National Institute of Technology Raipur

Applied Machine Learning
Term Project

Decision Trees using Random Forest Algorithm

By

Kunal Sachdeva

Roll No. 19115045

(6th Semester Computer Science & Engineering 2020-21)

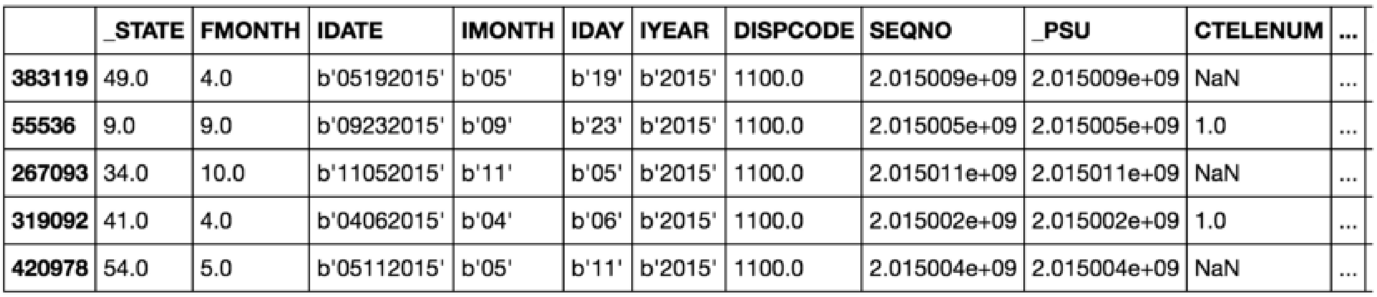
Behavioral Risk Factor Surveillance System

The goal of the Behavioral Risk Factor Surveillance System (BRFSS) is to gather standard, state-specific data on preventive health practices and risk behaviors in the adult population that are associated to chronic diseases, injuries, and preventable infectious diseases. Tobacco use, health-care coverage, HIV/AIDS education or prevention, physical activity, and fruit and vegetable eating are among the factors assessed by the BRFSS. A telephone survey is used to obtain data from a random sample of adults (one per household).

The Behavioral Risk Factor Surveillance System (BRFSS) is the nation's most comprehensive system of health-related telephone surveys, collecting data on U.S. citizens' health-related risk behaviors, chronic health issues, and use of preventative services from across the country. BRFSS began collecting data in 15 states in 1984 and has since expanded to include all 50 states, the District of Columbia, and three US territories. Each year, the BRFSS conducts over 400,000 adult interviews, making it the world's biggest continually conducted health survey system. [1]

Dataset

There are a few hundred columns in the dataset. The problem I'm working on is a binary classification challenge that aims to predict someone's health. Individuals' socioeconomic and lifestyle traits are the features, with a label of 0 indicating bad health and 1 indicating good health. The Centers for Disease Control and Prevention gathered this data. [1]



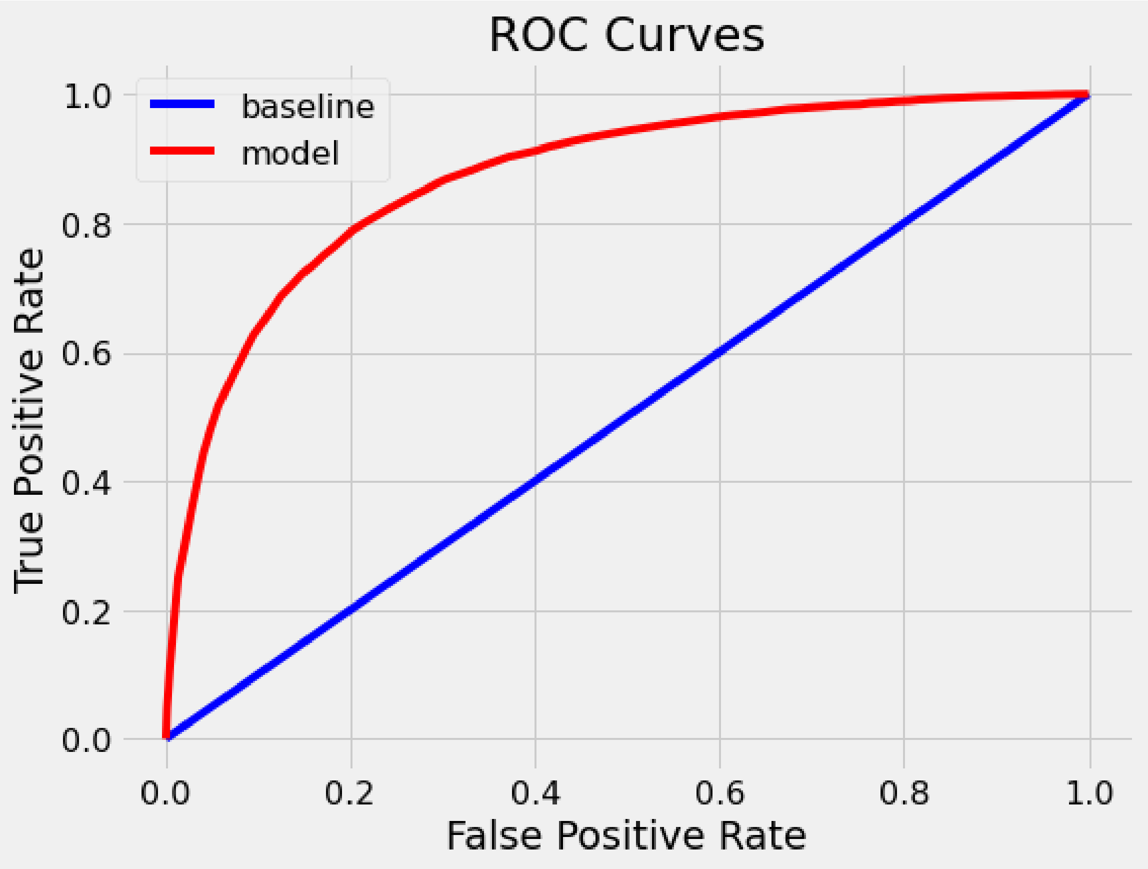
	_STATE	FMONTH	IDATE	IMONTH	IDAY	IYEAR	DISPCODE	SEQNO	_PSU	CTELENUM	...
383119	49.0	4.0	b'05192015'	b'05'	b'19'	b'2015'	1100.0	2.015009e+09	2.015009e+09	NaN	...
55536	9.0	9.0	b'09232015'	b'09'	b'23'	b'2015'	1100.0	2.015005e+09	2.015005e+09	1.0	...
267093	34.0	10.0	b'11052015'	b'11'	b'05'	b'2015'	1100.0	2.015011e+09	2.015011e+09	NaN	...
319092	41.0	4.0	b'04062015'	b'04'	b'06'	b'2015'	1100.0	2.015002e+09	2.015002e+09	1.0	...
420978	54.0	5.0	b'05112015'	b'05'	b'11'	b'2015'	1100.0	2.015004e+09	2.015004e+09	NaN	...

Sample of Data

Because this is an unbalanced classification problem, accuracy isn't the right statistic to use. Instead, use the Receiver Operating Characteristic Area Under the Curve (ROC AUC), a scale that ranges from 0 to 1 with a random estimate scoring of 0.5.

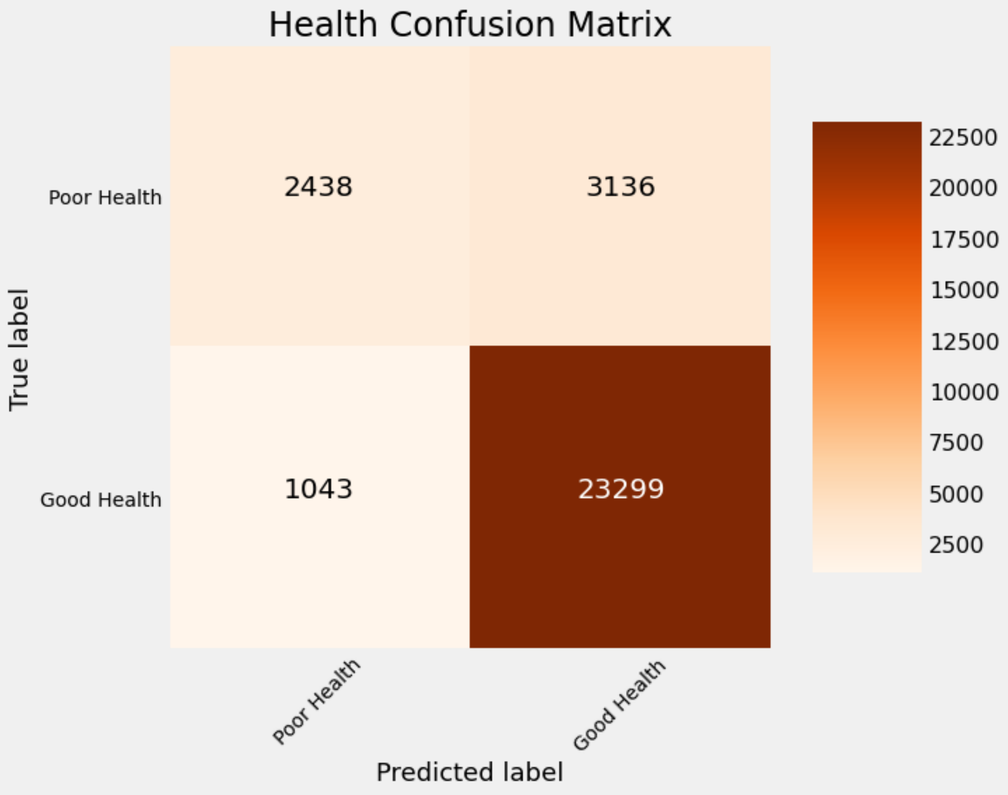
Results

The random forest's final testing ROC AUC was 0.87.



Random Forest ROC Curve

We may also plot the confusion matrix for the testing predictions as a diagnostic metric of the model.



This diagram depicts the model's correct predictions in the top left and bottom right corners, as well as the model's failed forecasts in the lower left and upper right corners.

References

1. https://www.kaggle.com/cdc/behavioral-risk-factor-surveillance-system