Classification project:

Predicting covid test results

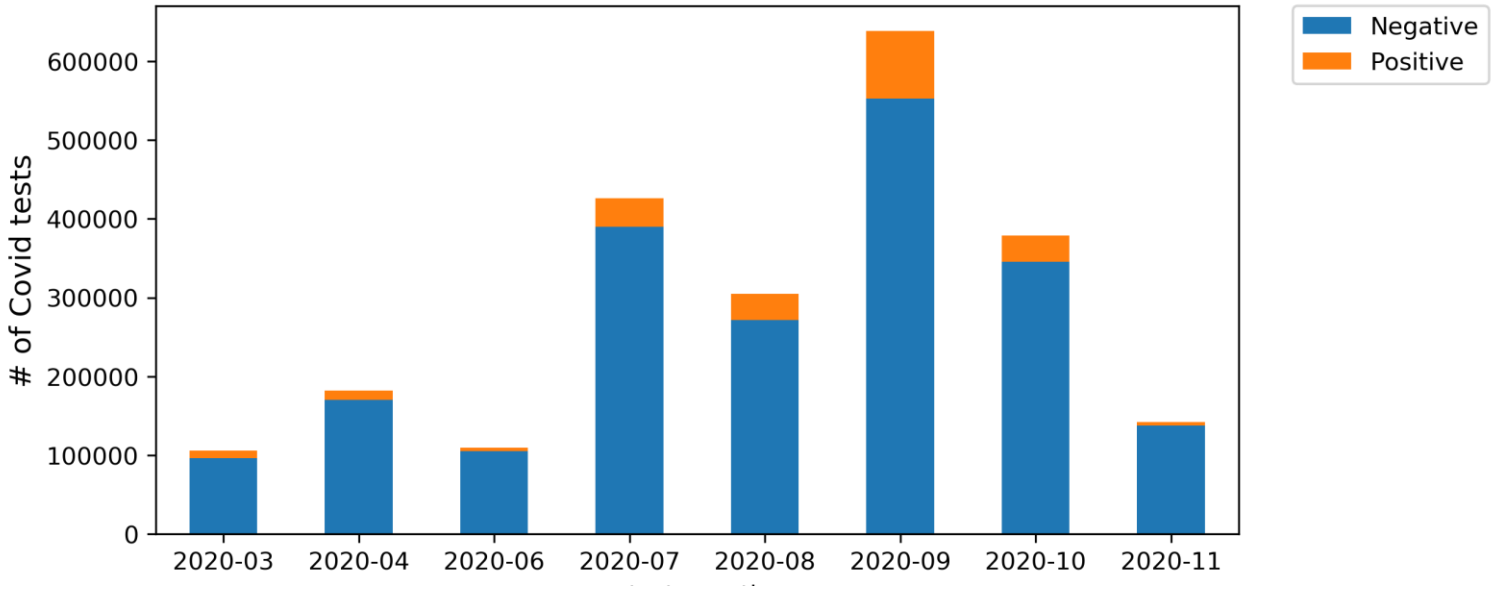
Ming Tang

May 14, 2021

Outline

- Background
  - Predict the covid test results (positive/negative)
  - Optimize the recall (=TP/(TP+FN)), i.e. minimize the false negatives (cases who have COVID-19 but are tested negative)
- Approach
  - Data collection: from this published article and GitHub.
  - Data exploration: pandas, numpy, matplotlib
  - Classification: scikit-learn, xgboost
  - Application: Streamlit, Heroku
- Conclusions
  - Models can achieve ~ 0.6 recall without sacrificing the overall accuracy

A quick look at the data



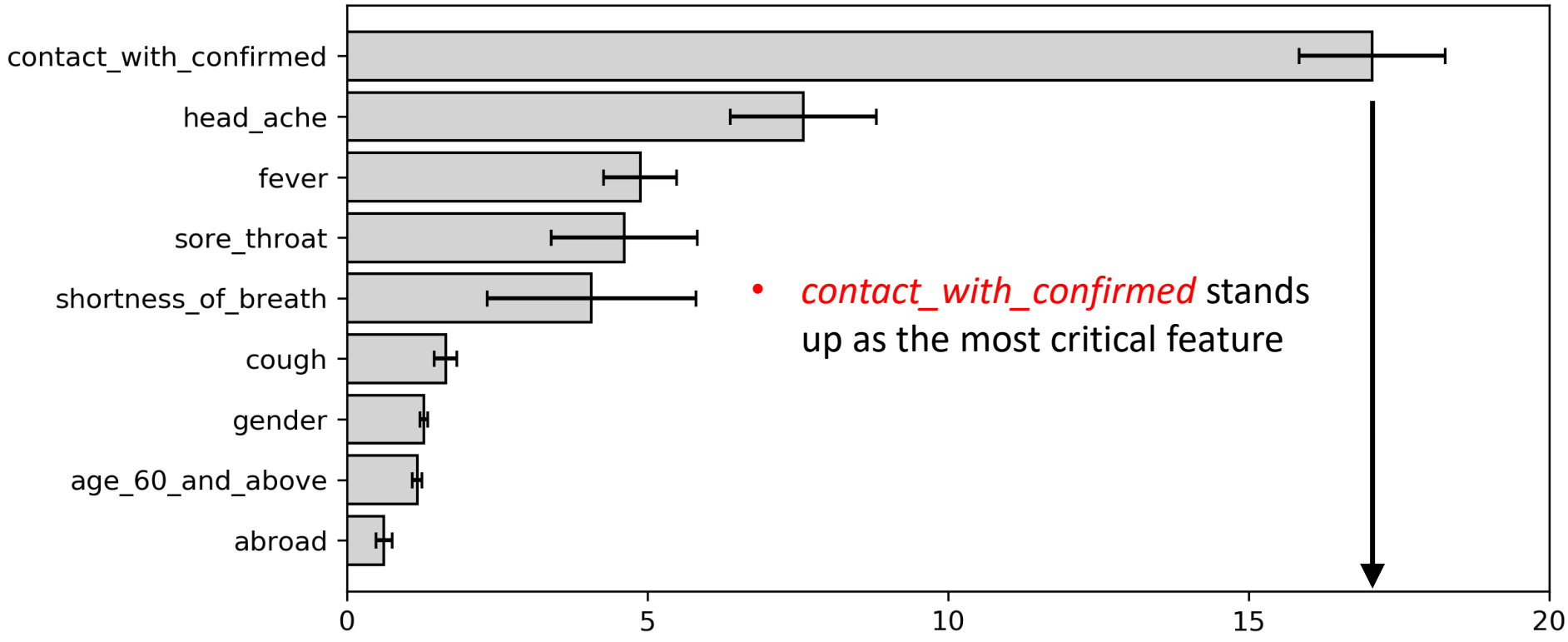
9 features (all binary)

1 target
(0-negative, 1-positive)

cough	fever	sore_throat	shortness_of_breath	head_ache	age_60_and_above	gender	contact_with_confirmed	abroad	corona_result
1	0	0	0	0	1	1	0	0	0
1	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	1	0
1	0	0	0	0	1	0	0	0	0

3 million rows

Feature importance

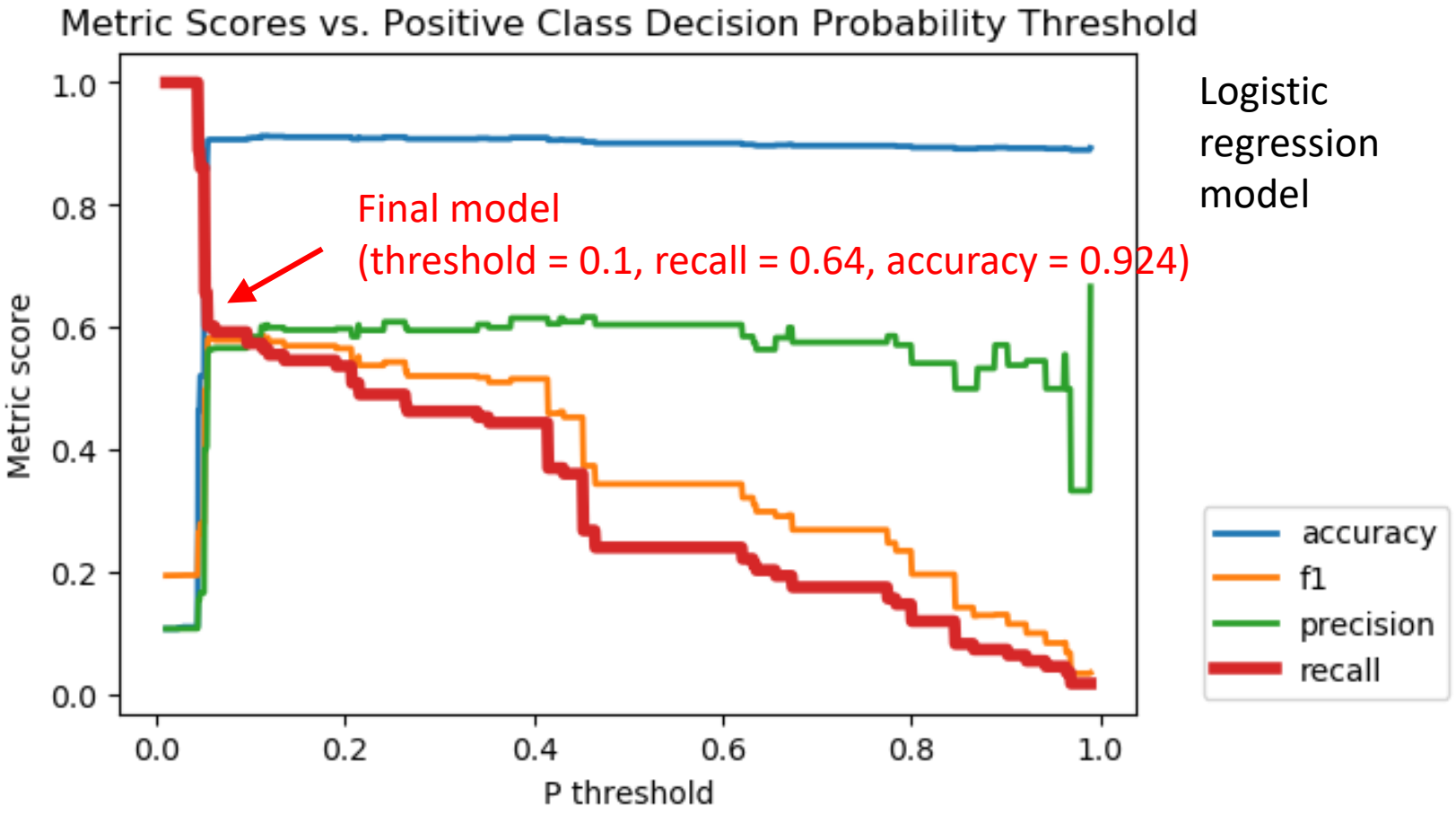


- *contact_with_confirmed* stands up as the most critical feature

A one unit change in *contact_with_confirmed* corresponds to 18 times higher in the odds of testing positive

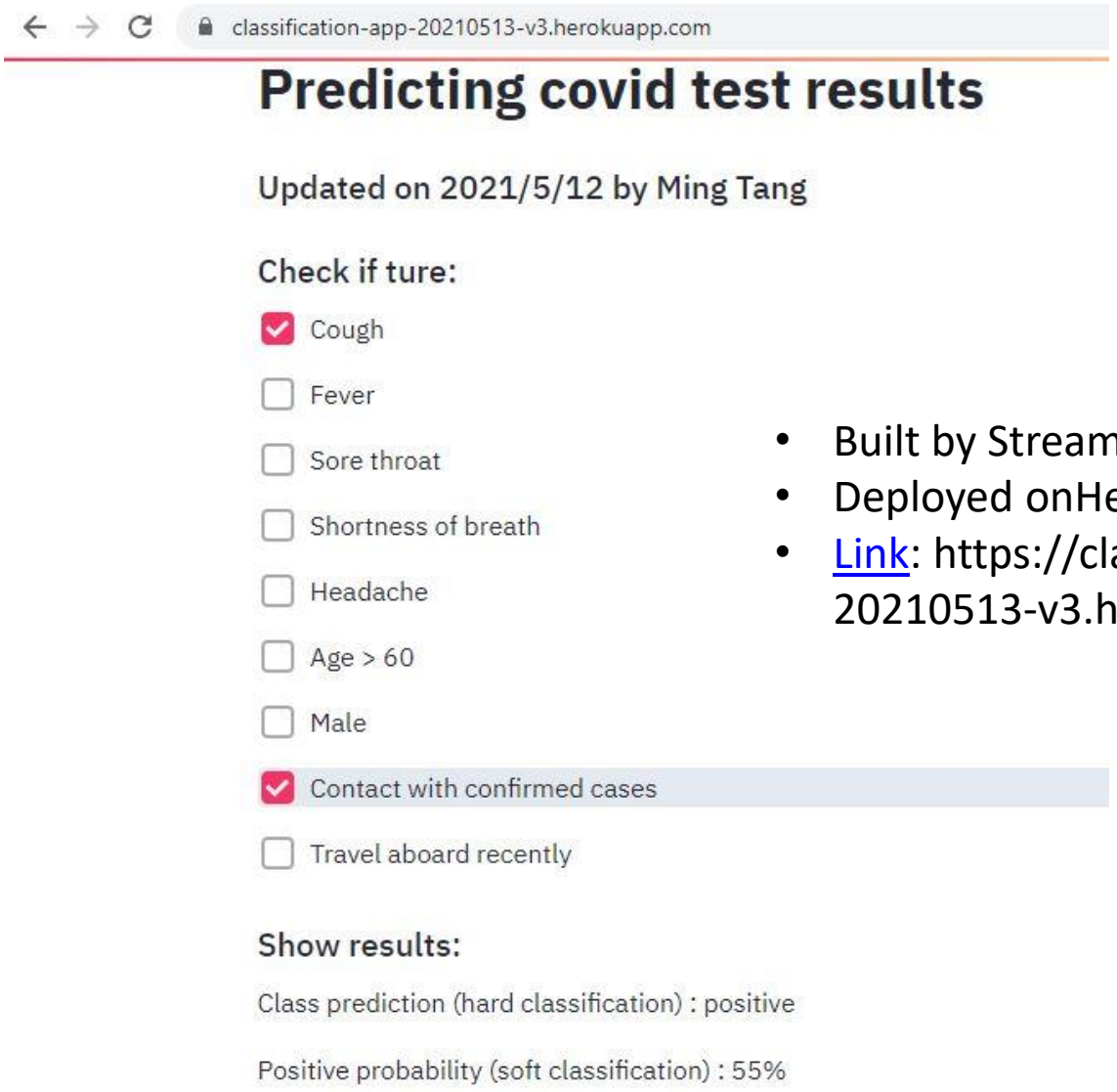
- Algorithm: Logistic Regression,
- Data: 100,000 samples, 1000 bag, each bag contains 100,000 rows with replacement
- Error bars: generated by bootstrapping and 95% confidence interval.

Model performance



- Model performance depends strongly on the threshold
- The predictability on the recall is rather limited, likely due to asymptomatic cases (90% of all cases are asymptomatic but still have 4% positive rate)
- The only way to get the perfect recall is to predict all cases as positive

Demo



- Built by Streamlit
- Deployed onHeroku
- Link: https://classification-app-20210513-v3.herokuapp.com/

Summary

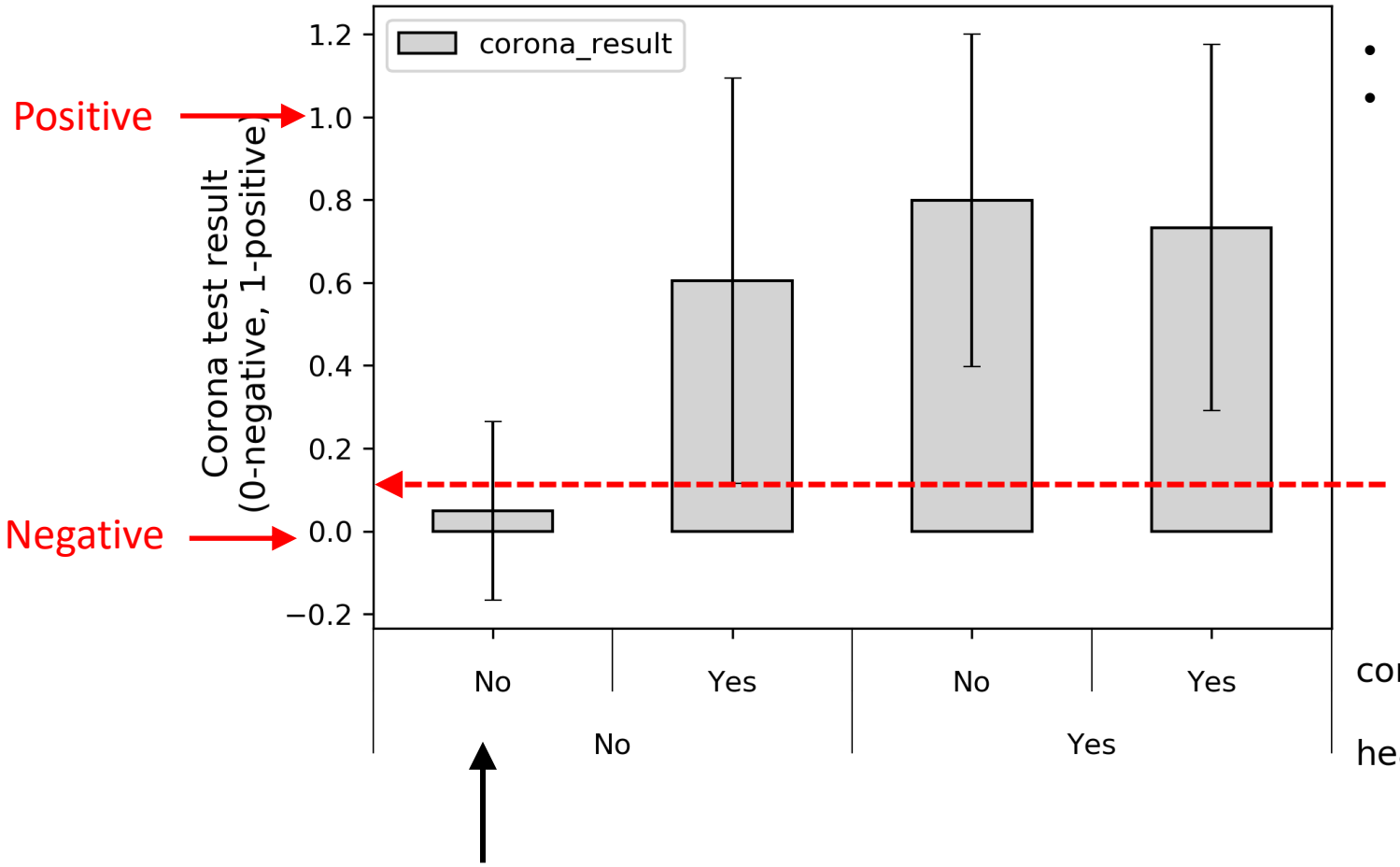
- Predicting covid test results is possible
- Recall is rather limited (~ 0.6), likely due to the asymptomatic cases

Thank you!

More detailed description and scripts are in my GitHub:
https://github.com/tangming2008/Project_ShortImmersive_Classification

Backup slides

Data exploration
(Why limited predicted recall?)



- Bar: average
- Error bar: standard deviation

Baseline on all observations: 10% positive

- ~ 2 million cases without head_ache of *contact_with_confirmed*.
- For this group, the positive test rate is as high as 5% (it is lower than the 10% baseline)