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.

Our data mining goal is stated as follow: using reported data regarding COVID-19 like confirmed cases and deaths, as well as other variables that may influence the pandemic, we will generate various predictive models to determine the most common variables in counties within the United States with low, medium, and high death rates. Once these factors have been identified, we will evaluate the Californian counties most likely to have low death rates if they were to be reopened. In addition, success criteria for achieving our data mining goal include the following factors: Building decision tree models that fit well, building regression models that generate well, upon assessing the performance and finding the best models, selecting rules that are great to formulate our strategic guidelines for reopening.

The biggest risk we face while completing this project is time constraint. The proposal must be delivered to stakeholders by mid-May which can result in insufficient data collection and/or insufficient iterations for tests run on the generated models. Also, reliance on publicly collected data can raise a risk of inaccuracy in data collected. As there is no guarantee that the data is accurate or relevant compared to firsthand collected and trusted data from nationally recognized sources, this will be an assumption to be taken into consideration. Additionally, data on COVID-19 is constantly changing each day. The current infection rates and fatality rates are persistently fluctuating as new cases are confirmed and patients are declared recovered or dead. Furthermore, assuming the state of California will continue to administer tests to those at risk, the data we use will be considered accurate and reported in a timely manner. The proposal will help to advise decision makers at the county level which will require data of the same granularity. If data is not provided at the county level, it may trigger additional assumptions to be made based on either state level data or national level data.

A successful completion of this goal will help in determining which counties may be eligible to reopen in hopes of stimulating the economy. The prediction model will give a valid starting point to explore which counties demonstrate as potential candidates with low death rates to be released from the “Stay at Home” order and those counties with potentially high death rates that should continue to stay under quarantine.

# Data Understanding

### Data Collection

As part of the initial data collection efforts, several relevant and accredited sources were reviewed and analyzed to compile an accurate data set regarding the pandemic in the United States. Each of the data sources analyzed allowed for different views and perspectives into the happenings within the states as a result of the pandemic. Additionally, the data reported gives an interesting insight in what could possibly be down the line in the fight against corona virus.

These sources include, but are not limited to:

<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

For an entire list of all sources and datasets reviewed, please refer to the appendix.

Initially we focused only on county level data in California. Although there were many data sources that provided plenty of variables to choose from, it was soon determined as being too small of a data set. With California only have 58 counties, there were simply not enough records to input into a predictive model and expect to have reliable results. With that, a wider net was cast for the initial data collection to ensure a more robust data set for the predictive model. Our second attempt was to review west coast states data including Washington, Oregon, and California. Again, there were simply not enough records. Although the proposal is specifically for California, the predictive models will be based off information compiled from all counties of the Unites States.

### Data Description

Prior to data collection, there were specific variables our team had already identified as points of interest for this project. These variables included number of cases reported, number of deaths reported, number of cases recovered, number of admitted ICU patients, population density, and some other socio-economic factors.

Numbers reported as a result of the pandemic were easily accessible however each state provided a daily update separately to its respective government site. This added an additional layer of complexity during the data collection process but allowed us to assume the data to be very reliant and accurate. The reported state data became more problematic after further review as we noticed each state provided data in different formats. For example, California posted information regarding admitted hospital patients by county, while other states reported on the state level. Another interesting and unexpected discovery was that not all states were providing the number for cases recovered. By the end of the initial review of state data, we moved forward in searching national data sources in hopes for more consistent data.

One great data source we reviewed was from the COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University found on the site github. The data here provided daily numbers of confirmed cases and reported deaths by county for the US. The one downside, however, was each day’s numbers were on separate files. Since we were not completely certain what date range we would be using or what would be valuable, all files were downloaded for each of the days provided. We expected we would have to combine the files and use either formulas, VLOOKUP's, or pivot tables in order to integrate and transform the data to be compatible for our final data set.

For variables regarding socioeconomic factors, the US census was the obvious choice to find reliable numbers. Additionally, the CDC’s Social Vulnerability Index provided several valuable variables including population of the disabled, education level, and median income. Although it was not in our initial data collection, we later found the County Health Rankings National Data compiled by the University of Wisconsin School of Medicine and Public Health.

As an overview of the data types, most if not all were numeric either being ratios, percentages, or counts.

### Data Exploration

We began initial exploration of data by analyzing readily available dashboards that used time series data and clustering analysis techniques to learn more about the variables relating to the project’s current target of death rate. These data sources utilized Microsoft Excel and Tableau to visually display large amounts of data in a more aesthetically pleasing and easily understandable way. Additionally, the interactive features of these dashboards allowed us to drill into the areas we were interested in.

In Figure 1, we can see the immediate impact of COVID-19 throughout the United States. Visually, it is obvious the east coast has been more severely impacted compared to the west coast. Thankfully, the now added scope of data from east coast states and counties will provide valuable insight on the factors impacting the pandemic.

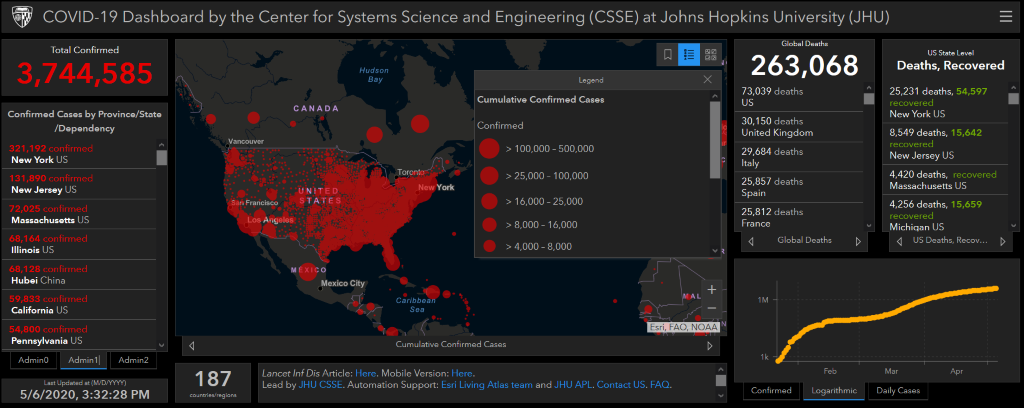


Figure 1: Johns Hopkins University as of May 6, 2020.

In Figure 2, we can see the gradual implementation of the Stay at Home order. In mid to late March, only a few states were declaring a state of emergency and mandating a stay at home order. By early to mid-April, most of the other states followed suit. The goal of the government mandated order was to prevent the spread of COVID-19. But did it work? After seeing Figure 2, we decided the number of days in quarantine would be valuable to the dataset as a possible factor to the target variable.

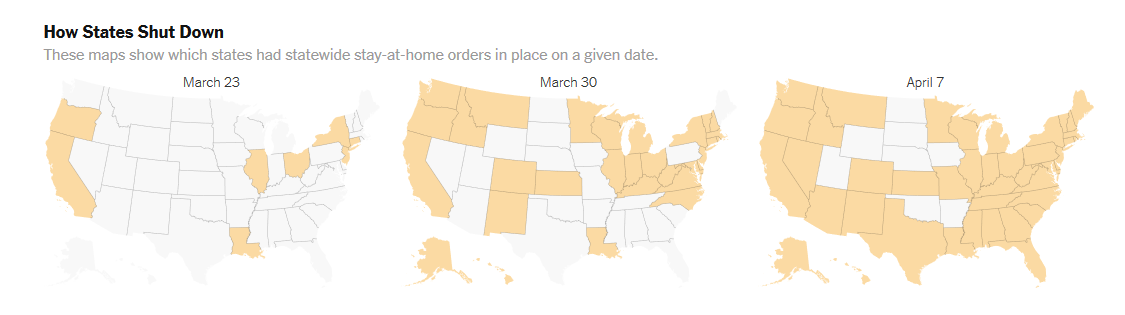


Figure 2: New York Times: How States Shut Down

Generally, similar information would be reported on COVID-19, like confirmed cases and new deaths. However, we did see some promising information on variables we did not initially think of. In Figure 3, there are two interesting pieces of information we decided to expand on, risk factors and risk info. Currently, symptoms of COVID-19 include coughing, shortness of breath, fever, and muscle pain, among other things. While there are not stats on what people currently suffer from, we looked for risk factors of COVID-19 that may make a person susceptible to contagion. In the end, we identified percentage of adult smokers, air pollution, percentage of insured people who have gotten the flu vaccination, and the amount of air pollution. Furthermore, we wanted to include other factors believed to affect any pathogen including average temperature, since people believe that heat will slow the spread, and housing issues, which can include overcrowding.

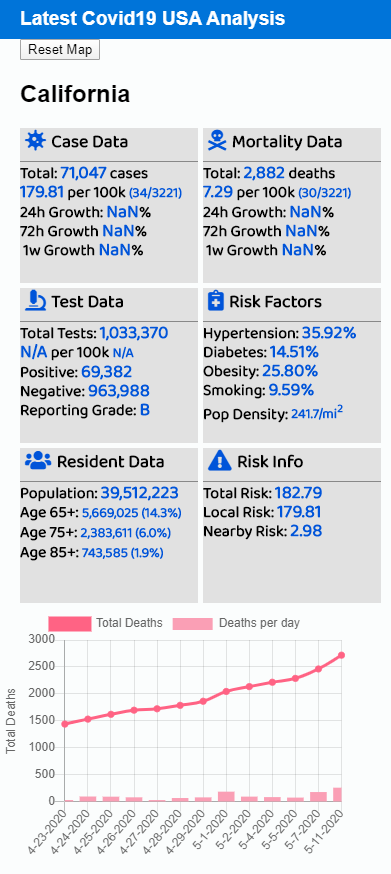


Figure 3: Latest COVID19 USA Analysis - California

### Data Quality Verification

As we began validating the metadata, we saw many similar fields across the different data sources. For example, the daily reports of confirmed cases documents population which could also be found on the US census. While we didn’t need the information from both sources, we did compare the numbers to validate the numbers. There were some variables we thought would be meaningful in the project like average temperature and crime rate, unfortunately, there were many counties missing these values. In the end, we decided there was not enough information for average temperature and had to remove it. For crime rate, we were able to find the crime rate quartile of each county based on the national level.

Between the US counties data and the Social Vulnerability Index, there was an abundant of information to choose from: education, gender, household totals, age, etc. However, we found data was separated into different columns. For gender, there were two columns, percentage of male and percentage of female; for education, there were four columns. Using all the fields for each category would result in a large dataset and variables highly correlated. To mitigate this, we would narrow each variable to the level we felt would have the highest impact. In our second attempt at data collection, we found the County Health Rankings data was extremely useful. It reported new information like percentages of smoking adults and air pollution all on the county level.

We soon found our initial plan of including cases recovered and ICU patients admitted would not be an option. The data for these variables were lacking on the county level and many were not even reported on the state level. Instead, we decided to use available ICU beds per county. Regarding confirmed new cases and reported death cases, our original plan to calculate the rates for the respective events did not go as plan either. The month of April, while provided much data, did not provide an accurate view of the current data. The pandemic earned great traction in mid-April. If we had focused on the data for the entire month, the numbers would be skewed in counties that had 0 cases on the first and then any amount after that. We adjusted our approach to provide a more accurate view of pandemic by using numbers from April 24th and April 30th to determine the infection rate as well as the death rate.

Regarding the physical quality of the data, we found missing values, incorrectly formatted data types, misspellings, and inconsistent data. Fortunately, we did not see any issues that could not be resolved during the next phase. This process will be described in the data transformation stage. Based on the data exploration, we have inferred that the information collected from the various sources should allow us to create a dataset suitable for modeling. As we design, build, and evaluate our model, the goal will be to determine which California counties may be eligible to be reopened in hopes of stimulating the economy while still be mindful and cautious of the COVID-19 threat.

# Data Preparation

### Data Transformation

As the Social Vulnerability Index had the most information, we decided to start there and were quickly able to identify factors that could possibly influence our data model. Variables like unemployment rate, poverty rate, minority rate, and disability rate were all previously identified, however, the Social Vulnerability Index also provided score for 4 major themes: socioeconomic status, household composition, minority status, and housing type, as well as an overall score. For socioeconomic status, the value considered factors like poverty, unemployment, and income. So instead of having these columns separately, we chose to go with the scores within each theme and the overall score.

The second dataset we reviewed was the counties dataset from John Hopkins University. This data was particularly large due to the levels of each category being divided into separate columns. As previously mentioned, looked at the percentages for the different levels and then determined which we would use. In the example for gender, since there are only two levels, we decided we would make this field a nominal field and have the variable reflect the dominant gender in the population, either male or female. In the case of education, we chose to only use the percentage of the population with high school or less. Similarly, we approached the values for ages and kept only the percentage of 65 plus.

Reviewing the daily cases dataset was the most tedious of all. While there were over 60 files, one confirmed cases file and one reported deaths cases file for each day, there were only a few fields that would provide us with some insight. We knew we couldn’t use data from every single day for each county, so we did not want to attempt and get all files into one document. We began by playing around and creating new fields that would summarize the data differently. As previously mentioned, we had originally planned to calculate the rate of confirmed cases and deaths over the period of April. However, many of the calculations resulted in errors. Since so many counties had 0 cases to begin with, there was no way to measure the rate of confirmed cases but deleting those records would have resulted in too much lost data for the model. After multiple iterations of what could be done with the cases confirmed and deaths, we ultimately change the data to calculate infection rate.

For infection rate per 10,000 people, we used the amount of cases reported on April 30th, divided by the population amount for the county, and multiplied by 10,000. For the death rate, we divided the number of reported deaths on April 30th with the amount of confirmed cases on the same day. As part of the data cleanup and to prevent skewed data, we removed all records with an infection rate of less than 0.1 and removed all records with less than 10 confirmed cases as of April 24th. At this point, the target variable of death rate had far too many values ranging from 0 to 30. To decrease the levels of the target variable, we classified each record with low, medium, or high based on the death rate. As an attempt to avoid skewed results, we labeled the lowest 25% of the death rates as low, 25% of the highest death rates as high, and those in between as medium.

After a few different iterations of datasets, we eventually added two additional formulated fields: infection rate increase and death rate increase. Using the counts for both reported on April 24th and April 30th, we determined if the number on April 30th was an increase or no increase.

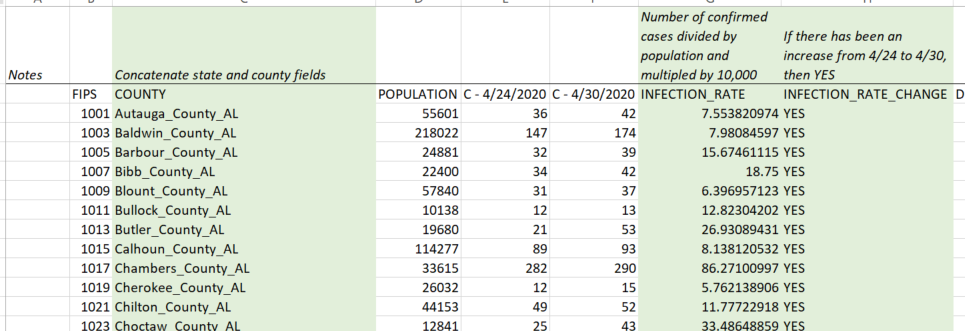
Some additional data cleanup performed was concatenating the county and state columns to have only one label for each record with the FIPS ID as the unique key. Records for the state level were listed in the dataset as well as other US territories. As we are only interested in county level data for US states, we removed all those records.

### Data Integration

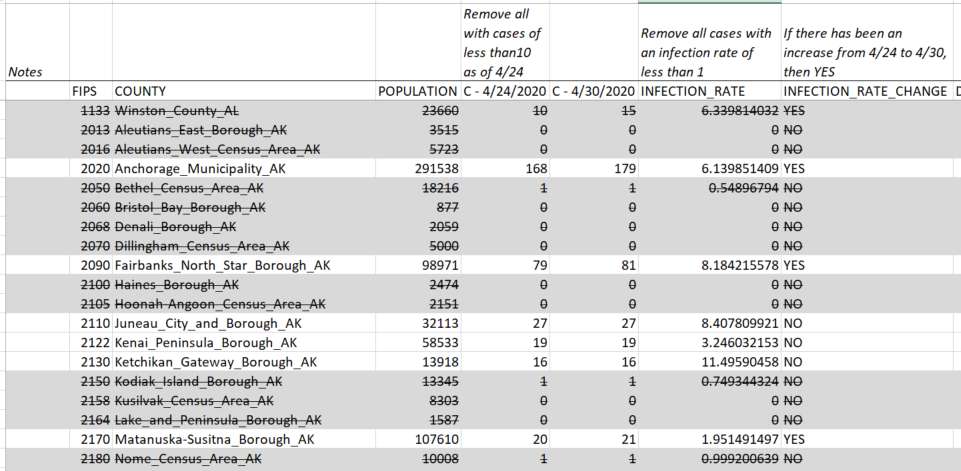
The data integration process was rather extensive and quite complex. As we utilized multiple different sources and each of those files were quite large on its own, compiling them together took some time. Luckily, the process was a bit easier as each county had a FIPS ID which served as the key for all the data collected. We created a large spreadsheet and imported the CSV files into a multi-sheet .xlsx file. Using VLOOKUP (Left Join) we pulled all the data from the fields we mentioned previously into a new sheet.

The last piece of information we needed to merge was number of days in quarantine. This data was only able to be obtained at the state level. However, since the stay at home order was mandated by the state, we will assume that the counties in the respective states had begun quarantine on the day the state had mandated it. Using excel formulas, we assigned each county the number of days in quarantine as the respective state.

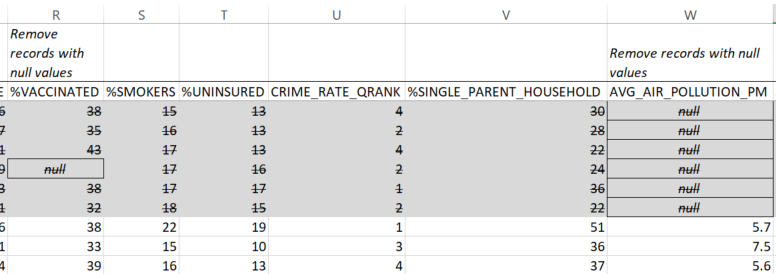
From the raw data set we created above, we first created aggregated variables and added the columns using the current the data. For example, we created variables like infection rate and death rate using formulas.

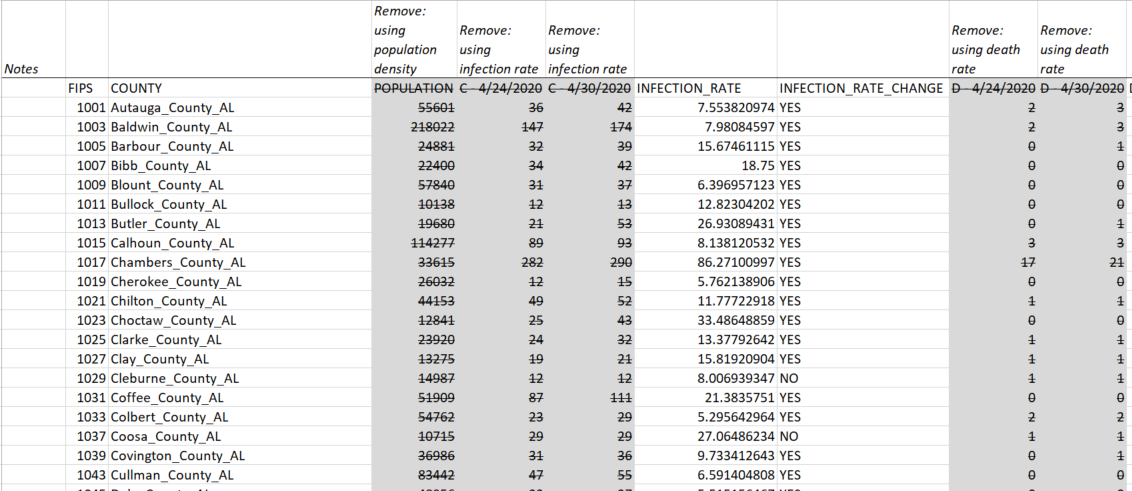


Next, we removed records that did not meet the criteria of 1. having an infection rate of at least 1 and 2. having more than 10 cases on April 24th.

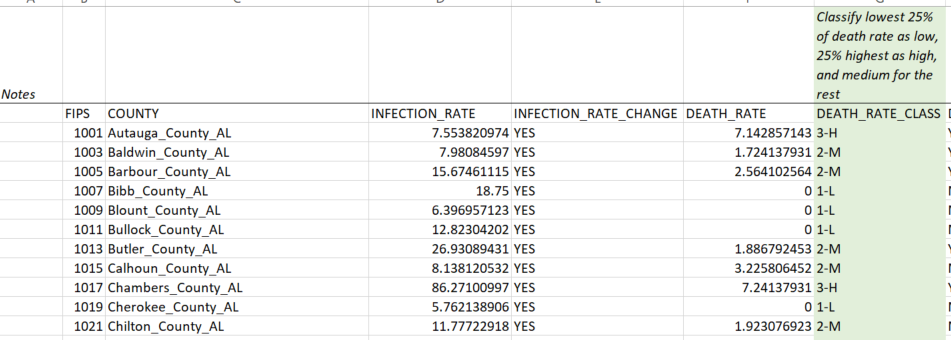


We then went on to remove the records with missing values as well as the columns we decided deemed not needed. For example, we removed the fields for number of cases on the outlined days as we have used this value to calculate the infection rate.





Lastly, we classified each record based on the corresponding death rate. The records with the lowest 25% of death rates were marked as low. The records with the highest 25% were marked as high. All records in between, about half the records, were marked as medium.



The entire process is documented in the attached spreadsheet, with each sheet highlighting the steps that were taken.

### Data Modeling Techniques

As the data modeling tool would be SAS Enterprise Miner, we formatted our dataset by originally creating a .csv file and ultimately saving as an xlsx file once all the transformations were done. As previously mentioned, we transformed the death rate from numeric to ordinal with values of low, medium, and high. However, we did not know what the impact of this change would cause, so we also left the numeric value of the death rate in the data set to reference.

The final dataset metadata is as shown below:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **NAME** | **DESCRIPTION** | **ROLE** | **LEVEL** | **LENGTH** | **LEVELS** | **VALUES** |
| \_FLU\_VACCINATED | Percentage of adults with flu vaccinations | INPUT | INTERVAL | 8 |  | Range of 15 - 66 |
| \_HOUSING\_ISSUES | Percentage of households with at least 1 of 4 housing problems: overcrowding, high housing costs, or lack of kitchen or plumbing facilities | INPUT | INTERVAL | 8 |  | Range of 6.5625 - 28.92285 |
| \_SINGLE\_PARENT\_HOUSEHOLD | Percentage of single parent households | INPUT | INTERVAL | 8 |  | Range of 9.430276 - 80 |
| \_SMOKERS | Percentage of adult smokers | INPUT | INTERVAL | 8 |  | Range of 5.908719 - 80 |
| \_UNINSURED | Percentage of uninsured adults | INPUT | INTERVAL | 8 |  | Range of 2.262724 - 31.20792 |
| AVG\_AIR\_ POLLUTION\_PM | Average daily amount of fine particulate matter in micrograms per cubic meter | INPUT | INTERVAL | 8 |  | Range of 3.9 - 19.7 |
| AVG\_TEMP | Average temperature in April | REJECTED | NOMINAL | 5 | 128+ |  |
| COUNTY | County name and state | REJECTED | NOMINAL | 30 | 128+ |  |
| CRIME\_RATE\_QRANK | Violent crime rate nationally ranked in quartiles | INPUT | NOMINAL | 8 |  | Range of 1 - 4 |
| DAYS\_IN\_ QUARATINE | Days county has been under the stay at home order | INPUT | INTERVAL | 8 |  | Range of 0 - 42 |
| DEATH\_RATE | Death rate as of April 30 | INPUT | INTERVAL | 8 |  | Range of 0 - 30 |
| DEATH\_RATE\_ CLASS | Death rate as classified as low, medium, or high | INPUT | ORDINAL | 3 | 3 | LOW/ MEDIUM/HIGH |
| DEATH\_RATE\_ INCREASE | Death rate increase from 4/24 to 4/30 | INPUT | NOMINAL | 3 | 2 | YES/NO |
| FIPS | County ID number | INPUT | INTERVAL | 8 |  | --- |
| HOUSEHOLD\_ COMPOSITION | Total house composition score from SVI | INPUT | INTERVAL | 8 |  | Range of 0 - 0.9962 |
| HOUSING\_ TRANSPORTATION\_TYPE | Total housing type and transportation score from SVI | INPUT | INTERVAL | 8 |  | Range of 0.0035 - 1 |
| ICU\_BEDS | Number of available ICU beds in county | INPUT | INTERVAL | 8 |  | Range of 0 - 2126 |
| INCOME\_RATIO | Ratio of household income at the 80th percentile to income at the 20th percentile | INPUT | INTERVAL | 8 |  | Range of 3.029475 - 9.206592 |
| INFECTION\_RATE | Rate of infection as of April 30 | INPUT | INTERVAL | 8 |  | Range of 1.017589 - 608.2343 |
| INFECTION\_RATE\_ INCREASE | Confirmed cases increase from 4/24 to 4/30 | INPUT | NOMINAL | 3 | 2 | YES/NO |
| MINORITY\_STATUS | Total minority status score from SVI | INPUT | INTERVAL | 8 |  | Range of 0.001 - 0.9997 |
| OVERALL\_SVI | Total overall SVI score | INPUT | INTERVAL | 8 |  | Range of 0 - 0.9984 |
| POP\_DENSITY | County population divided by total square miles | INPUT | INTERVAL | 8 |  | Range of 1.7 - 69468.4 |
| SOCIOECONOMIC\_STATUS | Total socioeconomic score from SVI | INPUT | INTERVAL | 8 |  | Range of 0.003 - 0.9994 |

# Modeling

### Data Mining Techniques

Our initial approach to this project is to create various decision tree models. Decision trees allow for a step by step breakdown of input variables as they affect the target variable. In doing so we test how well the independent variables affect the target variable. Since our target is ordinal, we will be using both the entropy model and Gini model. Additionally, we will run the models for two different target outcomes, one for low and one for high.

In addition to the decision trees, we decided a logistic regression model would also be beneficial. The main goal of this project is to identify the California counties that are potential candidates for the gradual reopening of California. Our target variable is an ordinal variable with three levels, low, medium, and high. Our dataset contains many independent variables whose relationship with the dependent variable needs to be determined. The logistic regression model is the obvious choice. Not only will it help to predict the dependent variable, it will meticulously analyze the relationship between our selected independent variables as well.

### Modeling Test Design

The final dataset used for the data modeling consists of 1,700 records. To ensure enough data for both the testing and validation models, we partitioned the data using the stratification method with allocations of 70% testing and 30% validation. In order to test multiple models, we used both modeling techniques available, entropy and gini. We originally used a default branch of 2 and used roughly 2.5% of the testing data to determine the leaf size of 30. Additionally, we will be testing the target variable as both low and high for the decision trees.

### Modeling

To begin creating the model, the dataset was imported into SAS Enterprise Miner using a file import node. Although there are no missing values in the dataset, we connected an impute node just to make sure that we can properly handle any missing values in our data set. Then, a data partition node is connected in order to partition the data 70 and 30 in both training and validation data set respectively. Subsequently, we completed the same steps as above, but this time, we changed the order of the target variable from the file import node from ascending to descending. This changed the target value to be high instead of low.

From each data partition node, we connect two decision tree nodes. All decision nodes had a maximum branch of 2 and leaf size of 30. Each path had two decision trees, one as entropy and one as gini. After assessing the initial models, we added an additional decision tree to each path, an entropy decision tree with a maximum of 4 branches. This allowed us to compare six different models.

For the logistic regression model, again we imported the file with the file import node and connected the impute and data partition node. This time, instead of a decision tree, a regression tree node will be used. For the regression tree node, we used the default model.

### Model Performance Assessment

To assess the model performance for our decision trees, we used the following performance measures: F1 Score: weight 0.30, threshold 0.10, Simplicity: weight 0.30, threshold 0.25, Lift: weight 0.3, Stability: weight 0.10. The best model will be based on the overall score, which is the highest of the resulted combination function.

Overall Score = wF1Score \*Score F1Score + wSimplicity\*ScoreSimplicity + wLift\*ScoreLift + wStability \*ScoreStability

For the regression model performance assessment, we chose five performance measures including: F1 Score, Precision, Recall, and Specificity. All of these measures will be calculated from the confusion matrix generated by the model and found in the output file in model displayed result. The best regression model will be based on the highest overall score of models determined by the following combination function.

Overall Score = wAccuracy\*Score Accuracy + wPrecision \*Score Precision + wRecall \*Score Recall + wSpecificity \*Score Specificity n + wF1Score \*Score F1Score

Please refer to the appendix for measure descriptions and function models.

# Evaluation

### Evaluate the model against business objectives and success measures

When initially running the decision trees, there were two trees for each path differentiated by the following splitting methods: entropy and gini. The decision trees and its’ results are documented in the appendix of this paper. Let’s start by looking at the target variable of low and the trees along this path. The entropy decision tree resulted in 9 leaves. The tree began splitting at population density for the first and second node. Subsequently, it split at minority status, minority, and days in quarantine. In looking at the gini decision tree, the results were drastically different. The gini tree only showed one leaf and no splits.

When running the second path with the target variable as high, the results were almost identical to the results from the low path. Since the four models were so similar, we added an additional decision tree to each path. This time, we changed the branch maximum from 2 to 4 in hopes to get different results. By changing the maximum branch, we were able to get different results. This tree also had 9 leaves, but the tree looked different. This tree first split at the ICU beds available to days in quarantine, to median household income. Besides days in quarantine, the 2 max entropy trees were completely different from the 4 max entropy trees.

It is important to point out that when the additional 4 max entropy trees were added to the target variable high path, the results were similar to the tree from the target variable low path. The overall scores of the decision trees were extremely low and we didn’t see much value. Therefore, we went back to the beginning and redid our dataset with more information. This happened quite a few times, and when we landed on the dataset using more health factors and used only the themed scores from the SVI dataset, we were finally able to get better results.

Over 20 decision tree models were produced and tested with multiple datasets and at the end of the entire process, the best decision tree produced. The best decision tree produced was an entropy model with a target variable for death rate as low, a maximum of 2 branches, and 30 leaves. This tree had a misclassification rate of 0.477, giving it an accuracy rate of 0.523. With 10 leaves, the simplicity score was 0.60, the lift score was 0.3117. Additionally, the tree was stable throughout in the 30th percentile. When reviewing the node rules, there were three nodes that produced a high confidence level for the death rate target level of 2-M: node 11, 15, and 25. Only node 13 produced a high confidence level for death rate target level of 3-H. None of the nodes had a high confidence for the death rate target level of 1-L. These results can be further reviewed from the appendix.

The initial logistic regression model with the original data set ran with 5 iterations. In relation to the success criteria, the model had an accuracy score of .2592, precision score of .1667, recall score of .0077, specificity score of .9871, and an F-1 score of .0147. The scores were completely terrible and resulted in a confusion matrix with no false positives and no true positives. We went through the same process as with the decision trees, going through all the different datasets until we landed on the final dataset. For the final dataset, the first run of the logistic regression model produced a much higher (compared to the very first run) score for precision of 0.53.

To optimize the model even further, we removed all variables with insignificant p-value scores from the results of the first run. If the p-value scores were higher than 0.05, then they were removed. This left the model with five variables: DAYS\_IN\_QUARANTINE, DEATH\_RATE\_INCREASE – NO, INCOME\_RATIO, MINORITY\_STATUS, and %UNINSURED. Although the second run resulted in only an increase of one true positive case, the overall score for the second run did increase for a total score of 0.524. For more details on the results, please refer to the appendix

Analyzing the coefficients of the parameters, we get plenty of insight with how the model works. The parameter with the highest effect on the target variable of DEATH\_RATE\_CLASS LOW is MINORITY\_STATUS at 1.1850 followed by DEATH\_RATE\_INCREASE NO at 0.9726. The two parameters have such a high effect, that when comparing them with a different parameter with a negative effect, they will always be at least twice as likely to result in a LOW outcome. For example, we can assume that for those records that have a high-income ratio number and has a death rate increase of no, then it is twice as likely that the record will have result in a low death rate. From this, we can conclude that the two major parameters that will affect the death rate with an outcome of low is MINORITY\_STATUS and DEATH\_RATE\_INCREASE NO.

After all, as shown in the summary table in the appendix, our best decision based on the overall score is E\_low\_2 with the overall score of 0.544 calculated based on our combination function. Furthermore, node 15 of this model has the best rules stated as follow. If POP\_DENSITY >= 93.4, AND MINORITY-STATUS >= 0.6681, AND DEATH\_RATE\_INCREASE IS OF: YES, then DEATH\_RATE\_CLASS = 2-M IS predicted with a least 78% confidence level. So, based on the rules, the decision makers do not have to rush to open counties with high population density, high rate of minority, and finally, where the death rate is still increasing.

### next steps

The main objective of this project is to determine a guideline to strategically lift the Stay at Home orders. Using data collected from various sources, we successfully generated multiple predictive models to determine the most common variables in counties with both low and high death rates. After multiple iterations, in-depth analysis of results by the prediction models, and given the time constraints as well as limited access to data, we are confident in selecting the logistic regression model as the basis of the proposal.

We recommend to the stakeholders of this document to identify the California counties based on following factors: DAYS\_IN\_QUARANTINE, DEATH\_RATE\_INCREASE – NO, INCOME\_RATIO, MINORITY\_STATUS, and %UNINSURED.

By focusing on these elements, we can predict the counties with low death rates and compare them with the county’s actual numbers. We can infer that those with comparable numbers will likely have low death rates should the county were to reopen. These results will provide decision makers with a pool of candidates to further analyze before choosing any county is reopened.

# Lessons learned (10 points)

Looking back at the entire project, we did not realize how very time-consuming and tedious the data collection and preparation process was. The amount of data available and the various ways it could be interpreted really gives way to producing an infinite number of datasets. The application of the CRISP-DM methodology, although quite extensive, is truly comprehensive. It forces the owner to look at data mining problems from a broad lens and ensures, if done correctly, the outcome will deliver what was intended. We only wished we had more time to go through the iterative process of collecting more data and building more models. While we are happy with our results, we know that with more time, it could only get better and more accurate.

# 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>

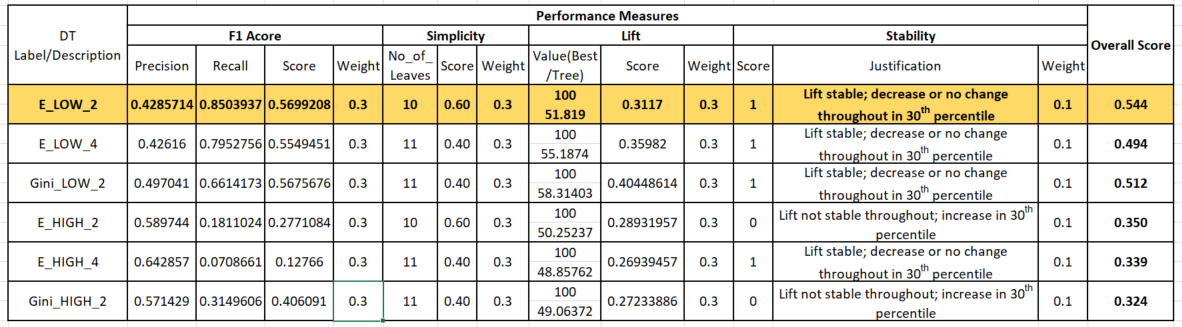
<https://data.census.gov/cedsci/>

### Decision Trees

1 Evaluation

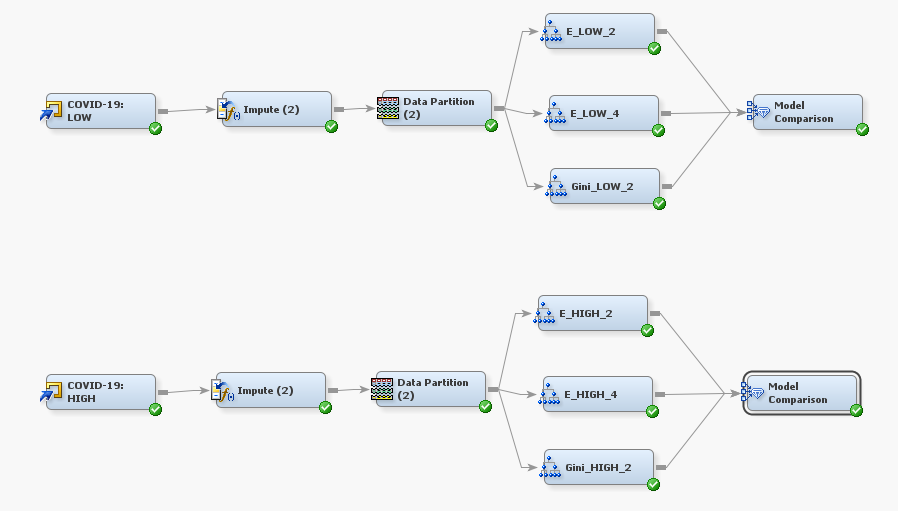
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Measure** | **Description** | **Definition of Value Function** | **Weight** | **Threshold** |
| F1 Score |  | 2(Precision\*Recall)/ (Precision+Recall) | 0.3 | >0.10 |
| Simplicity | Number of leaves for the tree when compared with an ideal tree with 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

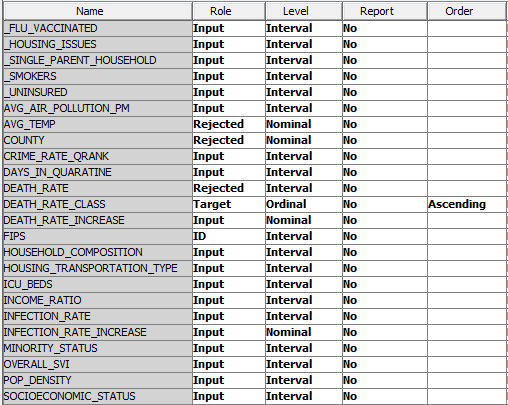


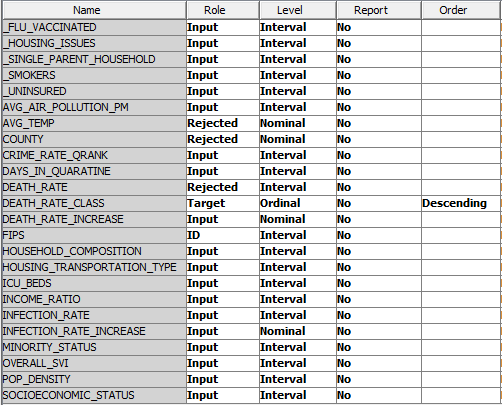
3 Evidence of Experimentation

3.1 Process Flow Diagram

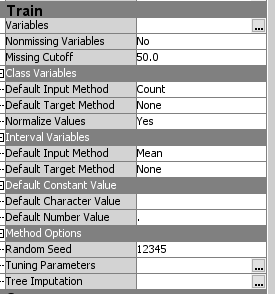


3.2 Variables

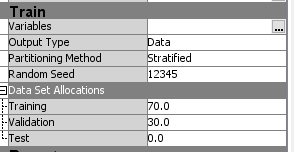


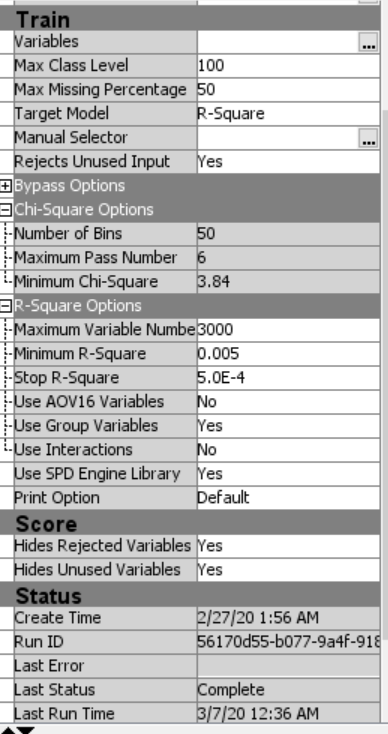


3.3 Impute Node

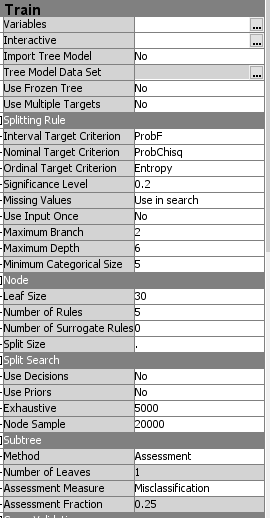
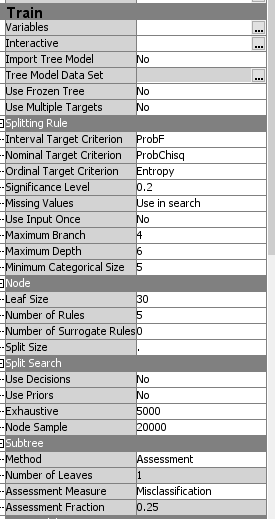


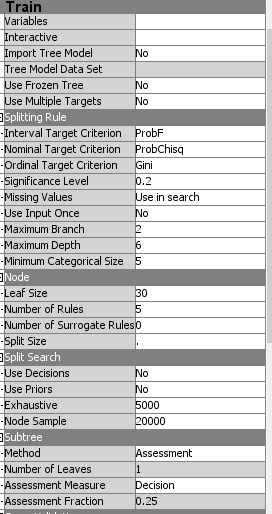
3.4 Data Partition Node





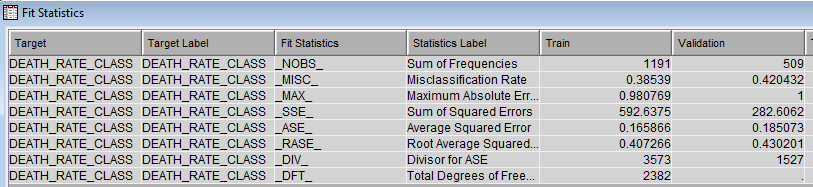
3.5 Decision Tree Nodes

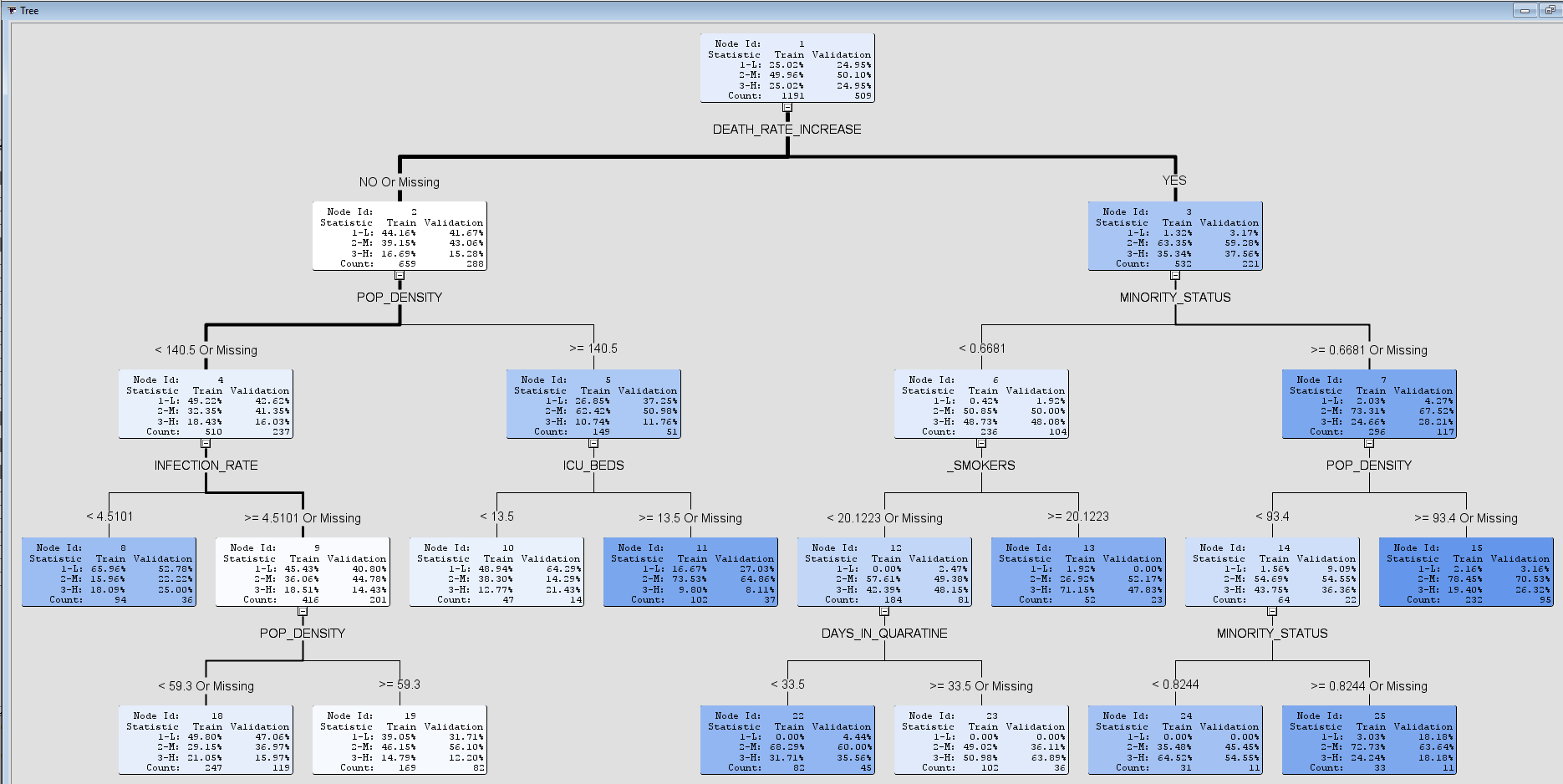


3 Best Results E\_LOW\_2

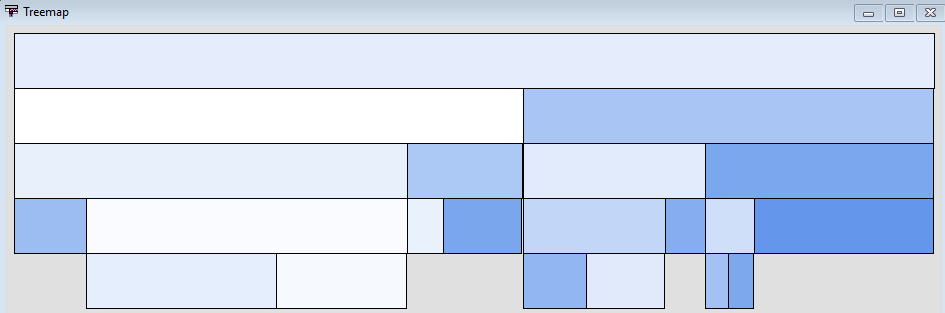
3.5.1 Fit Statistics



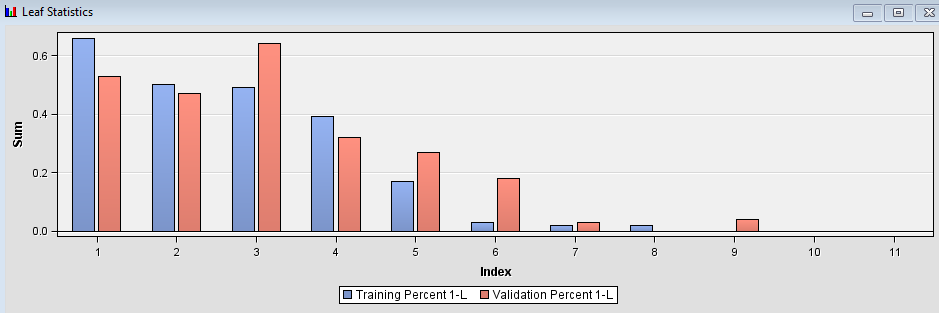
3.5.2 Tree



3.5.3 Treemap



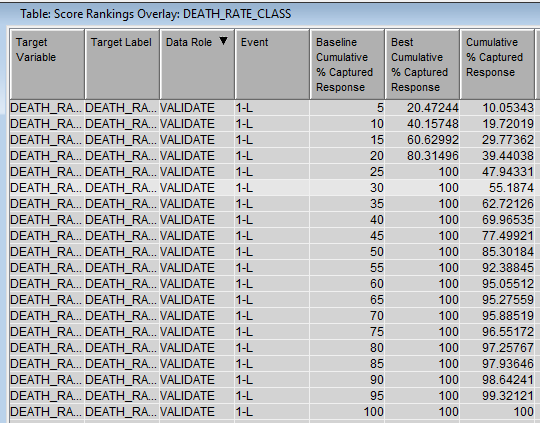
3.5.4 Leaf Statistics



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

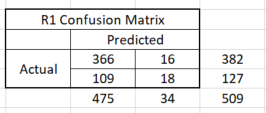
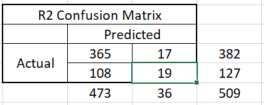
Predicted: DEATH\_RATE\_CLASS=1-L = 0.03

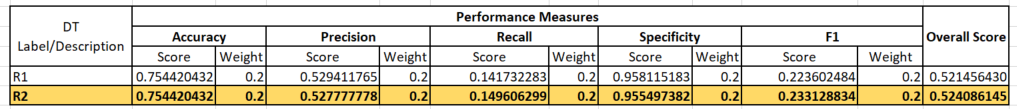
### Logistic Regression

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 | | | |

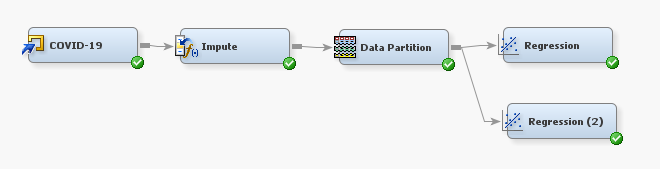
2 Summary of Results

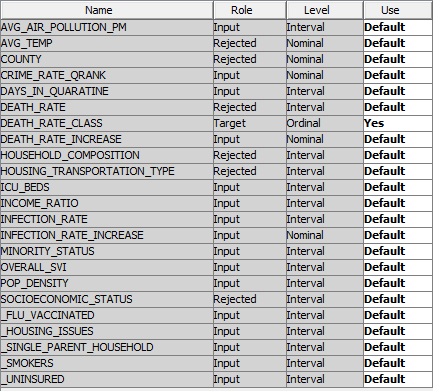


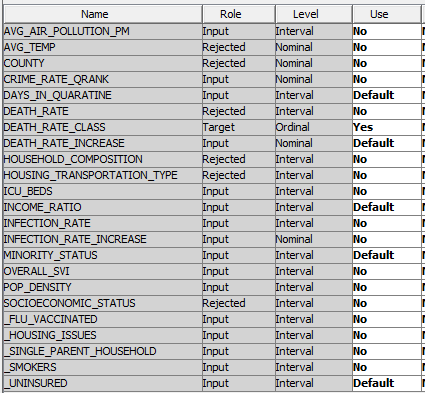
2 Evidence of Experimentation

2.1 Process Flow Diagram

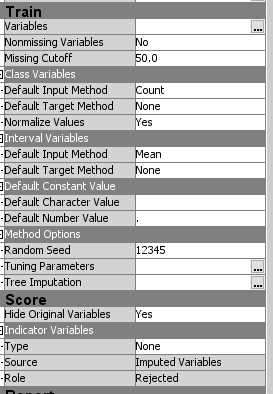


2.2 Variables

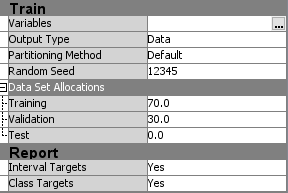




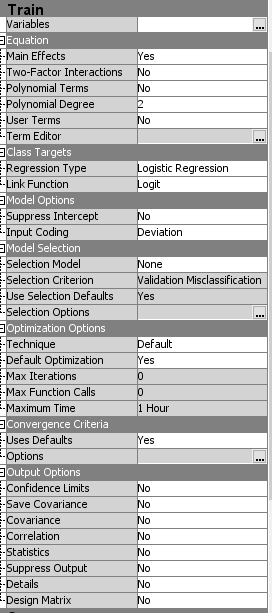
2.3 Impute Node



2.4 Data Partition Node

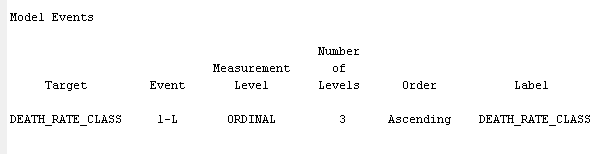


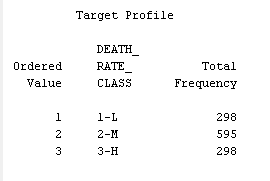
2.5 Regression Node

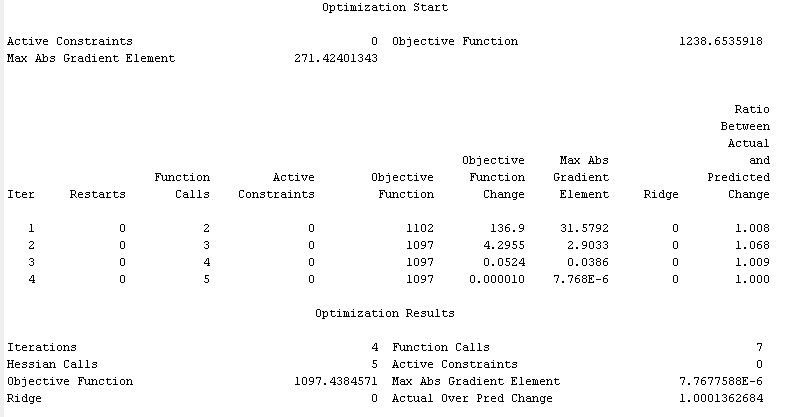


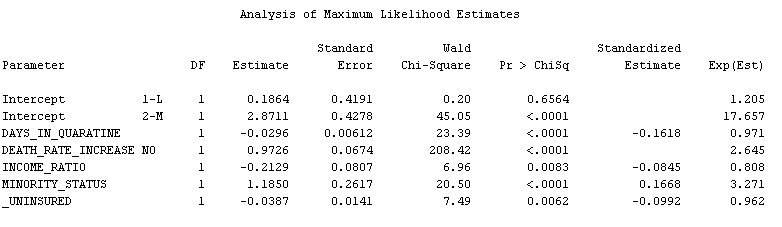
3 Results

3.1 Output

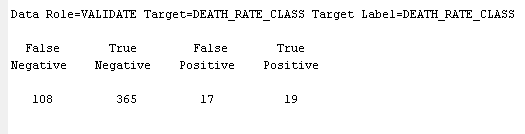




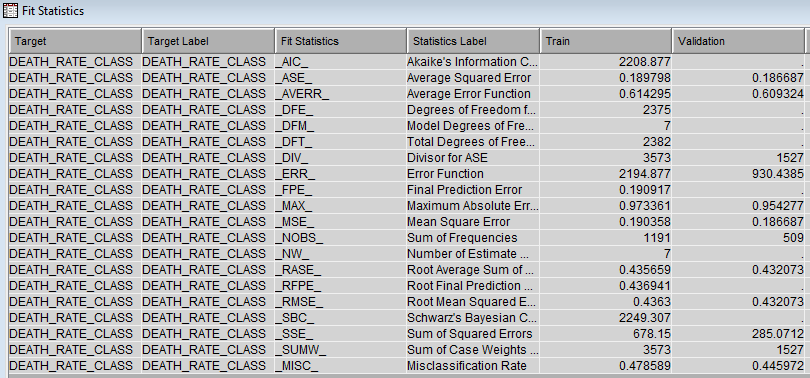




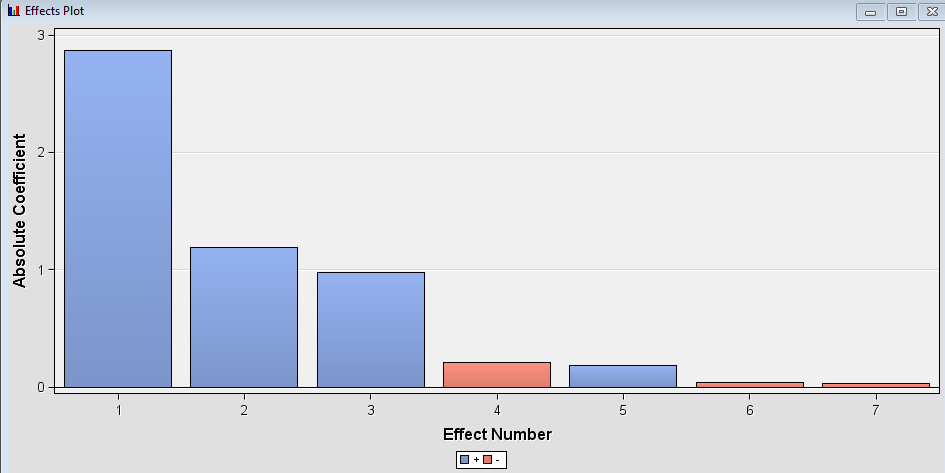
3.2 Confusion Matrix



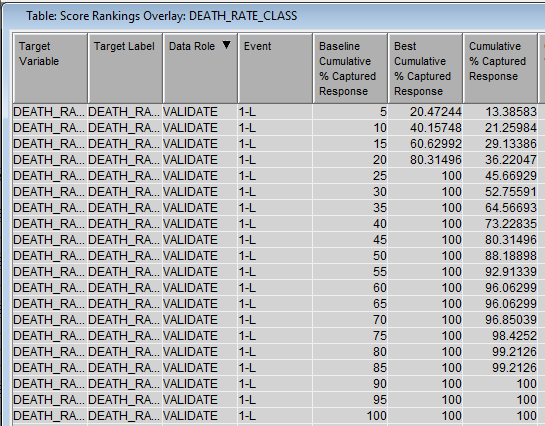
3.3 Fit Statistics



3.3 Effect Plot



3.4 Table: Score Rankings Overlay



3.5 Graph: Score Rankings Overlay %Response

