DS4A Team 1: Identifying Key Factors to Predict Prevalence

of Heart Disease in the United States

1. **Introduction and Background**

Up until 2020, prior to COVID 19, heart disease was the leading cause of death in the United States, followed by cancer, accidents (unintentional injuries), and chronic lower respiratory diseases. According to the Centers for Disease Control and Prevention (CDC), about 659,000 deaths[[1]](#footnote-0) are caused by heart diseases each year, which equates to 1 in every 4 deaths. Heart disease is defined as a variety of conditions that affect the heart’s structure and function, such as coronary artery disease, arrhythmias (heart rhythm problems, congenital heart defects, heart valve disease, disease of heart muscle, and heart infection.

1. **Business Case / Scope**

What key factors can be used to help predict the prevalence of heart disease by county in the United States?

The leading risk factors to heart disease are:

* High blood pressure
* High cholesterol
* Cigarette smoking
* Diabetes
* Unhealthy diet and physical inactivity
* Obesity / Overweight

Since heart disease is prevalent, especially in some US Counties compared to others, this could put an economic strain on the government. It would be beneficial for the Federal and State government to identify where exactly they can enact adequate healthcare and safety measures for all populations & counties (e.g. adoption of the Affordable Care Act in all 50 US states). By considering other factors such as healthcare delivery and social and economic factors and their interaction effects on the risk factors, the key factors responsible for heart diseases can be mitigated when the government and the populace understand the effect of each factor.

1. **Data Description / Cleaning**

We pulled data from the Centers for Disease Control and Prevention (CDC). The data was exported from the following categories: Health Care Delivery & Insurance, Social and Economic Data, Risk Factors, Heart Disease & Stroke Data, and Healthcare Costs. Each category included sub variables, representing one factor or another, with information aggregated by county. In total, we downloaded and merged 38 csv files. The data in each file had the following properties:

Number of Rows: 3226

Number of Columns: 4 (cnty\_fips, display\_name, value, theme\_range)

File Size:156kb

Data Set Link: [https://nccd.cdc.gov/DHDSPAtlas/?state=Count](https://nccd.cdc.gov/DHDSPAtlas/?state=County)y

It’s important to note that each row represented a different U.S. county (as indicated by the cnty\_fips code).

**DATA CLEANING**

In order to effectively clean our data using Python, we imported the following programs: OS, pandas, numpy, matplotlib.pyplot, and seaborn.

Next, we updated the ***display\_name*** column (ex. “Arlington, (VA)”) into two separate columns, ***county*** and ***state***. This was performed by separating the column information by “,”, removing the quotation marks, and then dropping the ***display\_name*** column.

Each dataset has a ***value*** column. We extracted the ***value*** feature from all .csv files, while merging the files into the dataframe (df). The ***value*** columns were renamed as the specific file name to represent the value description. The values are represented as columns to create a dataframe of 3226 rows and 40 columns. The datasets were merged based on the ***cnty\_fips***. County FIPS is a 5-digit Federal Information Processing Standard Publication (FIPS) code in the United States that uniquely identifies U.S counties. The ***theme\_range*** column was dropped as it is not relevant to the project scope.

A picture containing graphical user interface

Description automatically generated

Updated column headers for ease of recognition and analysis (e.g., “Pharmacies and Drug Stores” to “no\_of\_pharmacies”).

Next, we reviewed missing data in the dataframe. If we simply dropped all the Not A Numbers (NaNs), then we would have lost over 80% of the dataset. As such, we systematically manipulated the data step-by-step. First, we calculated the number of missing rows by category to understand which variables are missing the most data compared to which variables are missing the least. Five variables (cardiac rehab participation, cardiac rehab sessions, cardiac rehab completion, number of cardiac physicians, number of neurologists, and number of neurosurgeons) are all missing data for over 27% of U.S. counties. Four variables are only missing data for two counties, so we investigated further. It was determined that the data for the two counties were associated with American Samoa and Mariana Islands. We decided to drop the two counties in the database. Now, 6 variables of 40 are complete with data. We investigated further.

Second, we investigated additional counties outside of the continental U.S. (U.S. territories), to view their impact of missing data on the remaining 34 variables. It was concluded that we should remove the following territories from the dataframe: St. Croix, St. John, St. Thomas, Guam, and Puerto Rico. With the removal of these territories, 30 of 40 variables are complete with data.

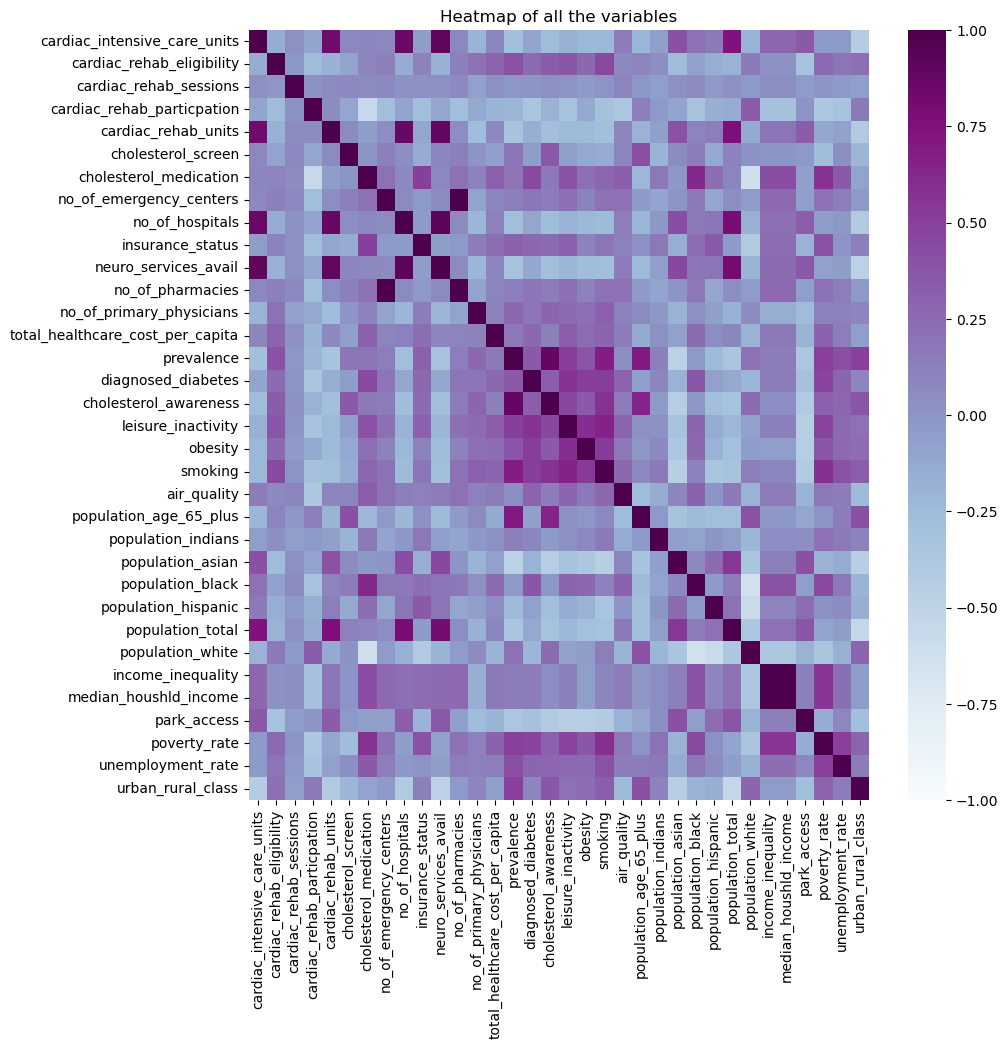
We continued to investigate further. Of the 10 variables that still have missing county data, the lowest category of missing rows is “Air Quality”, with 24 counties missing. The 24 counties missing “Air Quality” are all located in Alaska. Instead of dropping these counties, we populated the missing data with the average of the remaining Alaska counties. For “Cholesterol Medication”, which is missing data for 83 counties, we used the same approach and populated the missing data with the average data per state. Now, 32 variables are complete with data.

We decided to drop four variables that still had over 50% of counties missing data (cardiac rehab completion, number of cardiac physicians, number of neurologists, and number of neurosurgeons), decreasing our dataframe from 40 variables to 36 in the process.

Lastly, there are variables remaining with a significant number of counties missing data. We agreed that it would be best to keep the remaining variables “as is” to alleviate the possibility of bias into the dataset. Therefore, we dropped the remaining missing counties in the dataset. This whole process caused only 30% data loss instead of the initial 80%. Once we completed our data cleaning, we were left with a dataset consisting of 36 variables and 2310 counties to perform our Exploratory Data Analysis.

1. **Exploratory Data Analysis**

We performed Exploratory Data Analysis (EDA) to examine the dataset before building a model. This provided an opportunity to detect any errors and outliers, as well as understand patterns within the dataset using statistical graphs.



First, based on the figure above, we developed a heatmap to show the correlation of all variables. As you can see, we have positive and negative correlations. It would be interesting to examine each aspect of the heatmap above; however, our scope of this project is specifically on the prevalence for heart disease. We examined the positive and negative correlations of the variables in comparison to prevalence.

| ***Positive Correlation*** | ***Negative Correlation*** |
| --- | --- |
|  |  |

For ease of comparison, we group these variables. We decided to keep the same groupings as listed by the CDC: (1) Risk Factors; (2) Healthcare Opportunity; and (3) Socioeconomic Factors.

**INVESTIGATING RISK FACTORS**

Risk factors for heart disease are specific habits, behaviors, or conditions that can increase the probability of a person to develop heart disease, which include lack of exercise, unhealthy eating, smoking, diabetes, age and family history. In our dataset, the risk factors are categorized as cholesterol awareness, diagnosed diabetes, obesity, leisure inactivity (lack of exercise), and smoking. We created a heatmap to visually show the correlation of these variables to prevalence.

**A picture containing graphical user interface

Description automatically generated**

Each category in our Risk Factors grouping has a positive correlation, greater than .35, to prevalence.

In addition to correlation, we looked at the statistical distribution (minimum, maximum, count, mean, and intervals).

Graphical user interface, text, application

Description automatically generated

Note, the difference between minimum and maximum values for each category is not large, and the standard deviation is low. As such, we continued to look at the data through additional visualization techniques. We created histograms for this purpose.

|  |  |
| --- | --- |
|  |  |
|  |  |

From the graphs above, we can see that the variables are all normally distributed, although diagnosed diabetes is more noticeably right-skewed (this can also be seen from the summary statistics above).

Finally, we looked at the statistical significance of these correlations.

Graphical user interface, text, application

Description automatically generated

Since, all variables have a p-value less than 0.5, we could deduce that the correlations with prevalence have strong statistical significance. As a result, all variables are of interest for the model.

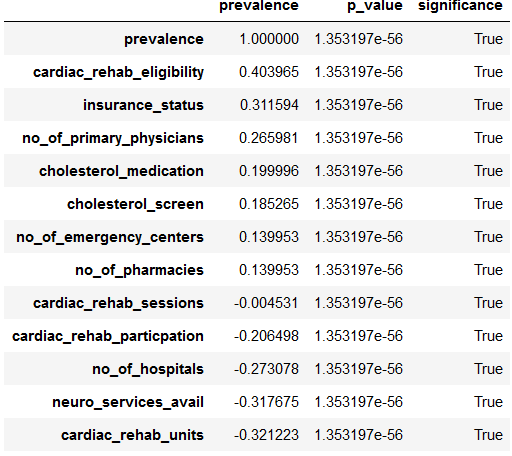
**INVESTIGATING HEALTH CARE OPPORTUNITY**

Healthcare access is important to note for heart disease. The inability to obtain healthcare services such as prevention services, diagnosis, treatment, and management of heart disease has a major impact on an individual’s health. We have the following variables listed under Healthcare Opportunity: Cholesterol Medication, Cholesterol Screening, Cardiac Rehab Eligibility, Cardiac Rehab Sessions, Cardiac Rehab Participation, Cardiac Rehab Units, Insurance Status, Number of Emergency Centers, Number of Hospitals, Neurological Services Available, Number of Pharmacies, and Number of Primary Physicians.

Calendar

Description automatically generated with low confidence

In relation to Prevalence, correlation is spread out from -.32 to +.32. In addition to correlation, we looked at the statistical distribution (minimum, maximum, count, mean, and intervals).



All variables have a p-value less than 0.5, meaning that their correlations with prevalence have strong statistical significance. As a result, all variables are of interest for the model.

We continued to look at the data through additional visualization techniques. We created histograms for this purpose. Again, we saw bell curves indicating normal distributions, although certain variables were skewed either left or right as shown below.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

How statistically significant are these correlations?

Graphical user interface, text, application

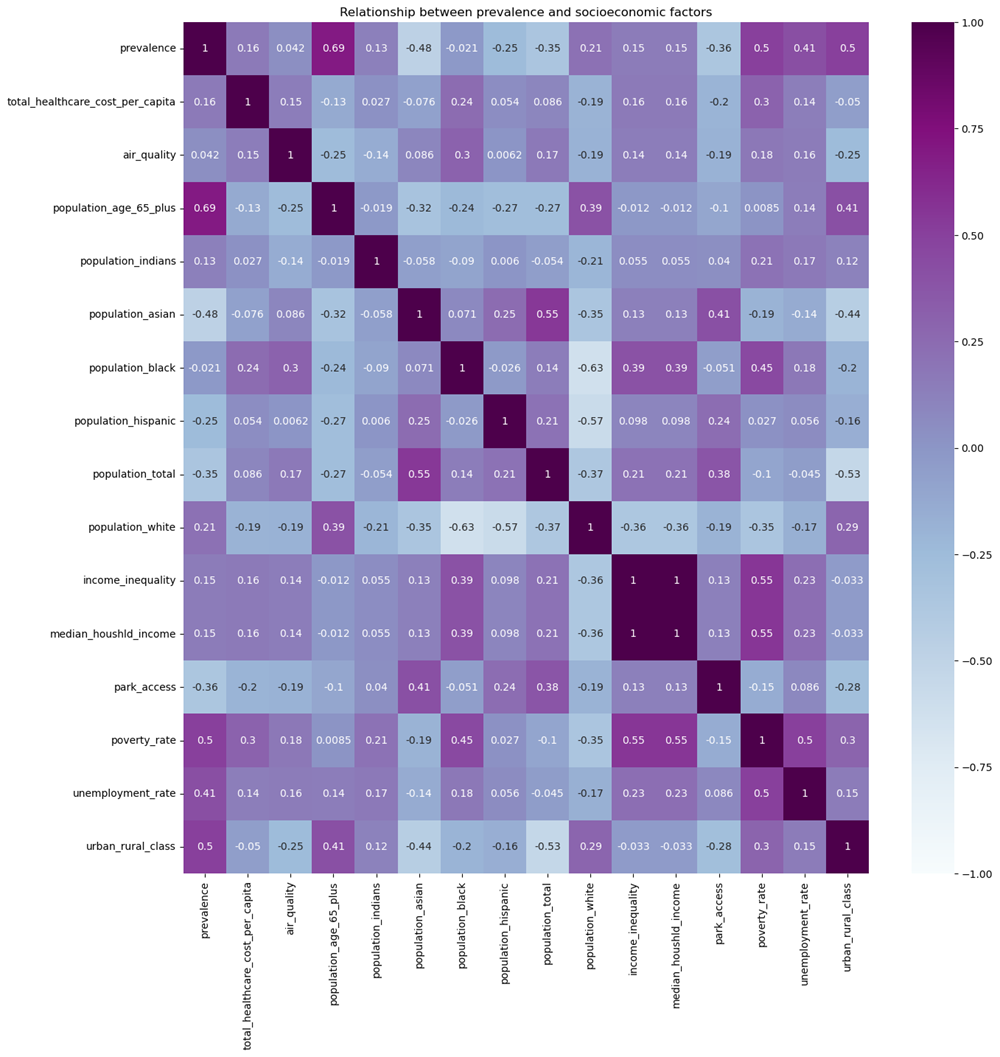
Description automatically generated

In reviewing all the visualizations, the following variables are of interest in moving forward to model for healthcare opportunities: insurance status, cardiac rehab eligibility, cholesterol medication, cholesterol screen, cardiac rehab sessions, cardiac rehab participation, and cardiac rehab units. The remaining variables of physicians, emergency centers, pharmacies, and hospitals were dropped because more qualifications of these variables would be needed to distinguish their true differences that the exported data is not providing. For example, we would need to know national rankings of hospitals, types of services provided, ratings for physicians, etc. Since we do not have that data, these variables were dropped.

The chosen variables all have a p-value less than 0.05, meaning that their correlations with prevalence have strong statistical significance. In our study, we do not interpret this to mean that these variables truly increase prevalence since they are supposed to be preventative factors which can reduce prevalence. Instead, we interpret this to mean that, the more people have access to the services represented by these variables, the more educated they are about heart diseases and the easier it is for heart diseases to be detected and treated.

**INVESTIGATING SOCIOECONOMIC FACTORS**

Socioeconomic factors influence cardiovascular disease prevalence. The lower the socioeconomic status, the greater the risk of cardiovascular disease. This is because the cost of healthcare is high while the access to the proper resources is low if a country's socioeconomic status is low.



The socioeconomic variables range in the correlation to prevalence (positive and negative). We reviewed the variables further through statistical distributions and histograms.

Graphical user interface, application, table, Excel

Description automatically generated

Below is another visual in a histogram.

|  |  |
| --- | --- |
|  |  |
|  |  |

Graphical user interface, application

Description automatically generated

Since all the variables have p-value less than 0.05, their correlations with prevalence have strong statistical significance

Further investigation of urban\_rural\_class shows that there is a tendency for heart diseases to be prevalent in non-metro areas (urban rural class category), while the reverse is the case for metro areas. Perhaps, this could be due to a decrease of activities in non-metro areas which can translate to leisure inactivity, and vice-versa for metro areas. However, investigation of the relationship between nonmetro, metro areas and leisure inactivity, indicates that this can be true. Similar observations can be made of park access except that the behavior is opposite to leisure inactivity, which one would expect.

There is a positive correlation between the population of age 65+ and prevalence, which we expected. However, surprisingly, the relationship between population total and prevalence is negative, suggesting that the more the population, the lower the prevalence of heart disease. Both variables capture contrasting effects of population (demographics). As such, we will let our model reflect both variables.

Income inequality and Median Household Income are 100% correlated with each other. Since there is multicollinearity between them, we decided to drop one of the variables to avoid misleading results. We chose Median Household Income. Unemployment Rate and Poverty Rate are fairly correlated, but we feel that Poverty Rate is a better variable that can be used in our study to communicate our findings on prevalence of heart disease (i.e. we would like to see how the prevalence in low income and poor regions compare with prevalence). Moreover, poverty rate is more correlated with prevalence than unemployment rate.

The final choices for socioeconomic factors are Total Healthcare Costs Per Capita, Air Quality, Park Access, Population Age +65, Population Total, Urban Rural Class, Median Household Income, and Poverty Rate.

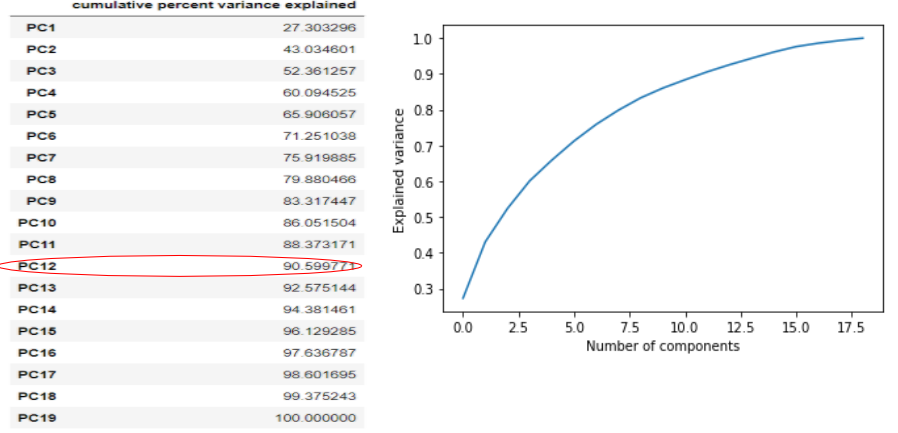
1. **Modeling**

On model development, we developed multiple linear regression models. The following four models were used: Principal Component Analysis (PCA), Exploratory Regression, Least Absolute Shrinkage and Selection Operator (Lasso), and Lasso Cross Validation (CV). The motivation for developing these four models was to determine the simplest and most efficient model (i.e. one that yields the maximum score and least root mean square test error with minimum number of predictors).

**LINEAR REGRESSION WITH PRINCIPAL COMPONENT ANALYSIS (MLR\_PCA)**

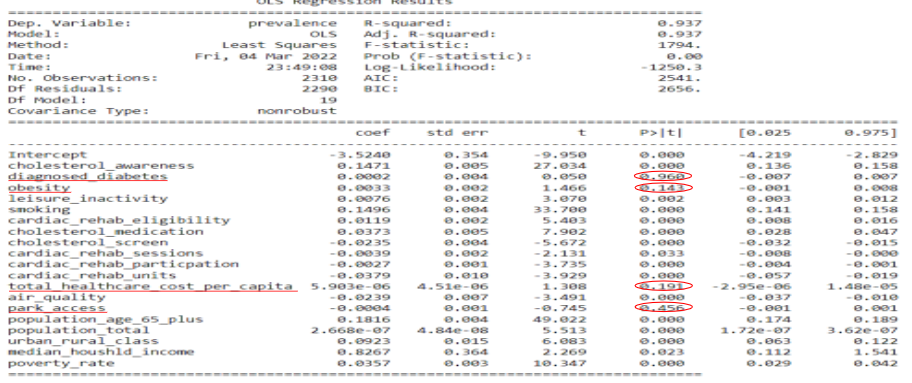
As a result of our Exploratory Data Analysis, we have chosen 18 variables reflecting the three categories of the original 38 variables. Even though we were able to decrease the variables by 20, we felt as though we should consider reducing the variables a little more by performing a Principal Component Analysis (PCA). PCA is a method used to reduce a large data set of variables to a smaller set of variables, called principal components, to simplify the analysis and modeling.

To perform the PCA, we used the Scikit-learn library. First, we normalized the data set and scaled the data to a univariate mode using Scikit-learn StandardScaler() class. We used a scaler to calculate the mean and standard deviation of each variable and then transform. Next, we applied the PCA() class, which is the decomposition submodule in Scikit-learn, to the 18 variables. Our explained variance ratio, attribute of PCA() class, returned a one-dimensional numpy array containing percentage values of variance captured by the principal components. As shown in the table below, 12 principal components were required to capture 90% of variability in the data. While this is not a significant reduction, we fit the principal components as predictors in a linear regression model (MLR\_PCA) to evaluate the performance in comparison with the regular and lasso regression models.



**EXPLORATORY REGRESSION BASED MODEL (MLR\_SIMPLE)**

The second model that we created was based on Exploratory Regression. We first created a regression model using all the data to identify predictors that have significance. We call this the exploratory phase. Thereafter, since we are more concerned with developing predictive models, it means there should be a way to validate the model. So after identifying important predictors from the exploratory phase, we removed the penalized variables, split the data into train-test split (75:25), built a new regression model (MLR\_SIMPLE) with the training data set, and validated the model with the testing set.



By inspecting the individual p-values of the coefficients of the predictors (as shown in the above table), it is observed that "diagnosed diabetes", "obesity", "total healthcare cost per capita", and "park access" with p-values 0.960, 0.143, 0.191 and 0.456 respectively, all appear to have larger p-values than alpha, considering alpha significant level of 0.05, therefore suggesting that their resulting coefficients are insignificant in the model. While it may be argued, by looking at the variables from the surface, that these four variables (especially diabetes) should significantly influence prevalence of heart disease as we believed a-priori, the observations from the exploratory model suggest that it is possible that some other variables (such as cholesterol in the case of diabetes), which are correlated with these four variables (are more correlated with prevalence such as cholesterol with 87.7% correlation), are accounting for the variations in the model significantly that the effects of these four variables become insignificant.

Based on these findings, we penalized these four variables and built a predictive linear regression as afore-explained. This model (MLR\_SIMPLE) predicts prevalence by an error of 0.4391 with 91.73% R-squared value of variations in prevalence explained by the variations in the risk factors, healthcare opportunities and socio-economic factors.

**LASSO MODEL (MLR\_LASSO)**

The third model created was the Lasso Regression Model. Lasso regression is a type of linear regression with shrinkage of parameters (i.e. pernalizing some predictors). Shrinkage is where data values are shrunk towards a central point as the mean, and the lasso procedure encourages simple, sparse models (i.e. models with fewer parameters)[[2]](#footnote-1). In this linear regression model (MLR\_LASSO), the subset of predictors that minimizes prediction error of the target variable, prevalence, is identified.

The Lasso regression requires a lambda parameter, called lasso parameter, which determines how the coefficients of the variables should be penalized. To determine the best Lasso parameter, we had to tune this parameter by training different Lasso model and select the one with the best 𝑅-squared value.

As a result of the Lasso Model, twelve predictors were penalized as insignificant, and the model predicts prevalence by an error of 0.432 with 91.99% R-squared value of variation in prevalence.

**LASSO CROSS VALIDATION MODEL (MLR\_LASSOCV)**

The fourth and final model is the Lasso Model with Cross Validation (MLR\_LassoCV). The LassoCV is similar to the Lasso Regression except that the former includes a k-fold cross validation process. Here, we used 5-fold cross validation to choose the tuning parameter lambda that applies penalty to the predictor coefficients.

As a result of the Lasso Cross Validation Model, only three factors were penalized as insignificant, and the model predicts prevalence by an error of 0.434 with 91.93% R-Squared value of variation in prevalence.

1. **Insights**

Diagram, venn diagram

Description automatically generated

In the exploratory regression model, lasso model, and lasso cross validation model, each model penalized factors. In the diagram above, “diagnosed diabetes” and “park access” were penalized in all three models.

| **Penalized (Insignificant) Predictors** | | |
| --- | --- | --- |
| **Exploratory Model** | **Lasso Model** | **Lasso Cross Validation** |
| 1. Diagnosed diabetes  2. Obesity  3. Park access  4. Total healthcare cost   per capita | 1. Diagnosed diabetes  2. Obesity  3. Cholesterol screen  4. Cardiac Rehab sessions  5. Cardiac rehab units  6. Total healthcare cost per capita  7. Air quality  8. Park access  9. Population total  10. Median household income  11.Urban rural class large fringe metro  12. Urban rural class small metro | 1. Diagnosed diabetes  2. Park access  3. Urban rural class small   metro |

The summary of the model result for all four models are shown in the table below:

Table

Description automatically generated

We can see that the Lasso has an R-squared value of 0.9199 which is higher than the simple exploratory regression of 0.9173, principal component analysis of 0.8889, as well as the Lasso cross validation with a value of 0.9193. Likewise, the prediction test error (0.432) of the Lasso model is better (i.e. lower) than the simple exploratory regression model of 0.439, principal component analysis of 0.501, as well as the Lasso cross validation with a value of 0.43397.

As such, we conclude that the best model is the Lasso Model to predict prevalence in US counties. This model predicts prevalence by an error of 0.4391 with 91.73% of variations in prevalence explained by the variations in the risk factors, healthcare opportunities and socio-economic factors.

The final nine factors that seem to influence prediction of prevalence of heart disease are: (1) high cholesterol awareness (2) physical inactivity (3) smoking (4) cardiac rehabilitation eligibility (5) cardiac rehabilitation participation (6) cholesterol medication (7) population of +65 (8) poverty rate (9) urban rural classification. It is important to note that we came to this conclusion based on the given dataset used from the CDC website. Different results could be obtained based on a different dataset as well as it is equally possible to get different results for another heart disease other than Coronary Heart Disease which is considered in our study. It is also important to note that assumptions made in the analysis may not necessarily be the best and further assessment may be required to validate them.

1. Centers for Disease Control and Prevention. (2022, January 13). *FASTSTATS - deaths and mortality*. Centers for Disease Control and Prevention. Retrieved March 27, 2022, from https://www.cdc.gov/nchs/fastats/deaths.htm [↑](#footnote-ref-0)
2. -, Great Learning Team, et al. “A Complete Understanding of Lasso Regression.” *GreatLearning Blog: Free Resources What Matters to Shape Your Career!*, 26 Dec. 2021, https://www.mygreatlearning.com/blog/understanding-of-lasso-regression/. [↑](#footnote-ref-1)