A Project Presentation on

“HEART DISEASE”

Bachelor of Technology

in

Computer Science and Engineering

SRM INSTITUTE OF SCIENCE & TECHNOLOGY
(Deemed to be University u/s 3 of UGC Act, 1956)

Submitted by:

SYED SAIF ALI ALVI(RA1711003030261)

SHARFUL HAQUE(RA1711003030296)

DIVYANSHU MISHRA(RA171100303057)

INTRODUCTION

- Of all the applications of machine-learning, diagnosing any serious disease using a black box is always going to be a hard sell. If the output from a model is the particular course of treatment (potentially with side-effects), or surgery, or the *absence* of treatment, people are going to want to know **why**.

- This dataset gives a number of variables along with a target condition of having or not having heart disease. Below, the data is first used in a simple random forest model, and then the model is investigated using ML explainability tools and techniques.

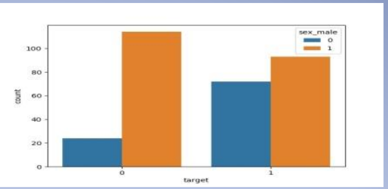
THE DATA

- Next, load the data,
- In [2]:dt = pd.read_csv("../input/heart.csv")

  Let's take a look,
- In [3]:dt.head(10)
- Out[3]:

```
In [36]: runfile('C:/Users/SH/Downloads/Visulalization File.py', wdir='C:/Users/SH/
Downloads')
   age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  ca  thal
target
0   63    1   3       145   233    1        0      150      0      2.3      0   0     1
1
1   37    1   2       130   250    0        1      187      0      3.5      0   0     2
1
2   41    0   1       130   204    0        0      172      0      1.4      2   0     2
1
3   56    1   1       120   236    0        1      178      0      0.8      2   0     2
1
4   57    0   0       120   354    0        1      163      1      0.6      2   0     2
1
5   57    1   0       140   192    0        1      148      0      0.4      1   0     1
1
6   56    0   1       140   294    0        0      153      0      1.3      1   0     2
1
7   44    1   1       120   263    0        1      173      0      0.0      2   0     3
1
8   52    1   2       172   199    1        1      162      0      0.5      2   0     3
1
9   57    1   2       150   168    0        1      174      0      1.6      2   0     2
```

It's a clean, easy to understand set of data. However, the meaning of some of the column headers are not obvious. Here's what they mean,



1. **age**: The person's age in years
2. **sex**: The person's sex (1 = male, 0 = female)
3. **cp:** The chest pain experienced (Value 1: typical angina, Value 2: atypical angina, Value 3: non-anginal pain, Value 4: asymptomatic)
4. **trestbps:** The person's resting blood pressure (mm Hg on admission to the hospital)
5. **chol:** The person's cholesterol measurement in mg/dl
6. **fbs:** The person's fasting blood sugar (> 120 mg/dl, 1 = true; 0 = false)
7. **restecg:** Resting electrocardiographic measurement (0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria)
8. **thalach:** The person's maximum heart rate achieved
9. **exang:** Exercise induced angina (1 = yes; 0 = no)
10. **oldpeak:** ST depression induced by exercise relative to rest ('ST' relates to positions on the ECG plot.)
11. **slope:** the slope of the peak exercise ST segment (Value 1: upsloping, Value 2: flat, Value 3: downsloping)
12. **ca:** The number of major vessels (0-3)
13. **thal:** A blood disorder called thalassemia (3 = normal; 6 = fixed defect; 7 = reversable defect)
14. **target:** Heart disease (0 = no, 1 = yes)

To avoid HARKING (or Hypothesizing After the Results are Known) I'm going to take a look at online guides on how heart disease is diagnosed, and look up some of the terms above.

Diagnosis: The diagnosis of heart disease is done on a combination of clinical signs and test results. The types of tests run will be chosen on the basis of what the physician thinks is going on , ranging from electrocardiograms and cardiac computerized tomography (CT) scans, to blood tests and exercise stress tests .

Looking at information of heart disease risk factors led me to the following: **high cholesterol, high blood pressure, diabetes, weight, family history and smoking** . According to another source , the major factors that can't be changed are: **increasing age, male gender and heredity**. Note that **thalassemia**, one of the variables in this dataset, is heredity. Major factors that can be modified are: **Smoking, high cholesterol, high blood pressure, physical inactivity, and being overweight and having diabetes**. Other factors include **stress, alcohol and poor diet/nutrition**.

I can see no reference to the 'number of major vessels', but given that the definition of heart disease is **"...what happens when your heart's blood supply is blocked or interrupted by a build-up of fatty substances in the coronary arteries"**, it seems logical the *more* major vessels is a good thing, and therefore will reduce the probability of heart disease.

Given the above, I would hypothesis that, if the model has some predictive ability, we'll see these factors standing out as the most important.

Let's change the column names to be a bit clearer,

In [4]:dt.columns = ['age', 'sex', 'chest_pain_type', 'resting_blood_pressure', 'cholesterol', 'fasting_blood_sugar', 'rest_ecg', 'max_heart_rate_achieved', 'exercise_induced_angina', 'st_depression', 'st_slope', 'num_major_vessels', 'thalassemia', 'target']

In [5]: dt['sex'][dt['sex'] == 0] = 'female' dt['sex'][dt['sex'] == 1] = 'male'

dt['chest_pain_type'][dt['chest_pain_type'] == 1] = 'typical angina'

dt['chest_pain_type'][dt['chest_pain_type'] == 2] = 'atypical angina'

dt['chest_pain_type'][dt['chest_pain_type'] == 3] = 'non-anginal pain'

dt['chest_pain_type'][dt['chest_pain_type'] == 4] = 'asymptomatic'

dt['fasting_blood_sugar'][dt['fasting_blood_sugar'] == 0] = 'lower than 120mg/ml'

dt['fasting_blood_sugar'][dt['fasting_blood_sugar'] == 1] = 'greater than 120mg/ml'

dt['rest_ecg'][dt['rest_ecg'] == 0] = 'normal' dt['rest_ecg'][dt['rest_ecg'] == 1] = 'ST-T wave abnormality' dt['rest_ecg'][dt['rest_ecg'] == 2] = 'left ventricular hypertrophy'

dt['exercise_induced_angina'][dt['exercise_induced_angina'] == 0] = 'no'

dt['exercise_induced_angina'][dt['exercise_induced_angina'] == 1] = 'yes'

dt['st_slope'][dt['st_slope'] == 1] = 'upsloping' dt['st_slope'][dt['st_slope'] == 2] = 'flat'

dt['st_slope'][dt['st_slope'] == 3] = 'downsloping' dt['thalassemia'][dt['thalassemia'] == 1] = 'normal' dt['thalassemia'][dt['thalassemia'] == 2] = 'fixed defect'

dt['thalassemia'][dt['thalassemia'] == 3] = 'reversable defect'

Check the data types,
In [6]:dt.dtypes
Out[6]:

```
age                          int64
sex                         object
chest_pain_type             object
resting_blood_pressure       int64
cholesterol                  int64
fasting_blood_sugar         object
rest_ecg                    object
max_heart_rate_achieved      int64
exercise_induced_angina     object
st_depression              float64
st_slope                    object
num_major_vessels            int64
thalassemia                 object
target                       int64
dtype: object
```

In [7]:dt['sex'] = dt['sex'].astype('object') dt['chest_pain_type'] = dt['chest_pain_type'].astype('object') dt['fasting_blood_sugar'] = dt['fasting_blood_sugar'].astype('object') dt['rest_ecg'] = dt['rest_ecg'].astype('object') dt['exercise_induced_angina'] = dt['exercise_induced_angina'].astype('object') dt['st_slope'] = dt['st_slope'].astype('object') dt['thalassemia'] = dt['thalassemia'].astype('object')

In[8]: dt.dtypes

Out[8]:

```
age                           int64
sex                          object
chest_pain_type              object
resting_blood_pressure        int64
cholesterol                   int64
fasting_blood_sugar          object
rest_ecg                     object
max_heart_rate_achieved       int64
exercise_induced_angina      object
st_depression               float64
st_slope                     object
num_major_vessels             int64
thalassemia                  object
target                        int64
dtype: object
```

In[9]: dt = pd.get_dummies(dt, drop_first=True)

Now let's see,

In [10]: dt.head()

Out[10]:

```
   age  resting_blood_pressure  cholesterol  max_heart_rate_achieved  st_depression  num_major_vessels  target
st_slope_downsloping  st_slope_flat  st_slope_upsloping  sex_male  chest_pain_type_atypical angina  chest_pain_type_non-anginal
pain  chest_pain_type_typical angina  fasting_blood_sugar_lower than 120mg/ml  rest_ecg_left ventricular hypertrophy
rest_ecg_normal  exercise_induced_angina_yes  thalassemia_fixed defect  thalassemia_normal  thalassemia_reversable defect
0   63                     145          233                         150           2.3                0       1
0              0                   1         1                            0                                  0
0                                        0                                  0                1                            0
0                   1                              0
1   37                     130          250                         187           3.5                0       1
0              0                   1         1                            0                                  1
0                                        1                                  0                0                            0
1                   0                              0
2   41                     130          204                         172           1.4                0       1
1              0                   0         0                            1                                  0
0                                        1                                  0                1                            0
1                   0                              0
3   56                     120          236                         178           0.8                0       1
1              0                   0         1                            1                                  0
0                                        1                                  0                0                            0
1                   0                              0
4   57                     120          354                         163           0.6                0       1
1              0                   0         0                            0                                  0
1                                        1                                  0                0                            1
1                   0                              0
```

Permissions: **RW** End-of-lines: **CRLF** Encoding: **ASCII** Line: **63** Column: **17** Memory: **48 %**

THE MODEL

The next part fits a support vector machine model to the data,

In [11]:X_train, X_test, y_train, y_test = train_test_split(dt.drop('target', 1), dt['target'], test_size = .271, random_state=37) #spillting the data

model = SVC(kernel='linear',gamma='scale',probability=True)

model.fit(X_train, y_train)

Out[11]:

```
SVC(C=1.0, cache_size=200, class_weight=None, coef0=0.0,
    decision_function_shape='ovr', degree=3, gamma='scale', kernel='linear',
    max_iter=-1, probability=True, random_state=None, shrinking=True, tol=0.001,
    verbose=False)
```

In[12]:y_predict = model.predict(X_test)

y_pred_quant = model.predict_proba(X_test)[:, 1]

y_pred_bin = model.predict(X_test)

Assess the fit with a confusion matrix,

In [16]: confusion_matrix = confusion_matrix(y_test, y_pred_bin)

total=sum(sum(confusion_matrix))

Out[16]:

```
array([[28,  7],
       [ 3, 23]])
```

Diagnostic tests are often sold, marketed, cited and used with **sensitivity** and **specificity** as the headline metrics. Sensitivity and specificity are defined as,

$$Sensitivity = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

$$Specificity = \frac{True\ Negatives}{True\ Negatives + False\ Positives}$$

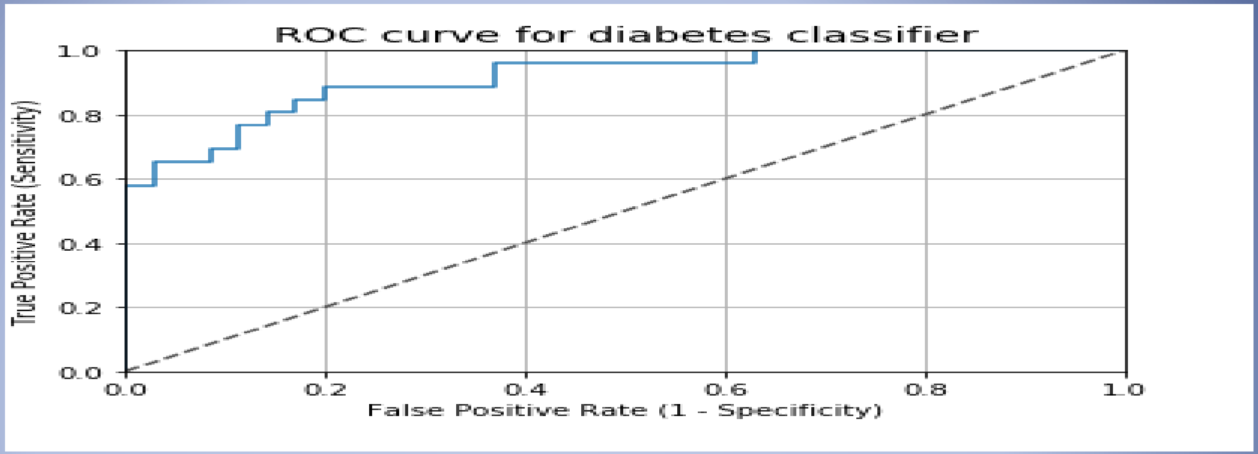
```
In[17]:total=sum(sum(confusion_matrix))
sensitivity = confusion_matrix[0,0]/(confusion_matrix[0,0]+confusion_matrix[1,0])
print('Sensitivity : ', sensitivity )

specificity = confusion_matrix[1,1]/(confusion_matrix[1,1]+confusion_matrix[0,1])
print('Specificity : ', specificity)
```

```
Sensitivity :  0.8717948717948718
Specificity :  0.8863636363636364
```

That seems reasonable. Let's also check with a **Receiver Operator Curve (ROC)**

IN[18]: fpr, tpr, thresholds = roc_curve(y_test, y_pred_quant) fig, ax = plt.subplots() ax.plot(fpr, tpr) ax.plot([0, 1], [0, 1], transform=ax.transAxes, ls="--", c=".3") plt.xlim([0.0, 1.0]) plt.ylim([0.0, 1.0]) plt.rcParams['font.size'] = 12 plt.title('ROC curve for diabetes classifier') plt.xlabel('False Positive Rate (1 - Specificity)') plt.ylabel('True Positive Rate (Sensitivity)') plt.grid(True)



Another common metric is the **Area Under the Curve**, or **AUC**. This is a convenient way to capture the performance of a model in a single number, although it's not without certain issues. As a rule of thumb, an AUC can be classed as follows,

- 0.90 - 1.00 = excellent
- 0.80 - 0.90 = good
- 0.70 - 0.80 = fair
- 0.60 - 0.70 = poor
- 0.50 - 0.60 = fail

Let's see what the above ROC gives us,

In [19]: auc(fpr, tpr)
Out[19]: 0.9131868131868132

OK, so it's working well.

THE EXPLANATION

Now let's see what the model gives us from the ML explainability tools.

Permutation importance is the first tool for understanding a machine-learning model, and involves shuffling individual variables in the validation data (after a model has been fit), and seeing the effect on accuracy. Learn more here.

Let's take a look,

In [20]:perm = perm = PermutationImportance(model, random_state=1).fit(X_test, y_test)

diag = eli5.explain_weights(perm, feature_names = X_test.columns.tolist())

Out[20]:

	feature	weight	std
0	num_major_vessels	4.819277e-02	0.040321
1	chest_pain_type_typical angina	3.855422e-02	0.020728
2	thalassemia_fixed defect	2.650602e-02	0.009016
3	resting_blood_pressure	2.409639e-02	0.007620
4	st_depression	1.927711e-02	0.022346
5	thalassemia_reversable defect	1.686747e-02	0.012287
6	max_heart_rate_achieved	1.686747e-02	0.012287
7	st_slope_flat	1.445783e-02	0.009016
8	chest_pain_type_atypical angina	1.445783e-02	0.009016
9	thalassemia_normal	1.204819e-02	0.000000
10	age	7.228916e-03	0.009639
11	st_slope_upsloping	4.819277e-03	0.005902
12	exercise_induced_angina_yes	4.819277e-03	0.005902
13	st_slope_downsloping	4.819277e-03	0.009639
14	chest_pain_type_non-anginal pain	0.000000e+00	0.000000
15	fasting_blood_sugar_lower than 120mg/ml	0.000000e+00	0.000000
16	rest_ecg_left ventricular hypertrophy	0.000000e+00	0.000000
17	rest_ecg_normal	0.000000e+00	0.000000
18	cholesterol	0.000000e+00	0.000000
19	sex_male	-2.220446e-17	0.013198

So, it looks like the most important factors in terms of permutation is a thalessemia result of 'reversable defect'. The high importance of 'max heart rate achieved' type makes sense, as this is the immediate, subjective state of the patient at the time of examination (as opposed to, say, age, which is a much more general factor).

Let's take a closer look at the number of major vessles using a Partial Dependence Plot (learn more here). These plots vary a single variable in a single row across a range of values and see what effect it has on the outcome. It does this for several rows and plots the average effect. Let's take a look at the 'num_major_vessels' variable, which was at the top of the permutation importance list

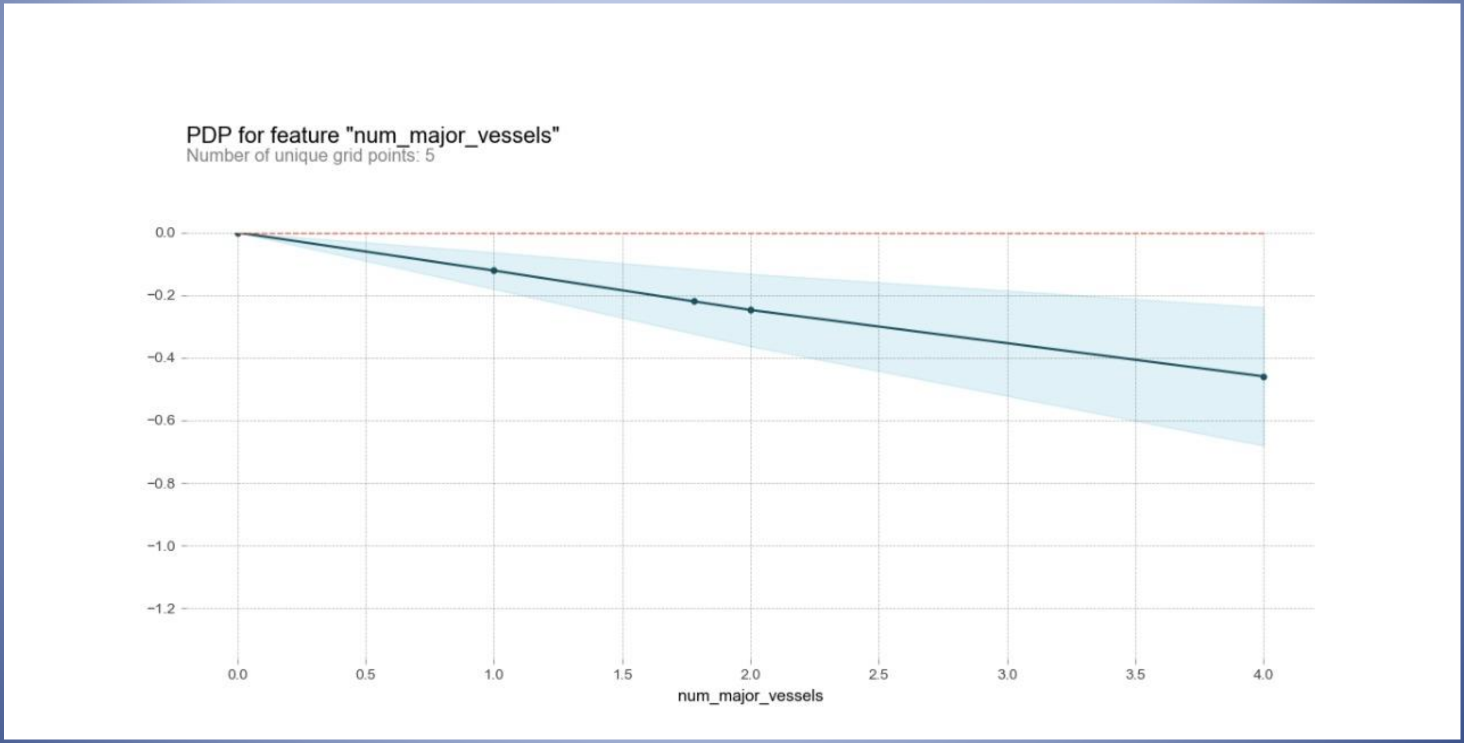
```
In[21]:cr = classification_report(y_test,y_predict)
    print(cr)
```

Out[21]:

	precision	recall	f1-score	support
0	0.87	0.87	0.87	39
1	0.89	0.89	0.89	44
accuracy			0.88	83
macro avg	0.88	0.88	0.88	83
weighted avg	0.88	0.88	0.88	83

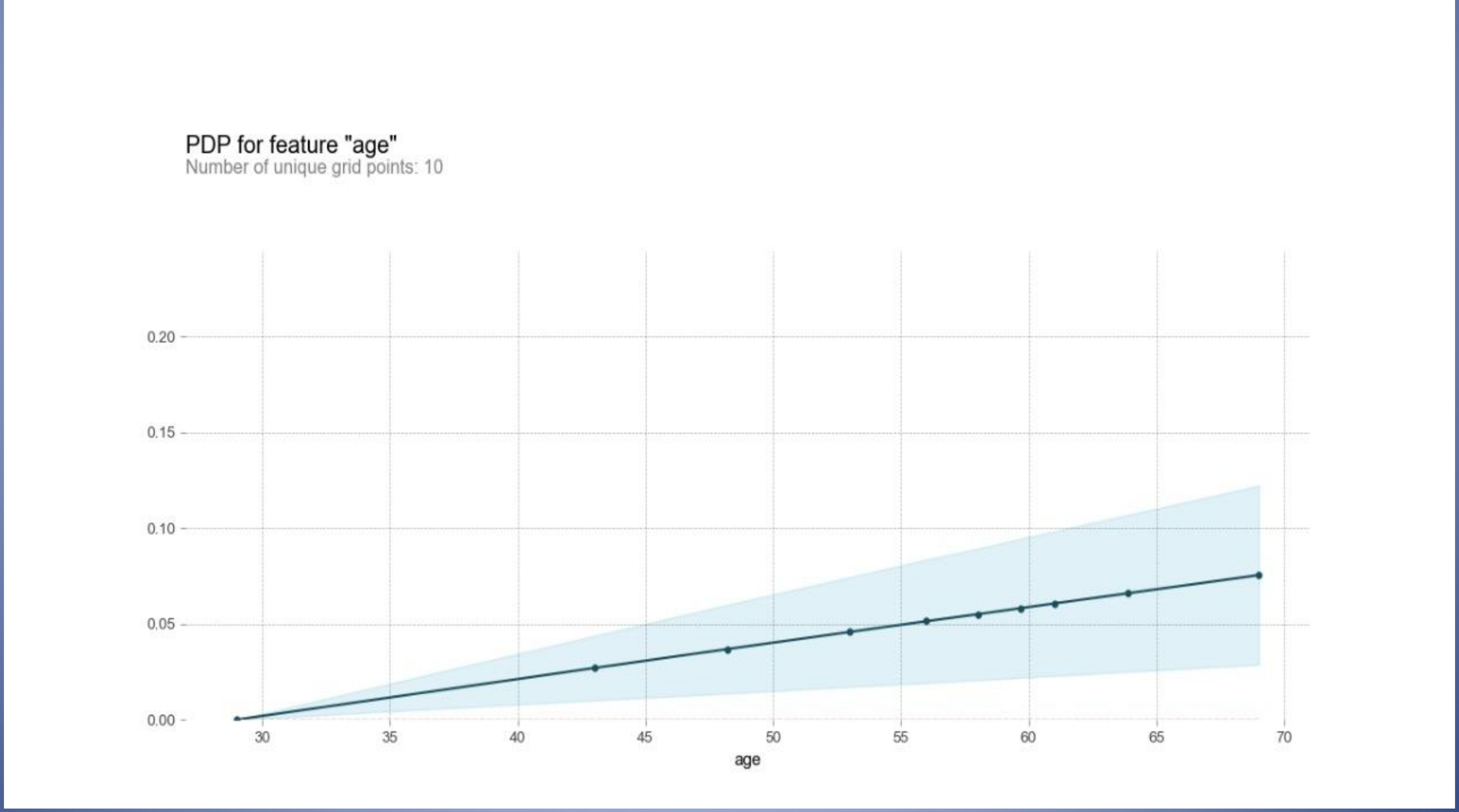
```
In[21]:base_features = dt.columns.values.tolist()
    base_features.remove('target')
    feat_name = 'num_major_vessels'
    pdp_dist = pdp.pdp_isolate(model=model, dataset=X_test, model_features=base_features
    feature=feat_name)
    pdp.pdp_plot(pdp_dist, feat_name)
    plt.show()
```

Out[21]:



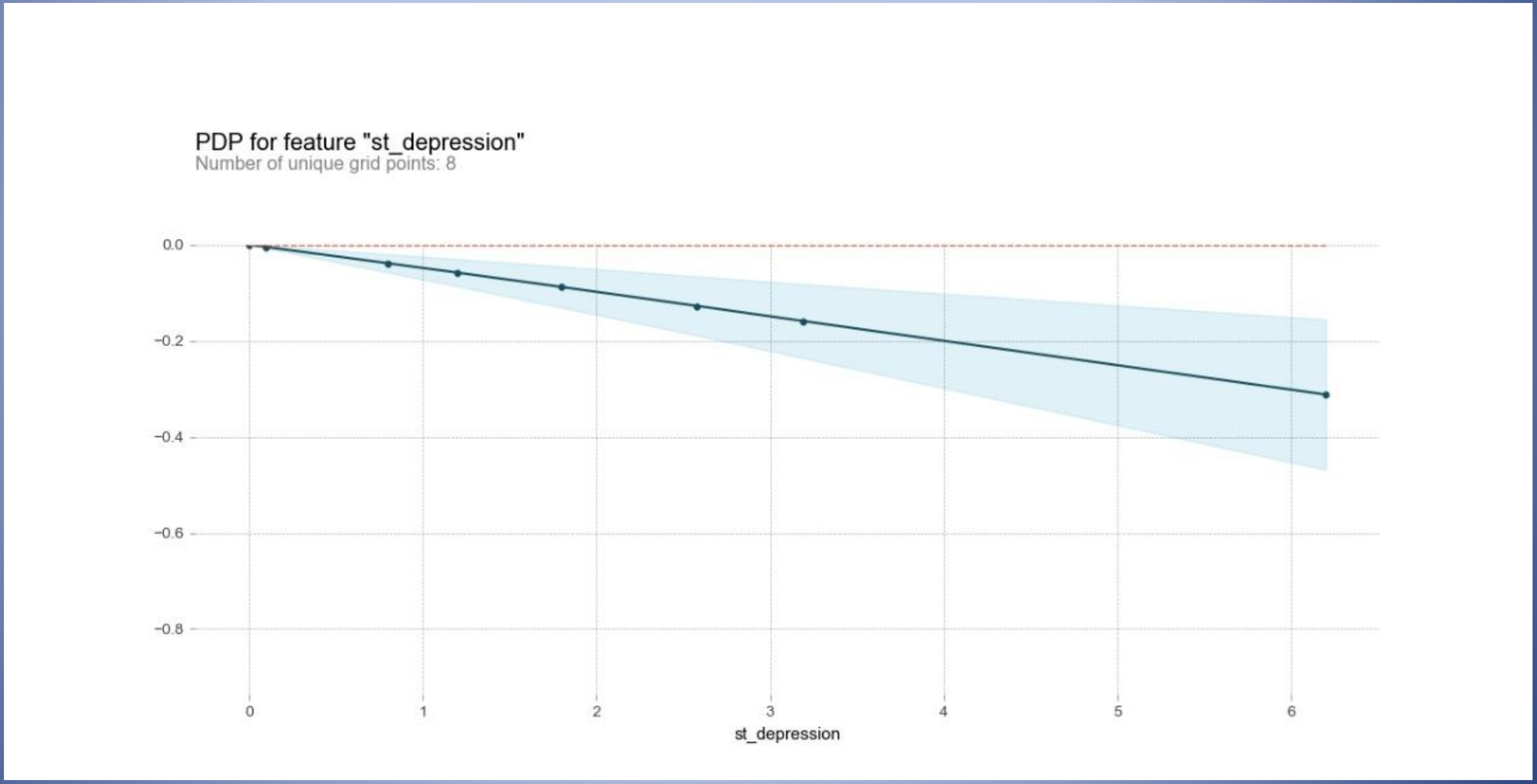
```
In[22]:feat_name = 'age'
    pdp_dist = pdp.pdp_isolate(model=model, dataset=X_test, model_features=base_features, feature=feat_name)
    pdp.pdp_plot(pdp_dist, feat_name)
    plt.show()
```

Out[22]:

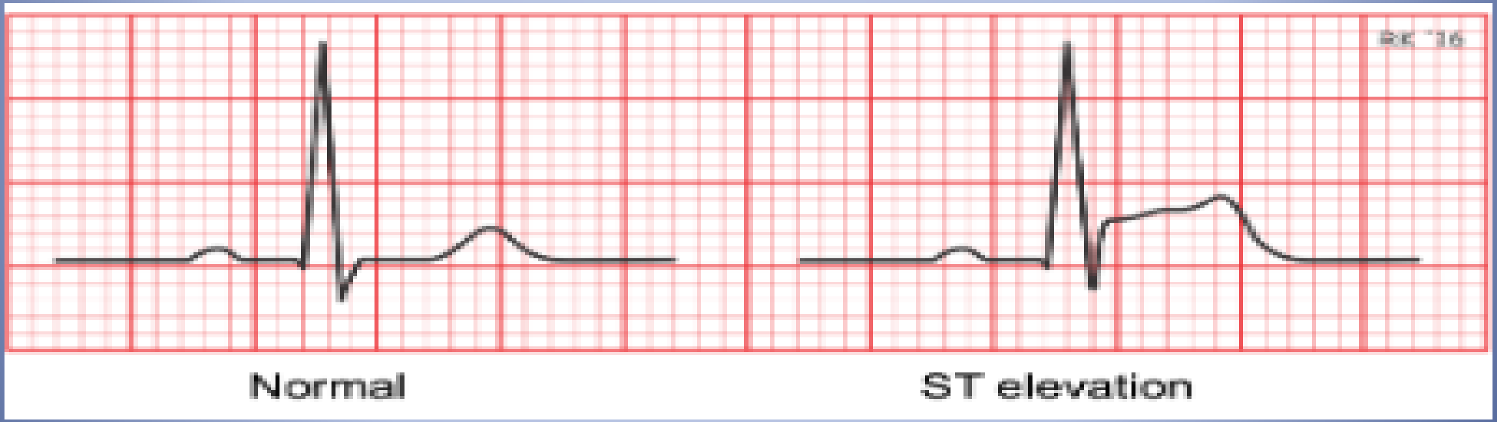


```
In[23]:feat_name = 'st_depression'
    pdp_dist = pdp.pdp_isolate(model=model, dataset=X_test, model_features=base_features,
    feature=feat_name)
    pdp.pdp_plot(pdp_dist, feat_name)
    plt.show()
```

Out[23]:



The ST segment represents the heart's electrical activity immediately after the right and left ventricles have contracted, pumping blood to the lungs and the rest of the body. Following this big effort, ventricular muscle cells relax and get ready for the next contraction. During this period, little or no electricity is flowing, so the ST segment is even with the baseline or sometimes slightly above it. The faster the heart is beating during an ECG, the shorter all of the waves become. The shape and direction of the ST segment are far more important than its length. Upward or downward shifts can represent decreased blood flow to the heart from a variety of causes, including heart attack, spasms in one or more coronary arteries (Prinzmetal's angina), infection of the lining of the heart (pericarditis) or the heart muscle itself (myocarditis), an excess of potassium in the bloodstream, a heart rhythm problem, or a blood clot in the lungs (pulmonary embolism)."



So, this variable, which is described as 'ST depression induced by exercise relative to rest', seems to suggest the higher the value the higher the probability of heart disease, but the plot above shows the opposite. Perhaps it's not just the depression amount that's important, but the interaction with the slope type? Let's check with a 2D PDP,

APPLICATION

```
Lets check your heart!!!!!
Enter you age?
21
The persons resting blood pressure (mm Hg on admission to the hospital)
121
The persons cholesterol measurement in mg/dl
180
The persons maximum heart rate achieved
102
ST depression induced by exercise relative to rest  1- 4
0
The number of major vessels (0-3)
3
the slope of the peak exercise ST segment (Value 0: upsloping, Value 1: flat, V
0
The persons sex (1 = male, 0 = female)
1
The chest pain experienced (Value 0: typical angina, Value 1: atypical angina,
2
The persons fasting blood sugar (> 120 mg/dl, 1 = true; 0 = false)
0
Resting electrocardiographic measurement (0 = normal, 1 = having ST-T wave abnc
0
Exercise induced angina (1 = yes; 0 = no)
0
A blood disorder called thalassemia (0 = normal; 1 = fixed defect; 2 = reversab
0
Congratulations you DO NOT HAVE HEART DISEASE
```

CONCLUSION

- This dataset is old and small by today's standards. However, it's allowed us to create a simple model and then use various machine learning explainability tools and techniques to peek inside. At the start, I hypothesised, using (Googled) domain knowledge that factors such as cholesterol and age would be major factors in the model. This dataset didn't show that. Instead, the number of major factors and aspects of ECG results dominated. I actually feel like I've learnt a thing or two about heart disease!
- I suspect this sort of approach will become increasingly important as machine learning has a greater and greater role in health care.