We are group 1 and I’m Qirui, I will go over our project about heart attack analysis and prediction. Other members are Christina Chi, Yujia Dai, Zexin Lin, Coco Liu, and Minhao Zhang. Our dataset contains 13 features representing different measures and one label with 1 meaning more chance of heart attack and 0 meaning less chance of heart attack.

With this dataset, we come up with these questions, which risk factors for heart attacks are discernible? Is there any patterns could help us find viable strategies for lowering heart attack risk? Could we train a model to predict it? And how confident are these results?

Firstly, we did exploratory data analysis. Using histogram and kernel density estimation, we find that all of the continuous variables do not skew too much and roughly follow a normal distribution, here is the result. And we calculate correlation matrix, to find any apparent relations between variables and also for feature selection. This is the heatmap and correlation scores between features and output. We find all correlation score has absolute value smaller than 0.7, so we do not need to remove any variables.

Before the modeling, we have data processing including handling null values and feature scaling for our data. We choose to delete null values but all data points are complete so we delete none of them. We also normalized data to prevent learnable coefficients to be too sensitive to large samples.

Then we used cross validation to select best models. We sorted the features based on absolute value of correlation scores between features and output in descending order, here is the models with corresponding features. Then we perform 5-fold CV on models. This results in the best model with these 8 features and achieve validation accuracy about 85%.

With the selected model, we used boostrapping to estimate the accuracy with B = 10000. Here is the result, we get 85% mean accuracy and 87% mean F-score which means the model is pretty good at prediction.

Also we build 95% confidence interval for feature parameters, and for example we can see that the CP contribute more to the results.

With our question about finding any viable strategies to prevent heart attack disease, we also used bootstrapping to estimate correlation between age, sex, Chest pain, Maximum heart rate achieved and output. Here is the result and we can see that they are not very correlated with the output though chest pain and max heart rate’s score is slightly higher.

To conclude, there is not very effective strategies we could use at home to diagnose the heart attack disease, and it is better to get medical examination. But measuring exercise induced angina and slope of peak exercise ST segment would be good for monitoring such disease.

However, for better prediction, larger dataset would be more representative of the population. Our small dataset results in problems like ineffectiveness of normalization. Collecting more data and using data augmentation would be good strategies for this. That’s all for the presentation, thanks for listening