Predicting Bank Loan Defaults with Machine Learning *

Samuel Chan *Independent Scholar*

The following report describes the methodology and exploratory process in the prediction of loan defaults on an applicant level using indicators gathered from a standard loan application form.

Keywords: bank loan defaults, socioeconomic, finhacks, financial

Executive Summary

The following report describes the methodology and exploratory process in the prediction of bank loan defaults on applicant level using indicators gathered from a standard loan application form. The data is compiled and provided by FinHacks 2018.

Having explored the data using descriptive statistics and correlation tables, the author begin by inspecting relationships between variables using a series of visualization. The visualization is done using Hadley Wickham's `ggplot2` library, and helps to unpack hidden relationships between the different socioeconomic variables in relation to heart disease mortality.

After exploring the data, a classification model is constructed to predict bank loan default from the socioeconomic features given. The author also concluded from his analysis that while many factors are helpful indicators, features with the most significant role in the prediction of bank loan defaults are socioeconomic factors and to a lesser extent, historical factors.

Gender, unemployment rate, racial distribution, and demographic variables does not seem to offer any statistically significant correlation with the target variable and on its own, contribute little information to an applicant's likelihood of loan default.

To maintain independence, the following sections have been developed with a fictional, unrelated case of "health disease risk prediction". The report should serve as a template or reference - participants are encouraged to apply creativity in developing their own analysis / reports.

*Replication files are available on the author's Github account (http://github.com/onlyphantom).

The Dataset

The analysis and resulting project is based on 3,198 observations of county-level data, each containing specific characteristics of the county collected over the span of two years (denoted `a` and `b` in the `yr` variable). Each county is not identified by name, but carries a unique `row_id`. The variables in the dataset have names of the form `category__variable`, where `category` is the high level category of the variable (e.g. econ or health) and `variable` is what the specific column contains.

The dataset contains 34 variables, with the last, `heart_disease_mortality_per_100k` being the response variable (target). Among the predictor variables are four categorical variables:

- `area__rucc`: Rural-Urban Continuum Codes "form a classification scheme that distinguishes metropolitan counties by the population size of their metro area, and nonmetropolitan counties by degree of urbanization and adjacency to a metro area. Each county in the U.S. is assigned one of the 9 codes."

- `area__urban_influence`: Urban Influence Codes "form a classification scheme that distinguishes metropolitan counties by population size of their metro area, and nonmetropolitan counties by size of the largest city or town and proximity to metro and micropolitan areas."

- `econ__economic_typology`: County Typology Codes "classify all U.S. counties according to six mutually exclusive categories of economic dependence and six overlapping categories of policy-relevant themes. The economic dependence types include farming, mining, manufacturing, Federal/State government, recreation, and nonspecialized counties."

- `yr`: One of the two years for that particular record

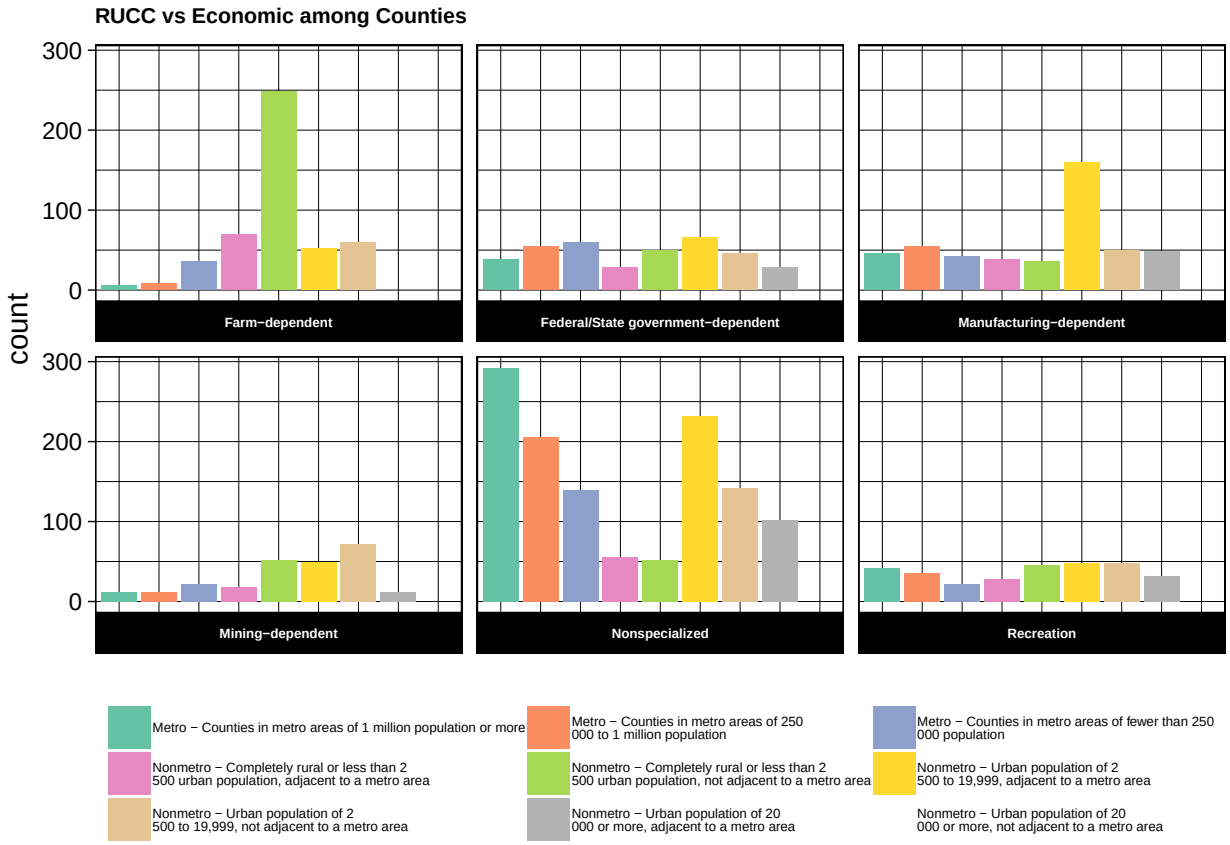

Figure 1: Visualizing Categorical Variables in our Dataset

The remaining, non-categorical, variables are numerical:

Numeric.variables	First.6.values
econ__pct_civilian_labor	0.408,0.556,0.541,0.5,0.471,0.501
econ__pct_unemployment	0.057,0.039,0.057,0.061,0.05,0.048
econ__pct_uninsured_adults	0.254,0.26,0.07,0.203,0.225,0.212
econ__pct_uninsured_children	0.066,0.143,0.023,0.059,0.103,0.055
demo__pct_female	0.516,0.503,0.522,0.525,0.511,0.516
demo__pct_below_18_years_of_age	0.235,0.272,0.179,0.2,0.237,0.207
demo__pct_aged_65_years_and_older	0.176,0.101,0.115,0.164,0.171,0.121
demo__pct_hispanic	0.109,0.41,0.202,0.013,0.025,0.022
demo__pct_non_hispanic_african_american	0.039,0.07,0.198,0.049,0.008,0.046
demo__pct_non_hispanic_white	0.829,0.493,0.479,0.897,0.953,0.903
demo__pct_american_indian_or_alaskan_native	0.004,0.008,0.013,0.007,0.003,0.002
demo__pct_asian	0.011,0.015,0.085,0.001,0,0.006
demo__pct_adults_less_than_a_high_school_diploma	0.194,0.164,0.159,0.182,0.122,0.138
demo__pct_adults_with_high_school_diploma	0.424,0.234,0.238,0.407,0.413,0.295
demo__pct_adults_with_some_college	0.227,0.342,0.186,0.249,0.307,0.281
demo__pct_adults_bachelors_or_higher	0.154,0.259,0.417,0.163,0.157,0.287
demo__birth_rate_per_1k	12,19,12,11,14,11
demo__death_rate_per_1k	12,7,6,12,12,8
health__pct_adult_obesity	0.297,0.288,0.212,0.285,0.284,0.283
health__pct_adult_smoking	0.23,0.19,0.156,NA,0.234,0.22
health__pct_diabetes	0.131,0.09,0.084,0.104,0.137,0.112
health__pct_low_birthweight	0.089,0.082,0.098,0.058,0.07,0.089
health__pct_excessive_drinking	NA,0.181,0.195,NA,0.194,0.067
health__pct_physical_inacticity	0.332,0.265,0.209,0.238,0.29,0.272
health__air_pollution_particulate_matter	13,10,10,13,9,13
health__homicides_per_100k	2.8,2.3,9.31,NA,NA,3.8
health__motor_vehicle_crash_deaths_per_100k	15.09,19.79,3.14,NA,29.39,13.74
health__pop_per_dentist	1650,2010,629,1810,3489,2439
health__pop_per_primary_care_physician	1489,2480,690,6630,2590,1540

The variables in our dataset adopt a naming scheme that takes the form of `category__variable`, where category is the high level category of the variable (e.g.`area`, `econ`, `demo` or `health`) and variable is what the specific column contains.

From looking at the first 6 values of these variables, it is clear that there are missing values (`NA`) in some of these socioeconomic indicators, and that our analysis will need to include the necessary preprocessing steps to deal with them. When we take a closer look at the variables (Figure 2), it also seems like some rows (county) have a higher proportion of missing values compared to the other counties; Some variables, such as the `health_homicides_per_100k` also seems to contain more missing observations than others.

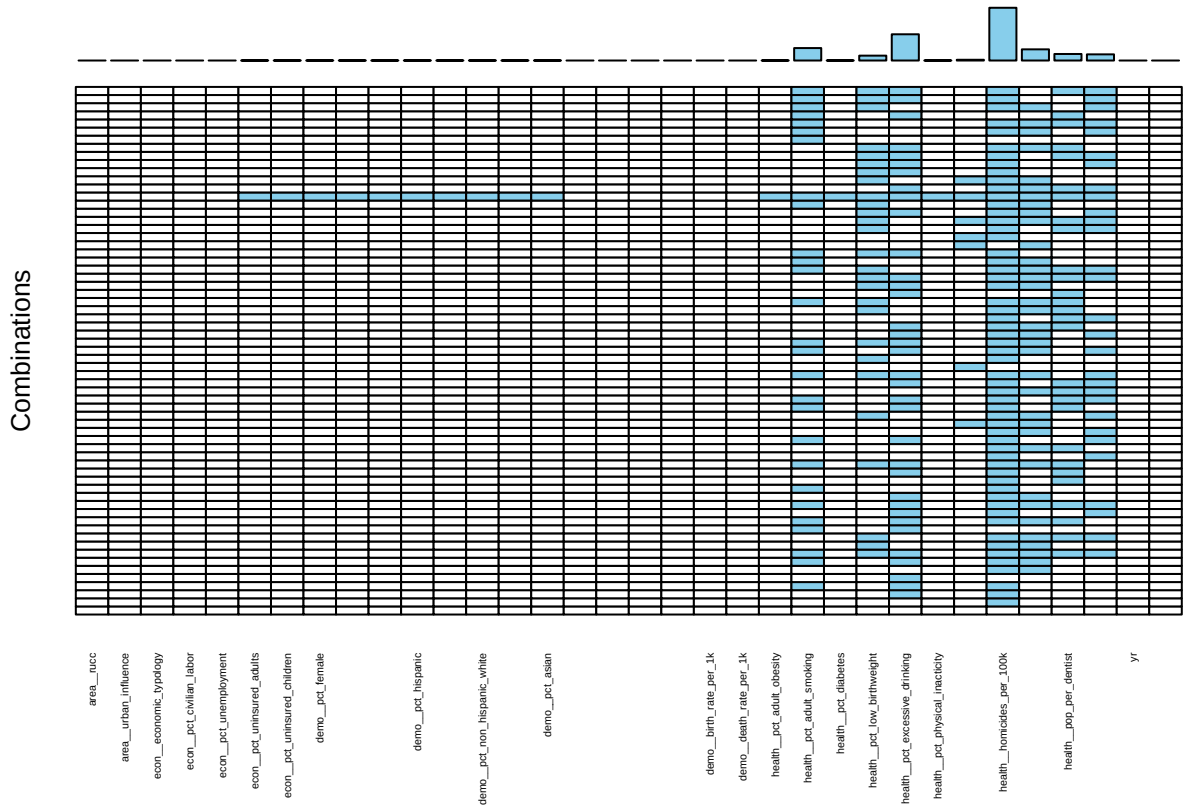

Figure 2: Inspecting the Missing Values in our Dataset

Target Variable

The target is `heart_disease_mortality_per_100k`, a **positive integer** that indicates the rate of heart disease per 100,000 individuals across the United States at the county-level. Within the data, we can observe that the heart disease mortality rate has a pretty wide spread, ranging from 109-per-100,000 to 512-per-100,000. Using the socioeconomic variables from the data, this report aims to highlight the identified contributors and any existing correlations between the target variable and underlying socioeconomic factors.

```
> summary(project_full$heart_disease_mortality_per_100k)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  109.0   237.0   275.0   279.4   317.0   512.0
```

Data Cleansing and Feature Engineering

An explanation of the process and techniques used to analyze the data, including data cleansing, calculation of statistics, visualization and exploration, modeling, and testing.

Imputation by the EMB algorithm

> Missing data is a ubiquitous problem in social science data. Respondents do not answer every question, countries do not collect statistics every year, archives are incomplete, subjects drop out of panels. Most statistical analysis methods, however, assume the absence of missing data, and are only able to include observations for which every variable is measured[1]

If we were to use the complete-case approach by omitting the rows with incomplete measurement, we would effectively eliminate 66.57% of the original data, leaving us with too little to base our research on. In statistics, this method is synonymous with "listwise deletion" where a record is excluded from analysis if any single variable is missing.

Instead, the approach we'll take is multiple imputation, an approach that is versatile and have some great advantages over the other options. Specifically, multiple imputation have been shown to reduce bias compared to the complete-case / listwise deletion approach, as it retains all existing information in the data instead of excluding potentially significant "eviedence" from the analysis. Ad-hoc methods of imputation, such as imputation using the mean or median, may be too limited by its simplicity and introduce strong biases in variances as well as covariances in the data.

The application of multiple imputation on our dataset involves the generation of m values for each missing cell in our data matrix and thus resulting in m number of "complete" datasets. Across these different datasets, the observed values are the same but the missing values are filled in with a distribution of imputations that reflect the uncertainty about the missing data. To keep this project viable while respecting the laid out deadline for this project, we will pick a random set from m and assume that the imputation on this set is a good estimate of the missing values. The package we're using is the second version of Amelia (Amelia II), which itself implements the ideas developed by Honaker and King[2].

The algorithm used to construct the imputation model is termed as EMB, or "Expectation-Maximization with Bootstrapping". One of the assumptions it make is that the complete data are multivariate normal:

$D \sim \mathcal{N}(\mu, \Sigma)$

Which states that dataset D has a multivariate normal distribution with mean vector μ and covariance matrix Σ. This distribution is often a crude approximation of the true distribution of the data, however there are evidence to suggest that the model works just as well when applied on categorical or mixed data[3].

The imputation model also makes the assumption that the data are missing at random (MAR), which is to say that the pattern of missingness only depends on the observed data D^{obs} and not the unobserved data D^{mis}.

These assumptions are plausible with our dataset: The 3,198 observations do seem to meet the missing-at-random assumption, and that it is plausibly multivariate normal.

[1] Honaker, J., King, G., Blackwell, M., 2018, "AMELIA II: A Program for Missing Data"

[2] Honaker, J., King, G., 2010, "What to do about missing values in time series cross-section data", American Journal of Political Science 54(2):561-581

[3] Schafer, J., 1997, Analysis of incomplete multivariate data. London: Chapman & Hall.

Feature Engineering

To facilitate the exploratory process, I've taken the post-imputated data and engineered new features on top of its existing ones. Features in our data have a direct influence in the predictive models that we will obtain through the means of machine learning, and proper investment into this process can be quite crucial to discovering important insights in later phases of the analysis.

The four variables we've created are:

- `largeold`: A **categorical** variable indicating if the county has a large (more than 20%) population aged 65 and above
- `metronot`: A **categorical** variable indicating if the county is in a Metro area
- `nonwhiteasian`: A **numeric** variable that aggregates the percentage of population that identifies as American Indian, Alaskan Native, Hispanic, or African American
- `healthissue`: A **numeric** variable that aggregates the percentage of population that meet the clinical definition of obesity, the percentage of population that smoke, the percentage of population with diabetes and the percentage of population that is physically inactive

An example of creating summaries using the newly engineered features:

```
, , metronot = metro

                                  largeold
econ__economic_typology            largeold normal
  Farm-dependent                          4     38
  Federal/State government-dependent      0     98
  Manufacturing-dependent                 3     85
  Mining-dependent                        2     32
  Nonspecialized                          8    424
  Recreation                             19     45

, , metronot = nonmetro

                                  largeold
econ__economic_typology            largeold normal
  Farm-dependent                       258    182
  Federal/State government-dependent    41    251
  Manufacturing-dependent               40    366
  Mining-dependent                      31    189
  Nonspecialized                       119    715
  Recreation                           138    110
```

Exploratory Data Analysis

1. Average heart disease mortality per 100,000 by Economic Type, Population Age and Area Type (Aggregated):

Table 2: Manufacturing and Mining -dependent counties observed higher heart disease mortality rate

	econ___economic_typology	largeold	metronot	heart_disease_mortality_per_100k
2	Manufacturing-dependent	largeold	metro	336.3333
15	Mining-dependent	largeold	nonmetro	307.6452
21	Mining-dependent	normal	nonmetro	303.0053
22	Nonspecialized	normal	nonmetro	302.3049
20	Manufacturing-dependent	normal	nonmetro	300.3115
14	Manufacturing-dependent	largeold	nonmetro	290.6000
9	Mining-dependent	normal	metro	288.2188
19	Federal/State government-dependent	normal	nonmetro	284.3586
16	Nonspecialized	largeold	nonmetro	279.7311
8	Manufacturing-dependent	normal	metro	273.5529
18	Farm-dependent	normal	nonmetro	273.0934
6	Farm-dependent	normal	metro	270.7895
7	Federal/State government-dependent	normal	metro	270.1837
4	Nonspecialized	largeold	metro	268.5000
10	Nonspecialized	normal	metro	263.4599

7 out of the top 10 counties by heart disease mortality are counties with **Manufacturing** or **Mining** dependent economic activities. Considering that there are a total of 6 possible economic type classes and that counties with Manufacturing or Mining dependent economic activities account for less than 23.4% of all counties, it's rather telling that these two economic classes is so dominant among the top 10 counties by heart disease mortality.

Table 3: Economic Typology by Proportion

Economic Typology	Frequency
Farm-dependent	0.1507192
Federal/State government-dependent	0.1219512
Manufacturing-dependent	0.1544715
Mining-dependent	0.0794246
Nonspecialized	0.3958724
Recreation	0.0975610

2. Comparing Aggregated Population Age with Economic Typology across County Area

A cross-tabulation of the `largeold` variable and the `econ__economic_typology` variable within the data will reveal that between metro and non-metro, the distinction in economic typology varied by population age is most stark in regions with **Farm-dependent** activities and **Recreation** activities:

```
, , metronot = metro

                                  largeold
econ__economic_typology            largeold normal
  Farm-dependent                          4     38
  Federal/State government-dependent      0     98
  Manufacturing-dependent                 3     85
  Mining-dependent                        2     32
  Nonspecialized                          8    424
  Recreation                             19     45

, , metronot = nonmetro

                                  largeold
econ__economic_typology            largeold normal
  Farm-dependent                        258    182
  Federal/State government-dependent     41    251
  Manufacturing-dependent                40    366
  Mining-dependent                       31    189
  Nonspecialized                        119    715
  Recreation                            138    110
```

In metro areas, all counties regardless of economic activities have a larger proportion of relatively young population. For Farm-dependent counties, that number is 38 to 4, and for Recreation counties that number is 45 to 19.

In non-metro areas, counties with farm-dependent economies and recreation boast a larger number than their younger counterparts, outnumbering them by 252 to 182, and 138 and 110 respectively. That is in stark contrast to the 4 to 38 ratio in metro counties.

3. Heart Disease Mortality broken down by Economic Typology, Population Age and other Health Indicators

Recalling that the `healthissue` feature we created earlier point to the aggregated mean of percentages of population that are obese, smoking, having diabetes, or physically inactive. When we bin this variable into "high" or "low" depending on whether or not the county exceed the sample average in this dataset, the result is very telling.

Among the top 12 counties with highest rate of heart disease mortality, 11 of them observed a `healthissue` rating that is greater than the sample mean. Obesity, smoking, diabetes or rate of population that is physically inactive turns out to be strong indicators of heart disease mortality independent of area (metro or rural) and independent of the proportion of senior residents:

Table 4: 11 out of top 12 Counties by Heart Disease Mortality have Poorer Than Average Health Indicators (Obesity, Diabetes, Smoking and Physically Inactive)

	econ___economic_typology	largeold	healthindicators	heart_disease_mortality_per_100k
10	Mining-dependent	normal	High	330.2368
11	Nonspecialized	normal	High	327.1744
3	Manufacturing-dependent	largeold	High	323.7391
9	Manufacturing-dependent	normal	High	318.4211
8	Federal/State government-dependent	normal	High	318.2750
4	Mining-dependent	largeold	High	312.7619
5	Nonspecialized	largeold	High	308.6154
7	Farm-dependent	normal	High	304.3491
2	Federal/State government-dependent	largeold	High	291.6923
16	Mining-dependent	largeold	Low	289.0833
12	Recreation	normal	High	279.0000
6	Recreation	largeold	High	276.7000

Key Findings and Visualizations

1. Across area types (metro or rural counties), economic activities and counties age demography (large proportion of senior population or without), health indicators such as obesity and physical inactivity among its residents are good predictors of heart disease mortality rate

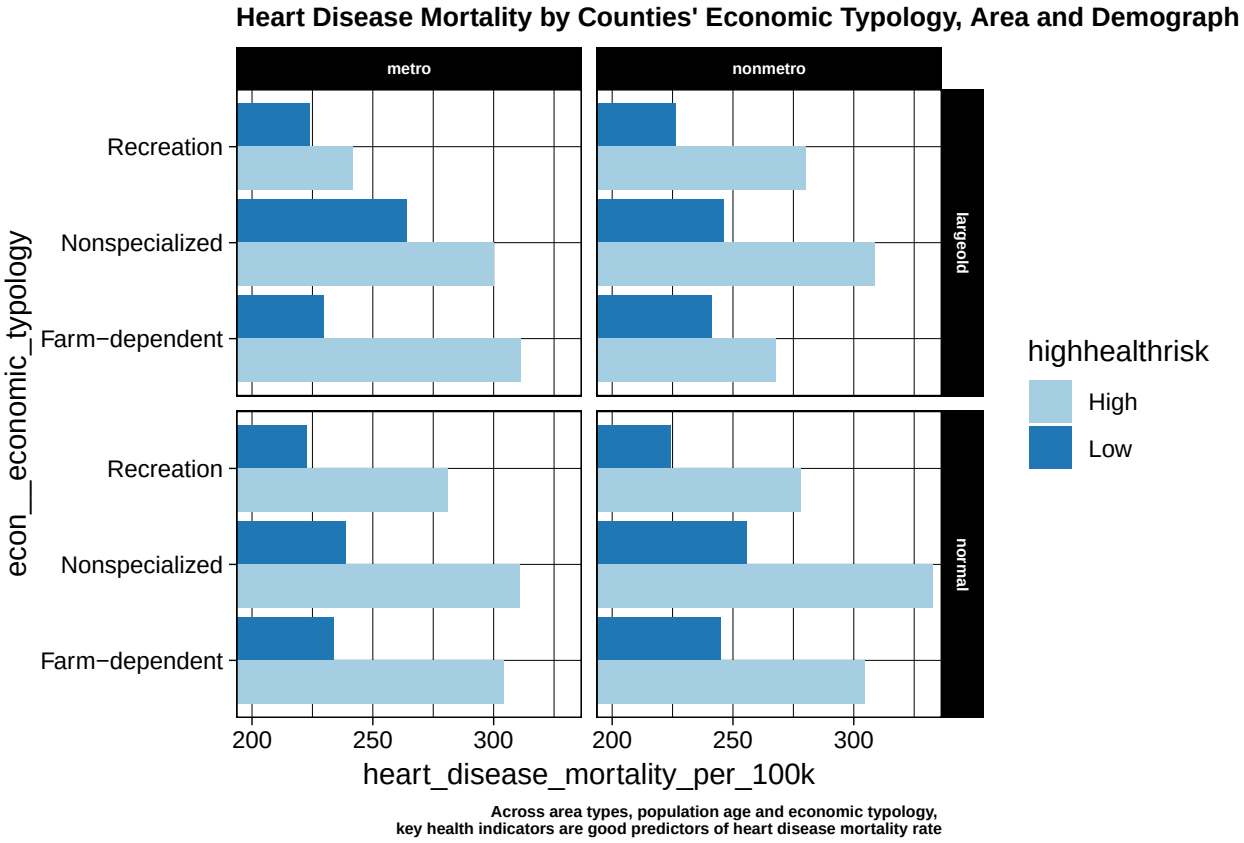

Figure 3: Key health indicators such as obesity and physical inactivity good predictors of heart disease mortality rate

2. Without knowledge of the counties' key health indicators (smoking, physical inactivity, obesity and diabetes), air pollution itself doesn't seem to offer a clear cut correlation with heart disease mortality. When key health indicators are taken into account, it seems that counties with high rate of health issues (smoking, physical inactivity, obesity and diabetes) do observe a stronger correlation between air pollution levels and heart disease mortality rate.

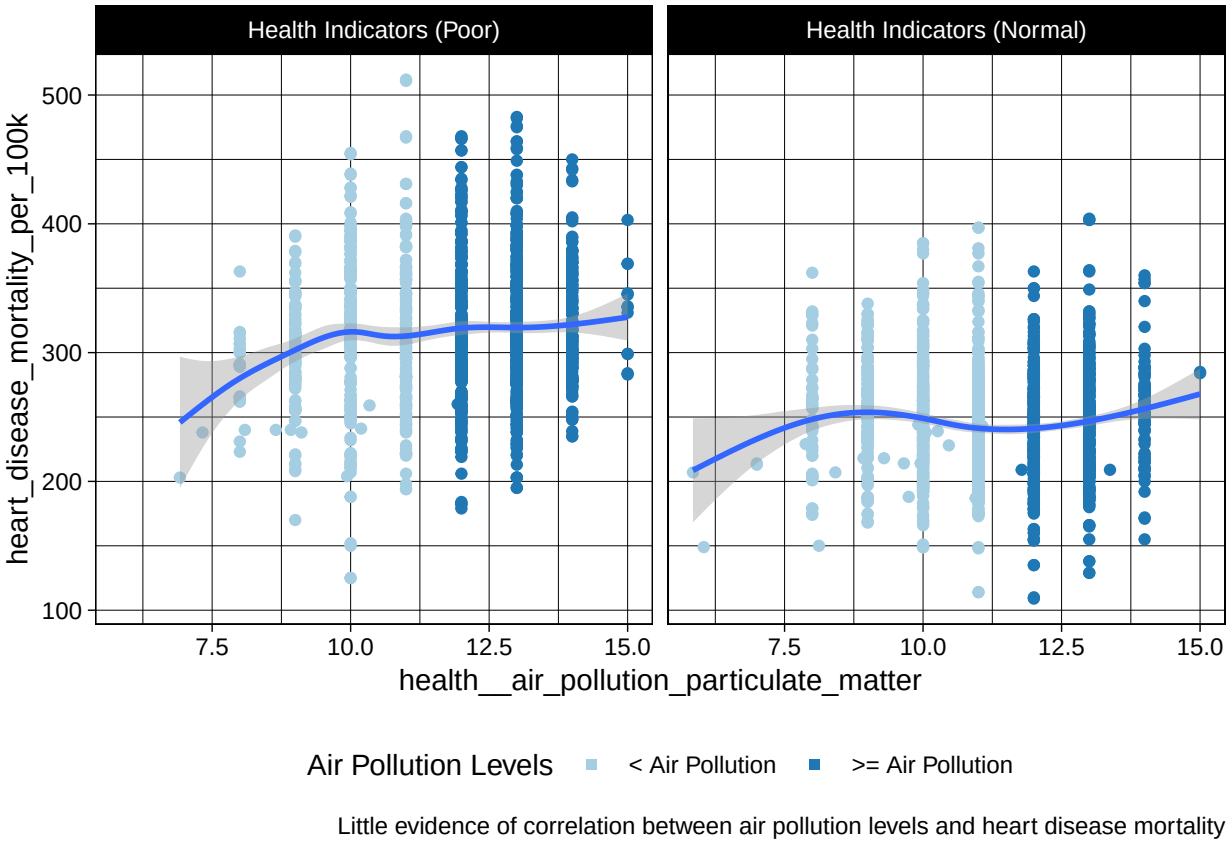

Figure 4: Comparing heart disease mortality rate using key health indicators and air pollution levels

Table 5: Correlation: Air Pollution Particulate Matter vs Heart Disease Mortality Rate

	Air Pollution Particulate Matter	Heart Disease Mortality Rate
health__air_pollution_particulate_matter	1.0000000	0.1610833
heart_disease_mortality_per_100k	0.1610833	1.0000000

3. Between counties with a higher proportion of female residence as compared to male, some deviations can be observed in the heart disease mortality rate but far stronger predictor is still key health indicators. In fact, the general rule seems to be that poor health indicators is a universal predictor of an above-average (indicated by the dashed line) heart disease mortality rate, with only one exception: Recreation-dependent counties with more female than male population.

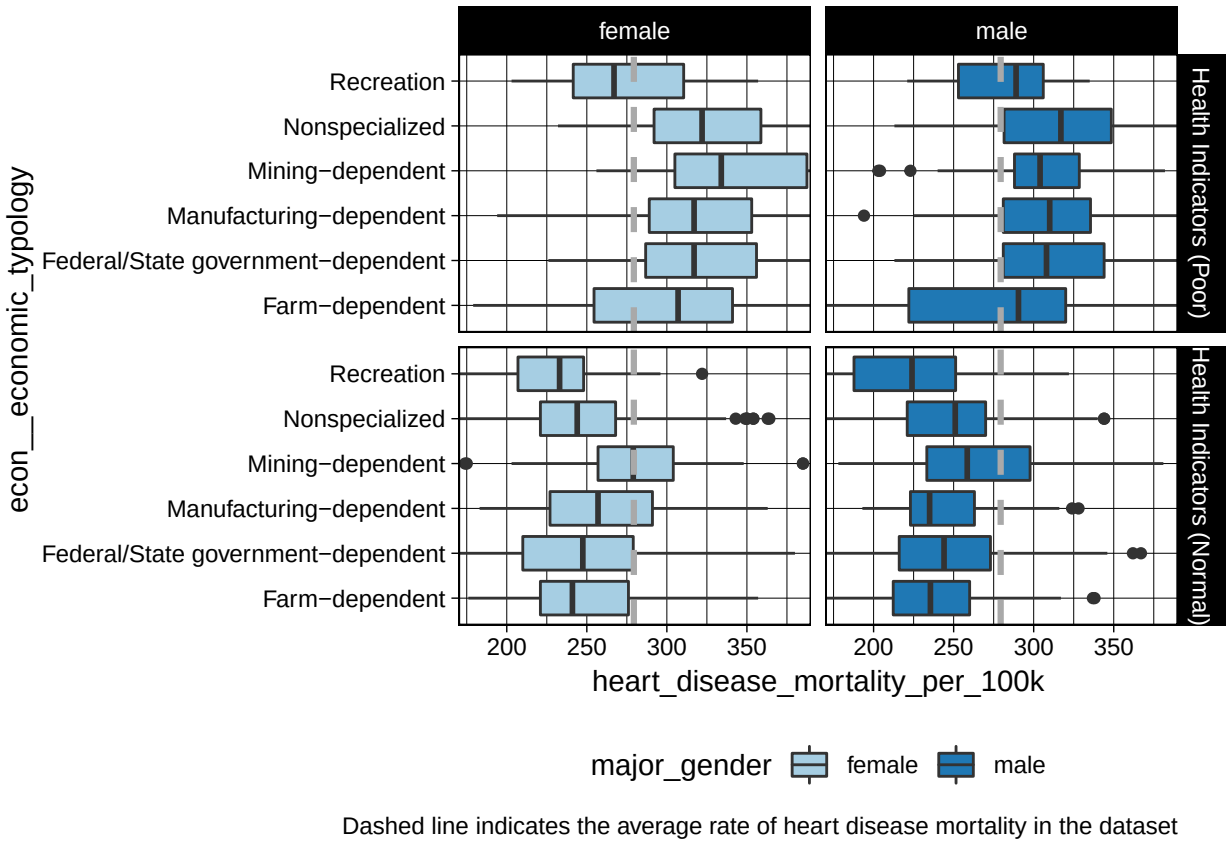

Figure 5: Counties that have poor key health indicators universally predicts an above-average heart disease mortality rate

Predicting Heart Disease Mortality Rate

In the final part of this report, I will detail the steps and methodology employed in the MPP Capstone Challenge. A high-level sequence of action is as follow:

1. Reading the data: Reading the data matrix and concatenate the predictor label (`heart_disease_mortality_per_100k`) by column

2. Exploratory Data Analysis: Finding correlations in the data and getting a sense of which variables among them constitute good predictors

3. Feature Engineering: Creating `largeold`, `metronot`, `nonwhiteasian` and `healthissue` variables out of existing ones

4. Mutliple imputation[4] on both the train and test datasets

[4]Refer to the **Imputation by the EMB algorithm** section in this report

5. Random Forest regression on each imputated dataset

6. Using model blending to find the most likely estimate for each row in the test data

Step 1 to 4 has been sufficiently documented, so the following sub-sections would aim to highlight the choice of algorithm and its implementation.

Prediction Model

Random forest is an ensemble-based state-of-the-art algorithm built on the decision tree method we learned about above and is also known for its versatility and performance. Among the family of ensemble-based classifier include a technique called boosting and it works by combining the performance of weak learners to gain an overall boosted performance.

The idea of ensembling is largely in principle and doesn't necessarily reference any particular algorithm. They describe any meta-algorithms that combine several machine learning techniques into one predictive model in order to decrease variance, reduce bias, or improve predictions.

When we apply the ensemble-based approach on a decision tree model, the trees we built are usually trained using resampled data. In the prediction phase, these trees then vote for a final prediction. Another way to apply ensemble methods on our tree model is known as bagging (bootstrap aggregation). Bagging proposes the idea of creating many subsets of training sample through random sampling (with replacement). Then each of these sets of training sample are used to train one unit of decision tree. This leads us to an "ensemble" of trees, and we'll use the average of all the predictions from these different trees in the prediction phase.

Random Forest extends the idea of bagging by taking one more measure: in addition to creating subsets from the training set, each of the tree is also trained using a random selection of features (rather than using all features). Because each tree is built with a random set of predictors and training samples, the collective of it is called a Random Forest, which is a lot more robust as a model compared to a single tree.

Among many of its advantages, random forest can be used to solve for both regression and classification tasks, handles extremely large datasets well (since the ensemble approach means it only use a small sampled subset from the full dataset), would solve for the dimensionality problems through implicit feature selection while treating noisy data (missing values and outlier values) out of the box. I've chosen random forest as the algorithm the heart disease mortality rate because of its suitability for this (regresion) task.

"section concatenated for brevity and relevance"

Implementation Details

"section concatenated for brevity and relevance"

Final Notes on Model's Performance

"section concatenated for brevity and relevance"