PREDICTING HEART DISEASE MORTALITY

CARLOS ALBERTO NAVA FONSECA[1]

CONTENTS

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

The present report documents the techniques and analysis of the rate of heart disease (per 100,000 individuals) across the United States at the county-level from socioeconomic indicators. The data is taken from the United States Department of Agriculture Economic Research Service (USDA ERS)

and the University of Wisconsin Population Health Institute. County Health Rankings & Roadmaps.

There are 33 variables in this dataset. Each row in the dataset represents a United States county that is unidentifiable in the dataset. While all the variables were studied, the discoveries lead to detect outliers, lack of data and wrong values. During the data processing, education and health variables were the most significantly correlated to the heart disease mortality. And, after using categories for some variables, both area information about the county and the economic typology play a major role in the prediction of heart disease.

When developing the Machine Learning model, only one model couldn't get an effective prediction. Therefore Stacking Regressions were used to improve accuracy.

INTRODUCTION

This chapter describes the dataset used to develop a Machine Learning model to predict Heart Disease. These datasets (training and test sets) were obtained from the United States Department of Agriculture Economic Research Service (USDA ERS).

Data Exploration: Target Variable

From the dataset provided, there are 33 variables, which are divided into four categories:

- Area: variables that contain information about the county
- Economic Indicators: they are categories of economic dependence, labour force, unemployment and insurance
- Health: indicators about obesity, smoking, diabetes and other characteristics of the county's population
- Demographics: information from the county's characteristics such as age percentage distribution, education, and others.

Also, additional information was given from the article from Jones (2009) [1] on relevant factors of heart disease linked with the rural population with certain economic characteristics.

The variable selected for prediction is: heart_disease_mortality_per_100k. It is defined as the rate of heart disease (per 100,000 individuals) across the data set. A brief summary of the variable is presented in Table 1.

From Table 1 it can be seen that the target variable main statistics. Plotting this variable it is observable in Figure 1 that is near a normal distributed variable and negatively skewed (slightly pointing to the left).

Table 1: Statistical Summary of Heart Disease Mortality per 100k

	count	mean	std	min	max
heart_disease_mortality_per_100k	3198.0	279.37	58.95	109.0	512.0

[1] *navafoseca.carlos@gmail.com*

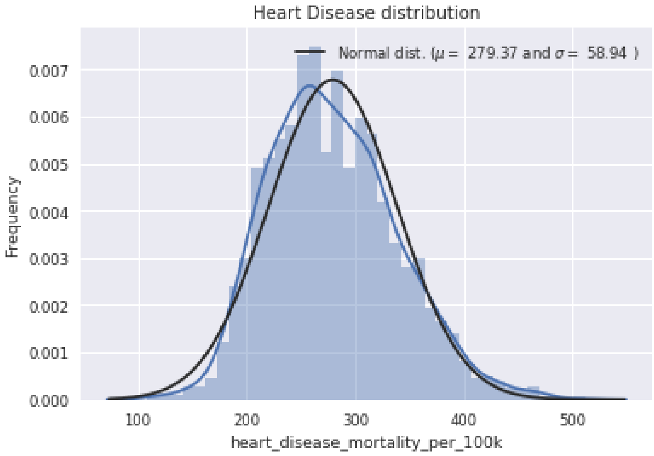

Figure 1: Target Variable Distribution vs Theoretical Normal Distribution (*Own elaboration made with: Python Seaborn*),

The Q–Q (quantile-quantile) plot helps to compare the two probability distributions by plotting their quantiles against each other. As seen in Figure 2 it is presented the graphically the properties of the normal distribution and the variable distribution. This Q–Q plot compares a sample of data on the vertical axis to a statistical population on the horizontal axis. The points follow a linear pattern, suggesting that the data are not distributed as a standard normal X $N(0, 1)$. Fortunately, there is little offset between the line and the points suggest that the data follows a X $N(\mu, \sigma)$ distribution.

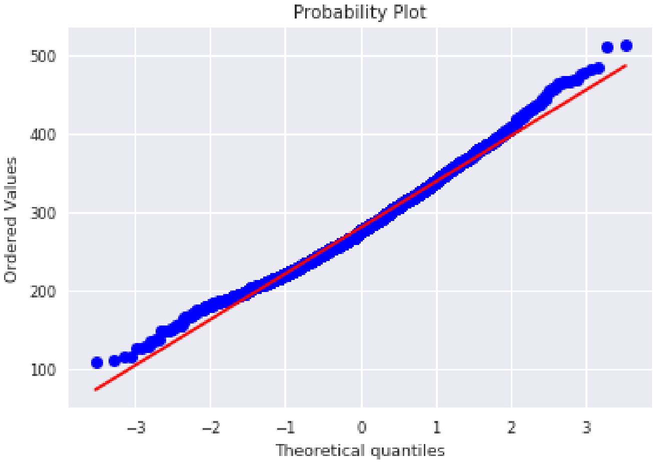

Figure 2: Probability Distribution vs Theoretical Probability Distribution (*Own elaboration made with: Python Seaborn*),

Since the target variable is left-skewed, it was transformed this variable and make it more normally distributed using the $\log(1 + x)$ function to be accurate in the floating-point accuracy. This skewed data is being normalised by adding one (in order to transform the 0 since $\log 0$ is not defined) and taking the natural logarithm. The data can be nearly normalised using this transformation technique. Actually, many of the algorithms to be tested assume that the data is normal and calculate with this assumption. So, if data is closer to normal, the better it fits linear models.

After the logarithmic transformation of the target variable, it was obtained Figure 3 with almost no skew suggesting that the data appears more normally distributed.

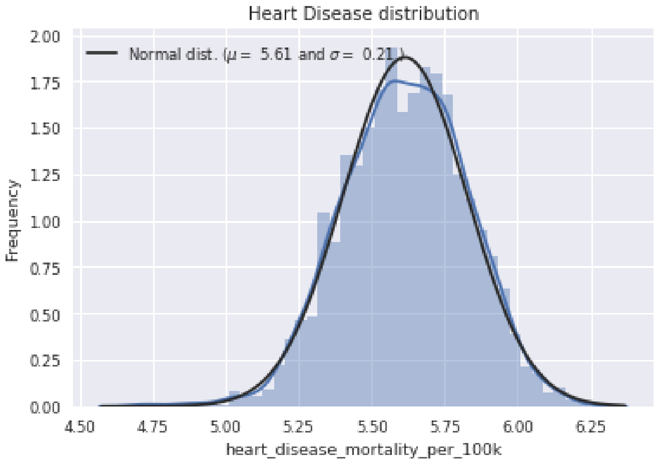

Figure 3: Target Variable Distribution vs Theoretical Normal Distribution (*Own elaboration made with: Python Seaborn*),

In Figure 4, it details the gap between the theoretical an data distribution are closer confirming that the logarithmic transformation normalized the data making it suitable to use linear modelling.

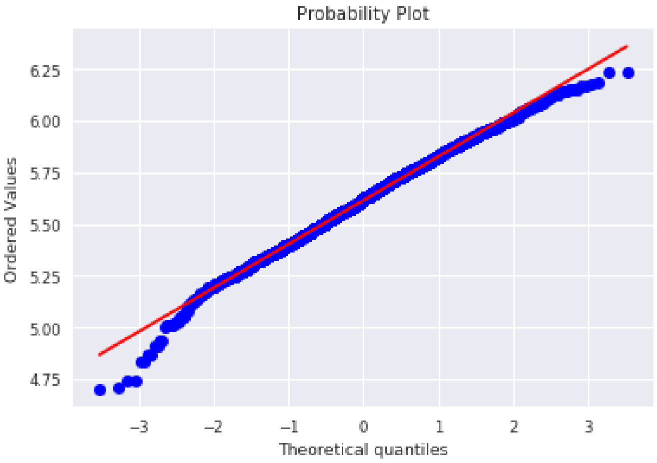

Figure 4: Probability Distribution vs Theoretical Probability Distribution (*Own elaboration made with: Python Seaborn*),

Data Exploration: Train Variables

From the data there are 3198 samples, which can be used to train model; also 33 features and 1 target variable. A brief summary of non-categorical data is presented in Table 2. It can be seen that most data is normalized since it is presented in terms of ratio per one hundred or percentage. Some data from the Health category (air pollution, homicides per 100,000 inhabitants, vehicle crash per 100,000 inhabitants, population per dentist and population

Table 2: Statistical Summary of Variables Description

	count	mean	std	min	max
econ__pct_civilian_labor	3198	0.47	0.07	0.21	1
econ__pct_unemployment	3198	0.06	0.02	0.01	0.25
econ__pct_uninsured_adults	3196	0.22	0.07	0.05	0.5
econ__pct_uninsured_children	3196	0.09	0.04	0.01	0.28
demo__pct_female	3196	0.5	0.02	0.28	0.57
demo__pct_below_18_years_of_age	3196	0.23	0.03	0.09	0.42
demo__pct_aged_65_years_and_older	3196	0.17	0.04	0.04	0.35
demo__pct_hispanic	3196	0.09	0.14	0	0.93
demo__pct_non_hispanic_african_american	3196	0.09	0.15	0	0.86
demo__pct_non_hispanic_white	3196	0.77	0.21	0.05	0.99
demo__pct_american_indian_or_alaskan_native	3196	0.02	0.08	0	0.86
demo__pct_asian	3196	0.01	0.03	0	0.34
demo__pct_adults_less_than_a_high_school_diploma	3198	0.15	0.07	0.02	0.47
demo__pct_adults_with_high_school_diploma	3198	0.35	0.07	0.07	0.56
demo__pct_adults_with_some_college	3198	0.3	0.05	0.11	0.47
demo__pct_adults_bachelors_or_higher	3198	0.2	0.09	0.01	0.8
demo__birth_rate_per_1k	3198	11.68	2.74	4	29
demo__death_rate_per_1k	3198	10.3	2.79	0	27
health__pct_adult_obesity	3196	0.31	0.04	0.13	0.47
health__pct_adult_smoking	2734	0.21	0.06	0.05	0.51
health__pct_diabetes	3196	0.11	0.02	0.03	0.2
health__pct_low_birthweight	3016	0.08	0.02	0.03	0.24
health__pct_excessive_drinking	2220	0.16	0.05	0.04	0.37
health__pct_physical_inacticity	3196	0.28	0.05	0.09	0.44
health__air_pollution_particulate_matter	3170	11.63	1.56	7	15
health__homicides_per_100k	1231	5.95	5.03	-0.4	50.49
health__motor_vehicle_crash_deaths_per_100k	2781	21.13	10.49	3.14	110.45
health__pop_per_dentist	2954	3431.43	2569.45	339	28130
health__pop_per_primary_care_physician	2968	2551.34	2100.46	189	23399

per physician) are not normalized. This represents an advantage now that the target data is normalized.

Further analysis of the data presented in Table 2 suggest that the variables might be correlated with each other. To graphically test this hypothesis, it is visualized as the correlations between variables in Figure 5.

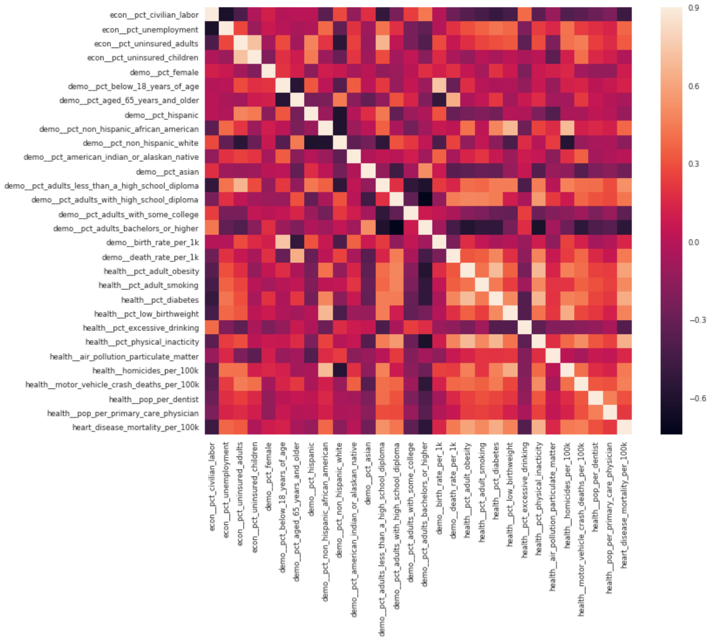

Figure 5: Correlation Heat Map (*Own elaboration made with: Python Seaborn*),

From this Figure, the assumption is that the lower right side of the graph holds more correlation to the target variable. There are also negative correlations, mainly from labour and education. Further detail on variables correlated with more than |0.5| with the target variable was plotted in the following Figure 6.

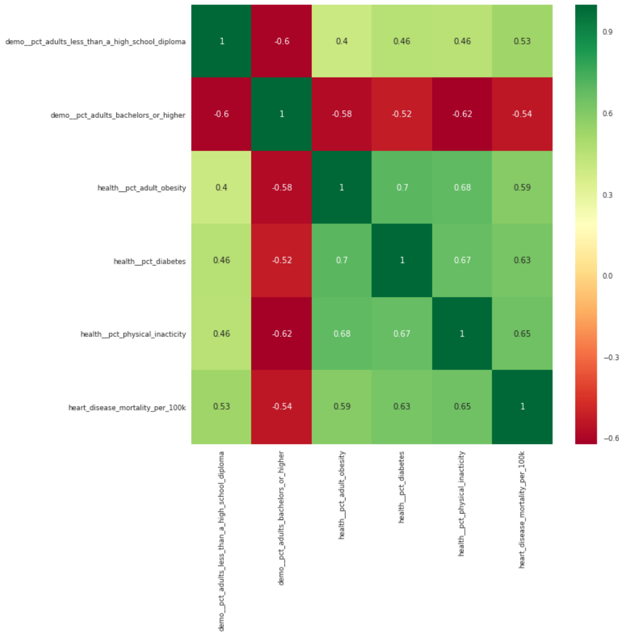

Figure 6: Detailed Correlation Heat Map (*Own elaboration made with: Python Seaborn*),

These findings suggest that the lifestyle factors that come with income can reduce or increase the risk for heart disease. With this in mind its logic to think that factors like education, income and wealth play an important role in overall health. Social position can influence a person's behaviour, impacting decisions related to diet, exercise and smoking.

Exploring the data in more detail, the analysis proceeds to the data set distribution. Presented in Figure 7 that data has many outliers and is not normally distributed.

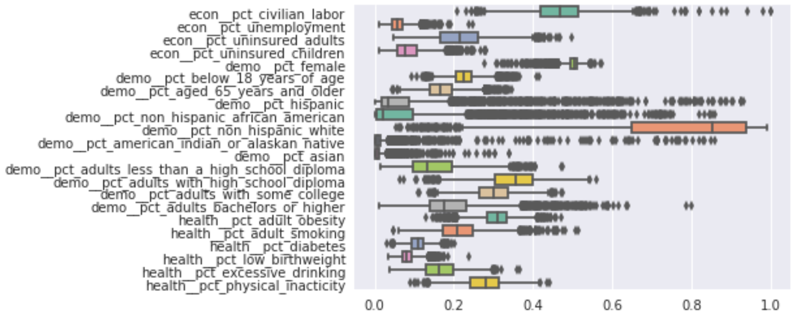

Figure 7: Percentage Variables (*Own elaboration made with: Python Seaborn*),

A closer look to data which is < |0.5| correlated to the target variable (see Figure 8) follows the same pattern. Therefore the model must be aware of the skewness of each variable and some of these variables might have outliers. From this analysis, the conclusion is that the mean of each variable presented in Table 2 as part of the statistical description, is sensitive to these outliers if they are removed or transformed.

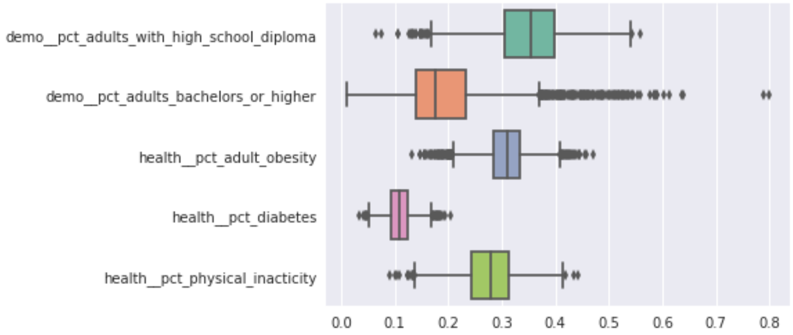

Figure 8: Percentage Variables with < |0.5| correlation (*Own elaboration made with: Python Seaborn*),

Missing Data and Skewness

Missing values in the training data set can affect the prediction or classification of a model negatively. Also, some machine learning algorithms can't accept missing data. Take for example Support Vector Machines or Linear Regression.

Having understood a general perspective on how data variables have some correlation role with the target variable, and a brief description of them, the following procedures are made to the the data variables. After the review the data composition in Table 3 that there is lack of data ranging from 60.48% to 0.16% which must be treated in order to try different models without errors.

From the previous section, it is known that data mean is sensitive. Therefore, it must take into account that the gaps in the data set must consider

Table 3: Missing Data Percentage

	Missing Ratio
health__homicides_per_100k	60.481
health__pct_excessive_drinking	29.277
health__pct_adult_smoking	13.794
health__motor_vehicle_crash_deaths_per_100k	12.138
health__pop_per_dentist	7.04
health__pop_per_primary_care_physician	6.594
health__pct_low_birthweight	5.241
health__air_pollution_particulate_matter	1.051
health__pct_adult_obesity	0.159
health__pct_diabetes	0.159
health__pct_physical_inacticity	0.159
...	...
demo__pct_aged_65_years_and_older	0.159

both the training and testing sets and go through the same procedure. Having known that the mean is sensitive, the median was considered to fill the missing values on variables who were below 1% of missing data.

The other variables received different treatment. Since data were missing between 5 to 60.5 per cent, the approach was to use normalized data within the range of the mean and one standard deviation. In other words, a sample random variable X with $N(\mu, \sigma)$ was used to fill the missing values.

Additionally, categorical variables must be mapped to account them in the analysis. The following data variables were mapped according to their characteristics:

- *area__rucc* will be mapped between Metropolitan Counties (Value: 1) and Non-Metropolitan Counties (Value: 0).
- *area__urban_influence* will be mapped between Large influence (Value: 1) and Small influence (Value: 0).
- *econ__economic_typology* will be mapped between economic activities that are classified as stressful (manufacturing, government dependent and mining) with value of 1, others will be non-stressful (Value:0).
- *yr* will be mapped according to *a* as 1 and *b* as 0.

Also, based on the correlation analysis and the mapped variables some elements were dropped since they could not be transformed in normalized data or had not a direct approach to heart disease. Homicides, crash deaths, and population per dentist and physicians were left out.

To make all data normally distributed, the skew for each data variable was calculated. Values for the first 10 variables are presented in Table 4.

The assumption of normality at the beginning of this analysis leads to modelling that is simple, mathematically tractable, and powerful compared to tests that do not make the normality assumption. Unfortunately, the skewed data set is in fact not approximately normal. However, an appropriate transformation of a data set can often yield a data set that does follow approximately a normal distribution. This increases the applicability and usefulness of models based on the normality assumption.

The Box-Cox transformation is a particularly useful family of transformations. It is defined as:

$$T(Y) = (Y^{\lambda} - 1)/\lambda \tag{1}$$

Table 4: Skew Value in Data

	Skew
demo__pct_american_indian_or_alaskan_native	7.78
demo__pct_asian	7.338
demo__pct_hispanic	3.182
demo__pct_non_hispanic_african_american	2.282
area__urban_influence	2.104
demo__pct_adults_bachelors_or_higher	1.515
health__pct_low_birthweight	1.23
econ__pct_unemployment	1.195
econ__pct_uninsured_children	1.19
demo__birth_rate_per_1k	0.951
...	...

where Y is the response variable and λ is the transformation parameter. For $\lambda = 0$, the natural log of the data is taken instead of using the Formula 1. Variables with an absolute value of skewness greater than 0.75 were transformed with a $\lambda = 0.15$.

The process that leads to normalizing the data in the training set was also used in the test set so it can predict with the assumptions made for the variables.

MACHINE LEARNING MODEL

Breiman (1996) [2] presents stacking regressions is a method for forming linear combinations of different predictors to give improved prediction accuracy. The idea is to use cross-validation data and least squares under non-negativity constraints to determine the coefficients in the combination. The idea was first presented by Wolpert (1992) [3]. Under this idea, a model was developed.

Base Models

The following models were tested:

1. LASSO Regression.
2. Elastic Net Regression.
3. Gradient Boosting Regression.
4. Extreme Gradient Boosting (or XGBoost).

The models were evaluated by their performance using cross-validation of the Root Mean Squared Logarithmic Error (RMSLE) error. The following results were obtained:

Now that the procedure has the Mean and Standard Deviation, the first approach is averaging base models. Obtaining the following result:
Averaged base models score: Mean = 0.1113 Std. = (0.0034).

When stacking the averaged models as a meta-model, it got:
Stacking Averaged models score: Mean = 0.1000 Std. = (0.0036).

Finally, after and creating a stacked classification model we used the exponential function on the predicted model. Obtaining the results in the original scale for submission.

Table 5: Base Models Results

	Mean	Std.
Lasso score:	0.1194	(0.0030)
ElasticNet score:	0.1193	(0.0030)
Kernel Ridge score:	0.1262	(0.0040)
Gradient Boosting score:	0.0991	(0.0032)
Xgboost score:	0.1067	(0.0032)

RESULTS AND CONCLUSIONS

The analysis presented in this report acknowledge the following findings:

- Filling the missing values with mean/median/mode for a high rate of present data (over 99%) do not generate outliers and is appropriate for data close to a normal distribution.
- In some variables, random sampling was used to predict missing values using a random variable with X $N(\mu, \sigma)$ distribution.
- Categorical variables were mapped in binary values to include them in the model.
- After filling the missing data and mapping the categorical variables, the skewness of the data was modelled with the Box-Cox technique to adjust them into normalized data.
- After the classification with base models, a stacking regression model was used to predict the values of the target variable.

Based on the Root Mean Squared Logarithmic Error (RMSLE), the model constructed for predicting the heart disease using a stacking regression showed to perform better than the base models alone.

Further work can be implemented with the categorical data and missing data with other techniques such as Decision Trees based on correlated variables.

REFERENCES

[1] Charlotte A. Jones, Arjuna Perera, Michelle Chow, Ivan Ho, John Nguyen, and Shahnaz Davachi. Cardiovascular disease risk among the poor and homeless - what we know so far. *Curr Cardiol Rev*, 5(1):69–77, Jan 2009. CCR-5-69[PII].

[2] Leo Breiman. Stacked regressions. *Machine Learning*, 24(1):49–64, Jul 1996.

[3] David H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241 – 259, 1992.