Cardiovascular Disease Prediction Exploratory Data Analysis

Bhavesh Shah, Griffin Coccari, Jamie Chen
DSCI 303

Overview

- Cardiovascular diseases (CVD’s) have been the #1 cause of death globally for the past few years. According to the World Health Org, close to 18 million people die annually from CVD’s.

- That being said, many studies show that almost 80% of CVD’s can indeed be prevented, including heart disease and stroke.

- Our Goal: Create a tool to help doctors with predicting a patient’s chances of getting CVD’s with high accuracy, so that they can be prescribed necessary treatments/ medication early-on.

Previous Research

Machine Learning–Driven Models to Predict Prognostic Outcomes in Patients Hospitalized With Heart Failure Using Electronic Health Records: Retrospective Study

- **Goal**: Predict 1-year in-hospital mortality, use of positive inotropic agents, and 1-year all-cause readmission rate
- **Method**: Decision tree of mortality risk (after consideration of logistic regression, support vector machine, artificial neural network, random forest, and extreme gradient boosting models)
- **Data**: real-world electronic health records

Using Deep Learning to Identify High-Risk Patients with Heart Failure with Reduced Ejection Fraction | Published in Journal of Health Economics and Outcomes Research

- **Goal**: Predict hospitalizations, worsening HF events, and 30-day and 90-day readmissions in patients with heart failure with reduced ejection fraction (HFrEF)
- **Data**: Adult HFrEF patients from IBM® MarketScan® Commercial and Medicare Supplement databases (2015-2017)
- **Method**: Sequential model architecture based on bi-directional long short-term memory (Bi-LSTM) layers (also tested traditional ML models such as logistic regression, random forest, and eXtreme Gradient Boosting (XGBoost))
- **Results**: For all outcomes assessed, the DL approach outperformed traditional machine learning models

Data Overview

Link to data: https://www.kaggle.com/sulianova/cardiovascular-disease-dataset

Snapshot of data:

12 Features

	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
0	0	18393	2	168	62.0	110	80	1	1	0	0	1	0
1	1	20228	1	156	85.0	140	90	3	1	0	0	1	1
2	2	18857	1	165	64.0	130	70	3	1	0	0	0	1
3	3	17623	2	169	82.0	150	100	1	1	0	0	1	1
4	4	17474	1	156	56.0	100	60	1	1	0	0	0	0

Number of observations: 70000

Target (0=does not have CVD; 1= has CVD)

Feature Overview

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 70000 entries, 0 to 69999
Data columns (total 13 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   id           70000 non-null  int64
 1   age          70000 non-null  float64
 2   gender       70000 non-null  int64
 3   height       70000 non-null  int64
 4   weight       70000 non-null  float64
 5   ap_hi        70000 non-null  int64
 6   ap_lo        70000 non-null  int64
 7   cholesterol  70000 non-null  int64
 8   gluc         70000 non-null  int64
 9   smoke        70000 non-null  int64
 10  alco         70000 non-null  int64
 11  active       70000 non-null  int64
 12  cardio       70000 non-null  int64
dtypes: float64(2), int64(11)
memory usage: 6.9 MB
None
```

Objective (measured) Features

- **Age** (in days) [NUMERIC]
- **Gender** (1=female, 2=male) [BINARY]
- **Height** (in cm) [NUMERIC]
- **Weight** (in kg) [NUMERIC]

Examination Features

- **Systolic blood pressure/ ap_hi** (mm of mercury - mmHg) [NUMERIC]
- **Diastolic blood pressure/ ap_lo** (mm of mercury - mmHg) [NUMERIC]
- **Cholesterol** (1=norm; 2=above norm; 3=well above norm) [TERNARY]
- **Glucose** (1=norm; 2=above norm; 3=well above norm) [TERNARY]

Subjective Features

- **Smoking** (0=does not smoke; 1=smokes) [BINARY]
- **Alcohol Intake** (0=does not drink; 1=frequent drinker) [BINARY]
- **Physical Activity** (0=not very active; 1=active) [BINARY]

Ideally, there shouldn’t be too many subjective features. We may need to give slightly less weight to these features in our final model to decrease variance.

Observation Overview

Data is already clean!

Missing Values

```
id             0
age            0
gender         0
height         0
weight         0
ap_hi          0
ap_lo          0
cholesterol    0
gluc           0
smoke          0
alco           0
active         0
cardio         0
dtype: int64
```

No need to impute

Unique Elements

```
id             70000
age             8076
gender             2
height           109
weight           287
ap_hi            153
ap_lo            157
cholesterol        3
gluc               3
smoke              2
alco               2
active             2
cardio             2
dtype: int64
```

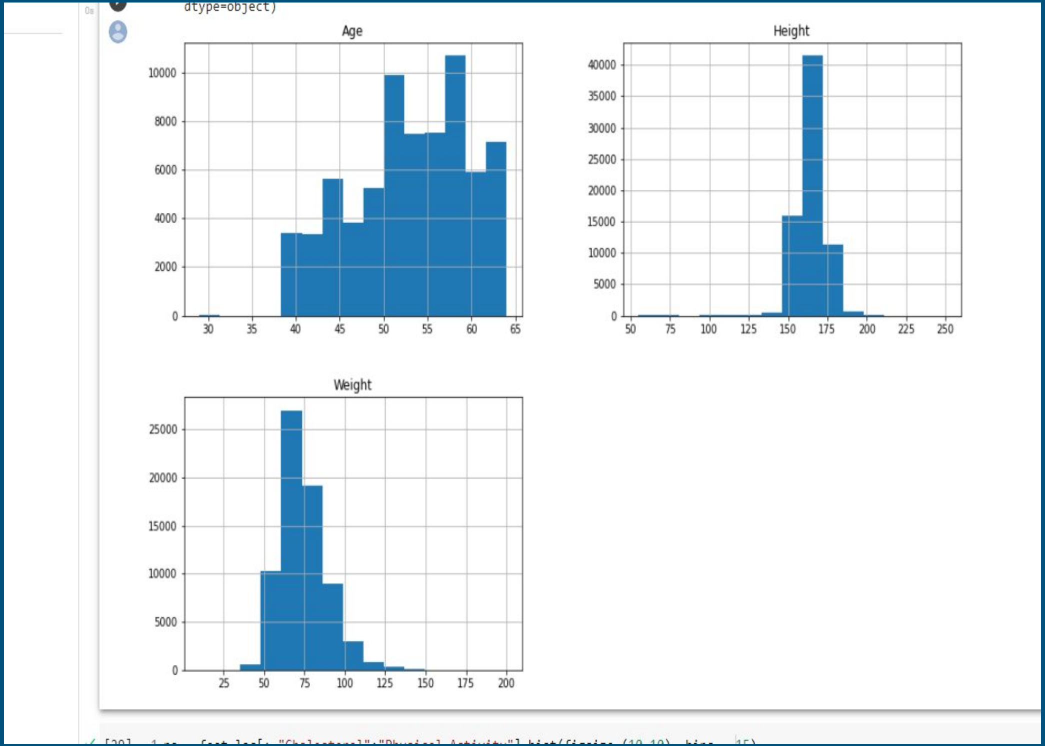
Binary and ternary variables are as expected

Duplicate Rows

```
0          False
1          False
2          False
3          False
4          False
           ...
69995      False
69996      False
69997      False
69998      False
69999      False
Length: 70000, dtype: bool
Total Number of Duplicates: 0
```

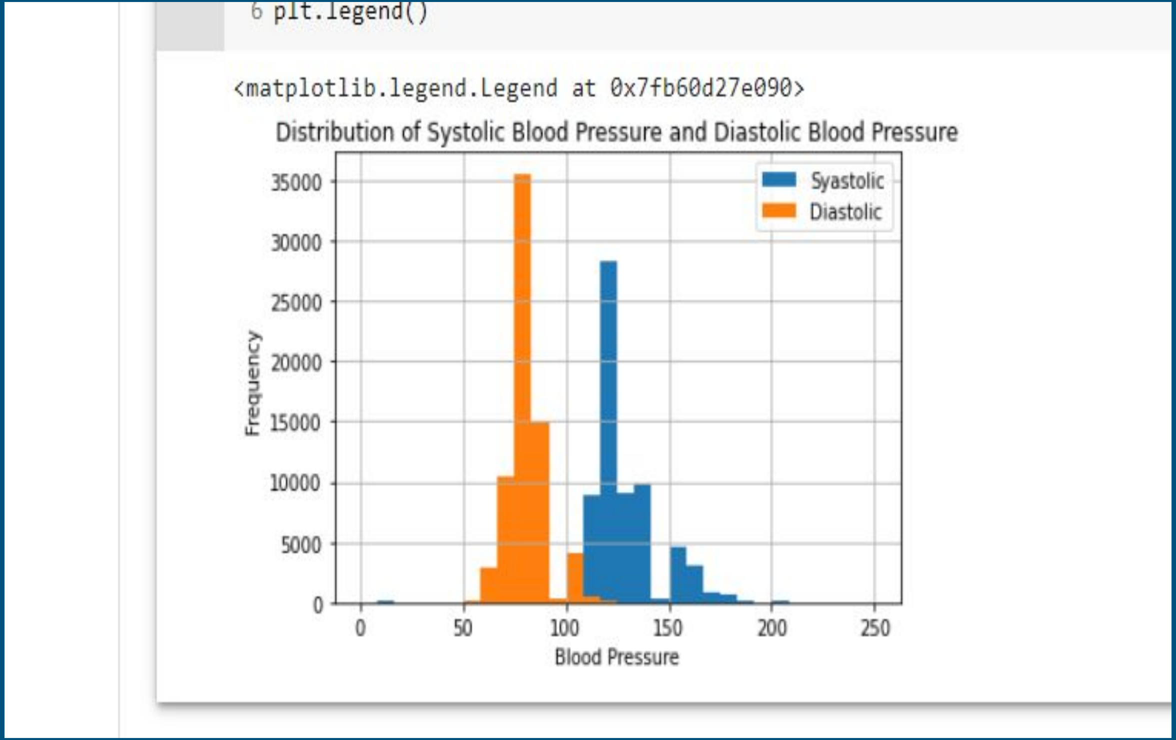
No need to remove tuples

Numerical Feature Distributions



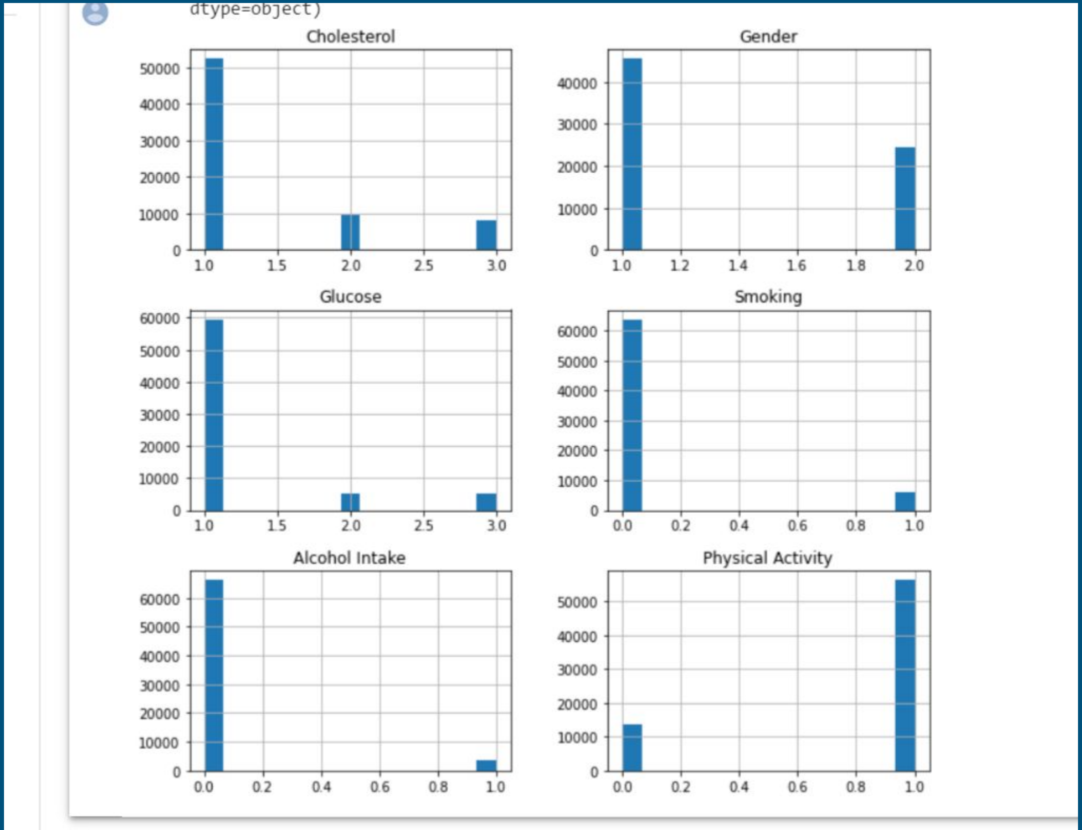
- Age is skewed to the right meaning we have mostly older participants in this dataset.
  - This indicates that our model will not be very applicable to a younger population.
- Height and weight appear to be normally distributed

Numerical Feature Distributions Cont.



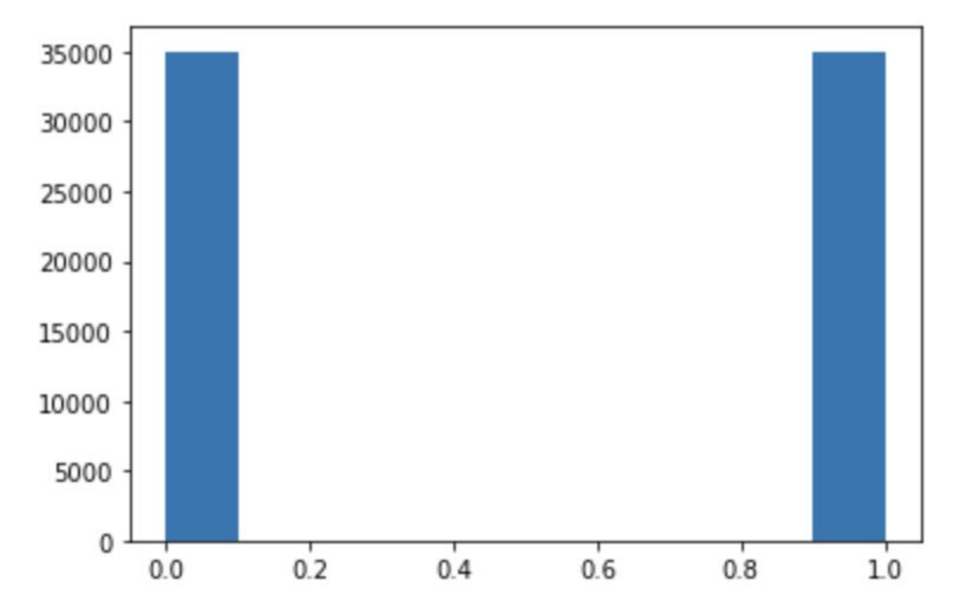
- Systolic Blood Pressure (pressure your heart exerts on your arteries when it pumps) and Diastolic Blood Pressure (pressure on your arteries between pumps of the heart) both appear to be normally distributed.

Categorical Feature Distributions



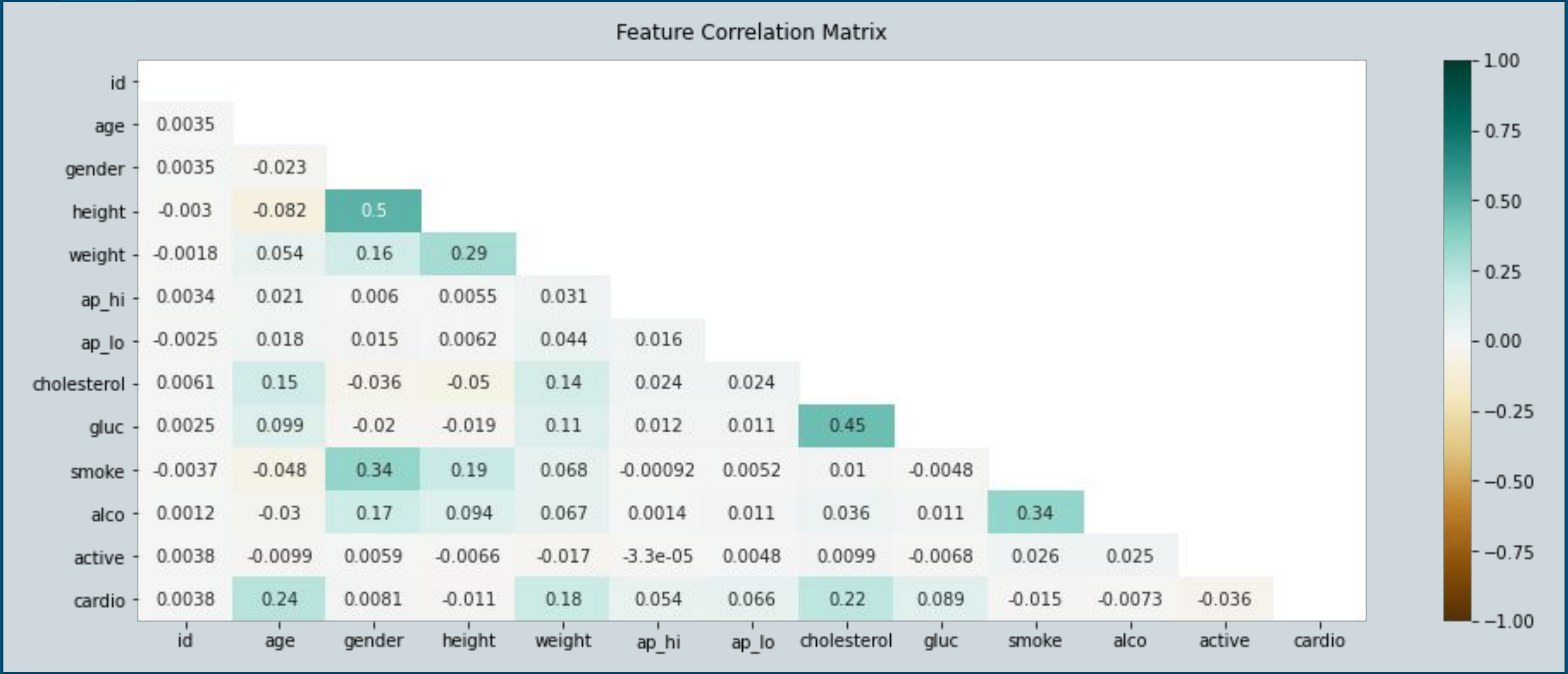
- Gender shows high bias because our sample deviates from the true distribution of gender in the population.
- Other variables such as cholesterol, glucose, smoking, alcohol intake, and physical activity are also significantly unbalanced.

Target Distribution



- Exactly equal number of observations for patients with cardiovascular disease and without cardiovascular disease.
- Confusion matrix will be a good way for us to measure our model performance.

Feature Correlation



Feature Correlation Continued

Most Correlated Features

1) Height and gender: .50
2) Cholesterol and glucose: .45
3) Alcohol and smoking: .34

Most Correlated with Target

1) Age: .24
2) Cholesterol: .22
3) Weight: .18

Don’t have multicollinearity since none of these values exceed .50 (usually considered minimum threshold for high dependence).

Not very linearly correlated (will definitely need to combine multiple features in our models)

Questions?