Framingham Heart Study

Jennifer Tin
August 21, 2019
University of Toronto - School of Continuing Studies
SCS 3253-029 Machine Learning

Project Overview

Goal: to create the model that will predict the likelihood of a patient having Chronic Heart Disease (CHD) in 10 years

Method: to create a model using a logistic regression

Data: from Kaggle - *Framingham Heart Study (this was 1 point in time)*

Why this topic?: I come from a health care background, and I thought it would be interesting to do this project on something that applicable to my field

Data Overview

- This dataset is from Kaggle, using the *Framingham Heart Study** dataset
  - Comprised of 4240 patients
  - A series of health related data in both continuous, nominal, yes/no variables are provided
  - There were some missing data

*The Framingham Heart Study in real life is a longitudinal dataset, starting in 1948, however, the data found on Kaggle only provided 1 point in time, which was a big limitation to the data model

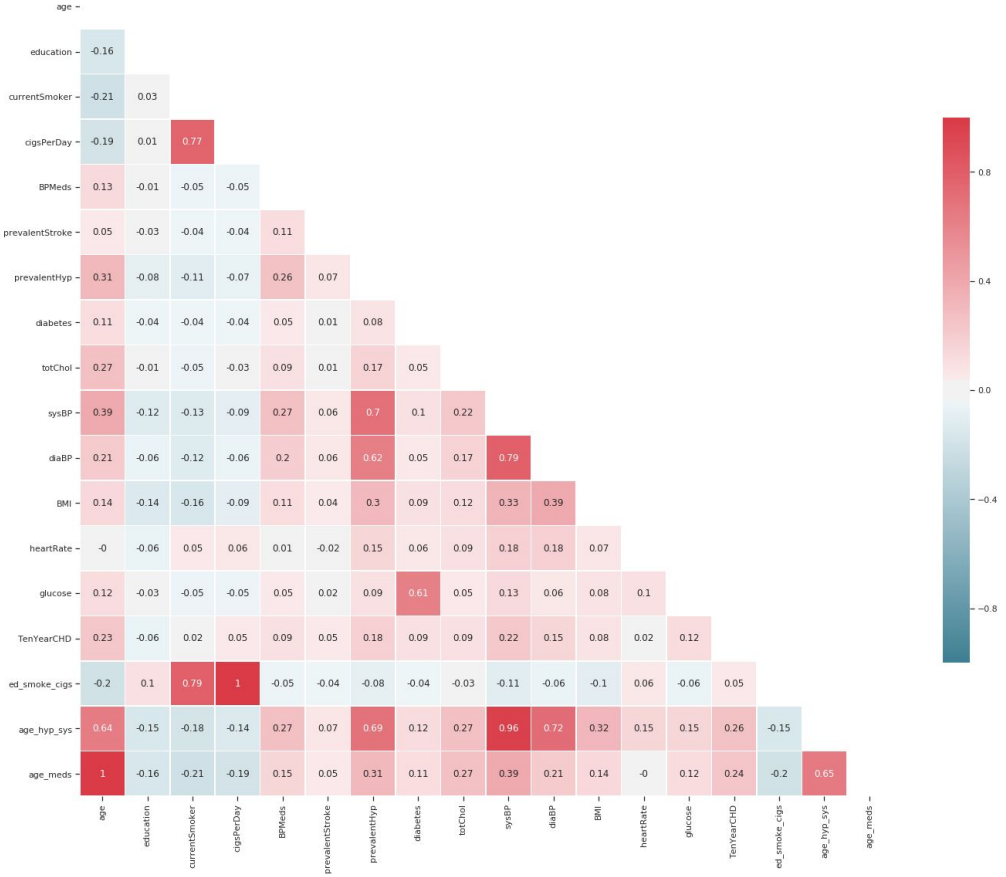
Data Overview

```
age                 0
education         105
currentSmoker       0
cigsPerDay         29
BPMeds             53
prevalentStroke     0
prevalentHyp        0
diabetes            0
totChol            50
sysBP               0
diaBP               0
BMI                19
heartRate           1
glucose           388
TenYearCHD          0
dtype: int64
```

- The data elements to the left is the dataset that was downloaded
- The numbers shown are the null values, and those were dropped during the data cleaning process
- After the dropping of null values, 3658 patients were remaining (down from 4240)

Creating The Model

- This model was built using a logistic regression
- Data feature engineering was tried, however, it was not successful



Creating The Model

```
[78] st.chisqprob = lambda chisq, df: st.chi2.sf(chisq, df)
     cols=heart_df_constant.columns[:-1]
     model=sm.Logit(data.TenYearCHD,heart_df_constant[cols])
     result=model.fit()
     result.summary()
```

```
---------------------------------------------------------------------------
PerfectSeparationError                    Traceback (most recent call last)
<ipython-input-78-112675ac4fff> in <module>()
      2 cols=heart_df_constant.columns[:-1]
      3 model=sm.Logit(data.TenYearCHD,heart_df_constant[cols])
----> 4 result=model.fit()
      5 result.summary()

                                  5 frames
/usr/local/lib/python3.6/dist-packages/statsmodels/discrete/discrete_model.py in _check_perfect_pred(self, params, *args)
    198                 np.allclose(fittedvalues - endog, 0)):
    199             msg = "Perfect separation detected, results not available"
--> 200             raise PerfectSeparationError(msg)
    201
    202     def fit(self, start_params=None, method='newton', maxiter=35,

PerfectSeparationError: Perfect separation detected, results not available
```

SEARCH STACK OVERFLOW

- The most correlated variables to “TenYearCHD” were selected for engineering, but the new variables did not make a difference in the model

- The variables with the least correlation were also dropped, however, this decreased the accuracy and AUC

The Final Model

```
[287] #test model accuracy
      sklearn.metrics.accuracy_score(y_test,y_pred)
```

```
0.8661202185792349
```

```
[288] sklearn.metrics.roc_auc_score(y_test,y_pred)
```

```
0.5368691817736404
```

Business Use - Cons

- Predictive models such as this one ***COULD*** be used in clinical practice, but it is also ***UNLIKELY*** due to:
  - There are huge risks with using a computer model in clinical practice due to the legal implications (i.e. malpractice, the sensitivity and specificity of the algorithms, the predictive accuracy, etc.)
  - To predict the likelihood of someone to have CHD in 10 years will require more data on a longitudinal basis, however, this model was built using point-in-time data at one interval

Business Use - Pros

- However, models such as this could be used for:
  - Research
  - Public health surveillance
  - Informing health policies

How To Improve The Model

- To improve this model, the data must be improved:
  - If we were to continue with the same data elements - Longitudinal data (collecting the patient's overtime) so there can be trends collected
  - To add more data elements - such as their lifestyles, eating habits, activity level, socio-demographic data → research shows there are lifestyle and genetic risk factors that increases the chances of someone having CHD, however, not all of these data points were collected in the dataset on Kaggle

With improved data, we can then try data feature engineering

Thank You