Sango Fon

Claremont graduate university

Reopening California

a proposal for the selection process of suitable counties for the gradual reopening of the California economy once the stay-at-home order is lifted

Table of Contents

[Business Understanding 2](#_Toc40896315)

[Data Understanding 3](#_Toc40896316)

[Data Collection 3](#_Toc40896317)

[Data Description 3](#_Toc40896318)

[Data Exploration 4](#_Toc40896319)

[Data Quality Verification 6](#_Toc40896320)

[Data Preparation 7](#_Toc40896321)

[Data Transformation 7](#_Toc40896322)

[Data Integration 9](#_Toc40896323)

[Data Modeling Techniques 10](#_Toc40896324)

[Modeling 12](#_Toc40896325)

[Data Mining Techniques 12](#_Toc40896326)

[Modeling Test Design 12](#_Toc40896327)

[Modeling 13](#_Toc40896328)

[Model Performance Assessment 13](#_Toc40896329)

[Evaluation 14](#_Toc40896330)

[Evaluate the model against business objectives and success measures 14](#_Toc40896331)

[next steps 15](#_Toc40896332)

[Lessons learned (10 points) 16](#_Toc40896333)

[Appendix 17](#_Toc40896334)

[Sources 17](#_Toc40896335)

[Decision Trees 18](#_Toc40896336)

[Logistic Regression 31](#_Toc40896337)

# Business Understanding

As the threat of COVID-19 continues to rise, the economy is quickly collapsing. The mandated “Stay at Home” order has resulted in even more devastating consequences. Business hours are becoming limited, workers are losing wages, small businesses are closing, and the population is becoming more dependent on government intervention and aid. The number of infections from COVID-19 continues to grow at a rapid pace. Due to these challenging times, the State of California must make a difficult decision on how to progressively and strategically lift the "Stay at Home" order for the population to return to a sense of normalcy. The objective is to determine a data-backed guideline to strategically choose the counties in California for reopening while continuing to prevent the spread of COVID-19. Furthermore, several success criteria for achieving the project objectives include the following: The project must be completed by May 20, 2020 to be taken into consideration. The outcome of this project should include a list data driven rules that decision makers should consider when crafting their strategic guidelines to lift the Stay at home orders.

PING ME IF YOU WANT A COPY OF THE PROJECT.

# Appendix

### Sources

<https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_daily_reports>

<https://raw.githubusercontent.com/JieYingWu/COVID-19_US_County-level_Summaries/master/data/counties.csv>

https://www.countyhealthrankings.org/explore-health-rankings/rankings-data-documentation

<https://svi.cdc.gov/>

<https://www.worldometers.info/coronavirus/>

<http://www.healthdata.org/covid/data-downloads>

<https://data.chhs.ca.gov/dataset/california-covid-19-hospital-data-and-case-statistics/resource/6cd8d424-dfaa-4bdd-9410-a3d656e1176e?view_id=b23b0158-a85d-4bf2-95b1-96f7556f7342>

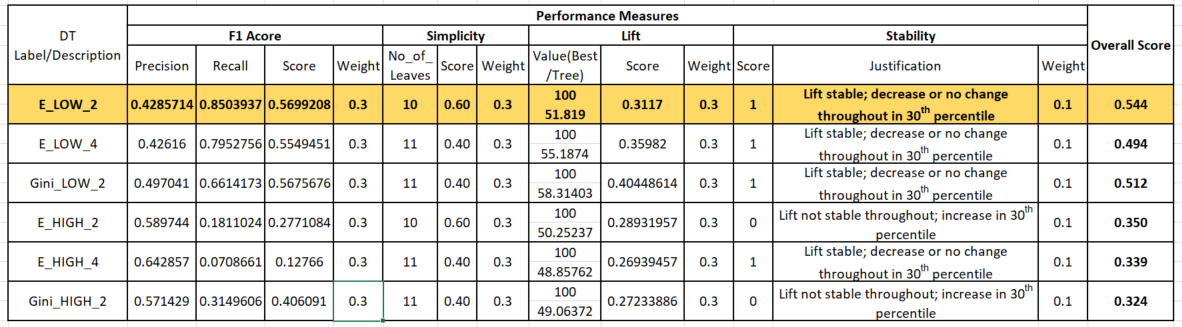
<https://data.census.gov/cedsci/>

### Decision Trees

1 Evaluation

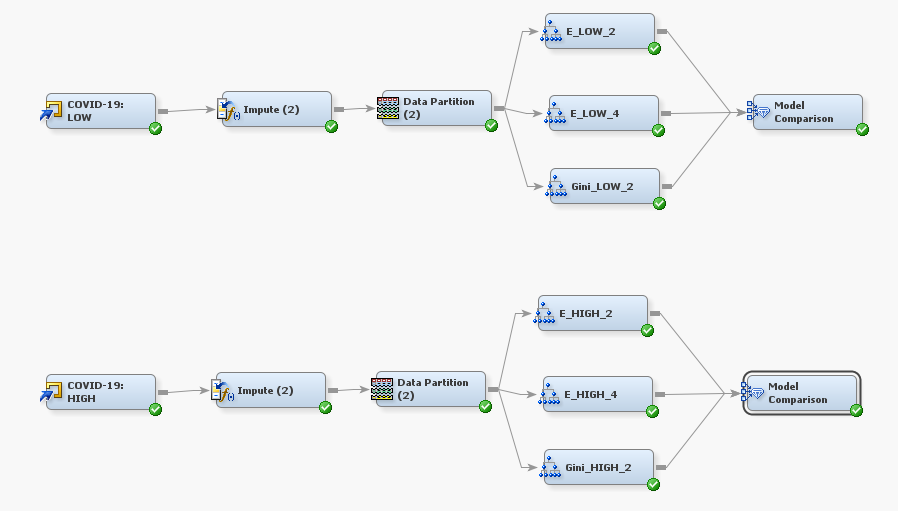
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Measure** | **Description** | **Definition of Value Function** | **Weight** | **Threshold** |
| F1 Score |  | 2(Precision\*Recall)/ (Precision+Recall) | 0.3 | >0.10 |
| Simplicity | Number of leaves for the tree when compared with an ideal tree with regard to the number of leaves | Defined by function ScoreSimplicity(NoOfLeaves)  If NoOfLeaves <= 2 or >=13; 0  If NoOfLeaves is >= 3 & <= 4;  (NoOfLeaves - 2)/5-2)  If NoOfLeaves is >= 5 & <= 8; 1  If NoOfLeaves >= 9 & <= 12  (13 -NoOfLeaves)/(13-8) | 0.3 | > 0.00 |
| Lift | Lift value (Cumulative%CapturedResponse) at 3rd percentile | (Tree – Baseline)/(Best – Baseline) | 0.3 | > 0.20 |
| Stability | Stability is binary as determined by the visual inspection of the non-cumulative %Response Lift Chart up to the 3rd percentile | 1 indicating stable model there the lift is stable or decrease throughout and 0 indicating an unstable model where lift increases up to the given nth decile. | 0.1 | > 0.00 |
| Combination Function | Overall Score = wF1Score \*Score F1Score + wSimplicity\*ScoreSimplicity + wLift\*ScoreLift + wStability \*ScoreStability | | | |

2 Summary of Results

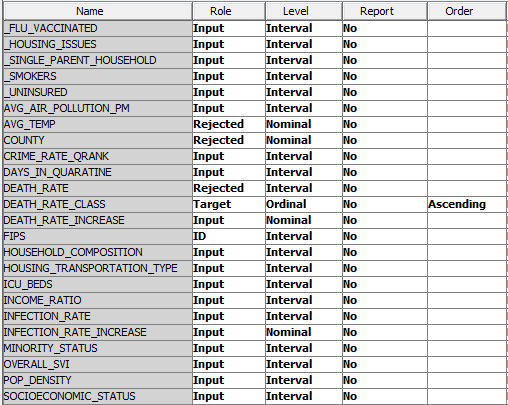


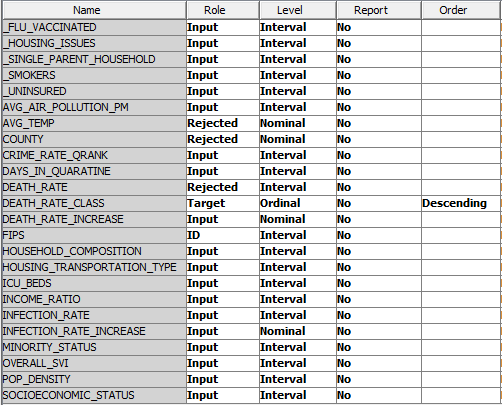
3 Evidence of Experimentation

3.1 Process Flow Diagram

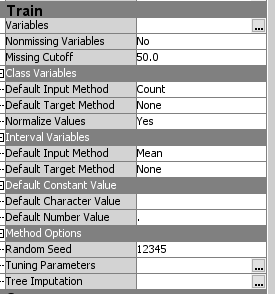


3.2 Variables

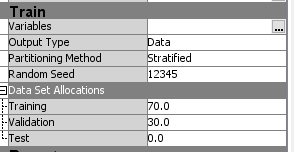


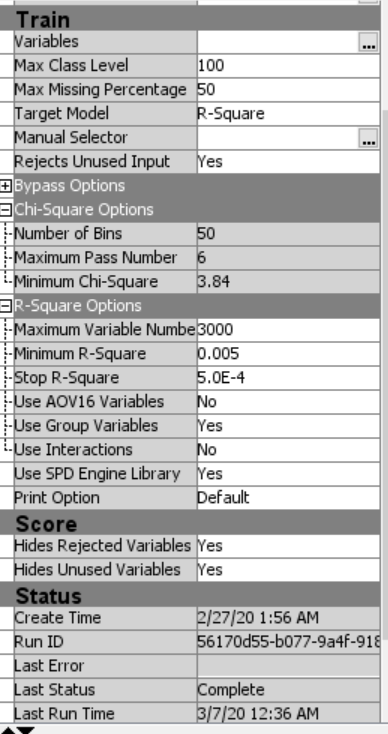


3.3 Impute Node

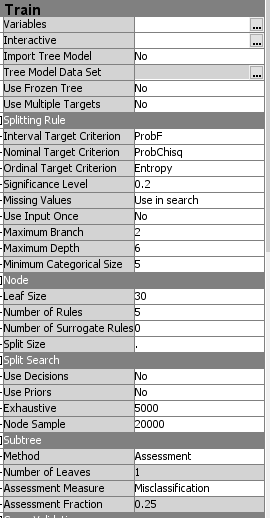
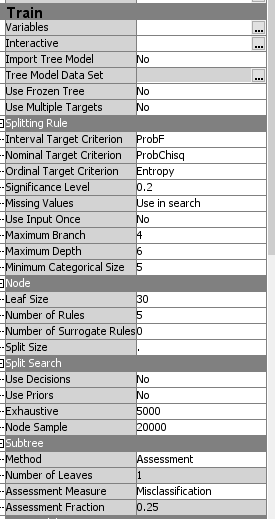


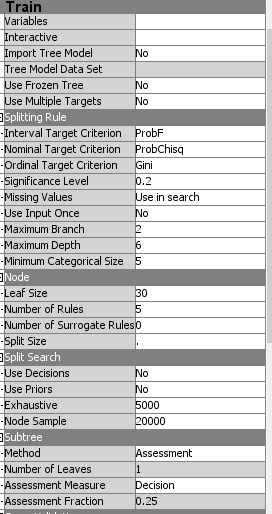
3.4 Data Partition Node





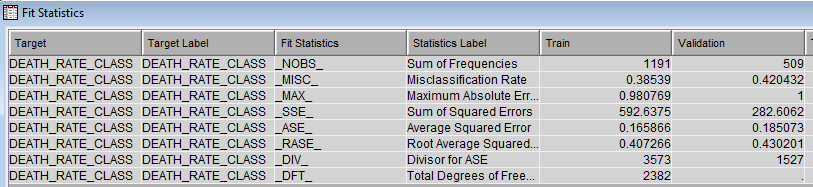
3.5 Decision Tree Nodes

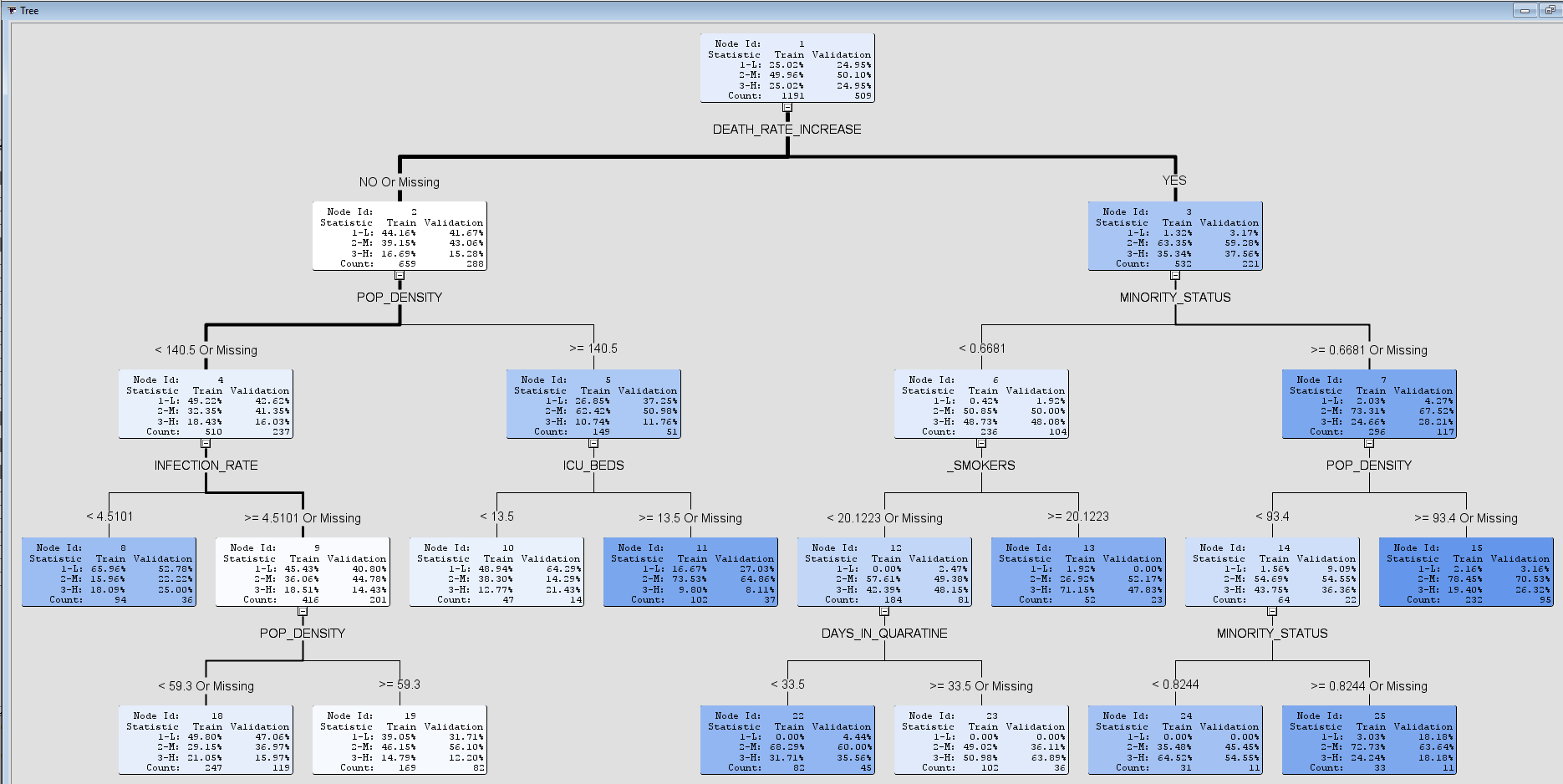


3 Best Results E\_LOW\_2

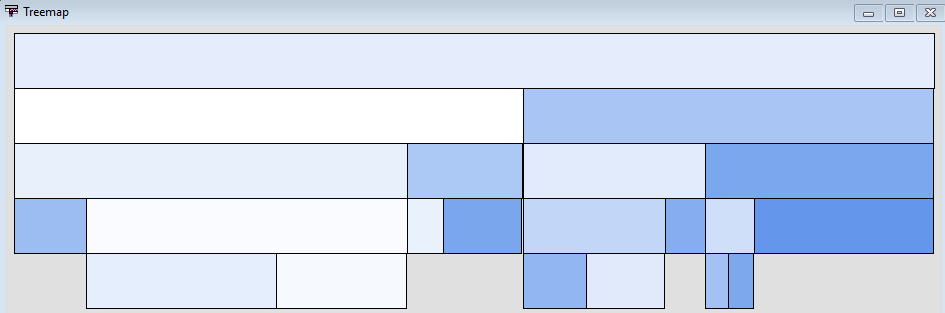
3.5.1 Fit Statistics



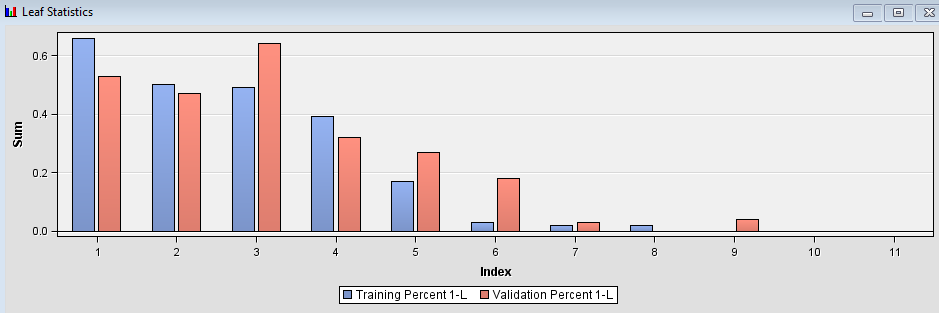
3.5.2 Tree



3.5.3 Treemap



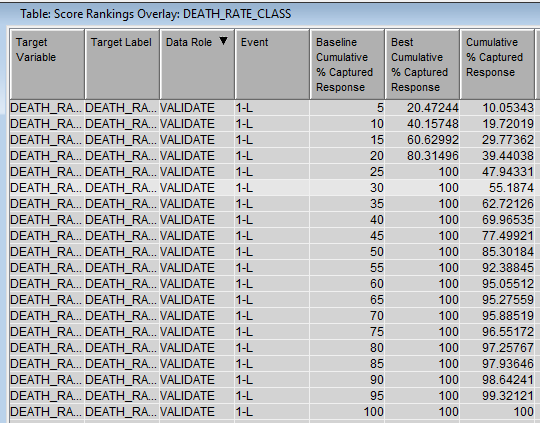
3.5.4 Leaf Statistics



3.5.5 Tree Result, Non-Cumulative %Response Chart



3.5.6 Table: Score Rankings Overlay



3.5.7 Node Rules

\*------------------------------------------------------------\*

Node = 8

\*------------------------------------------------------------\*

if POP\_DENSITY < 140.5 or MISSING

AND INFECTION\_RATE < 4.51006

AND DEATH\_RATE\_INCREASE IS ONE OF: NO or MISSING

then

Tree Node Identifier = 8

Number of Observations = 94

Predicted: DEATH\_RATE\_CLASS=3-H = 0.18

Predicted: DEATH\_RATE\_CLASS=2-M = 0.16

Predicted: DEATH\_RATE\_CLASS=1-L = 0.66

\*------------------------------------------------------------\*

Node = 10

\*------------------------------------------------------------\*

if POP\_DENSITY >= 140.5

AND ICU\_BEDS < 13.5

AND DEATH\_RATE\_INCREASE IS ONE OF: NO or MISSING

then

Tree Node Identifier = 10

Number of Observations = 47

Predicted: DEATH\_RATE\_CLASS=3-H = 0.13

Predicted: DEATH\_RATE\_CLASS=2-M = 0.38

Predicted: DEATH\_RATE\_CLASS=1-L = 0.49

\*------------------------------------------------------------\*

Node = 11

\*------------------------------------------------------------\*

if POP\_DENSITY >= 140.5

AND ICU\_BEDS >= 13.5 or MISSING

AND DEATH\_RATE\_INCREASE IS ONE OF: NO or MISSING

then

Tree Node Identifier = 11

Number of Observations = 102

Predicted: DEATH\_RATE\_CLASS=3-H = 0.10

Predicted: DEATH\_RATE\_CLASS=2-M = 0.74

Predicted: DEATH\_RATE\_CLASS=1-L = 0.17

\*------------------------------------------------------------\*

Node = 13

\*------------------------------------------------------------\*

if \_SMOKERS >= 20.1223

AND MINORITY\_STATUS < 0.6681

AND DEATH\_RATE\_INCREASE IS ONE OF: YES

then

Tree Node Identifier = 13

Number of Observations = 52

Predicted: DEATH\_RATE\_CLASS=3-H = 0.71

Predicted: DEATH\_RATE\_CLASS=2-M = 0.27

Predicted: DEATH\_RATE\_CLASS=1-L = 0.02

\*------------------------------------------------------------\*

Node = 15

\*------------------------------------------------------------\*

if POP\_DENSITY >= 93.4 or MISSING

AND MINORITY\_STATUS >= 0.6681 or MISSING

AND DEATH\_RATE\_INCREASE IS ONE OF: YES

then

Tree Node Identifier = 15

Number of Observations = 232

Predicted: DEATH\_RATE\_CLASS=3-H = 0.19

Predicted: DEATH\_RATE\_CLASS=2-M = 0.78

Predicted: DEATH\_RATE\_CLASS=1-L = 0.02

\*------------------------------------------------------------\*

Node = 18

\*------------------------------------------------------------\*

if POP\_DENSITY < 59.3 or MISSING

AND INFECTION\_RATE >= 4.51006 or MISSING

AND DEATH\_RATE\_INCREASE IS ONE OF: NO or MISSING

then

Tree Node Identifier = 18

Number of Observations = 247

Predicted: DEATH\_RATE\_CLASS=3-H = 0.21

Predicted: DEATH\_RATE\_CLASS=2-M = 0.29

Predicted: DEATH\_RATE\_CLASS=1-L = 0.50

\*------------------------------------------------------------\*

Node = 19

\*------------------------------------------------------------\*

if POP\_DENSITY < 140.5 AND POP\_DENSITY >= 59.3

AND INFECTION\_RATE >= 4.51006 or MISSING

AND DEATH\_RATE\_INCREASE IS ONE OF: NO or MISSING

then

Tree Node Identifier = 19

Number of Observations = 169

Predicted: DEATH\_RATE\_CLASS=3-H = 0.15

Predicted: DEATH\_RATE\_CLASS=2-M = 0.46

Predicted: DEATH\_RATE\_CLASS=1-L = 0.39

\*------------------------------------------------------------\*

Node = 22

\*------------------------------------------------------------\*

if \_SMOKERS < 20.1223 or MISSING

AND MINORITY\_STATUS < 0.6681

AND DEATH\_RATE\_INCREASE IS ONE OF: YES

AND DAYS\_IN\_QUARATINE < 33.5

then

Tree Node Identifier = 22

Number of Observations = 82

Predicted: DEATH\_RATE\_CLASS=3-H = 0.32

Predicted: DEATH\_RATE\_CLASS=2-M = 0.68

Predicted: DEATH\_RATE\_CLASS=1-L = 0.00

\*------------------------------------------------------------\*

Node = 23

\*------------------------------------------------------------\*

if \_SMOKERS < 20.1223 or MISSING

AND MINORITY\_STATUS < 0.6681

AND DEATH\_RATE\_INCREASE IS ONE OF: YES

AND DAYS\_IN\_QUARATINE >= 33.5 or MISSING

then

Tree Node Identifier = 23

Number of Observations = 102

Predicted: DEATH\_RATE\_CLASS=3-H = 0.51

Predicted: DEATH\_RATE\_CLASS=2-M = 0.49

Predicted: DEATH\_RATE\_CLASS=1-L = 0.00

\*------------------------------------------------------------\*

Node = 24

\*------------------------------------------------------------\*

if POP\_DENSITY < 93.4

AND MINORITY\_STATUS < 0.8244 AND MINORITY\_STATUS >= 0.6681

AND DEATH\_RATE\_INCREASE IS ONE OF: YES

then

Tree Node Identifier = 24

Number of Observations = 31

Predicted: DEATH\_RATE\_CLASS=3-H = 0.65

Predicted: DEATH\_RATE\_CLASS=2-M = 0.35

Predicted: DEATH\_RATE\_CLASS=1-L = 0.00

\*------------------------------------------------------------\*

Node = 25

\*------------------------------------------------------------\*

if POP\_DENSITY < 93.4

AND MINORITY\_STATUS >= 0.8244 or MISSING

AND DEATH\_RATE\_INCREASE IS ONE OF: YES

then

Tree Node Identifier = 25

Number of Observations = 33

Predicted: DEATH\_RATE\_CLASS=3-H = 0.24

Predicted: DEATH\_RATE\_CLASS=2-M = 0.73

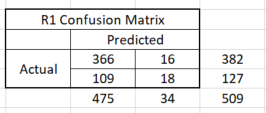
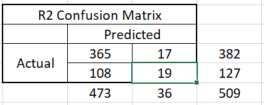
Predicted: DEATH\_RATE\_CLASS=1-L = 0.03

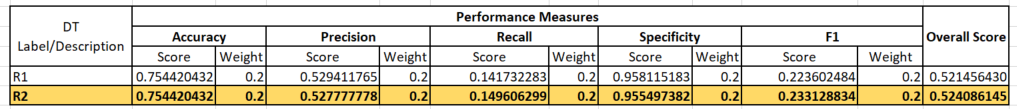
### Logistic Regression

1 Evaluation Approach

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Measure** | **Description** | **Definition of Value Function** | **Weight** | **Threshold** |
| Precision | Positive predictive value | TP/(TP+FP) | 0.2 | >0.5 |
| Recall | Ability of classifier to correctly detect true positive rate | TP/(TP+FN) | 0.2 | >0.2 |
| Specificity | Ability of classier to correctly detect true negative rate | TN/(TN+FP) | 0.2 | >0.0 |
| F1 Score |  | 2(Precision\*Recall)/ (Precision+Recall) | 0.2 | >0.2 |
| Combination Function | Overall Score = wAccuracy\*Score Accuracy + wPrecision \*Score Precision + wRecall \*Score Recall + wSpecificity \*Score Specificity n + wF1Score \*Score F1Score | | | |

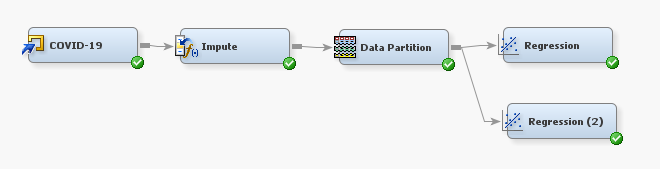
2 Summary of Results

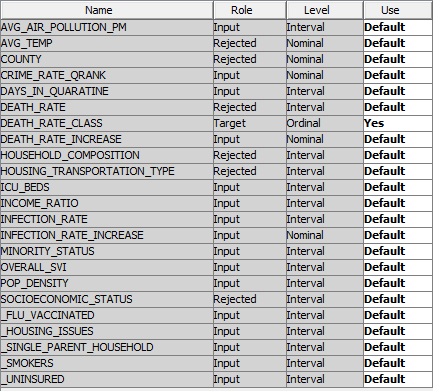


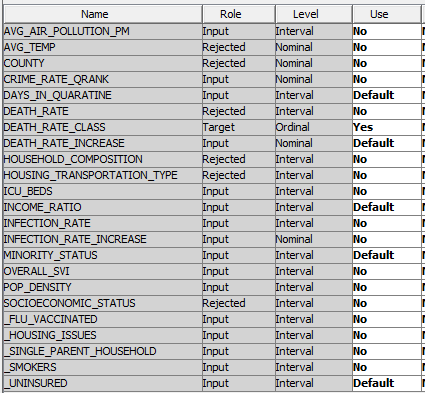
2 Evidence of Experimentation

2.1 Process Flow Diagram

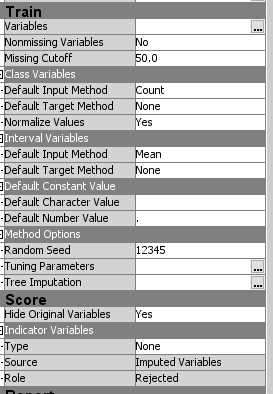


2.2 Variables

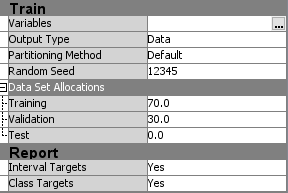




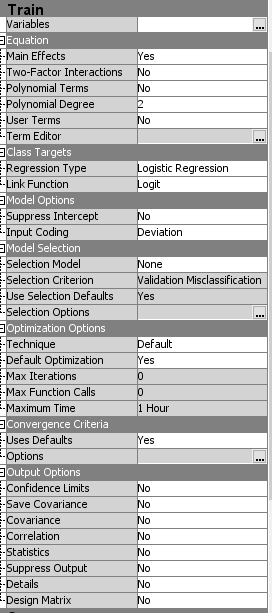
2.3 Impute Node



2.4 Data Partition Node

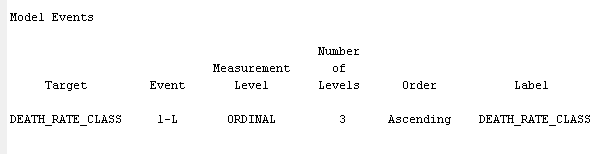


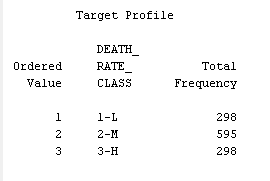
2.5 Regression Node

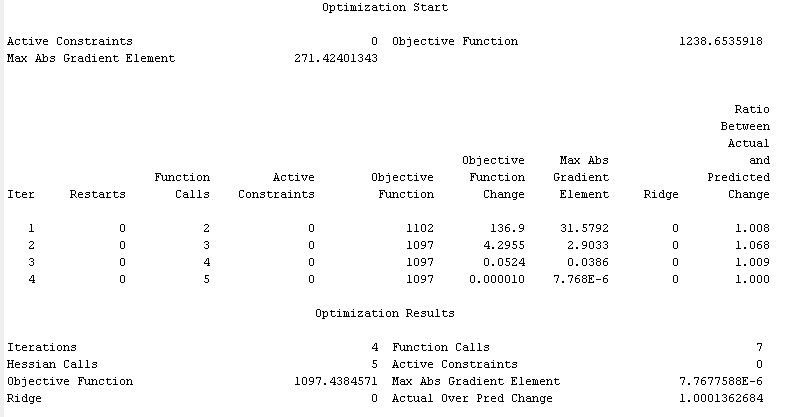


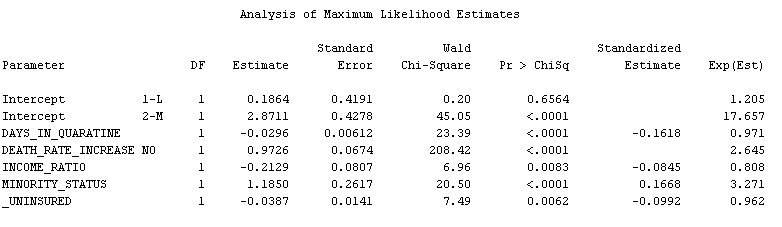
3 Results

3.1 Output

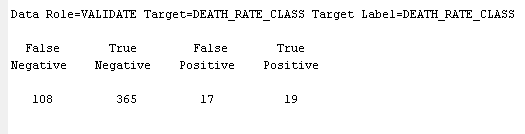




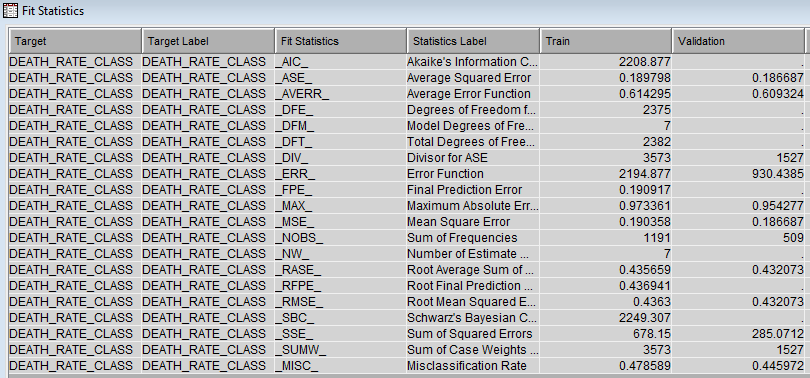




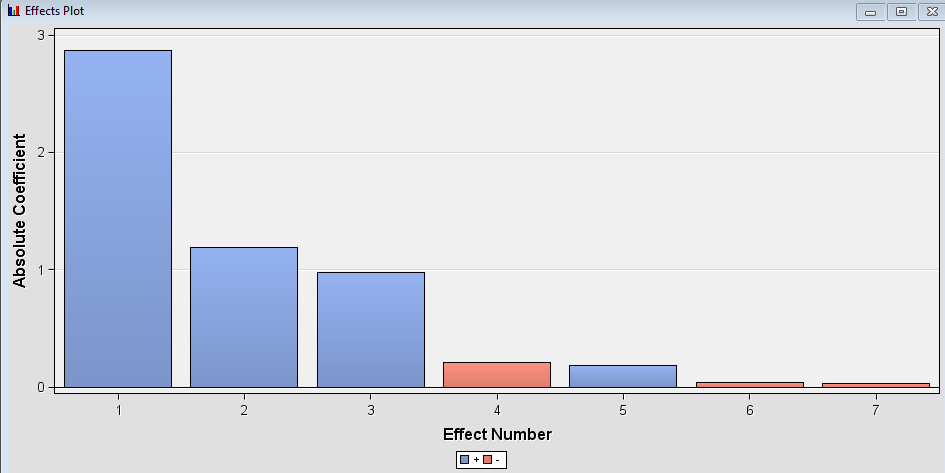
3.2 Confusion Matrix



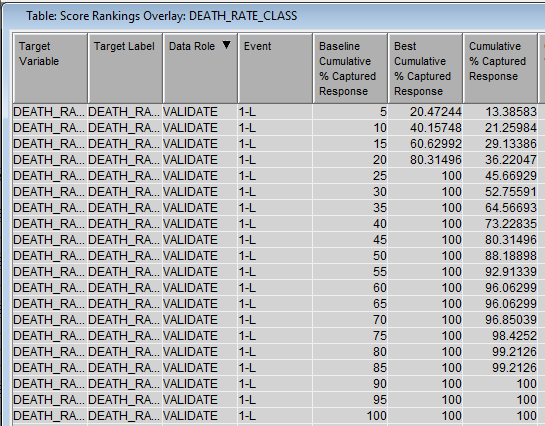
3.3 Fit Statistics



3.3 Effect Plot



3.4 Table: Score Rankings Overlay



3.5 Graph: Score Rankings Overlay %Response

