReddit?

For those that have tested positive I hope every single one of you feels better soon!

Covid19Positive

I see a lot of people on here buying around the same types of stuff (canned beans, rice, etc). Just remember to know how to cook each item you buy and

PandemicPrep

Is this post from PandemicPreps or Covid19Positive subReddit?

For those that have tested positive I hope every single one of you feels better soon!

Covid19Positive

I see a lot of people on here buying around the same types of stuff (canned beans, rice, etc). Just remember to know how to cook each item you buy and

PandemicPrep

Based off of everything I'm reading, hospitals are likely going to have to turn people away at some point.

Is this post from PandemicPreps or Covid19Positive subReddit?

For those that have tested positive I hope every single one of you feels better soon!

Covid19Positive

I see a lot of people on here buying around the same types of stuff (canned beans, rice, etc). Just remember to know how to cook each item you buy and

PandemicPrep

Based off of everything I'm reading, hospitals are likely going to have to turn people away at some point.

???

ROADMAP: Text Mining, Processing

Data Scarping

Text Normalization/Processing

- *Lower casing*
- *Removing puctuations*
- *Revoving special characters*
- *Handeling emojis and emoticons*
- *Stop word removals*
- *Common word removal*
- *Rare word removal*
- *Spelling correction*
- *Removing URLs and HTML tags*
- *Tokenization*
- *Stemming*
- *Lemmatization*

For those that have tested positive I hope every single one of you feels better soon!\n\n54

~~For those that have~~ tested positive ~~I~~ hope every single one ~~of you~~ feels better soon~~!\n\n54~~

test posit hope everi singl one feel better soon

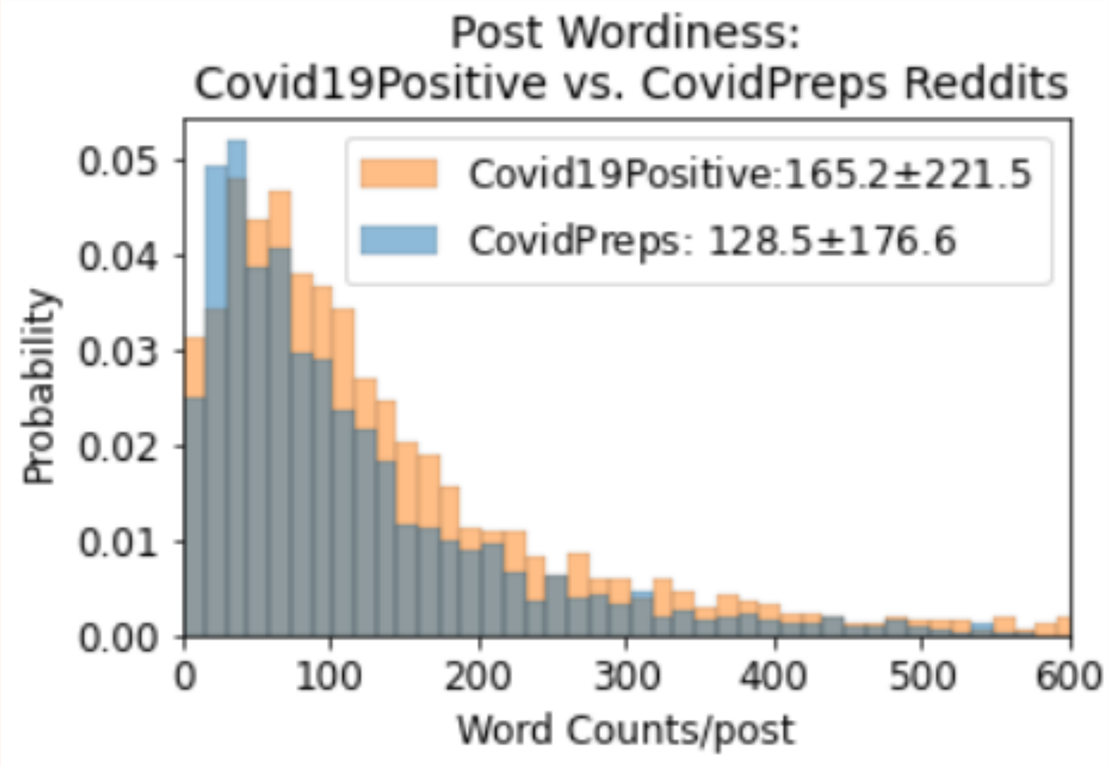
Count Characters

Raw	**23,000,000**
Processed	**12,000,000**

ROADMAP: Exploratory Data Analysis

Scarping

Text Normalization/Processing



Main outcome: Posts in Covid12Positive have more words.

Exploratory Data Analysis

ROADMAP: Modeling

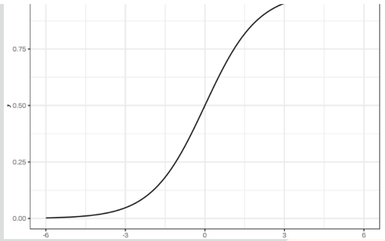
Scarping

Text Normalization/Processing

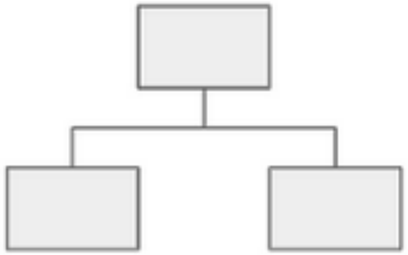
Exploratory Data Analysis

Supervised Learning Classification Models

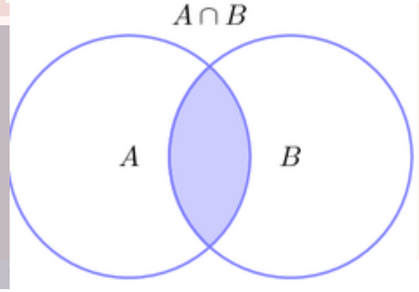
- **Regression**
  - Logistic Regression
  - ElasticNet
  - L1/Lasso
  - L2/Ridge



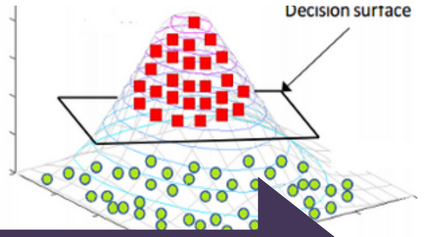
- **Ensemble**
  - Decision Tree
  - Extra Trees
  - Random Forest
  - Bagging



- **Bayes Theorem**
  - Multinomial Naive Bayes



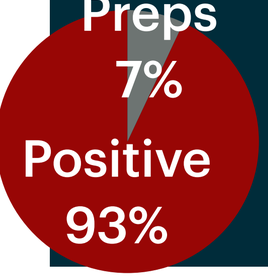
- **Support Vector Machine**



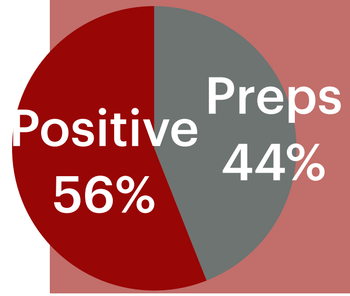
1. Post Classification Models
2. Sentiment Analysis

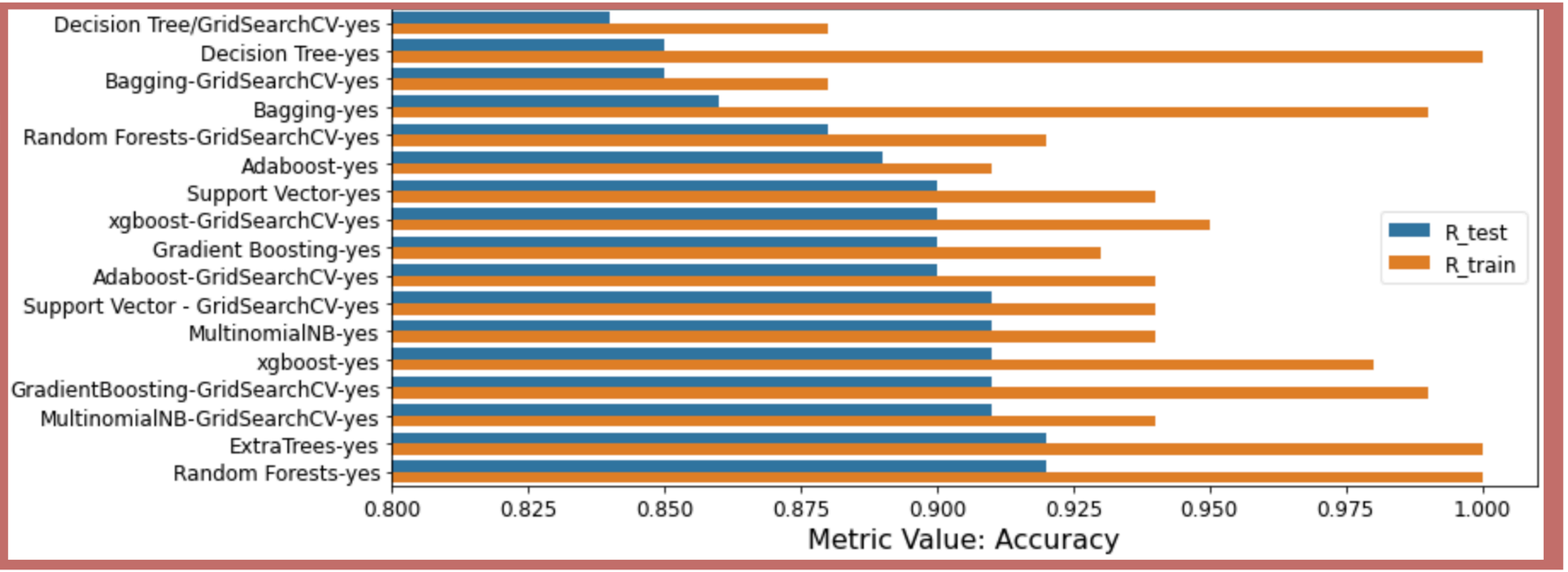
COMPARE DIFFERENT MODELS: ACCURACY

- Variance: LOW
- Bias: LOW
- RandomOverSampler included
- Covid19Positive: ~30,000 posts
  PandemicPreps: ~2,300 posts
- Features: ~19,500



- Variance: High
- Bias: LOW
- RandomOverSampler included
- Covid19Positive: ~2,900 posts
  PandemicPreps: ~2,300 posts
- Features: ~11,000





LOGESTIC REGRESSION GridSerachCV: OUTPUT

Baseline

Model: Logistic Regression
Input: CountVectorizer
Balanced: No / RandomOverSam

INPUTS:

Grids:
'C':
0.01,1.0, 10,50,60,65,70,
75,80,85,90,100 100, 150
'class_weight':
'balanced', 'None'
'penalty':
'l2','l1'

OUTPUT:

'C': 85,
'class_weight': 'balanced',
'penalty': 'l2',
'solver': 'lbfgs'

Recall:

What proportion of actual positives was identified correctly?

$$\text{Precision} = \frac{TP}{TP + FP}$$

Precision:

What proportion of positive identifications was actually correct?

$$\text{Precision} = \frac{TP}{TP + FP}$$

F1-Score:

A way of combining the precision and recall of the model

$$F1 = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

Accuracy:

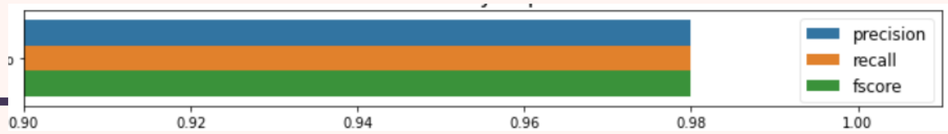
Number of correctly categorized examples divided by total

$$\text{accuracy} = \frac{\text{tp} + \text{tn}}{\text{tp} + \text{tn} + \text{fp} + \text{fn}}$$

$$precision = \frac{tp}{tp + fp}$$

$$recall = \frac{tp}{tp + fn}$$

$$F1 = 2 \times \frac{Precision * Recall}{Precision + Recall}$$



LOGESTIC REGRESSION GridSerachCV: OUTPUT

Model: Logistic Regression
Input: CountVectorizer
Balanced: No / RandomOverSampler

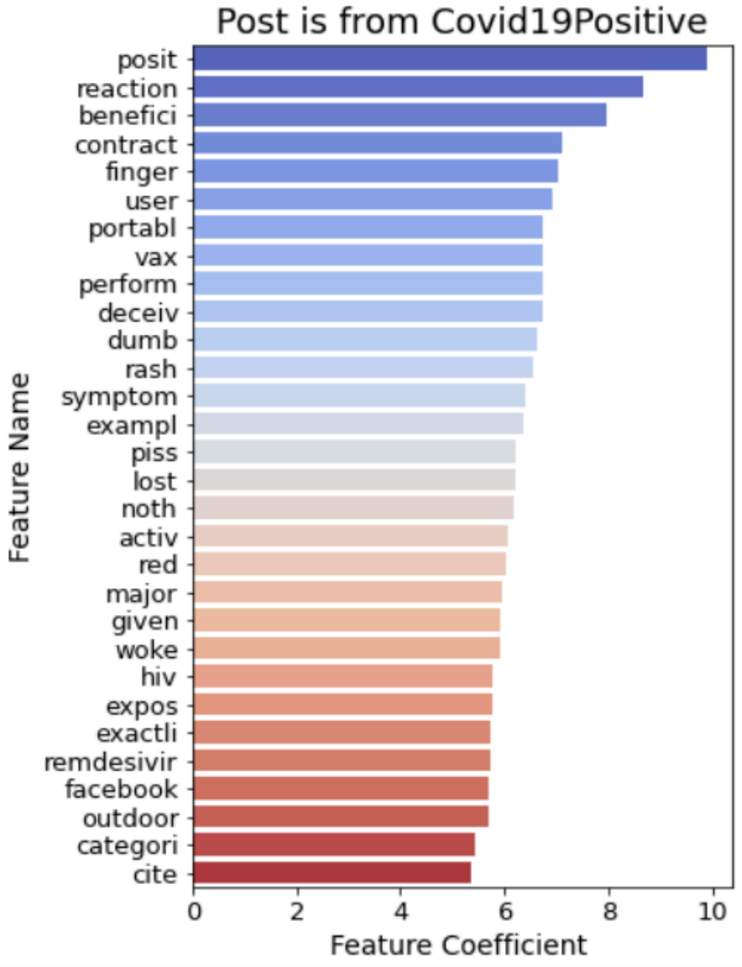
INPUTS:

Grids:

```
'C':
  0.01,1.0, 10,50,60,65,70,
  75,80,85,90,100 100, 150
'class_weight':
  'balanced', 'None'
'penalty':
  'l2','l1'
```

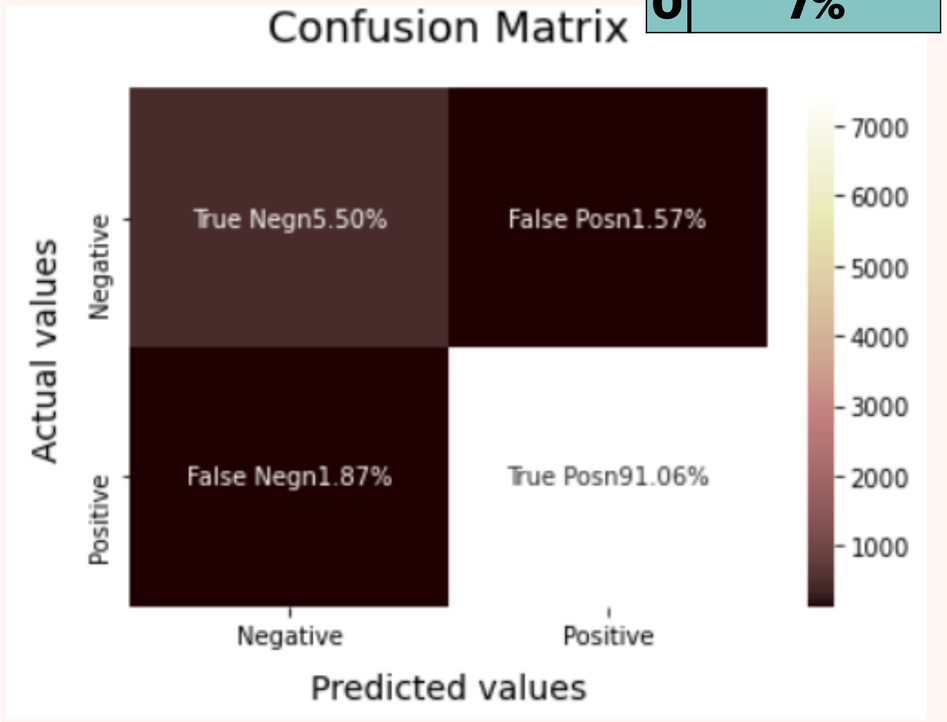
OUTPUT:

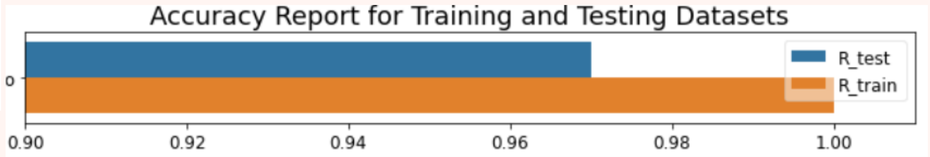
```
'C': 85,
'class_weight': 'balanced',
'penalty': 'l2',
'solver': 'lbfgs'
```

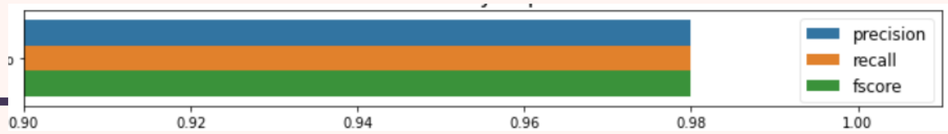


Baseline

1	93%
0	7%







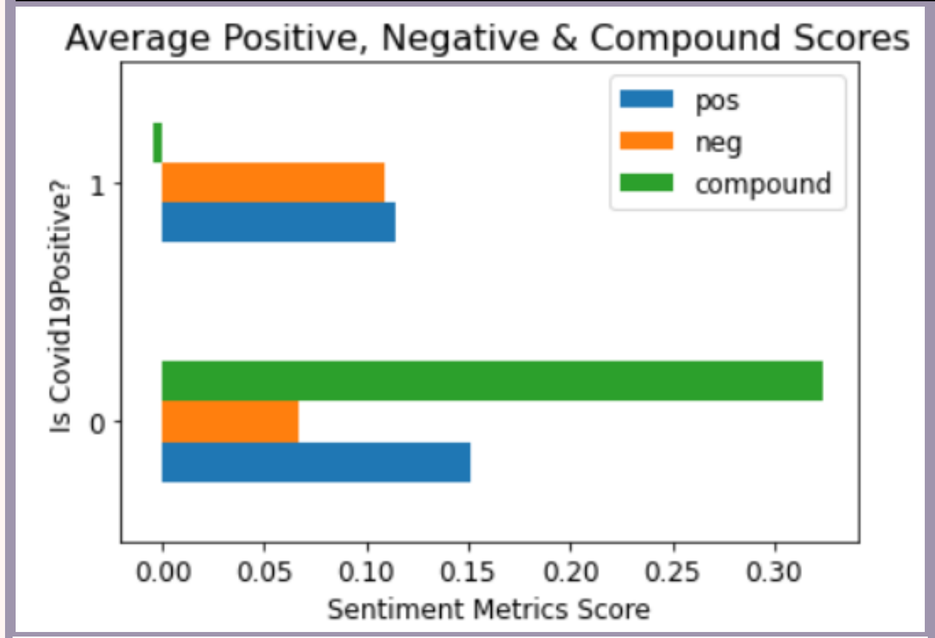
$$precision = \frac{tp}{tp + fp}$$

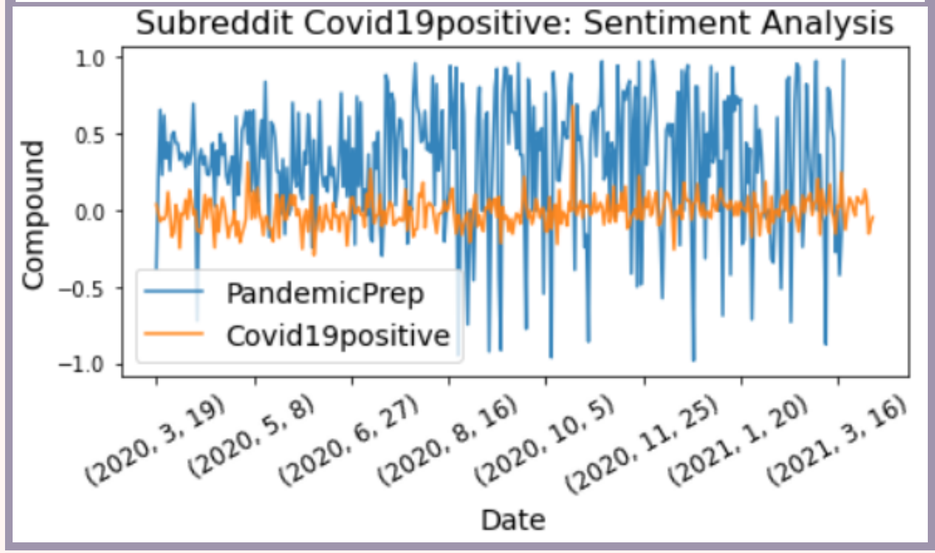
$$recall = \frac{tp}{tp + fn}$$

$$F1 = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

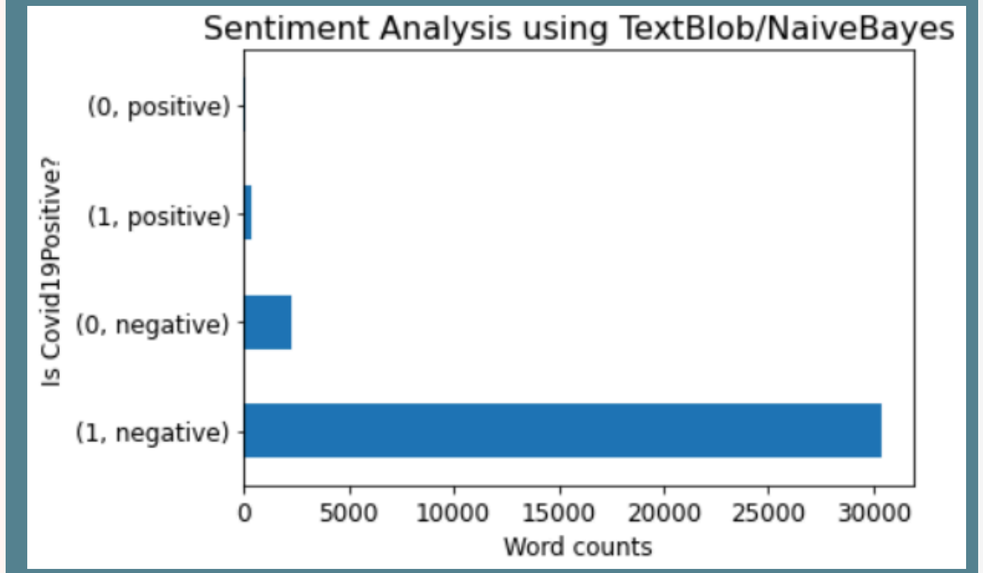
SENTIMENT ANALYSIS: 2020/3 — 2021/3

VADER Sentiment Analyzer





TextBlob/NaiveBayes




```
positive_vocab = ['awesome',
                  'outstanding',
                  'fantastic',
                  'terrific',
                  'good',
                  'nice',
                  'great',
                  'liked',
                  'happy']
```

```
negative_vocab = [
    'sick',
'positive',
'Rhinitis',
'Gastroenteritis',
'Coronaryheartdisease',
'Cough',
'Sputum',
'Runnynose',
'Nasalobstruction',
'Sneezing',
'Sorethroat',
'Dyspnea',
'Shortnessofbreath',
'Chesttightness',
'Chestpain',
'Palpitations',
'Fever',
'Chills',
'Fatigue',
'Myalgia',
'Lumbago',
'Jointpain',
'Headache',
'Dizziness',
'Vertigo',
'Abdominalpain',
'Diarrhea',
'Vomiting',
'Conjunctivalcongestion',
'Itchyeyes',
'Eyespain']
```

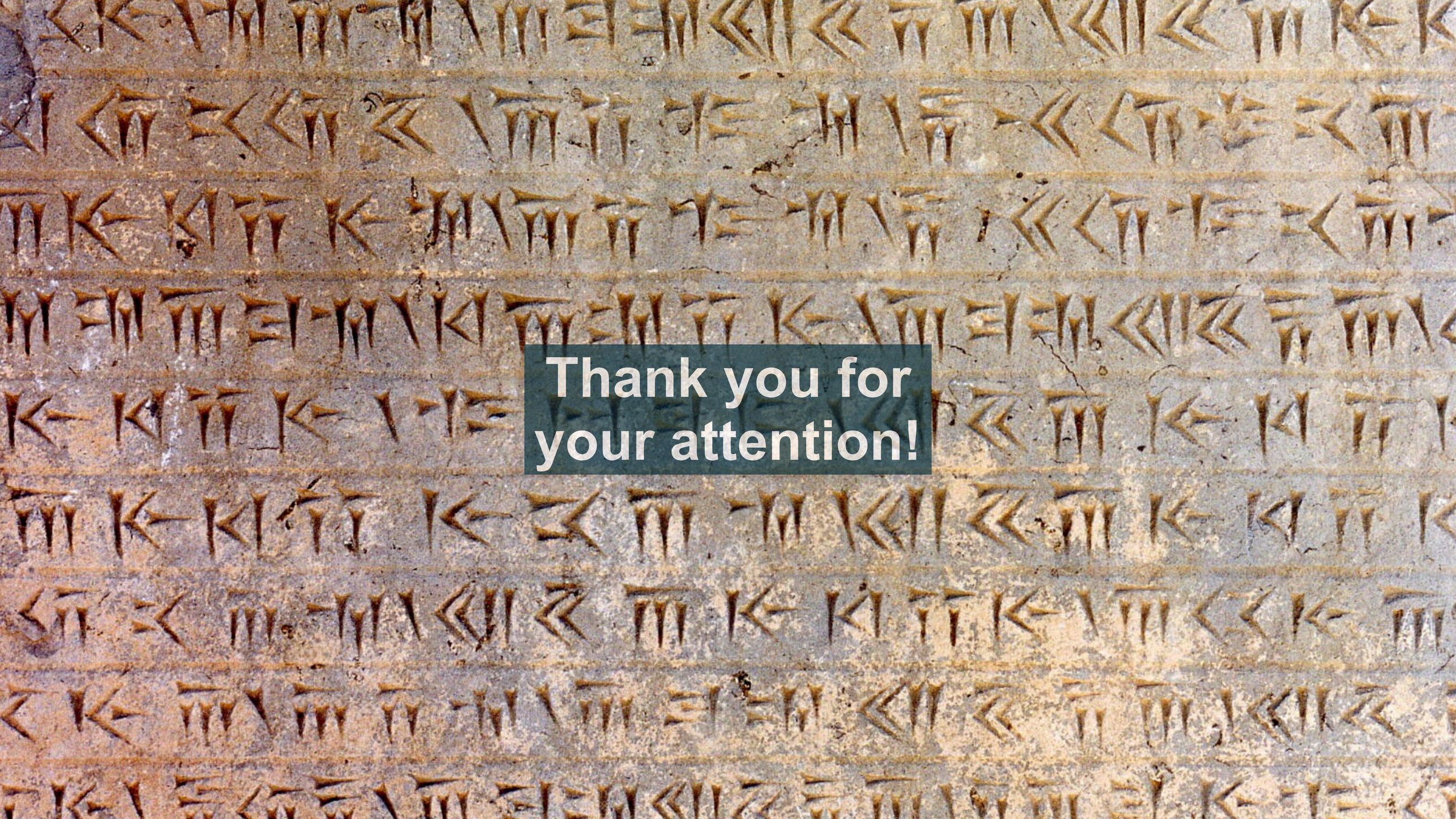
Conclusion

Logistic Regression is found to be the best model for classification because.....

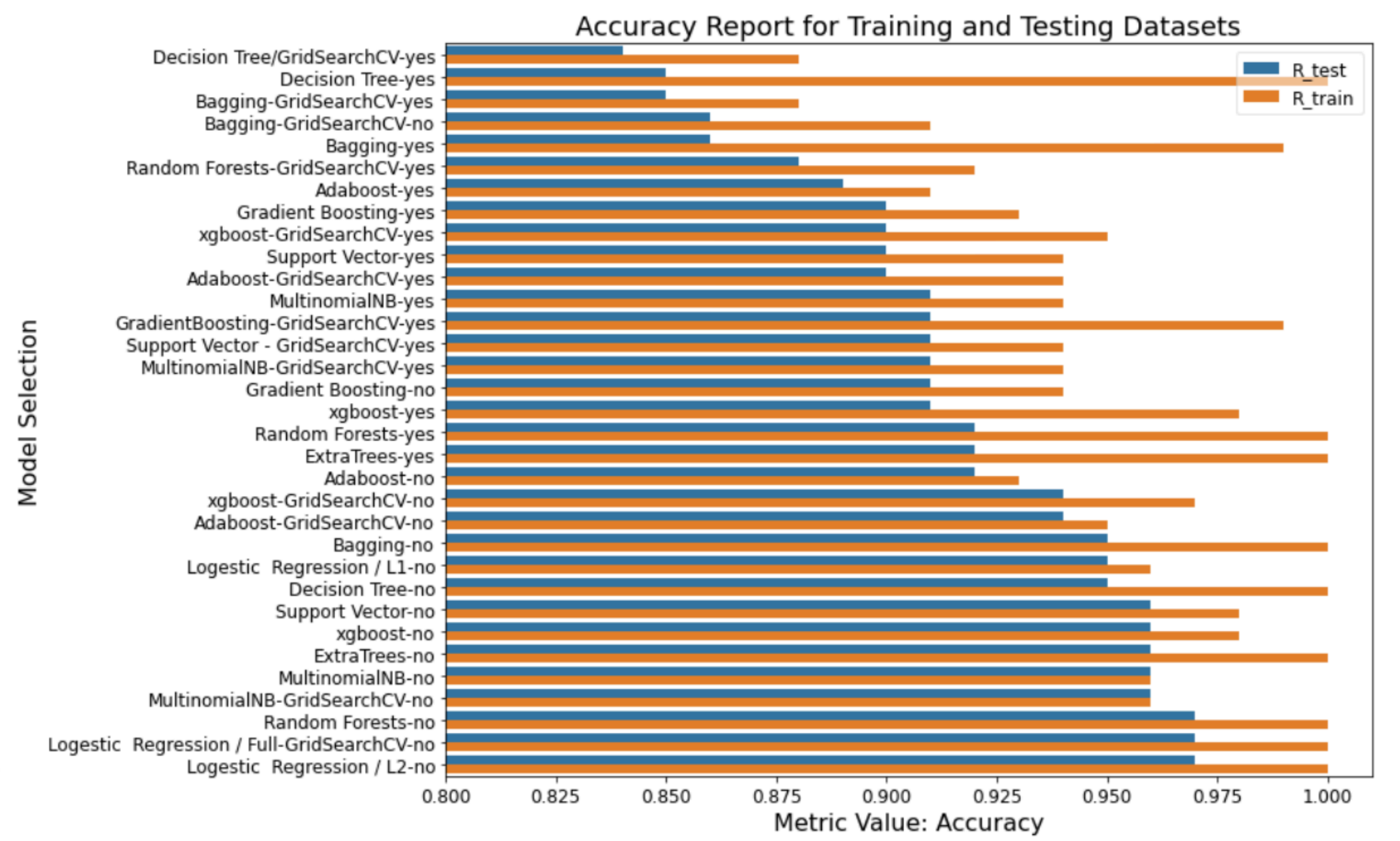
- ✓ Provides the highest accuracy scores, ~**99**% and ~**96**% for training and testing datasets
- ✓ Works great with ultra-imbalanced samples (~**93**% vs. ~**7**%)
- ✓ High rates for *true positive* (91.06% out of 93%) and for *true negative* (5.5% out of 7%)
- ✓ Low scores for *false positive* (1.57%) and *false negative* (1.87%)
- ✓ High scores for precision and recall (~98%)

In addition, this model is.....

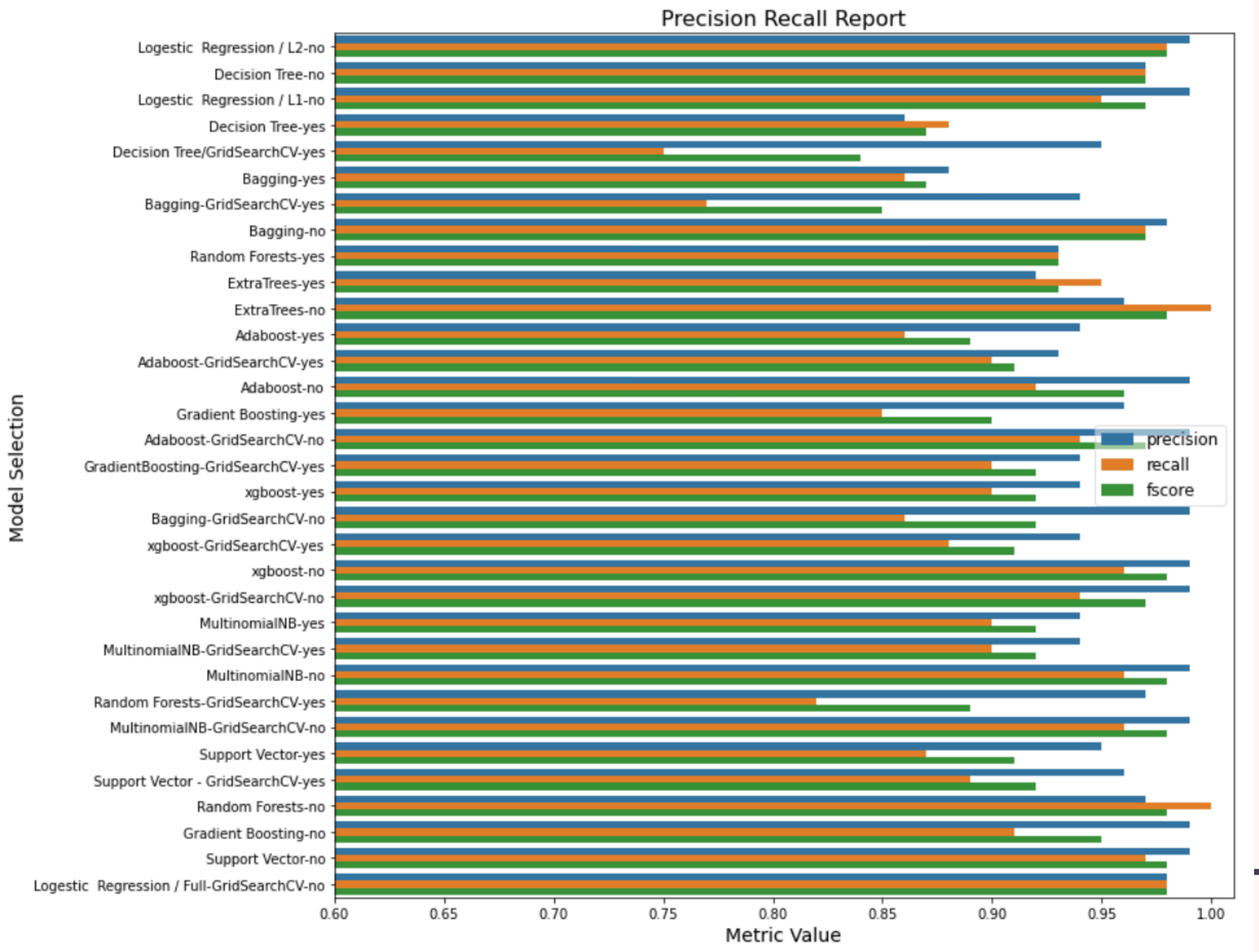
- ✓ Interpretable
- ✓ Optimizable coefficients to reduce variance and bias
- ✓ Capable to use different generalization methods i.e., Lasso, Ridge, ElasticNet
- ✓ Tunable parameters, solvers, and penalty functions for multiple cases
- ✓ Works best with both large and small datasets

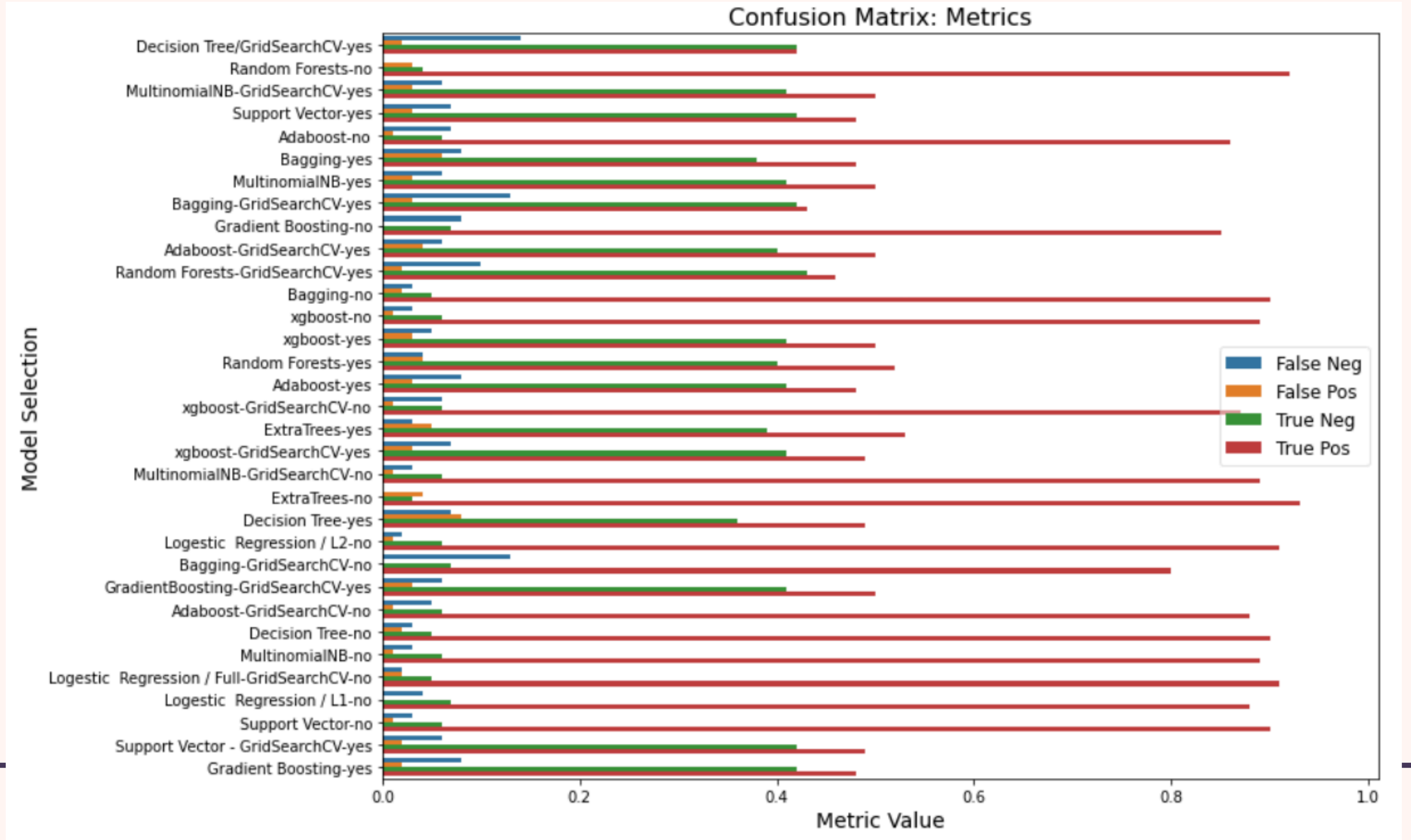

Thank you for
your attention!

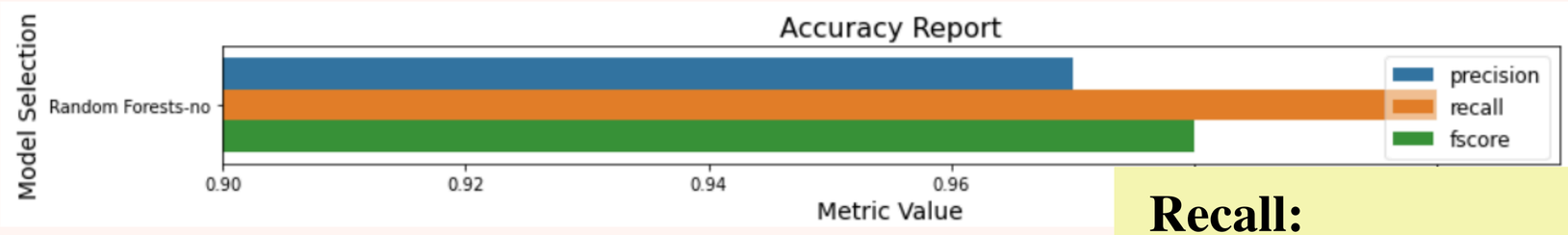
BACKUP 1: ACCURACY REPORT FOR ALL MODELS



BACKUP 2: PRECISION RECALL FOR ALL MODELS




Confusion Matrix: Metrics
Model Selection
Metric Value
False Neg
False Pos
True Neg
True Pos
Decision Tree/GridSearchCV-yes
Random Forests-no
MultinomialNB-GridSearchCV-yes
Support Vector-yes
Adaboost-no
Bagging-yes
MultinomialNB-yes
Bagging-GridSearchCV-yes
Gradient Boosting-no
Adaboost-GridSearchCV-yes
Random Forests-GridSearchCV-yes
Bagging-no
xgboost-no
xgboost-yes
Random Forests-yes
Adaboost-yes
xgboost-GridSearchCV-no
ExtraTrees-yes
xgboost-GridSearchCV-yes
MultinomialNB-GridSearchCV-no
ExtraTrees-no
Decision Tree-yes
Logestic Regression / L2-no
Bagging-GridSearchCV-no
GradientBoosting-GridSearchCV-yes
Adaboost-GridSearchCV-no
Decision Tree-no
MultinomialNB-no
Logestic Regression / Full-GridSearchCV-no
Logestic Regression / L1-no
Support Vector-no
Support Vector - GridSearchCV-yes
Gradient Boosting-yes
0.0
0.2
0.4
0.6
0.8
1.0



OM FORESTS-NO

Recall:

What proportion of actual positives was identified correctly?

$$\text{Precision} = \frac{TP}{TP + FP}$$

Precision:

What proportion of positive identifications was actually correct?

$$\text{Precision} = \frac{TP}{TP + FP}$$

F1-Score:

A way of combining the precision and recall of the model

$$F1 = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

Accuracy:

Number of correctly categorized examples divided by total

$$\text{accuracy} = \frac{\text{tp} + \text{tn}}{\text{tp} + \text{tn} + \text{fp} + \text{fn}}$$