C964 Data Story/Product Assessment

Alex Brown

11-02-2023

\*Remember to include 3 data visualization types for descriptive method\*

**Data Processing:** The focus of our data preprocessing started with analyzing the quality and completeness of the data provided. To our surprise, there were no issues with missing or invalid values present. Once we had verified the quality and completeness of the set, we moved on to finding appropriate classifications for each feature, the results of which can be found in detail in the provided Data Cleaning Log. After using scripts to classify each feature into groups, features with more than one group were split into multiple binary features. For example, the heart rate was split into three features including bradycardia, normal, and tachycardia. Each of these new features was given either a 0 to indicate no presence, or 1 to indicate the presence of that feature. For the target feature we chose Heart Attack Risk, which was classified as 0 or 1 in the original data. We interpret this to indicate the presence of a previous heart attack event, with a 0 meaning no previous heart attack, and 1 indicating at least one previous heart attack.

**Analysis of Data Features:**

The following coefficients were found when analyzing the weight of each feature on the target feature. In short, this shows the effect of each feature on the outcome of Heart Attack Risk, where positive numbers increase the chance, and negative numbers reduce the risk. The top 10 absolute highest traits have been highlighted.

'Age': 0.033533506639526006

'Sex': -0.012719081552371612

'Hypertension': 0.05412564402486295

'Diabetes': 0.07969274060029939

'Family History': -0.007750139319936583

'Smoking': -0.0959572578436689

'Alcohol Consumption': -0.06019421652431686

'Exercise Hours Per Week': 0.08808841053419869

'Previous Heart Problems': 0.013760943182008045

'Stress Level': -0.003606503125350712

'Physical Activity Days Per Week': -0.0784173597846879

'Sleep Hours Per Day': -0.034314157911411534

'Cholesterol\_0': -0.029389742174973473

'Cholesterol\_1': -0.03381475609972371

'Cholesterol\_2': 0.06296744043299052

'Heart Rate\_0': 0.00988814421932556

'Heart Rate\_1': -0.02019027471945716

'Heart Rate\_2': 0.010065072658453281

'Diet\_0': -0.020206379455390697

'Diet\_1': -0.033644259191475415

'Diet\_2': 0.05361358080518503

'Income\_0': -0.23419912900579049

'Income\_1': 0.06790682336256412

'Income\_2': 0.05400850338219674

'Income\_3': 0.022328074910097185

'Income\_4': 0.08971866950925203

'BMI\_0': -0.338139541445259

'BMI\_1': 0.17684630034004323

'BMI\_2': -0.080236126210229

'BMI\_3': -0.01200180195951802

'BMI\_4': 0.25329411143328173

'Triglycerides\_0': 0.01280148779438347

'Triglycerides\_1': 0.005107000484766213

'Triglycerides\_2': -0.018145546120816165

**Things to note:** Interestingly, a Diet of 2, or healthy diet, has a slight increased probability of heart attack risk. This could be explained retroactively by the fact that someone who has had a previous heart attack may be more likely to adopt a healthy lifestyle as a reaction to the event. Also notable is the fact that every income bracket makes the top 10, and having an income in the lowest bracket reduces the chances of heart attack.

Results from data product testing, revisions, and optimizations: After the data was cleaned and analyzed, we produced two trained models. After tuning hyperparameters and using the suggested hyperparameters, we produced two graphs, a cross-validation chart, which shows the mean-squared error over 5 fittings, and a Confusion Matrix, for each of the two models. The first model was trained using all 33 features produced in preprocessing, while the second model was reduced to using only the top 10 highest absolute coefficient features. Here are the results of the two models and their training:

Training 1, all 34 features:

A graph with red lines and numbers

Description automatically generated

A screenshot of a computer screen

Description automatically generated

At best, training with all features results in a Mean Square Error (MSE) of 0.35, and at worst 0.364. A confusion matrix unfortunately does not provide as much information as we would have liked. An imbalance in samples caused the prediction to side heavily with the prediction of no heart attack risk. There are 2240 examples in the sample where heart attack risk is 0, and 1296 where heart attack risk is present.

For the second training, where only the top 10 features were used, the following Cross-Validation graph was drawn:

A graph with red lines and dots

Description automatically generated

Using the top 10 features provided roughly the same MSE as all 33 features, with a best score of roughly 0.34 and worst at 0.37.

As a sidenote, other models were initially tested, including the Random Forest Classifier and MLP Neural Network, but neither of which could predict at the same level as the Logistic Regression model in initial tests, so we’re dropped. The following is the Cross-Validation of those models:

A graph with lines and numbers

Description automatically generated with medium confidence