Heart Disease Rate Prediction

Xiangbei Chen, Ming Lyu, Peicheng Tang, Yuankai Wang, Kaiyu Xie
May 17, 2019

Abstract

There are a lot of machine learning methods available to train and predict the result based on the given dataset. In our project. we tried to predict the heart disease rate in different counties in the US. We built seven models for the dataset provided by user nandvard on Kaggle[1]. We would like to compare all those models to see which model could be the best for our data set. Our final model is a regression model which achieved roughly 0.78 testing R square from an SVM regressor.

1. Introduction

In this project, we aimed to build models to predict the heart disease rate according to various features. Our dataset is from Kaggle, which collects data from counties in the United States. In all, there are 3199 data points and 33 variables in our dataset and they are from three different aspects. The first aspect which we included in our model is the area, which is categorized as rural area and urban area. The economic conditions of the counties are also considered. The economic conditions are put into six categories, including farming, mining, manufacturing, Federal/State government, recreation, and non-specialized counties. Our dataset also includes various health factors. To make the best prediction, we tried several methods, includes KNN, LinearRegression, RandomForest, Decision Tree, etc. After that, we compared the test score to determine the best model.

2. Preprocessing

In the dataset, we found out that there were three different CSV files, which were training values, training labels, and testing values. Since this dataset is used in a competition, we could not access the testing labels. Therefore, we decided to use only the training set. In this case, we need to split the dataset into training and testing with a ratio of 0.8 to 0.2. To import the dataset easier, we combined the training values and label files together to form a complete data file with both the features and the target.

In the dataset, there are a lot of missing data. We could not train the model with the NaNs left in the dataset. Therefore, we tried three methods to deal with it. The first one was to remove all the

lines with NaNs. However, if we used this approach, the data records left would be 1069, which only about a third of the original dataset. Therefore, we decided that we would not use this method. The second method was to fill all the missing values with 0, and the third method was to fill them with the mean value of that column. In this way, we could preserve more training data, but the data might be less reliable. This was a place where we did a trade-off between the number of data points and the accuracy on the data. We tried both fill methods on the models we built. It turns out that the two methods have almost no difference to the result. Therefore, we finally decided that we will use fill with mean values method to deal with the missing data.

3. Data Analysis

The heatmap shown in Figure 1 in the appendix shows the correlations of the features in the dataset. Bright color means positive linear relationship, and the deep color means a negative linear relationship. There is a clear negative linear between the percentage of adults with some college and features of health because there are a vertical and a horizontal dark line across at south and east of the center point. In the other words, the higher percentage of adults attended some colleges and percentage of bachelor or higher accompany with a lower percentage of the health problems, such as physical inactivity, vehicle crash death, obesity, smoking, and diabetes.

According to the heatmap, some variables have a strong linear relationship to the target. The percentage of civilian labor, percentage of bachelor or higher, and percentage of excessive drinking have strong negative correlations. A supposition is that people with high education and labor force focus more on their heart health. Therefore, they will go to the hospitals when the diseases are at an early stage and thus have a higher chance to get cured. On the opposite side, we also found out that the percentage of adults with less than a high school diploma have a positive correlation. They care less on their health and has a higher chance to get heart disease.

Obesity and diabetes have strong positive correlations with the target. An explanation is that obesity and does not directly kill people but they accompany with the heart disease, which causes death.

Surprisingly, the data shows that a higher percentage of excessive drinking leads to a lower percentage of heart disease mortality. It is counter-intuitive. The only way we can explain is that excessive drinking hurts other parts of the body more severely. People are more likely to pass away due to other reasons rather than heart disease.

The last most dominant feature that causes heart diseases is “physical inactivity”. We believe that this feature will highly likely to result in less exercise for the person. Less exercise will let the organs degenerate. So it will cause heart diseases.

4. Models

4.a Linear Regression

We used linear regression as our first approach to get a general idea about the dataset. The training result is reasonable. The R2 value is about 0.69. However, testing R2 is a negative number. Since the result is too bad, it is meaningless to further improve the linear regression model using feature engineering or feature selection.

4.b KNN Regressor

We also used KNN regressor to train and predict the results. Using grid search cross-validation, we got the test R2 = 0.716 while the optimal hyper-parameter, n_neighbors is 1. Considering the size of our features and the number of records we had for our training set, it was possible that the result might suffer from the curse of dimensionality. In addition, there are a lot of missing values in the original data set, which we filled with the average value of the corresponding feature. It is probable that not all the features play a significant role in predicting the heart disease rate. Thus, utilizing the twenty best features obtained from ridge/lasso regularization, we conducted the training process again. With only twenty features, we got the optimal hyper-parameter to be 21, while the final test R2 is 0.614, which is actually worse than using all the features.

Finally, we tried to use the principal component analysis to analyze the data. However, unfortunately, the outcome is rather terrible. Using only top twenty principal components as the new features, we only got 0.056 for our test R2.

4.c Decision Tree

We adopted a decision tree regressor to fit the target. Using the grid search and feature engineering, we found that the best depth of the tree is 5 and the highest validation accuracy is 0.568. Although the training accuracy purely increased, the validation accuracy decreased after the depth reached five as shown in figure 2 in the appendix. It shows that the decision tree is overfitting.

4.d Regularization Method

Using cross-validation and grid search, we found the best hyper-parameters alpha for Ridge and Lasso are 0.1 and 0.01. After we applied the model to the dataset with the best hyper-parameters, we tested it on our test dataset. The test R square is 0.668 for Lasso and 0.668 for Ridge.

As shown in figure 3 in the appendix, Inactivity increases heart disease ratio. It is reasonable to assume that people who do not play sports very often have a higher chance of getting heart diseases. However, it is rather difficult to explain why the increase in the percentage of 65 years old people decreases the heart disease ratio.

4.e Gradient boost

We applied the Gradient Boost method to our dataset. Using cross-validation and grid search, we found the best parameter learning rate = 0.11, max depth = 8 and n_estimators = 120. The search range is parameter learning rate = [10,200], max depth = [1,40] and n_estimators = [0.01,1]. The whole process took 50.3 minutes. The best validation R square is 0.661.

We built our Gradient Boost Model using the best parameters found. After fitting the training set, we tested our model on the test set. The training R square is 0.994 and the test R square is 0.693.

4.f Random Forest

After using the grid search, we got the best parameter of the n_estimatoors is 350. The validation R2 for this fit is 0.715. Then the n_estimator of 350 was used to fit the Random Forest model. Using the fitted model to make predictions, we obtained the test R2 to be 0.726. The change of R2 for different numbers of trees is shown in Figure 4 in the appendix.

4.g SVM Regressor

The final method we used was an SVM Regressor. We searched through the hyperparameters from 100 to 1000 with step 10. We also tried different kernels like rbf, linear, poly, and sigmoid. We found out that when the C is 300 and the kernel is rbf, it gave the best testing R square, which is 0.782. The training R square we got is 0.969.

5. Results

Table 1: Model Performance

	α	Training R Square	Testing R Square
Linear Regression	N/A	0.690	-3.264
KNN Regression	N_neighbors =1	0.595	0.716
Decision Tree	Depth = 5	0.693	0.568
Lasso	α = 0.01	0.70	0.668
Ridge	α = 0.1	0.70	0.668
Gradient Boost Tree	N_estimator = 120 Max Depth = 8 Learning_rate = 0.11	0.994	0.693
Random Forest	N_estimator = 470	0.721	0.729
SVM Regressor	C = 300 kernel = rbf	0.969	0.782

Support vector machine regressor had the best performance with R square 0.782. KNN regression, Lasso regression, Ridge regression, Gradient boost tree, and Random Forest had almost the same performance with R square 0.7. Multi-Linear Regression has the worst performance. The negative R square indicates that Multi-Linear Regression model is not appropriate this dataset.

6. Conclusion

The highest R2 we got is 0.782 which is a good test score for our question of interest. By analyzing the data through various models, we can get better ideas of what factors will affect the heart health condition. Our dataset is focusing on people in the United States, but we believe that our analysis is useful for people outside the United States. To make our model more reliable for people outside the United States, we need to collect more data from various regions.

References:

[1]. nandvard. Microsoft Data Science Capstone Predict Heart Disease Rate (Regression), Kaggle, 2019, https://www.kaggle.com/nandvard/microsoft-data-science-capstone#Training_values.csv.

Appendix:

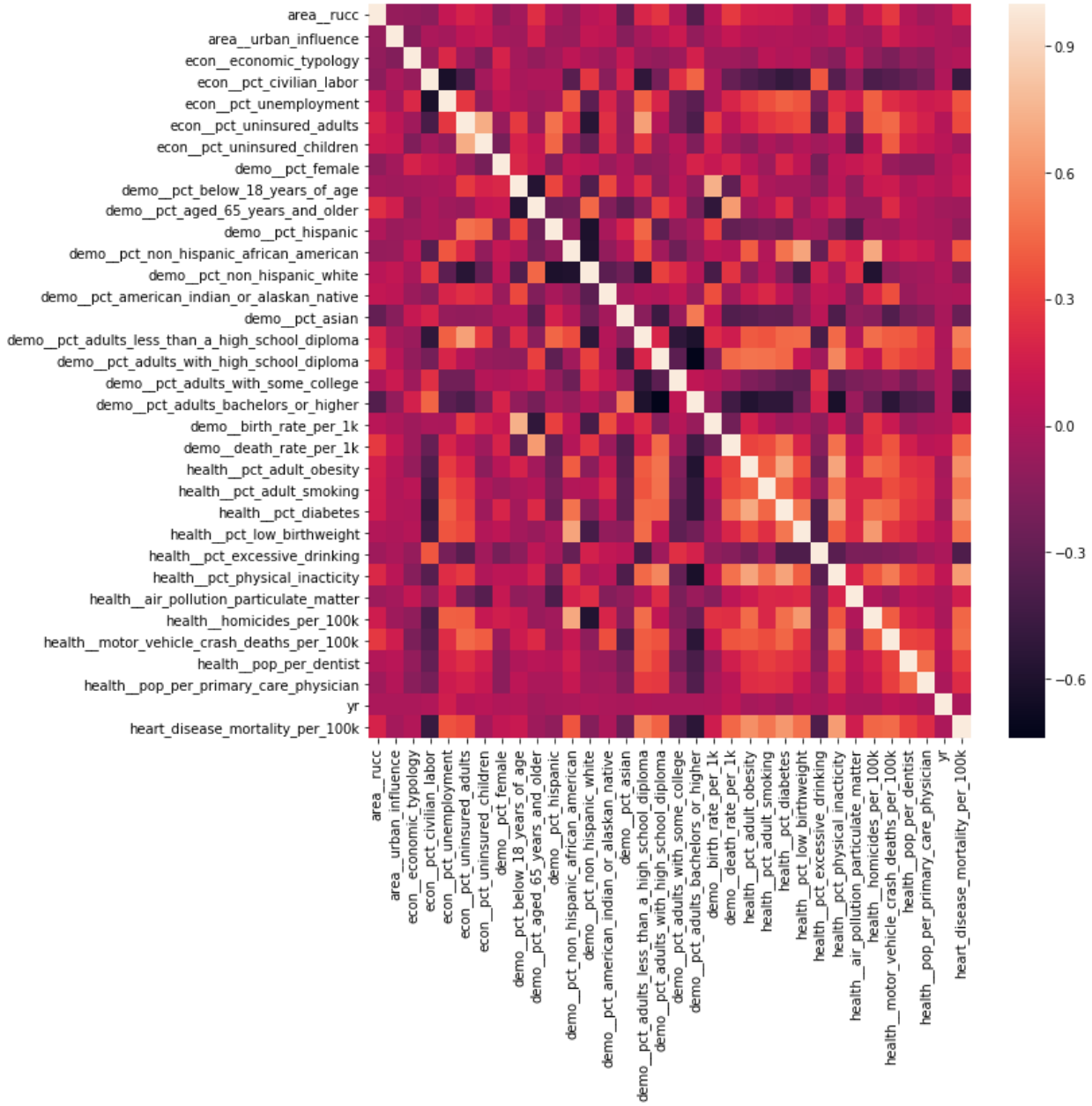

Figure 1: The heatmap for correlations of the features

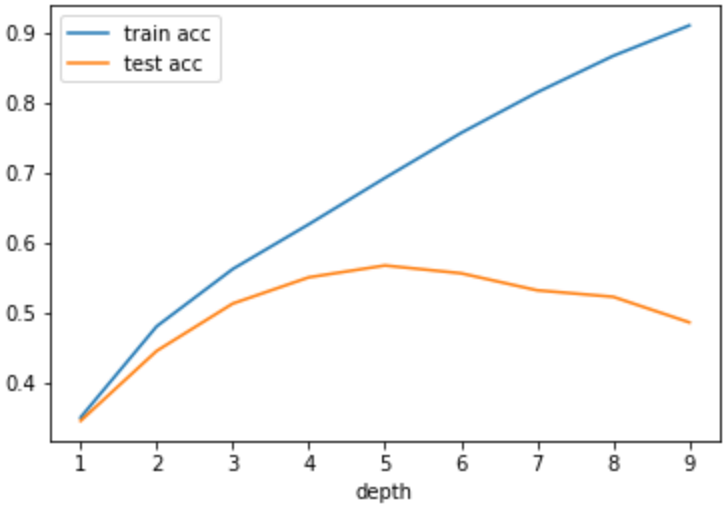

Figure 2: Decision Tree Training and Validation Score

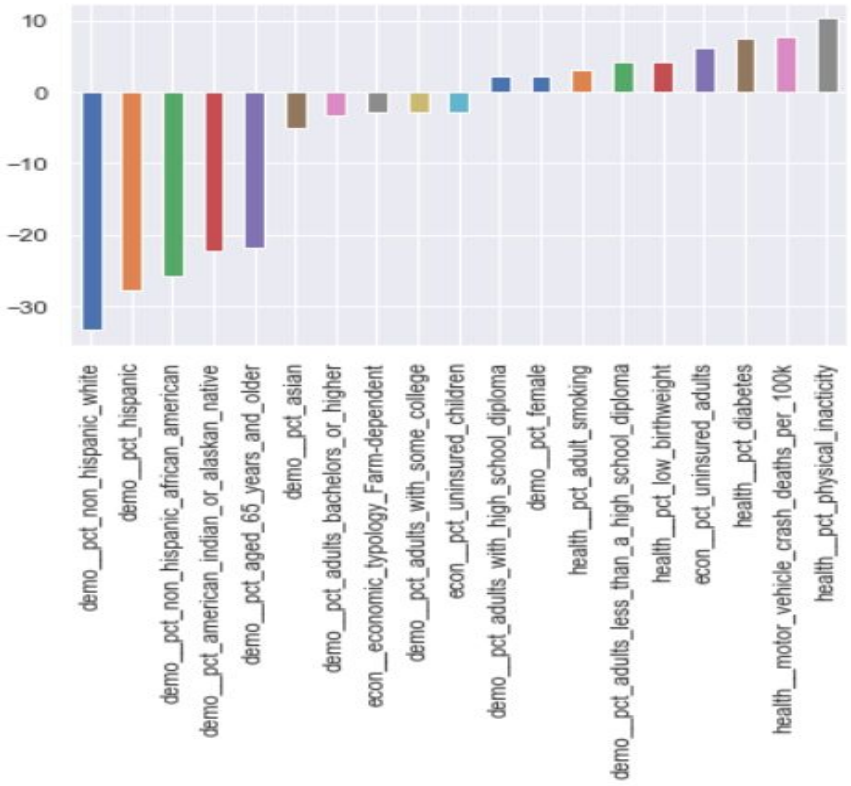

Figure 3: Features Coefficient for Ridge Regularization

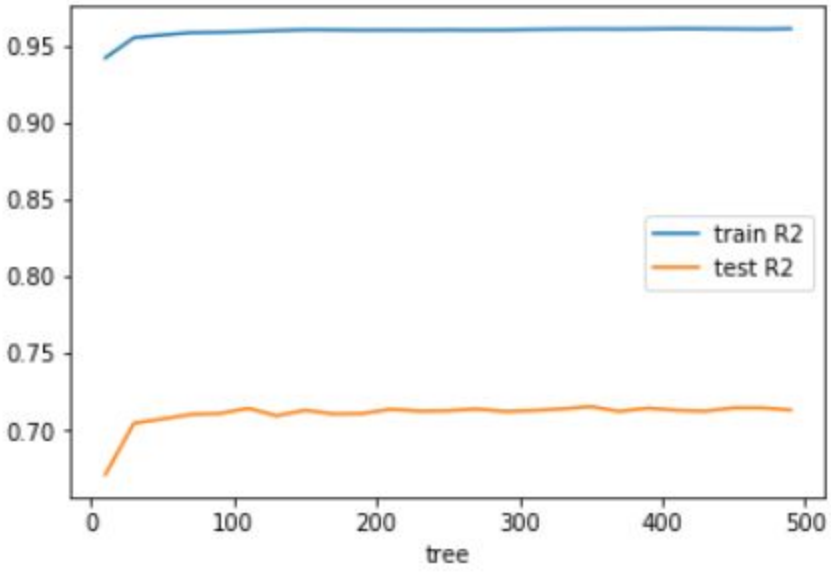

Figure 4: R2 values graph for Random Forest method using different numbers of trees