Machine Learning Model to Predict Risk of Heart Disease in Hospital Patients

Business Problem

Heart disease is the leading cause of death in the United States, killing 659,000 people and costing $363 billion in diagnosis and treatment annually. Is there a way to predict which patients will develop heart disease in a certain time to save lives and reduce diagnosis costs?

Background

Heart disease kills many people in the U.S. each year and is responsible for billions of dollars in treatment costs. The fact that the U.S. has high obesity and smoking rates (42.5% and 12.5%, respectively) has not helped. The billions that hospitals spend on blood tests and EKGs to identify heart disease alone could better be spent actually treating the disease. Fortunately, there are many key predictors that increase the risk of heart disease. With such predictors, a dataset can be constructed and a machine learning model can be trained on the dataset to predict who will develop heart disease.

This project aims to save lives and reduce diagnosis costs by predicting which patients will develop heart disease in the future. A machine learning model will draw on a dataset of patients to accurately identify those who will develop heart disease in the next 5 years, using accuracy and precision to evaluate model performance. This will allow hospitals to focus their resources more effectively.

Dataset

The dataset used is a CSV file containing patient information from the Behavioral Risk Factor Surveillance System (BRFSS), a survey of patients to get their health features. Each of the more than 300,000 rows represents one patient, and each column is a feature of their general health (age, BMI, exercise level, smoking, etc). The features present in the dataset have already been pre-selected; i.e., the features are not correlated with each other so no feature selection is needed.

To prepare the dataset for training a model, each feature was checked for NA values and non-numerical columns. There were no NA values, and any Boolean/multiclass columns had their strings replaced with integers (e.g., a five-category column received 0-4 as integers). Finally, because very few patients in the dataset had heart disease, those patients were oversampled to make the number of patients with and without heart disease equal.

Methods/Analysis

After cleaning the data, constructing the machine learning model will consist of two parts: running exploratory data analysis (EDA) then training/testing the actual model: a random forest classifier. A random forest classifier was chosen as the model as it performs very well when the dataset is very large and has a categorical target variable. This target variable, the presence of heart disease in the patient, has elements of either 0 (no) or 1 (yes).

The first step, EDA, was conducted to gain insight into the data and ballpark the range that the accuracy and precision might fall under. This involved constructing a few visualizations for insight. First, histograms of the numerical variables (and bar charts for categorical ones) were constructed to reveal any skewness in the data. Then, a bar chart was made of the age groups split by the presence of heart disease to show which ages might be most at risk. Finally, a correlation matrix was set up to ensure that the features were, in fact, not correlated (this was the only visual made after replacing strings with integers). It was revealed that older people are more likely to get heart disease and it was confirmed that there was little to no feature correlation.

The random forest classifier was very simple to train and test once data cleaning and EDA were completed. Random forest is a collection of several decision trees, each generating a class prediction. The majority prediction wins in a manner resembling democracy (see Figure 1). The data was first split into training and testing sets, with 20% going to testing. After training the model, predictions were generated from the testing set and compared to the actual values of the target variable. Because the target variable is categorical, the model's performance was evaluated with accuracy and precision. These values were 96.8% and 94.0%, respectively (see Figure 2). In this

case, a higher accuracy meant more of the model's predictions were correct, while a higher precision meant more patients predicted to develop heart disease actually developed it. Both values were above 90%, which means the model is safe to use in hospitals for predicting heart disease.

Conclusion

The project's goal was to build a model to accurately predict which patients will develop heart disease within 5 years. With scoring metrics above 90%, it succeeded in this endeavor. The model that is described within the project has the potential to transform the healthcare sector if its accuracy remains as high as the results. Hospitals will immediately know who will have heart disease in 5 years and shift their treatments to only those patients, saving time, costs, and most importantly, lives.

Assumptions

The random forest model was developed with a few assumptions in mind to allow the project to proceed. First, heart disease was assumed to be a single diagnosable disease. However, heart disease is an umbrella term for many similar diseases with similar risk factors; the model is outputting whether a patient will develop one of them. Second, every patient in the dataset was assumed to have answered the survey truthfully so the model wouldn't be incorrect in its real-world predictions.

Challenges/Limitations

The model building stage was not without its trip-ups. Logistic regression was the first logical choice for binary classification since it also handles large datasets, but it quickly proved to be undesirable, producing accuracy and precision scores of 72% and 73%, respectively. Medical models must be very accurate as patients' lives can be drastically affected by a false positive/negative. The logistic model was scrapped in favor of a Keras neural network which produced higher scores, but the network took many hours to run even after tuning hyperparameters. Finally, a random forest classifier was selected as it also performs well with large data and a categorical target variable. This classifier resulted in the high scores seen above, though its runtime still leaves

something to be desired. More computing power will be needed to solve this problem for future uses.

Future Uses/Recommendations

In addition to this model, similar models could find uses outside of healthcare, such as in marketing. If a similar survey is conducted asking buyers what products they find useful and what they will buy, then the similar model can predict who will buy a class of products and who won't buy any class from that store. It is recommended that a similar model for cell service providers be developed first, as cell customer information tends to have less features than other marketplaces' customers and can be obtained via survey format like the BRFSS (many cell providers already conduct surveys). Such a model could predict who will remain with the provider and who won't. If this algorithm is successful as well, then other companies can begin to use similar models for binary classification problems.

Ethical Assessment

Assessing ethics should always be foremost, and it was one of the first steps after identifying the project's business problem. While this project succeeded in predicting heart disease, there is one ethical implication. Because of the high accuracy, health insurance companies can use these predictions to increase the costs on those patients at risk of developing heart disease. This will lead to the elderly having disproportionately higher insurance premiums than the young, as the chance of developing heart disease is higher among older generations. Great care must be taken to keep this model from being abused.

Implementation Plan

See Figure 3.

Questions

1. When testing the different models, were there any others that could be tested?
2. On a larger data scale, is there enough computing power for hospital data?
3. Are there other ways of optimizing runtimes besides more computing power?
4. If heart disease is multiple diseases, could the model predict which disease a patient would develop in addition to a Yes/No answer?
5. The diseases that make up the heart disease umbrella are mentioned to have “similar risk factors” but are they similar enough that one model can predict all?
6. Will this model work for predicting other diseases with similar risk factors?
7. Would hyperparameter tuning have affected the values of the scoring metrics?
8. What other future uses are there for this model besides for cell phone providers?
9. Why was over-sampling chosen over under-sampling to balance the classes?
10. How can insurance companies be prevented from abusing this model?

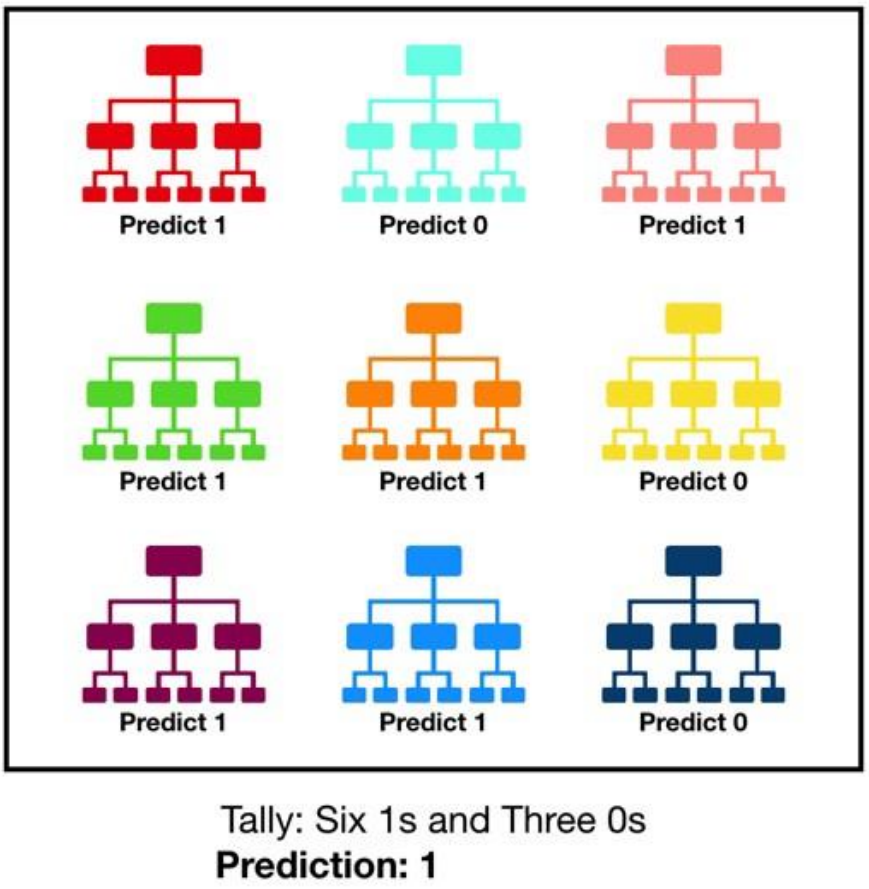

Figure 1: Diagram showing how a random forest classifier works. The decision trees making up the forest each output one class prediction, and the majority prediction is chosen as the forest's output, a "democratic" approach.

```
In [14]:
#Training and testing the model
rfc = RandomForestClassifier()
rfc.fit(X_train, y_train)
y_pred = rfc.predict(X_test)
```

```
In [15]:
#Getting the accuracy and precision of the model
print("Accuracy: {} percent".format(round(accuracy_score(y_test, y_pred), 3)*100))
print("Precision: {} percent".format(round(precision_score(y_test, y_pred), 3)*100))

#The accuracy and precision are bothvery high, indicating that this model is safe to use for heart disease prediction

Accuracy: 96.8 percent
Precision: 94.0 percent
```

Figure 2: Python code for training and evaluating the classifier model. Both accuracy and precision percentages were in the 90s, indicating the model is accurate for predicting heart disease.

Figure 3: The implementation plan for this project, in order: 1) The problem must be understood and the data must be suitable to solve the problem, 2) Any moral/ethical dilemmas must be dealt with, 3) It must be proven that the model has a benefit greater than the cost to build and deploy it, 4) The dataset must be cleaned, 5) The model will be built and performance evaluated, and 6) the model will be deployed accordingly. Questions must be answered at each step of the process; selected questions are shown above.

References

Allen, M. (2018). *Health Insurers and Vacuuming Up Details About You – And It Could Raise Your Rates.* ProPublica. Retrieved on April 30, 2022 from https://www.propublica.org/article/health-insurers-are-vacuuming-up-details-about-you-and-it-could-raise-your-rates.

Beckerman, J. (2021). *Heart Disease: Types, Causes, and Symptoms.* WebMD. Retrieved on April 30, 2022 from https://www.webmd.com/heart-disease/heart-disease-types-causes-symptoms.

Karabiber, F. (2022). *Binary Classification.* LearnDataSci. Retrieved on April 30, 2022 from https://www.learndatasci.com/glossary/binary-classification/.

Koehrsen, W. (2018). *Hyperparameter Tuning the Random Forest in Python.* Towards Data Science. Retrieved on April 30, 2022 from https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74.