A decorative graphic on the left side of the slide consisting of two overlapping parallelograms. The front parallelogram is blue and the back one is a light green. They are positioned diagonally, with the blue one in front of the green one.

Slide deck: Building a classifier for predicting heart failure



Week 2: Exploratory Data Analysis and Visualization

- Explored how different variables in the dataset were associated with each other.
- Created a bar chart relating fbs and gender: men more cases
- Created a bar chart showing how high and normal FBS levels correlate to people of different ages. There was an especially high proportion of people with high FBS levels for some ages in the range 50-65. On the other hand, the proportion of people with high FBS levels was negligible for age groups less than 40 and 75+.
- Created a scatter plot showing the relationship between each patient's resting blood pressure and maximum heart rate. The plot also showed whether the patient had a high fasting blood sugar or not.
- The scatter plot showed little to no correlation between the two variables, and the patients with a high FBS were also scattered throughout the plot. But we did notice that there was more variation in resting blood pressure for a given heart rate than there was variation in heart rate for a given blood pressure. The results were more vertically clustered.



Week 3: Exploratory Data Analysis and Visualization Part II

- We investigated serum creatinine and how it is affected by smoking, high blood pressure and diabetes.
- We graphed the relationship between serum creatinine levels and smoking habits among people who died by heart failure. The smokers had a higher level of serum creatinine. Since all the people in this sample were people who died, there may be a correlation between high serum creatinine levels and heart failure.
- We graphed the relationship between serum creatinine and high blood pressure. There was a positive association between the two variables. High blood pressure can lead to high serum creatinine levels, since it causes damage to arteries around the kidneys and restricts blood flow to the kidneys.
- There are some graphs showing how patients with different serum creatinine levels were split between diabetic and non-diabetic patients. There is then a graph more explicitly showing the relationship between serum creatinine and diabetes. This showed a greater proportion of diabetic patients having higher serum creatinine levels compared to non-diabetic patients. This showed a positive association between serum creatinine levels and diabetes.




Week 4: Regression

- Tried out linear regression with resting blood pressure variables and age, to see which one worked better (and so which variable had a higher association with age).
- Test and training datasets were built, and then a linear regression model was initialized and fitted to these datasets.
- Gathered some summary statistics for each of the 2 regression models: training data mean average error, test data mean average error, training data root mean squared error, test data root mean squared error.
- All of these error statistics were considerably higher for the regression model using blood pressure than for the regression model using heart rate.
- Also got each of the 2 models to predict the first 3 values in their testing datasets to see how close each one got. The model using heart rate gave predicted values that were much closer than the predicted values given by the model using blood pressure. So this led to the conclusion that the model using heart rate was better overall.



Week 5: Building a good classifier

- Tried to improve the accuracy of the logistical regression model predicting age by removing select variables.
- Found the correlation coefficient between age and all the variables in the dataset, to see which variables it might help to remove.
- Removing the two variables with the lowest absolute correlation coefficient values decreased the accuracy by ~3.3 percent.
- Removing the two variables that had negative correlation coefficient values made no difference to the accuracy of the logistical regression model.
- Removing only the variable 'maximum heart rate achieved', which had the most negative correlation coefficient increased the accuracy by ~1.6 percent.



Week 6: Building a good classifier to predict heart failure

- Tried different models and different sample sizes to find the classifier model with the highest accuracy.
- Using Principal component analysis with 3 components, the accuracy was ~76.7 percent. Changing the number of principal components did not change the accuracy.
- Using Gradient Boosting, the accuracy was 80 percent.
- The initial accuracy when using logistical regression was ~83.3 percent.
- Then to improve the logistical regression model, I changed the fit intercept to false, set the verbose value to 1, and changed the number of CPU scores used to a nonzero number. I also found the maximum iterations value which yielded the highest accuracy.
- The accuracy of the classifier model after all these changes and improvements was 87.8 percent.



Sources

Dataset: <https://archive.ics.uci.edu/ml/datasets/heart+disease>

Source about serum creatinine:

<https://jamanetwork.com/journals/jamainternalmedicine/fullarticle/648077>