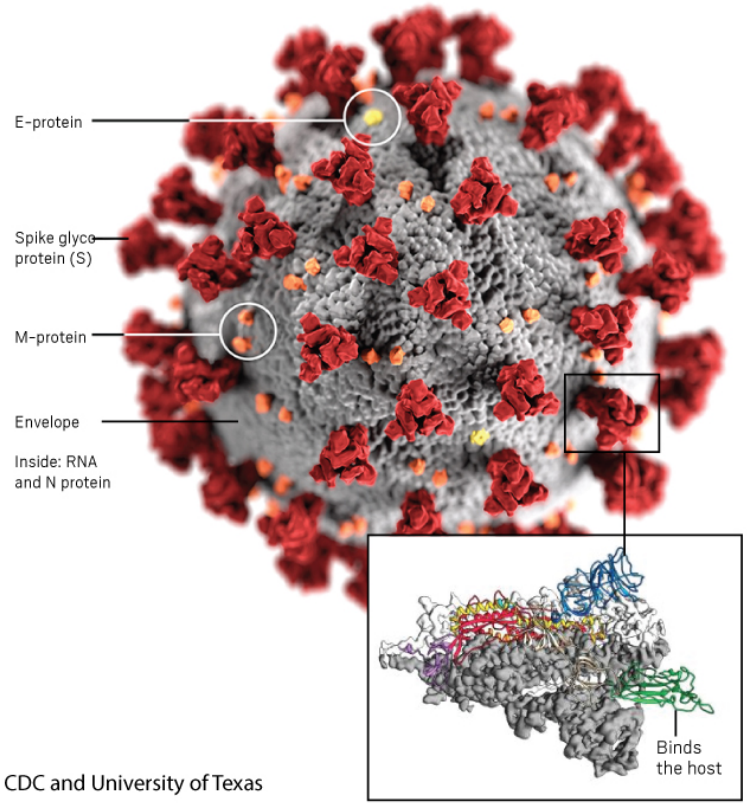
COVID-19 Diagnosis using Raman Spectroscopy



Author: Isaac Ghebregziabher

Capstone Project One

May 18, 2021

Springboard

Drop features with all values equal to 0

- No missing values in dataset
- 9 features wave-numbers with 0 intensity value
- Drop null features (treated as missing)

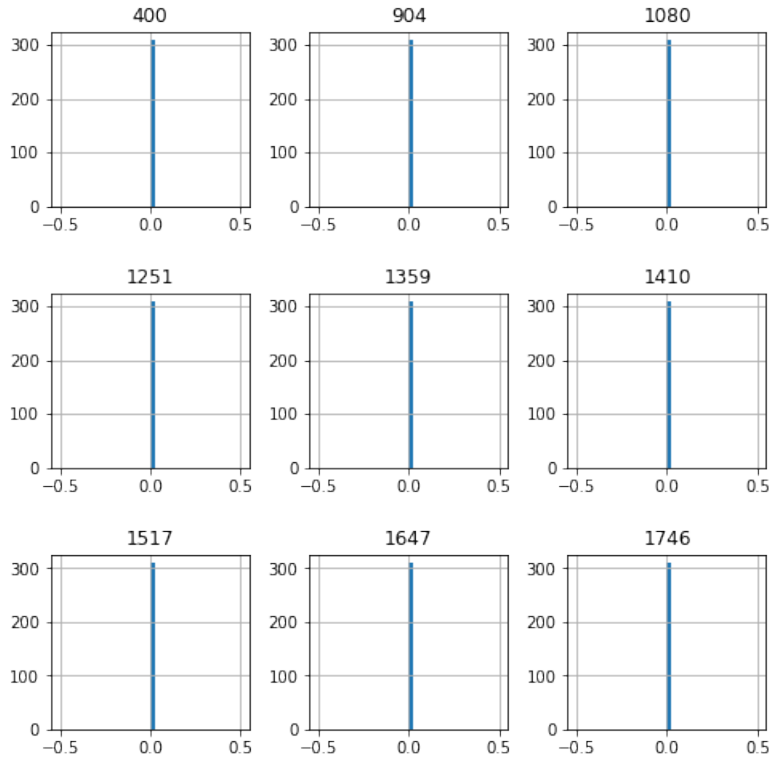


Figure: Single valued features (dropped).

Remaining features have normal distribution

- No concern on feature distributions.
- Most close to normal.
- little skew on several features.

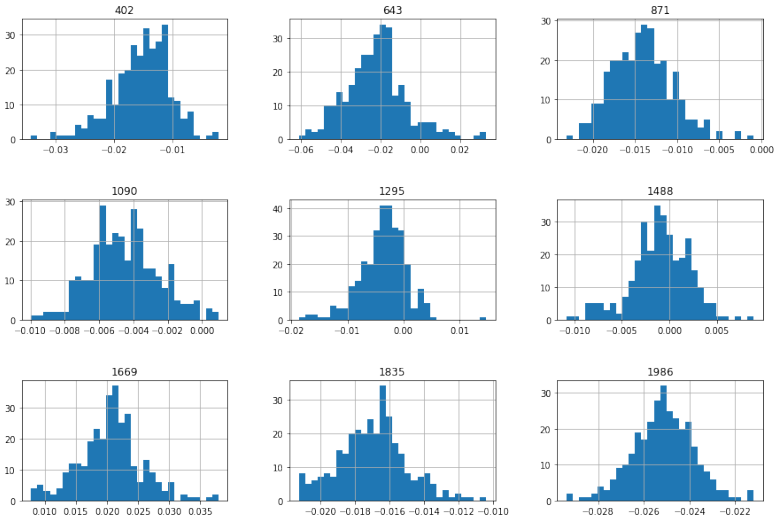


Figure: Close to normal feature distributions.

Dataset is balanced

- No class imbalance issues
- Dataset is balanced with ≈ 50:50 class ratio.

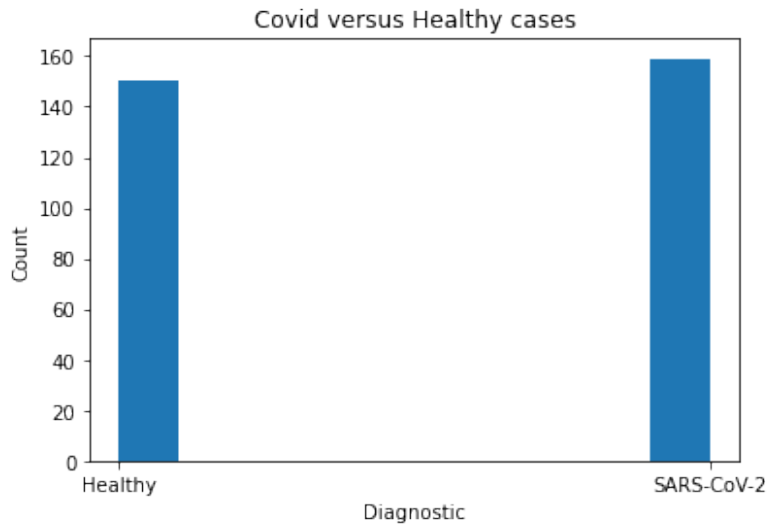
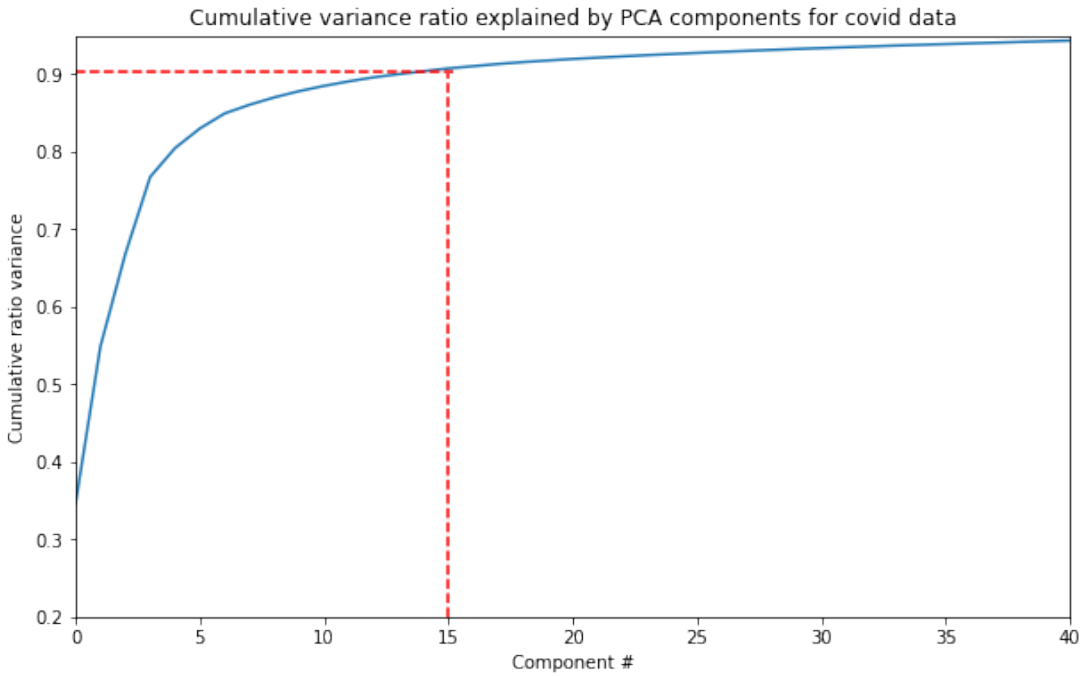


Figure: ≈ 50 : 50 COVID to Healthy class ratio.

Principal component analysis – feature reduction

- Over 90% data variance explained with 15 components.
- Feature reduction to 15 from 900!



Model training and accuracy

Three models considered:

- Decision tree
- Logistic regression
- Random forest

Model	Training accuracy
Decision Tree	1
Logistic regression	1
Random forest	1

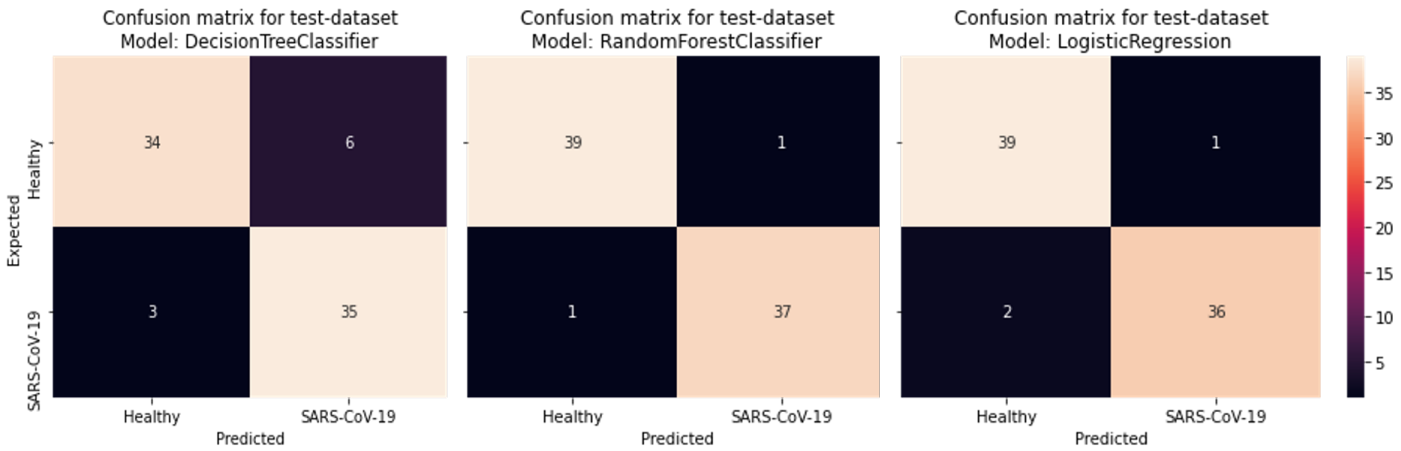
All models seem to over-fit.

Need to be tested with the test split.

Random forest classifier is the best performing model

Though all models persisted with good accuracy:

- Random fores performs the best
- 97% classification accuracy
- Random forest chosen for deployment



Accuracy = 0.88

Accuracy = 0.97

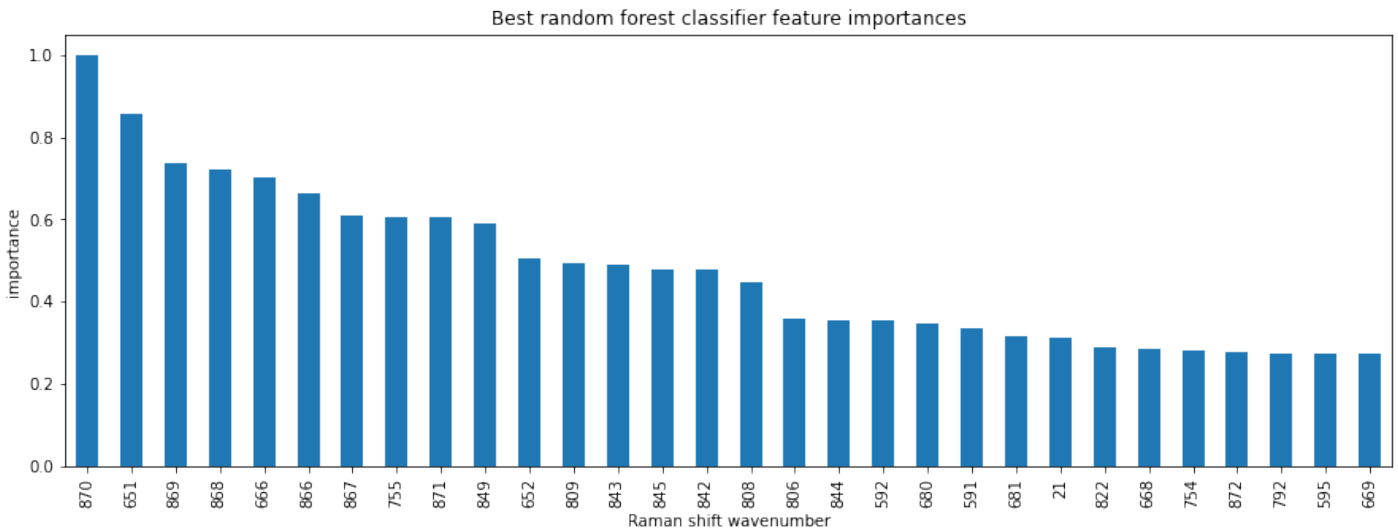
Accuracy = 0.96

Springboard

Important features with high predictive power

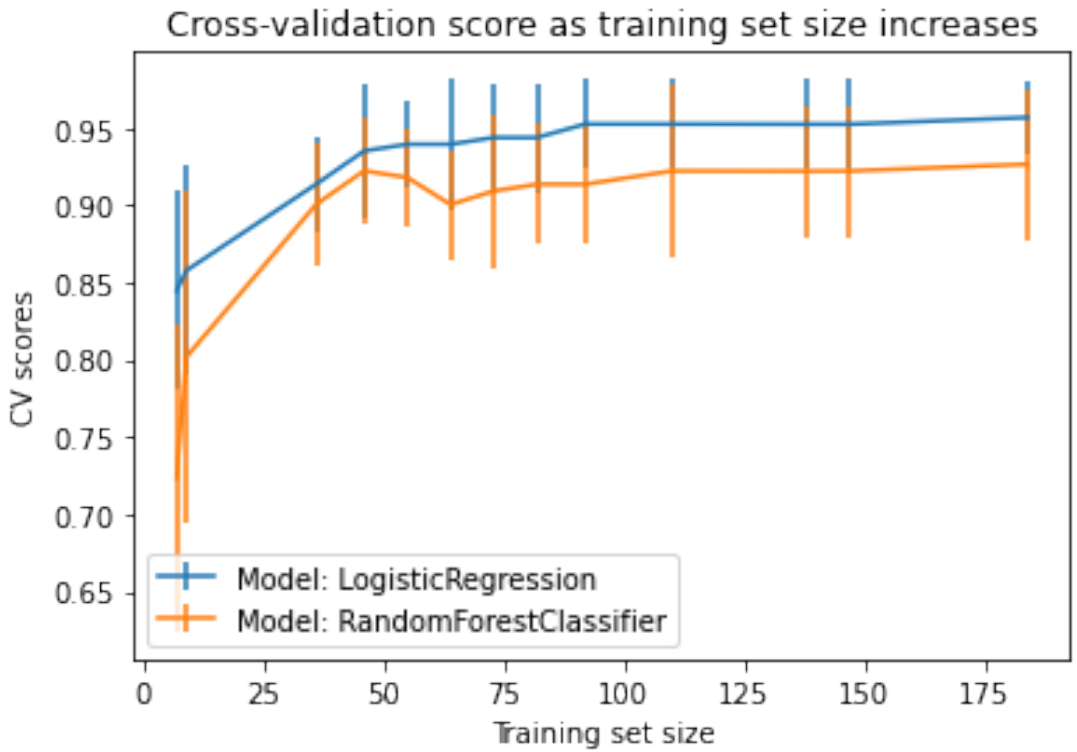
15 features out of 901 are the most important.

- Wavenumber in range [650, 870] has high predictive power
- feature 870 has the highest predictive power



Do we need more data to enhance model performance?

Model accuracy saturates well before the end of available data.



No need for more data.

Summary

- We obtained Raman spectroscopy for COVID detection experimental dataset from Kaggle.
- To get insight We applied data cleansing, wrangling, and exploring techniques.
- We compared and contrasted the performance of Logistic regression, decision tree, and random forest classification models
- We find Random forest to be the best with diagnostic accuracy of 97%

Acknowledgement

Springboard mentor: Yuxuan Xin

for time generous and insightful discussions