Analysis of Heart Disease Mortality Rate

Author: Adrian R Angkawijaya

Date: July 2018

Executive Summary

This document presents a summary analysis of the observations and the prediction of future heart disease mortality rate in the United States. The analysis was based on a wide range of sources that are made publicly by the United States Department of Agriculture Economic Research Service (USDA ERS) provided for the competition.

Our goal for the competition was to predict the mortality rate of heart disease (per 100,000 individuals) across the United States at a certain county areas, demographics and socioeconomic information.

The project is divided into several sections such as the following:

1. **Data Preparation** – the initial stage of preparing the data including understanding the dataset, seeing how many observations and variables, looking at the first few lines of the dataset, checking if there are any duplicates and merging the given train and test set for cleaning part on the next stage.

2. **Data Cleaning** – check for missing values and replace them with the median of each group, see if there are any features that can be transformed or make new features for easier analysis and better prediction model.

3. **Exploratory Data Analysis** – see the summary statistics of the numeric independent features and the dependent variable and their correlation with each other, create visualization plots to see the relationships between all the independent features with the dependent label variable.

4. **Prediction Model** – compare five different models by evaluating the indicated target performance metric of root mean squared error (RMSE), check the residual plots for the best model performance and see the most important features for the prediction model.

The results of the exploratory data analysis and prediction model is provided in this report. The final best prediction model was chosen to be the Light Gradient Boosting Model (LGBM) with a RMSE value of 30.72

General Overview from Exploratory Analysis

The data consists of 33 independent features (not including id) for both the training and test data with 3198 observations for the training set and 3080 observations provided for the test set. The datasets had no duplicates.

Note: *Only the training set was used for the analysis and prediction model since the test set was provided only to test the training model.*

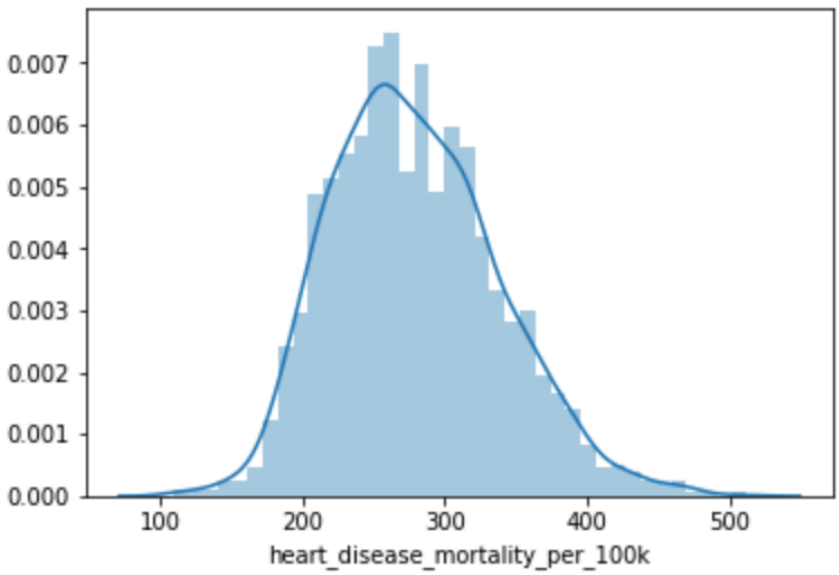
Numerical Relationships

There are 29 numeric features in the dataset and one numeric label with their summary statistics shown below:

Column	Min	Max	Mean	Median	Std Dev
econ__pct_civilian_labor	0.207	1	0.4671	0.468	0.0744
econ__pct_unemployment	0.010	0.248	0.059	0.057	0.022
econ__pct_uninsured_adults	0.046	0.496	0.217	0.216	0.067
econ__pct_uninsured_children	0.012	0.281	0.086	0.077	0.0398
demo__pct_female	0.278	0.573	0.498	0.503	0.0243
demo__pct_below_18_years_of_age	0.092	0.417	0.227	0.226	0.0342
demo__pct_aged_65_years_and_older	0.045	0.346	0.17	0.167	0.0436
demo__pct_hispanic	0	0.932	0.0901	0.035	0.1427
demo__pct_non_hispanic_african_american	0	0.858	0.091	0.022	0.147
demo__pct_non_hispanic_white	0.053	0.990	0.77	0.853	0.207
demo__pct_american_indian_or_alaskan_native	0	0.859	0.024	0.007	0.084
demo__pct_asian	0	0.341	0.0131	0.007	0.025
demo__pct_adults_less_than_a_high_school_diploma	0.01	0.473	0.148	0.133	0.068
demo__pct_adults_with_high_school_diploma	0.06	0.558	0.350	0.355	0.0705
demo__pct_adults_with_some_college	0.109	0.473	0.301	0.301	0.052
demo__pct_adults_bachelors_or_higher	0.01	0.798	0.199	0.176	0.089
demo__birth_rate_per_1k	4	29	11.676	11	2.739

demo__death_rate_per_1k	0	27	10.30	10	2.786
health__pct_adult_obesity	0.131	0.471	0.307	0.309	0.0432
health__pct_adult_smoking	0.046	0.513	0.212	0.207	0.0581
health__pct_diabetes	0.032	0.203	0.109	0.109	0.0232
health__pct_low_birthweight	0.033	0.238	0.083	0.080	0.0216
health__pct_excessive_drinking	0.038	0.367	0.164	0.163	0.042
health__pct_physical_inacticity	0.090	0.442	0.277	0.280	0.0529
health__air_pollution_particulate_matter	7	15	11.62	12	1.551
health__homicides_per_100k	-0.40	50.49	5.11	4.6	3.189
health__motor_vehicle_crash_deaths_per_100k	3.14	110.45	20.86	19.11	9.801
health__pop_per_dentist	339	28130	3371.39	2644.5	2478.27
health__pop_per_primary_care_physician	189	23399	2509.45	1969	2029.08
heart_disease_mortality_per_100k	109	512	279.36	275	58.95

Since heart disease mortality is of interest in this analysis, it was noted that the mean and median of this value are close to one another. The value of the standard deviation is also not that high indicating an approximately normal distribution in the mortality rate of heart disease as shown below.



Since the distribution is approximately normal, there was no need of further transformation. The relationship of heart disease mortality with the other variables were also noted. The table below shows the strength of the relationship with the numerical variables.

Column	Correlation with heart disease mortality rate
health__pct_physical_inacticity	0.649813
health__pct_diabetes	0.631337
health__pct_adult_obesity	0.593316
demo__pct_adults_less_than_a_high_school_diploma	0.527382
health__pct_low_birthweight	0.464391
health__pct_adult_smoking	0.463138
demo__death_rate_per_1k	0.444757
health__motor_vehicle_crash_deaths_per_100k	0.435633
demo__pct_adults_with_high_school_diploma	0.428137
demo__pct_non_hispanic_african_american	0.375537
econ__pct_unemployment	0.371620
econ__pct_uninsured_adults	0.334027
health__pop_per_dentist	0.292447
health__homicides_per_100k	0.292377
health__pop_per_primary_care_physician	0.217936
health__air_pollution_particulate_matter	0.147028
demo__birth_rate_per_1k	0.142176
demo__pct_below_18_years_of_age	0.121884
demo__pct_female	0.086765
demo__pct_american_indian_or_alaskan_native	0.004826
econ__pct_uninsured_children	-0.034209
demo__pct_aged_65_years_and_older	-0.056081
demo__pct_hispanic	-0.111976

demo__pct_non_hispanic_white	-0.157797
demo__pct_asian	-0.267016
health__pct_excessive_drinking	-0.300781
demo__pct_adults_with_some_college	-0.340764
econ__pct_civilian_labor	-0.476644
demo__pct_adults_bachelors_or_higher	-0.541385

It is shown that physical inactivity, diabetes, obesity and adults with less than a high school diploma have strong positive correlations with heart disease mortality rate while adults with bachelors or higher degree has a strong negative relationship. Some feature such as whether a person is American Indian or Alaskan native has close to no relationship with heart disease mortality.

Categorical Relationships

In addition to the numerical values, the observation also contain categorical features, including:

- ***area__rucc*** – indicating the Rural-Urban Continuum Codes with these elements:

  - Nonmetro - Urban population of 2,500 to 19,999, adjacent to a metro area
  - Nonmetro - Completely rural or less than 2,500 urban population, not adjacent to a met ro area
  - Metro - Counties in metro areas of 1 million population or more
  - Nonmetro - Urban population of 2,500 to 19,999, not adjacent to a metro area
  - Metro - Counties in metro areas of 250,000 to 1 million population
  - Metro - Counties in metro areas of fewer than 250,000 population
  - Nonmetro - Completely rural or less than 2,500 urban population, adjacent to a metro a rea
  - Nonmetro - Urban population of 20,000 or more, adjacent to a metro area
  - Nonmetro - Urban population of 20,000 or more, not adjacent to a metro area

- ***area__urban_influence*** – indicating the Urban Influence Codes with these elements:

  - Small-in a metro area with fewer than 1 million residents
  - Large-in a metro area with at least 1 million residents or more
  - Noncore adjacent to a small metro with town of at least 2,500 residents
  - Micropolitan adjacent to a small metro area
  - Micropolitan not adjacent to a metro area
  - Noncore adjacent to micro area and does not contain a town of at least 2,500 residents
  - Noncore not adjacent to a metro/micro area and does not contain a town of at least 2,5 00 residents

- Noncore adjacent to micro area and contains a town of 2,500-19,999 residents
- Noncore adjacent to a small metro and does not contain a town of at least 2,500 residen ts
- Noncore adjacent to a large metro area
- Micropolitan adjacent to a large metro area
- Noncore not adjacent to a metro/micro area and contains a town of 2,500 or more resi dents

- ***econ__economic_typology*** – indicating County Typology Codes with these elements:

  - Nonspecialized
  - Manufacturing-dependent
  - Farm-dependent
  - Federal/State government-dependent
  - Recreation
  - Mining-dependent

- ***yr*** – indicating two particular years with the elements a and b representing year a and year b.

Bar charts were created to observe the frequency of these features, summarized as follows:

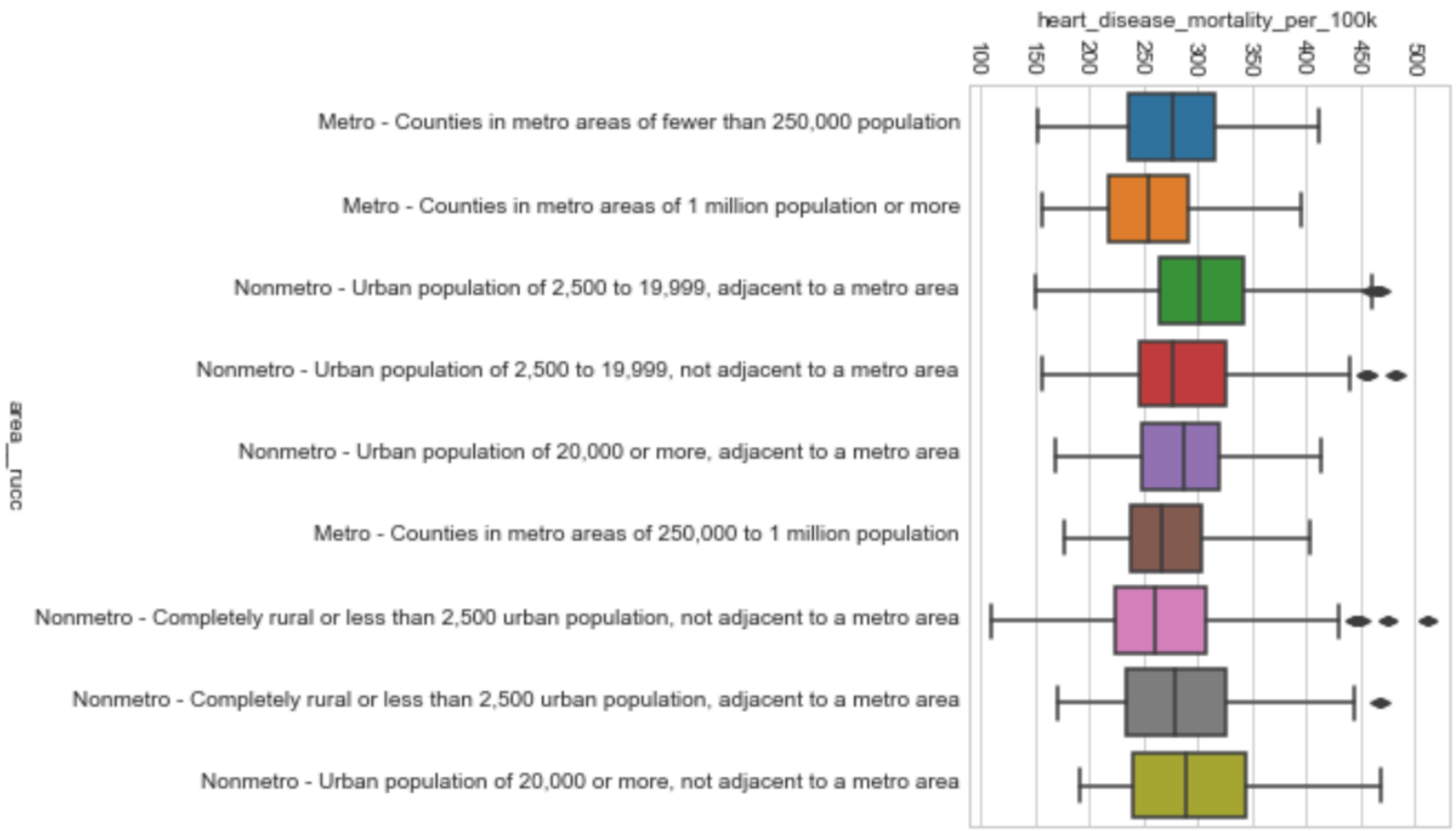
- The most common county was from the **nonmetro area with urban population of 2,500 to 19,999, adjacent to a metro area** while the least common one was from the **nonmetro area with urban population of 20,000 or more, not adjacent to a metro area**

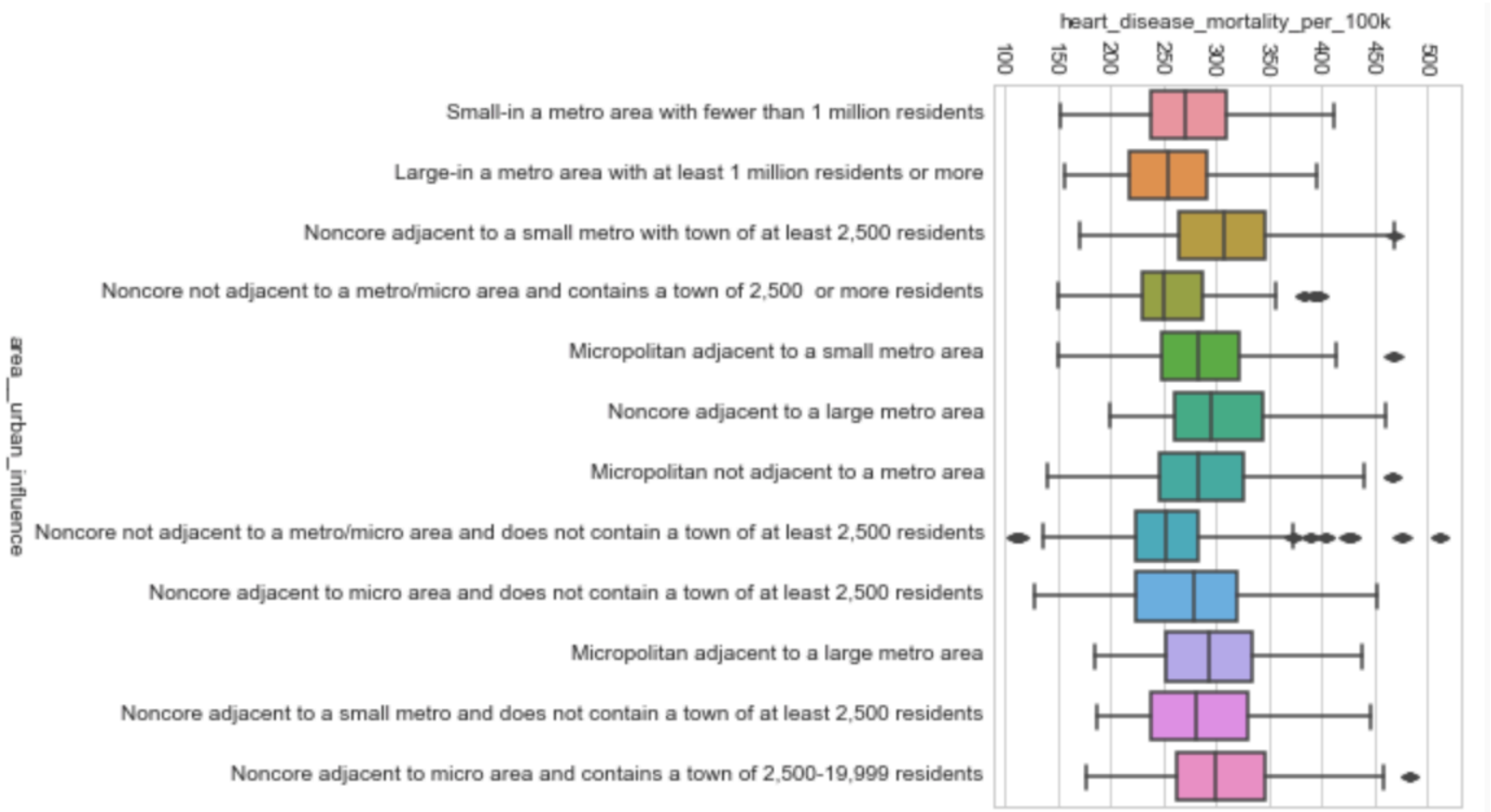
- The most common urban influence was from the **small-in a metro area with fewer than 1 million residents** while the least common one was from the **noncore not adjacent to a metr o/micro area and contains a town of 2,500 or more residents**

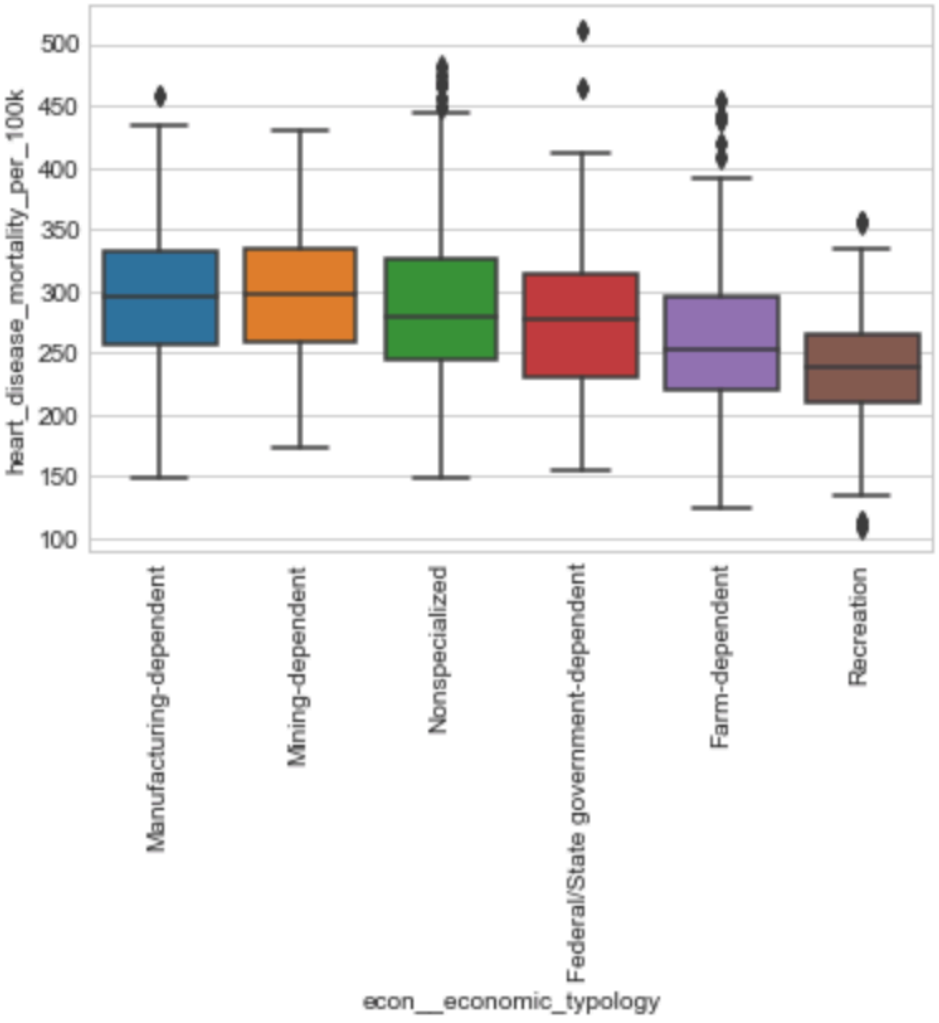
- The most common economic dependence was the **nonspecialized** type while the least com mon one was the **mining-dependent** type.

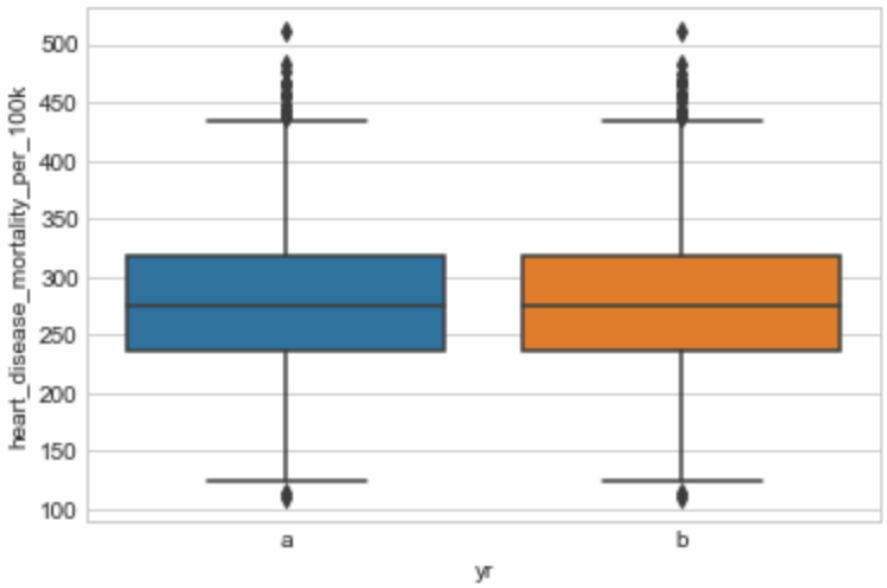
- Both year **a** and **b** have equal count of frequency.

The relationship between heart disease mortality with the above categorical features were also created. The following box plots show the categorical features relationship with heart disease mortality rate:


heart_disease_mortality_per_100k
100
150
200
250
300
350
400
450
500
area_rucc
Metro - Counties in metro areas of fewer than 250,000 population
Metro - Counties in metro areas of 1 million population or more
Nonmetro - Urban population of 2,500 to 19,999, adjacent to a metro area
Nonmetro - Urban population of 2,500 to 19,999, not adjacent to a metro area
Nonmetro - Urban population of 20,000 or more, adjacent to a metro area
Metro - Counties in metro areas of 250,000 to 1 million population
Nonmetro - Completely rural or less than 2,500 urban population, not adjacent to a metro area
Nonmetro - Completely rural or less than 2,500 urban population, adjacent to a metro area
Nonmetro - Urban population of 20,000 or more, not adjacent to a metro area


heart_disease_mortality_per_100k
100
150
200
250
300
350
400
450
500
area__urban_influence
Small-in a metro area with fewer than 1 million residents
Large-in a metro area with at least 1 million residents or more
Noncore adjacent to a small metro with town of at least 2,500 residents
Noncore not adjacent to a metro/micro area and contains a town of 2,500 or more residents
Micropolitan adjacent to a small metro area
Noncore adjacent to a large metro area
Micropolitan not adjacent to a metro area
Noncore not adjacent to a metro/micro area and does not contain a town of at least 2,500 residents
Noncore adjacent to micro area and does not contain a town of at least 2,500 residents
Micropolitan adjacent to a large metro area
Noncore adjacent to a small metro and does not contain a town of at least 2,500 residents
Noncore adjacent to micro area and contains a town of 2,500-19,999 residents


heart_disease_mortality_per_100k
500
450
400
350
300
250
200
150
100
Manufacturing-dependent
Mining-dependent
Nonspecialized
Federal/State government-dependent
Farm-dependent
Recreation
econ__economic_typology


heart_disease_mortality_per_100k
500
450
400
350
300
250
200
150
100
a
b
yr

Some key points to take away from the above relationship box plots include:

- The difference of heart disease mortality rate for the variety of counties seem to be simi lar having a similar range rate and median across all counties. However, the median mor tality rate for counties considered metro is lower than the counties considered as nonm etro.

- The urban influence feature also seem to have similar median and range mortality rate a cross all elements with some having noticeable distance high values or what classified as *outliers*.

- The economic dependence of Recreation type has the smallest range of mortality rate c ompared to the other economic dependence type. It also has the lowest median compar ed to the others that seemed to have similar rates.

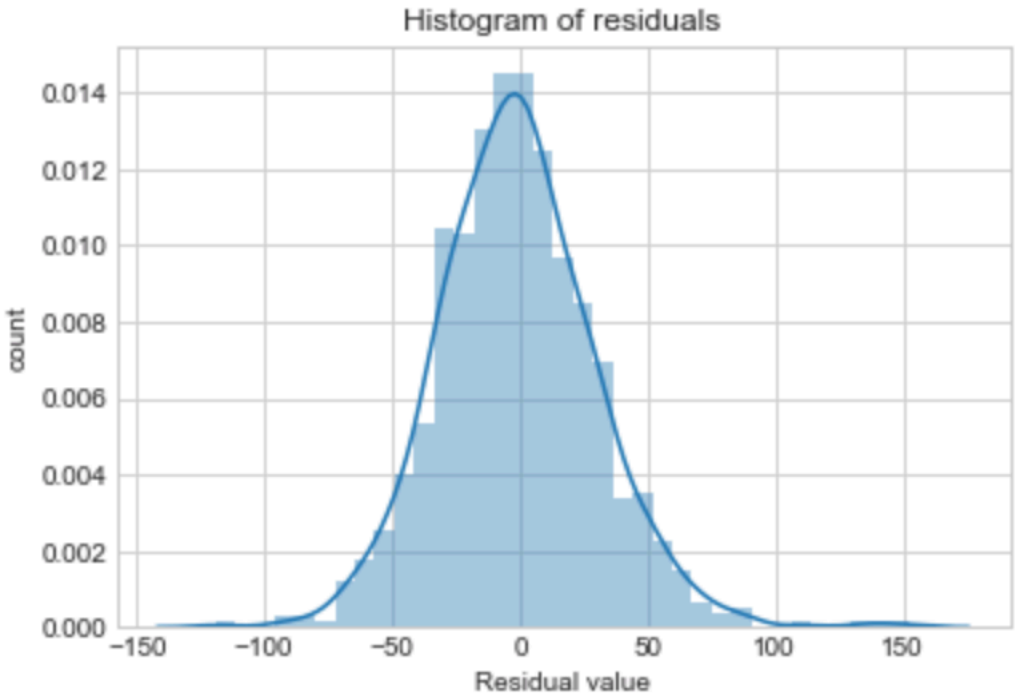
- The distribution between the two years, a and b to the mortality rate are exactly the sa me.

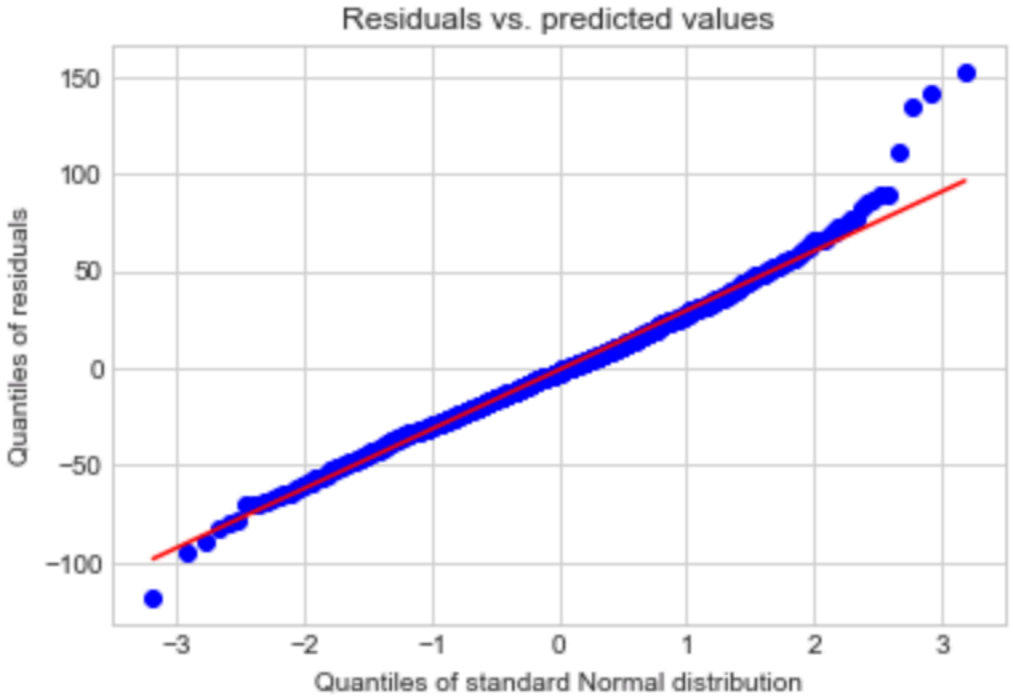
Prediction Model

The model was trained 70% of the data and evaluated with the remaining 30%. There are five r egression models that were built to predict the heart disease mortality rate. One model was ch osen by comparing their performance metric RMSE value and residual plots. The best model res ulted from using the Light Gradient Boosting Regressor algorithm having the lowest RMSE value and normally distributed residual plots.

Light Gradient Boosting Regressor (LGBM)

The LGBM model resulted in an R^2 value of 0.72 that is interpreted as 72% of the variability of the response data is explain by the model. The model has an RMSE value of 30.72 which was the lowest value out of the five models compared. The values of mean and median absolute errors were relatively close as well indicating a decent model performance. As mentioned earlier, the residual plots represents a normal distribution plot as shown on the next page.


Histogram of residuals
count
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
−150
−100
−50
0
50
100
150
Residual value


Residuals vs. predicted values
Quantiles of residuals
150
100
50
0
−50
−100
−3
−2
−1
0
1
2
3
Quantiles of standard Normal distribution

The top 10 most importance features of the model for predicting heart disease mortality rate were also noted. The list below shows the top 10 most important features for the LGBM model to predict mortality rate starting from the highest importance.

- demo__pct_aged_65_years_and_older
- demo__death_rate_per_1k
- demo__pct_adults_less_than_a_high_school_diploma
- demo__pct_non_hispanic_white
- demo__pct_american_indian_or_alaskan_native
- health__pop_per_primary_care_physician
- demo__pct_adults_with_some_college
- demo__pct_adults_bachelors_or_higher
- econ__pct_uninsured_adults
- health__pop_per_dentist

Conclusion

This analysis has shown that the different variety of features can be used to predict heart disease mortality rate in the United States. Particularly, the demographics having people aged 65 years and older, death rate, adults having less than a high school diploma have the most significant effect on determining the mortality rate in the United States.