Heart Disease Prediction

Previously, I have worked on the house price prediction project using and comparing different machine learning methods：

- **Cross Validation:** Using 12-fold cross-validation
- **Models:** On each run of cross-validation I fit 7 models (ridge, svr, gradient boosting, random forest, xgboost, lightgbm regressors)
- **Stacking:** In addition, I trained a meta StackingCVRegressor optimized using xgboost
- **Blending:** All models trained will overfit the training data to varying degrees. Therefore, to make final predictions, I blended their predictions together to get more robust predictions.

The house price prediction is well-established, so I want to challenge myself by starting a brand new project focusing more on data visualization and explainability. So that I could have the opportunities to work on both regression and classification issues on the course project. **I want to give a BIG THANK to Professor for giving me the opportunity to work on two datasets**

For this project, the goal is to predict if a person will have **heart disease** based on 303 observations and 14 available feature as follow:

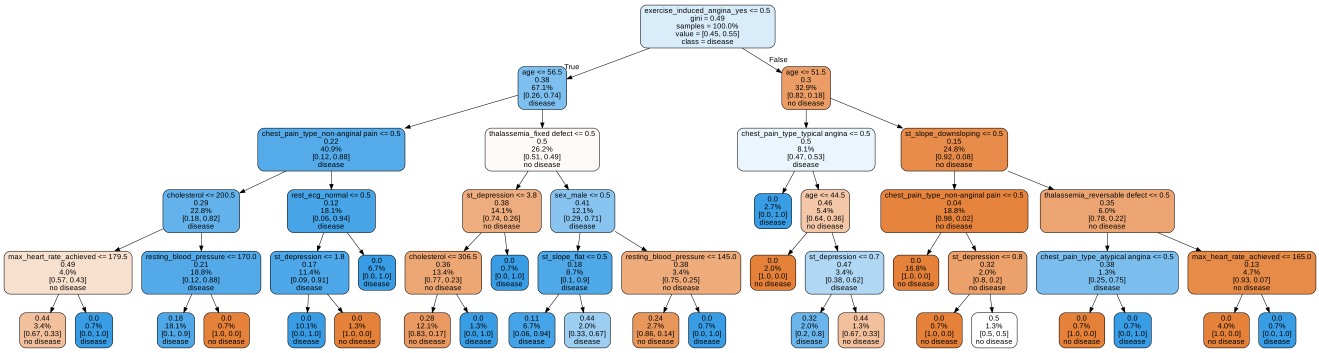
- **age**: The person's age in years
- **sex**: The person's sex (1 = male, 0 = female)
- **cp:** The chest pain experienced (Value 1: typical angina, Value 2: atypical angina, Value 3: non-anginal pain, Value 4: asymptomatic)
- **trestbps:** The person's resting blood pressure (mm Hg on admission to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

I used **random forest** as below for prediction.

```
model = RandomForestClassifier(max_depth=5,n_estimators=100,random_state=5)
model.fit(X_train,y_train)
```

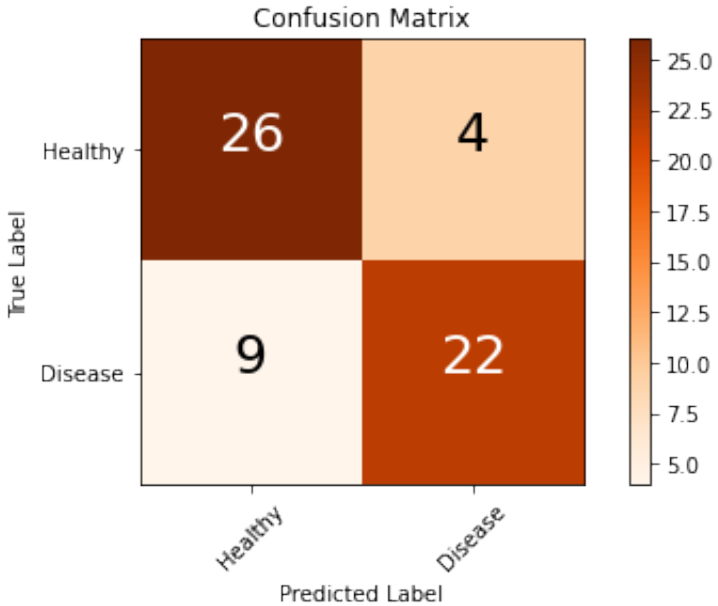
The primary purpose is to open the black box of the algorithm, and I tried 7 models and stacking / blending on the previous house prediction project. So I purposely didn't make too complicated models, so I could use different packages such as `pandas_profiling`, `graphviz`, `pydotplus`, `pdpbox`, `eli5` for data visualization, and `SHAP` for explainability on the random forest model.

Let's see how the random forest make decisions - using `graphviz`

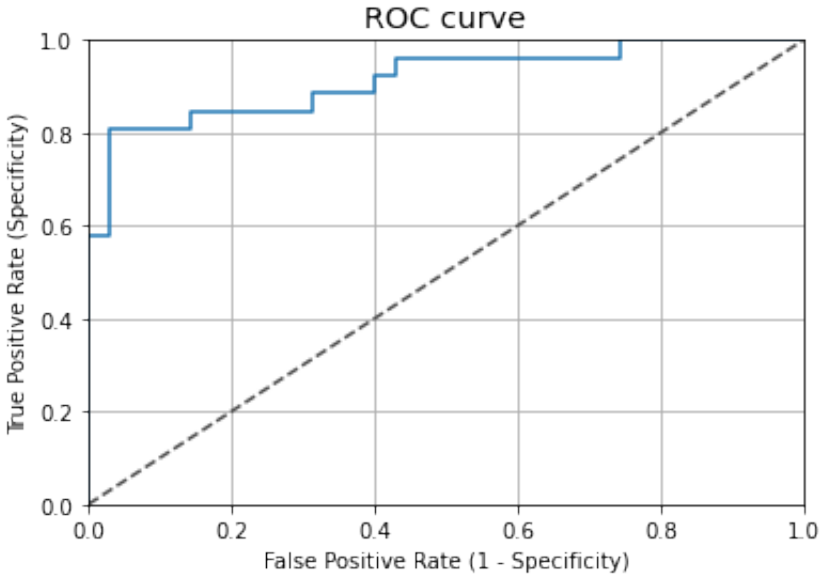


Pretty clear, we got some idea how the random forest model work...**Not a black box anymore!!!**

How accurate it is?



Not too bad...but the false positive is 9 of the 61 test sets...

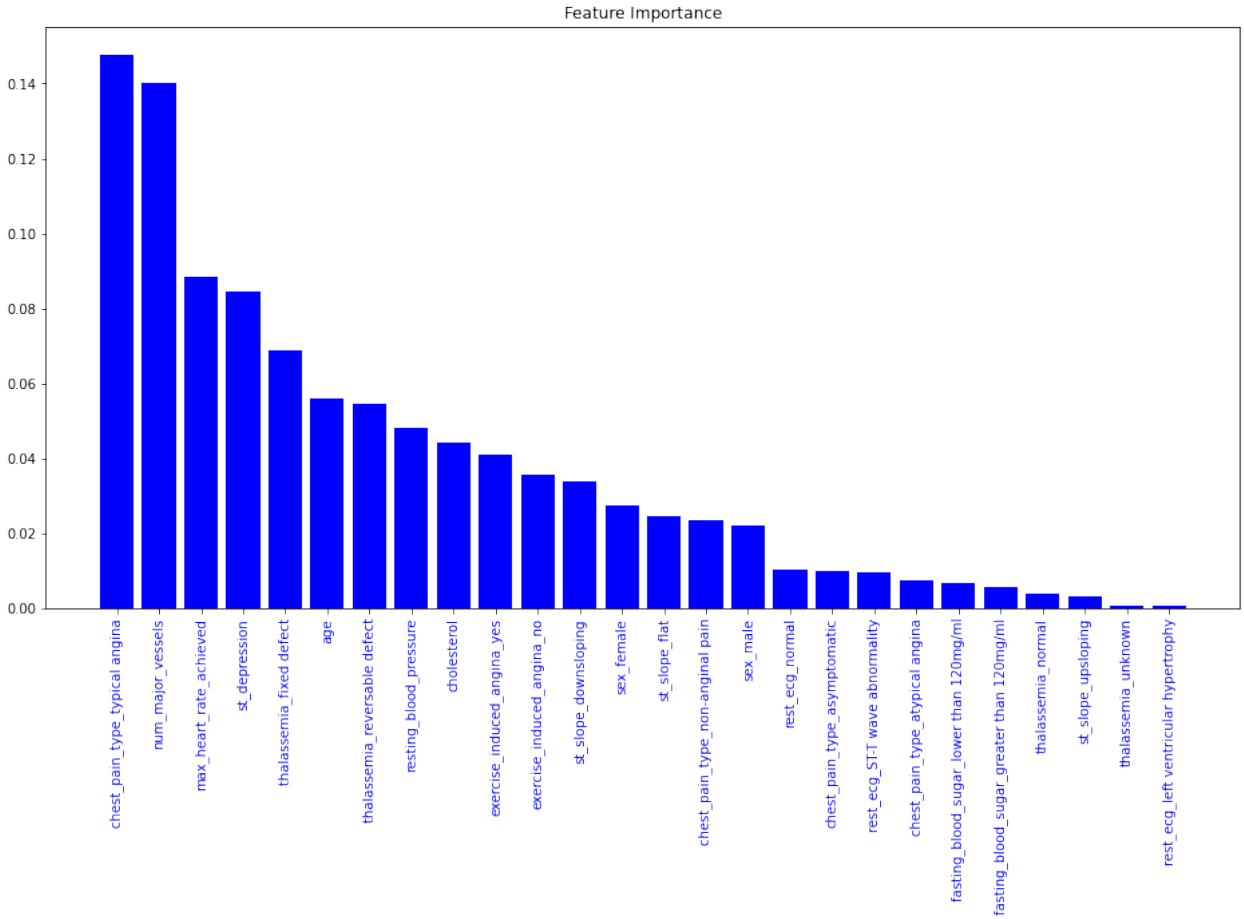


ROC curve seems pretty good!

Check how the features contribute to the results - using `eli5` :

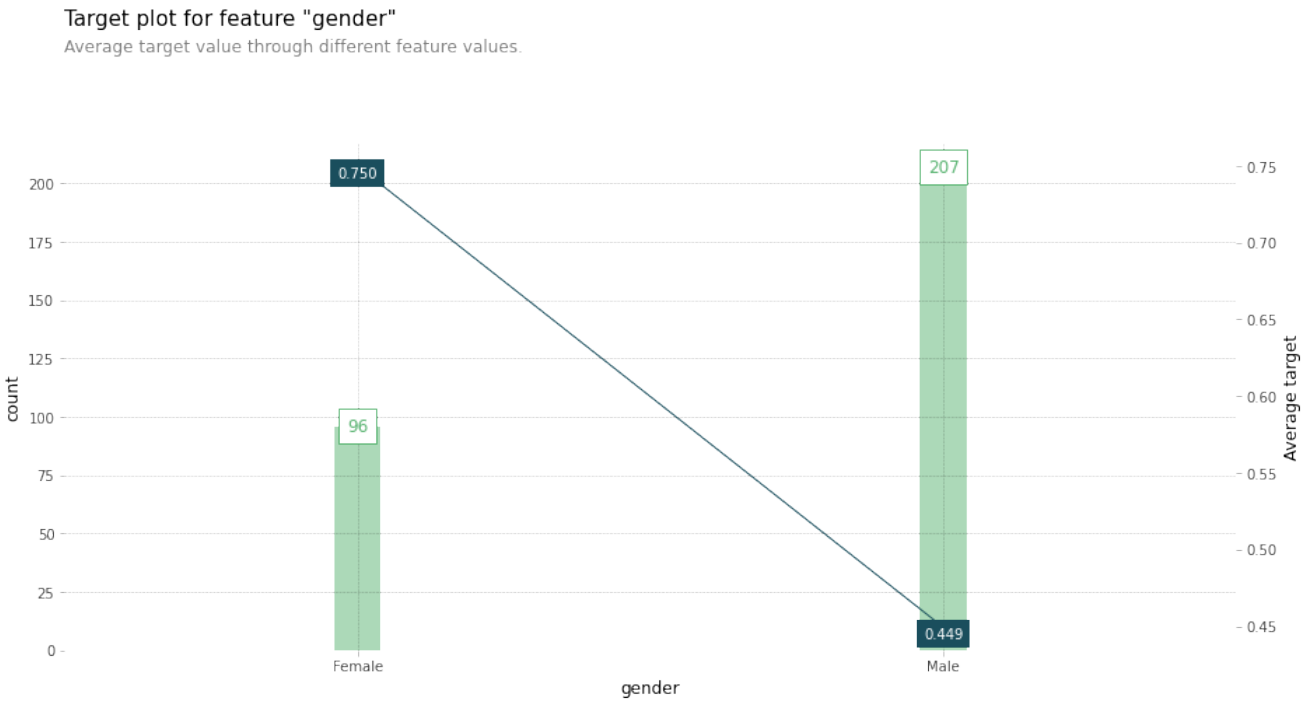
Weight	Feature
0.1477 ± 0.3141	chest_pain_type_typical angina
0.1402 ± 0.2353	num_major_vessels
0.0885 ± 0.1962	max_heart_rate_achieved
0.0847 ± 0.1604	st_depression
0.0690 ± 0.2245	thalassemia_fixed defect
0.0561 ± 0.1060	age
0.0544 ± 0.1846	thalassemia_reversable defect
0.0483 ± 0.0768	resting_blood_pressure
0.0441 ± 0.0855	cholesterol
0.0411 ± 0.1505	exercise_induced_angina_yes
0.0358 ± 0.1396	exercise_induced_angina_no
0.0338 ± 0.1212	st_slope_downsloping
0.0275 ± 0.0921	sex_female
0.0245 ± 0.0797	st_slope_flat
0.0235 ± 0.0788	chest_pain_type_non-anginal pain
0.0220 ± 0.0865	sex_male
0.0104 ± 0.0308	rest_ecg_normal
0.0098 ± 0.0470	chest_pain_type_asymptomatic
0.0097 ± 0.0353	rest_ecg_ST-T wave abnormality
0.0077 ± 0.0358	chest_pain_type_atypical angina

1. Make a plot

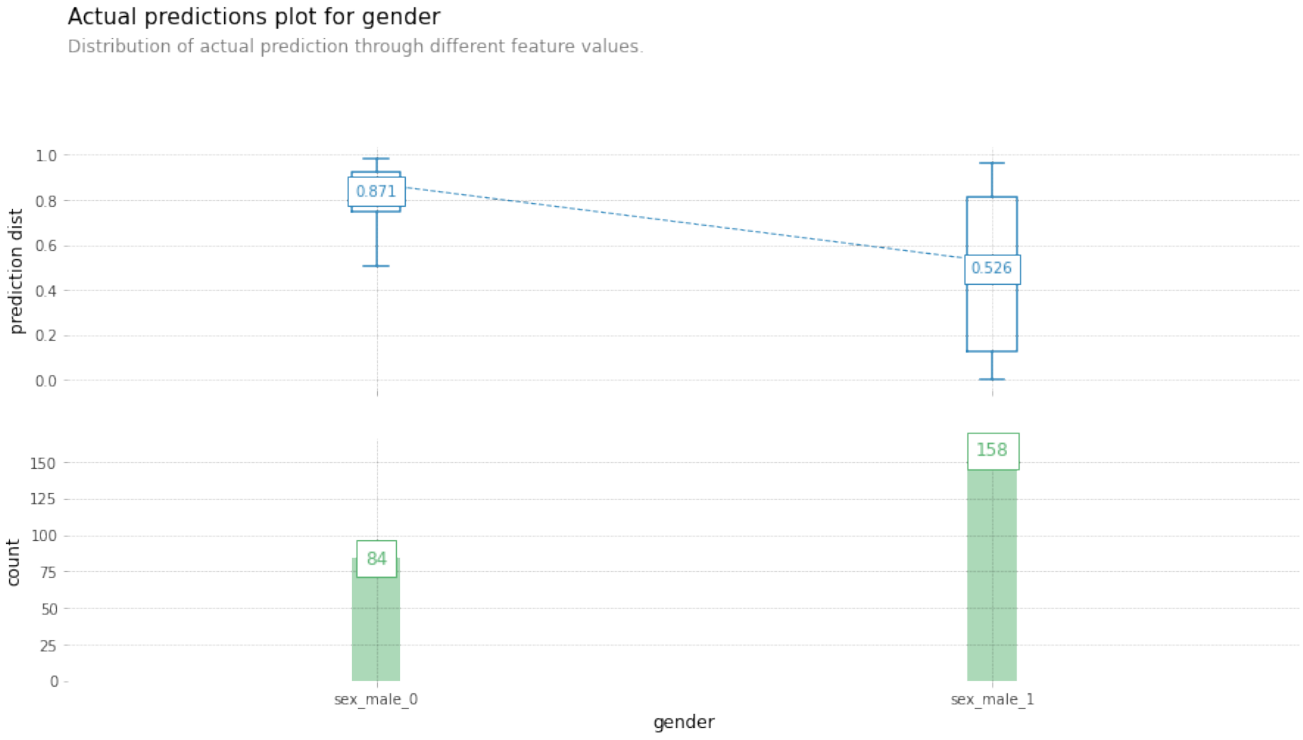


Now examine how feature 'sex' contribute to the prediction.

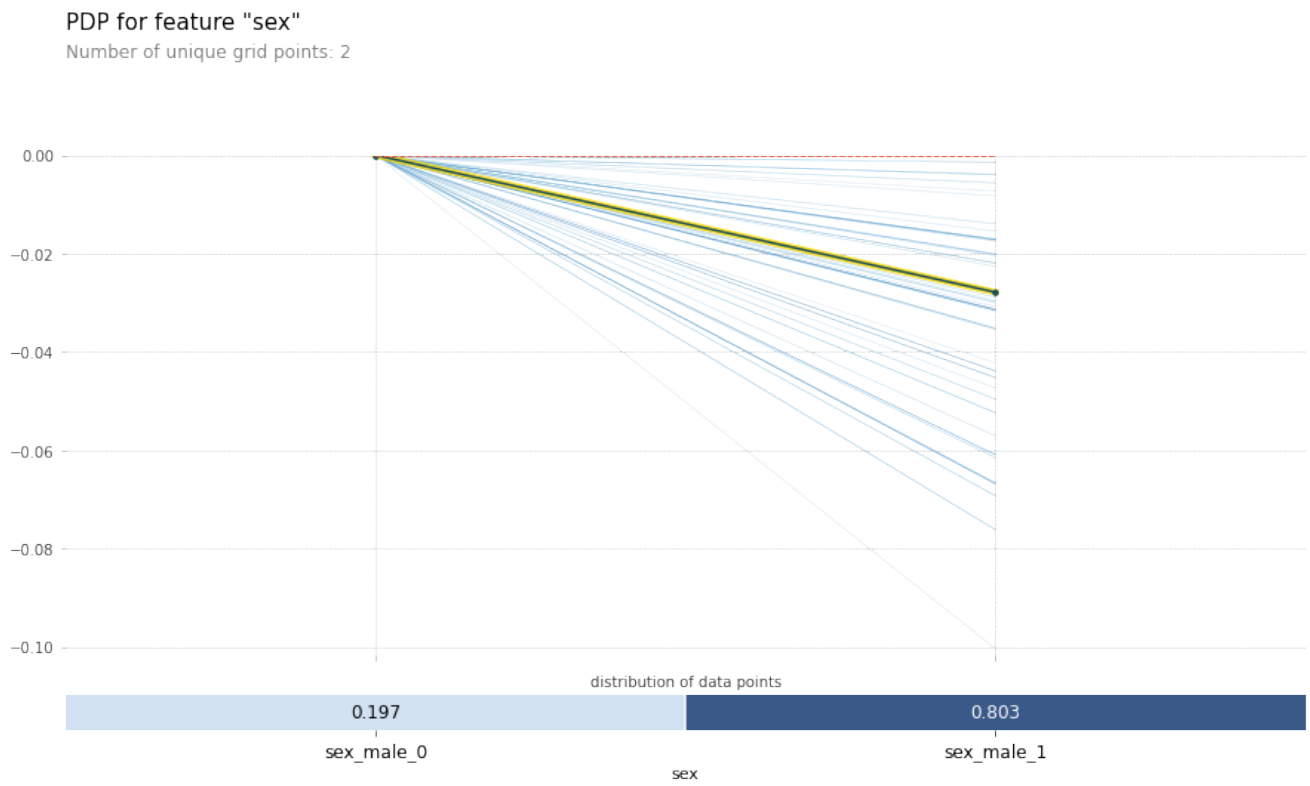
First, check the distribution for all data. It seems very clear that female have much larger potential to get heart disease. However, this is just a distribution for the data available



Now we examine the predictions based on the model. We confirm that women are more risky than men.



We can also check Partial Dependence Plot (PDP). In short, if the feature changes from female to male, the heart disease risk is dropping through different scenarios.

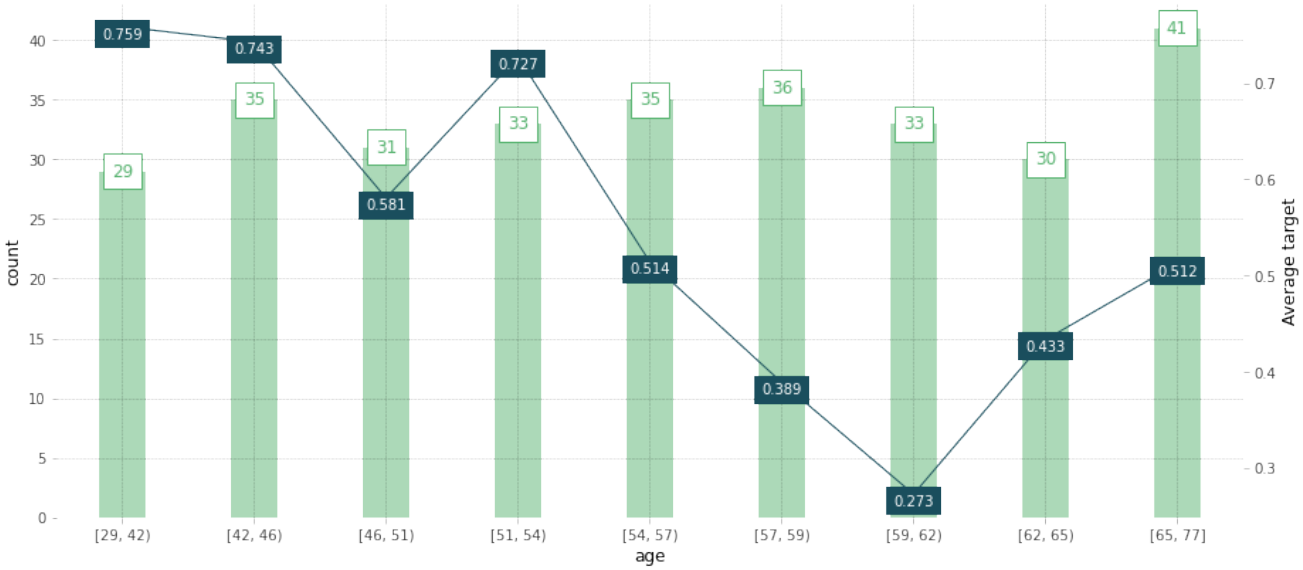


With the visualization assistance, we can know female does have more chance to get a heart disease under same conditions.

Let's check one feature with more groups - 'age'

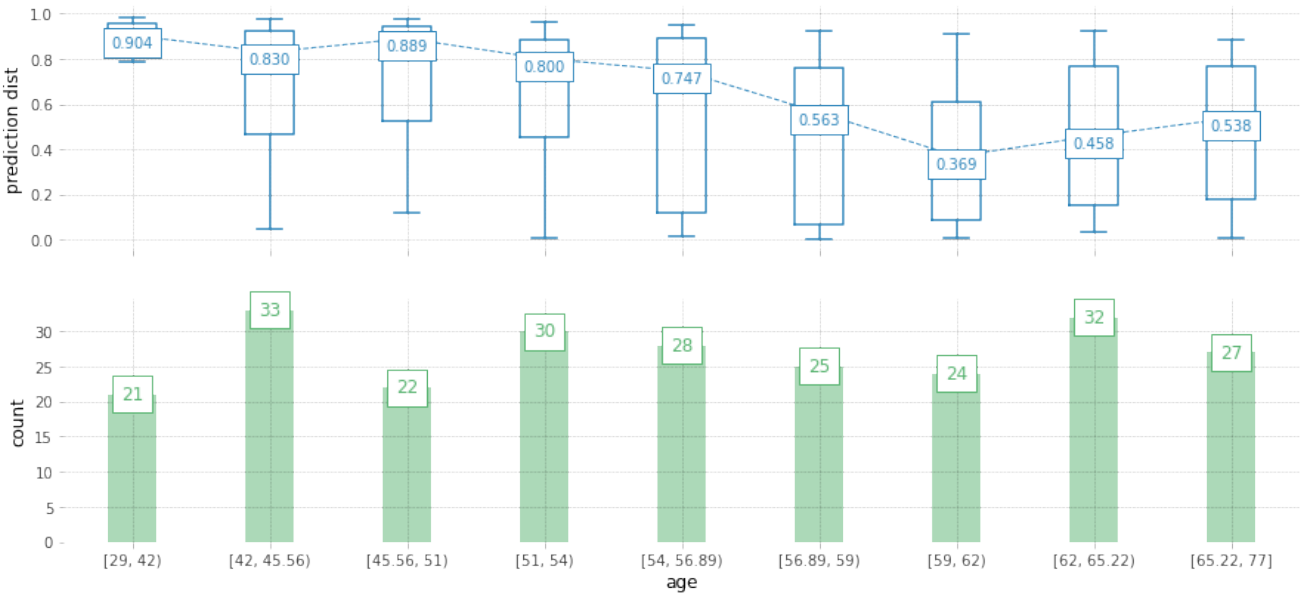
Target plot for feature "age"

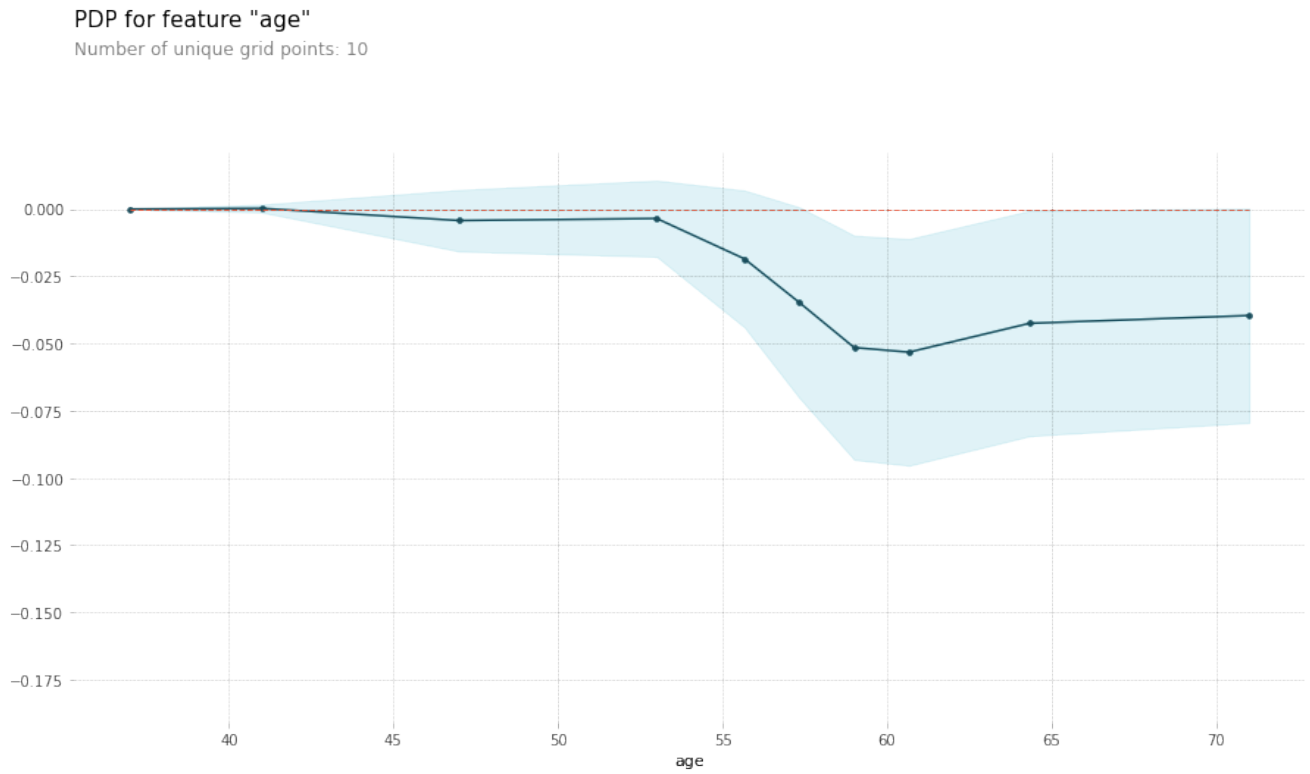
Average target value through different feature values.

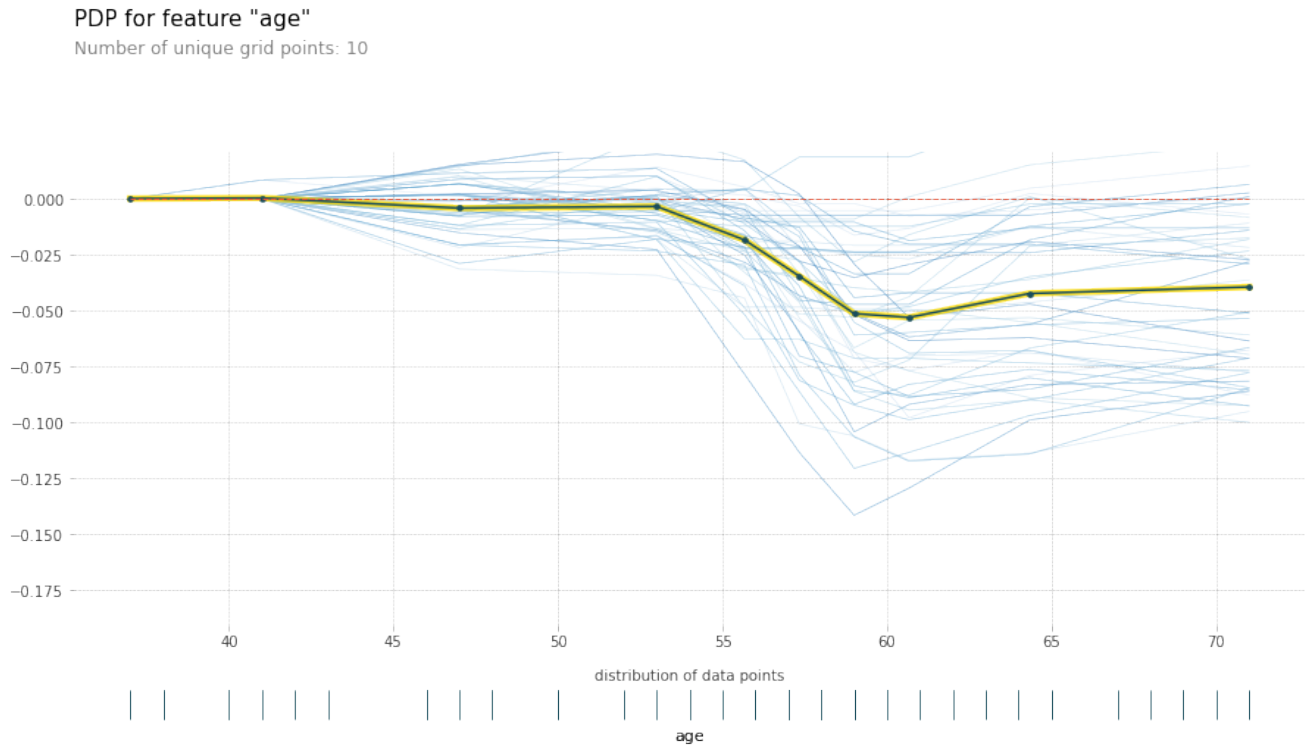


Actual predictions plot for age

Distribution of actual prediction through different feature values.

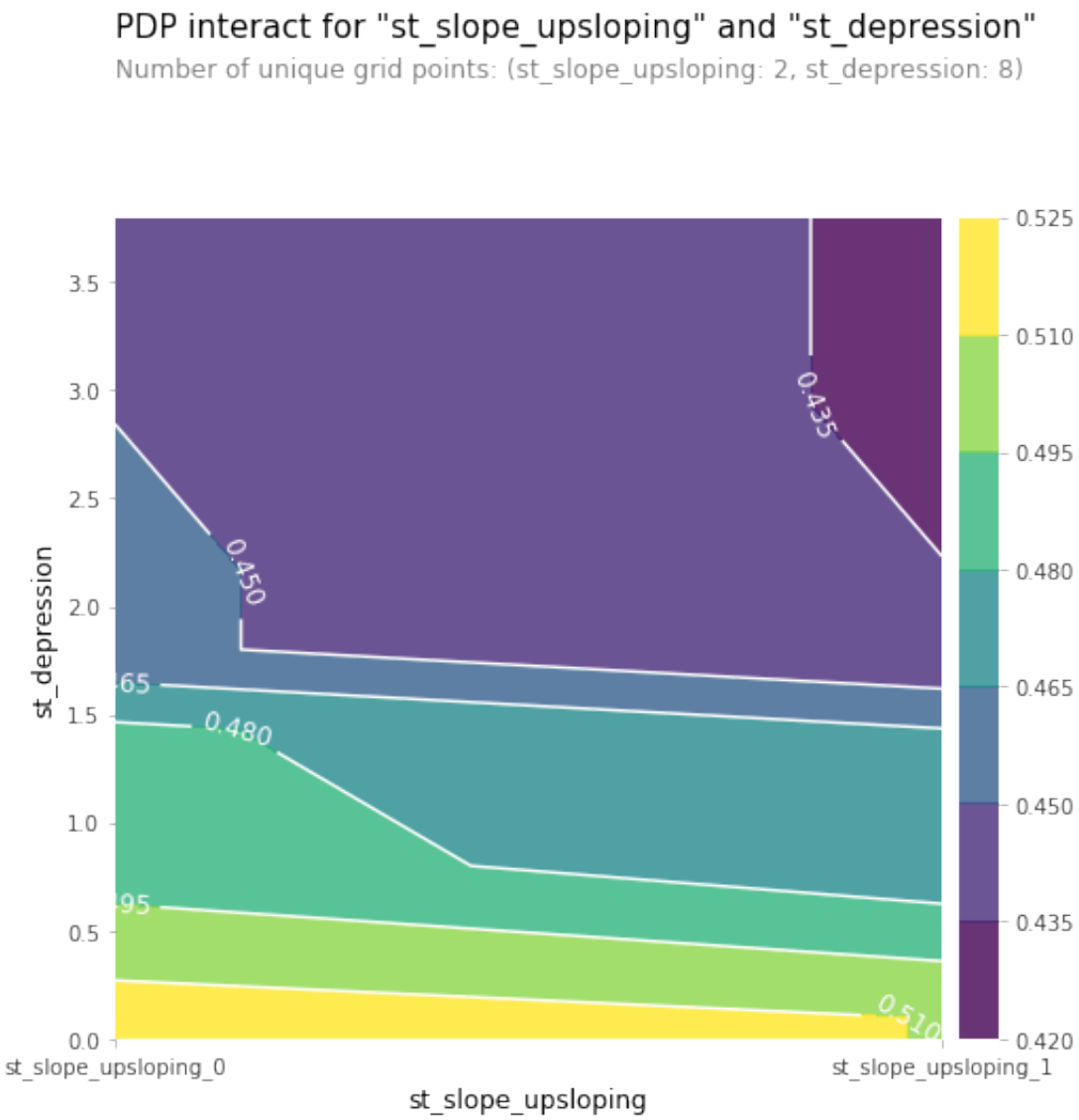


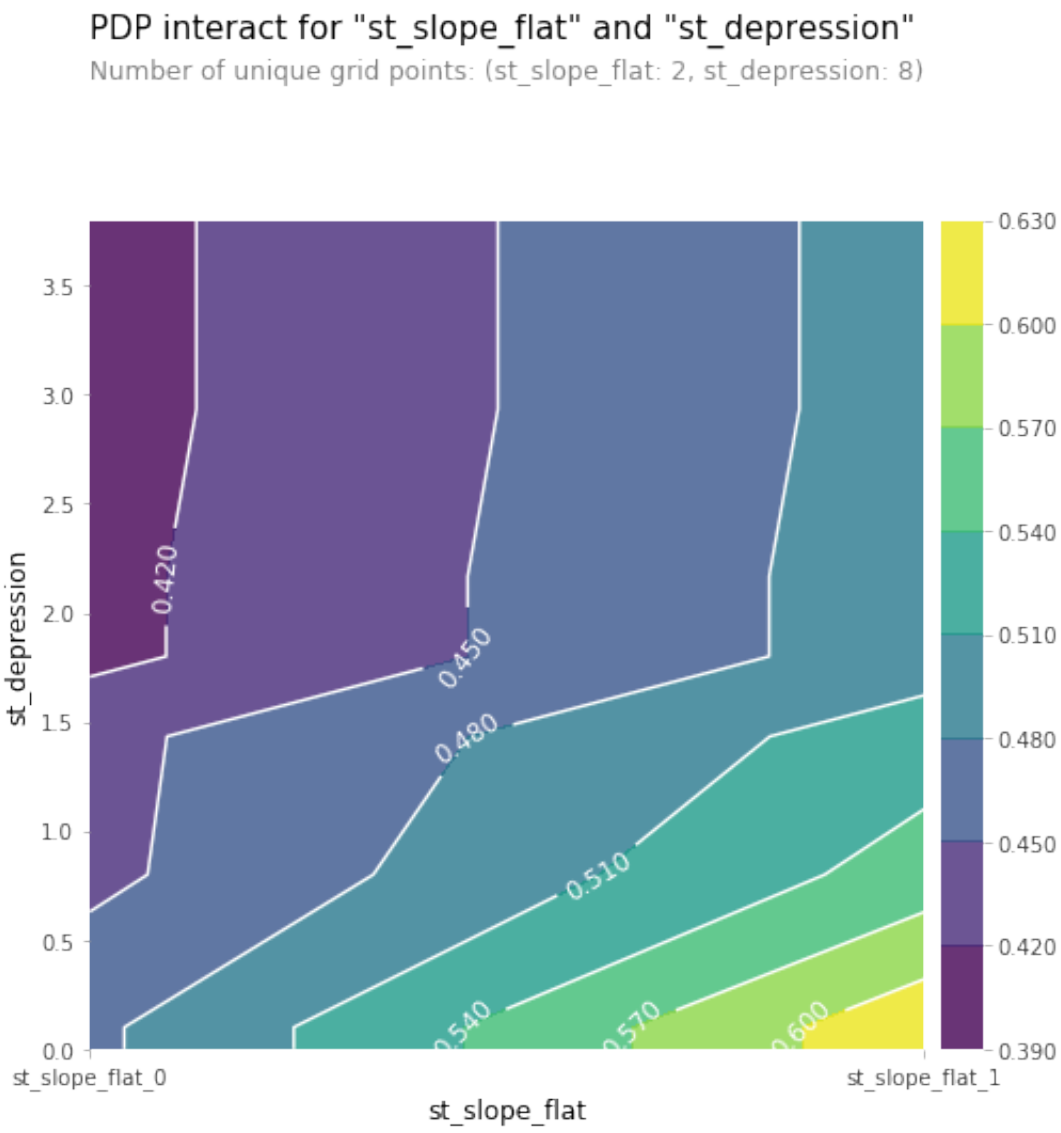




Similarly, we can yield that age from 59 to 62 are more risky for heart disease. It is actually pretty surprising, since we tend to think the older we are the more risky for heart disease. But again, this is based on a rather small dataset (303 observations).

We can also check the interactions between two features

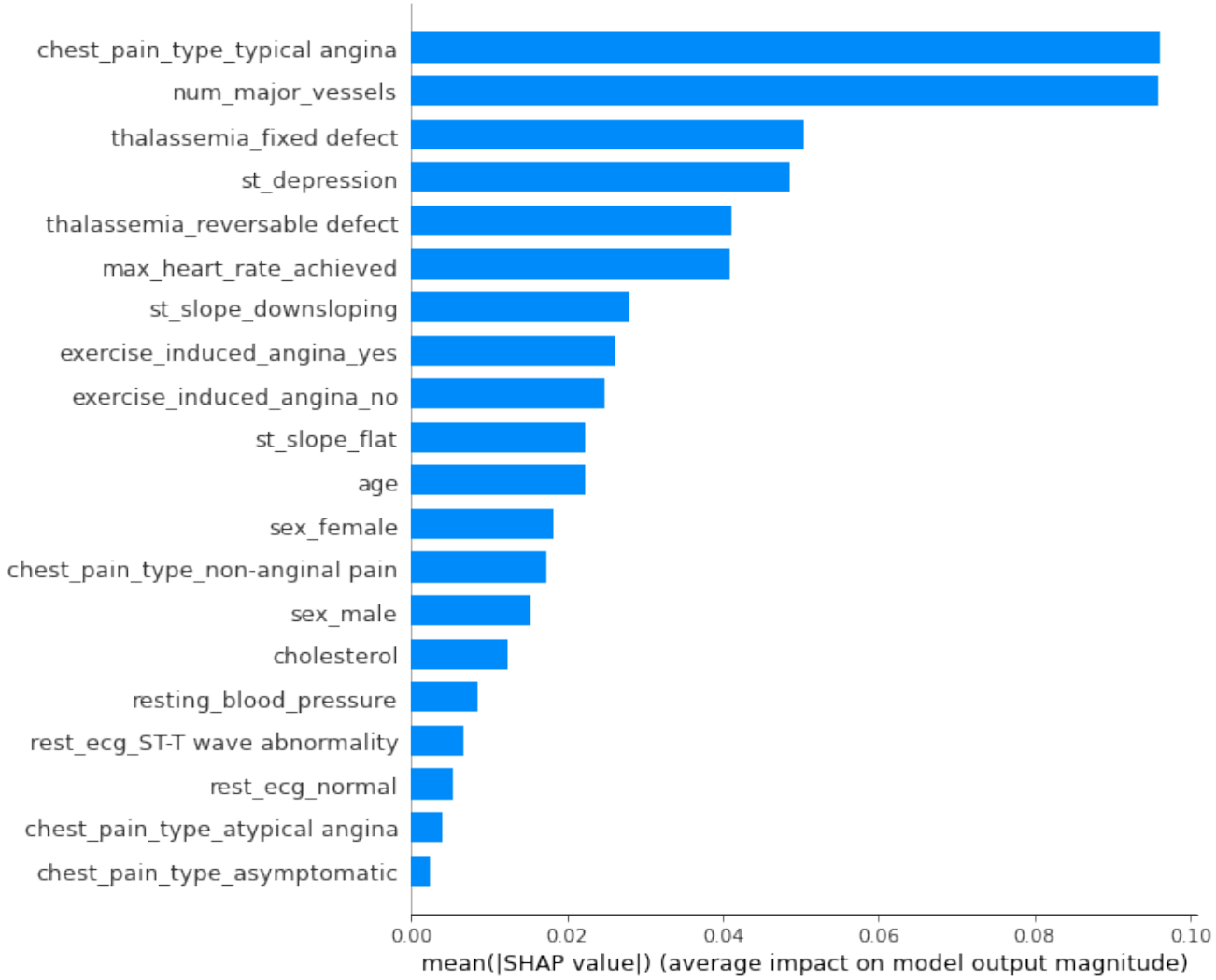




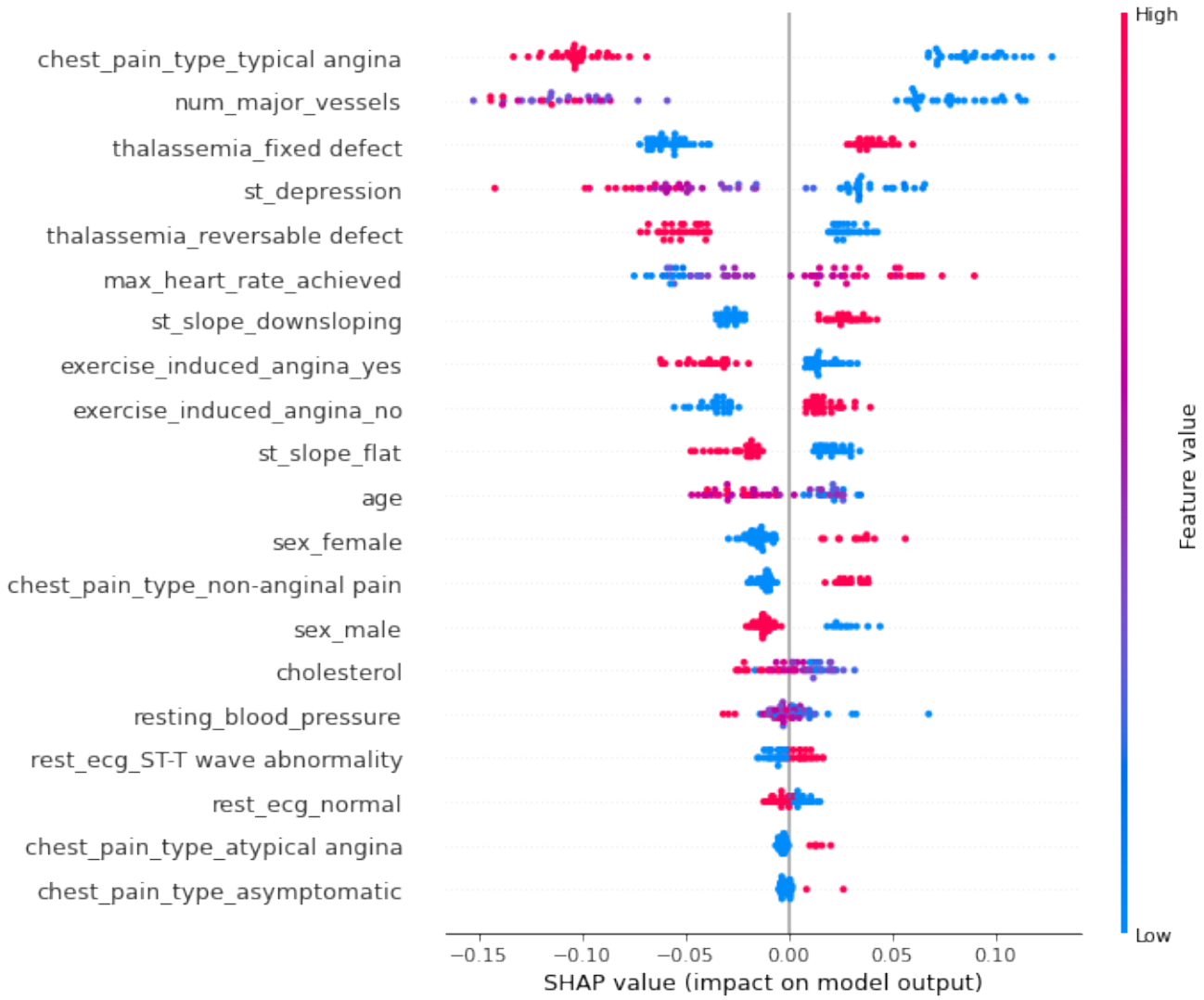
We can see how the two features interact with each other and how they contribute to the results.

Lastly, I applied SHAP explainer to open the black box

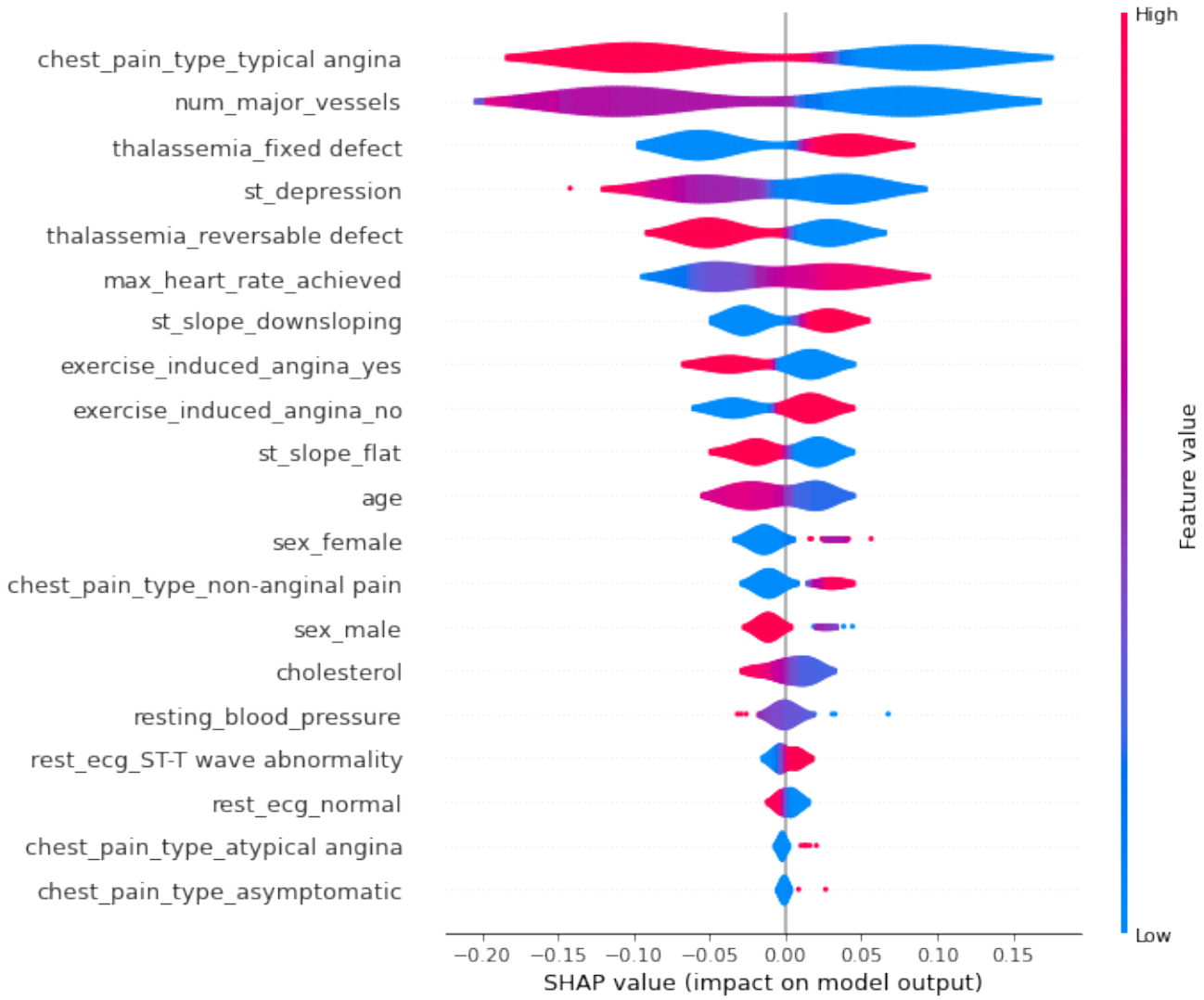
SHAP (SHapley Additive exPlanations) is a game theoretic approach to explain the output of any machine learning model. It connects optimal credit allocation with local explanations using the classic Shapley values from game theory and their related extensions (see papers for details and citations.



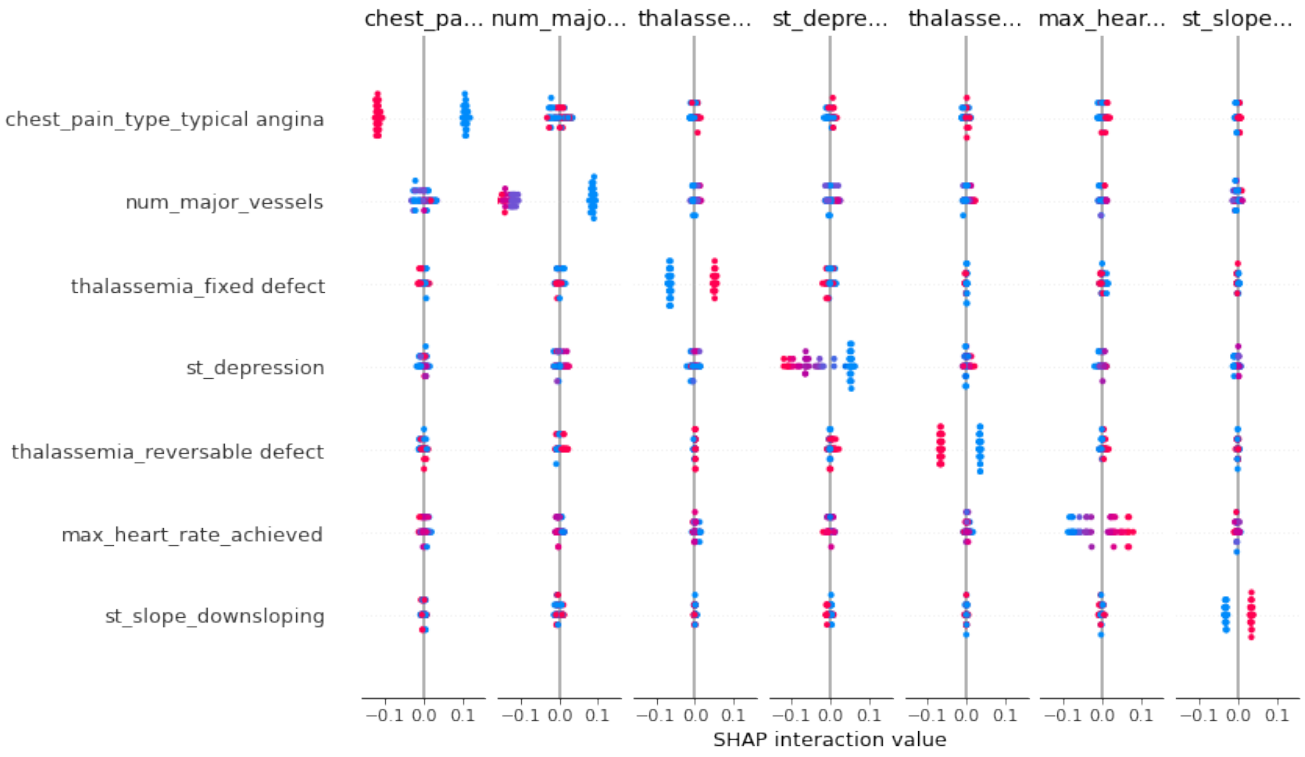
SHAP values are also used to evaluate the importance of the features. The top three features are typical angina chest pain, number of major vessels, and fixed defect thalassemia.



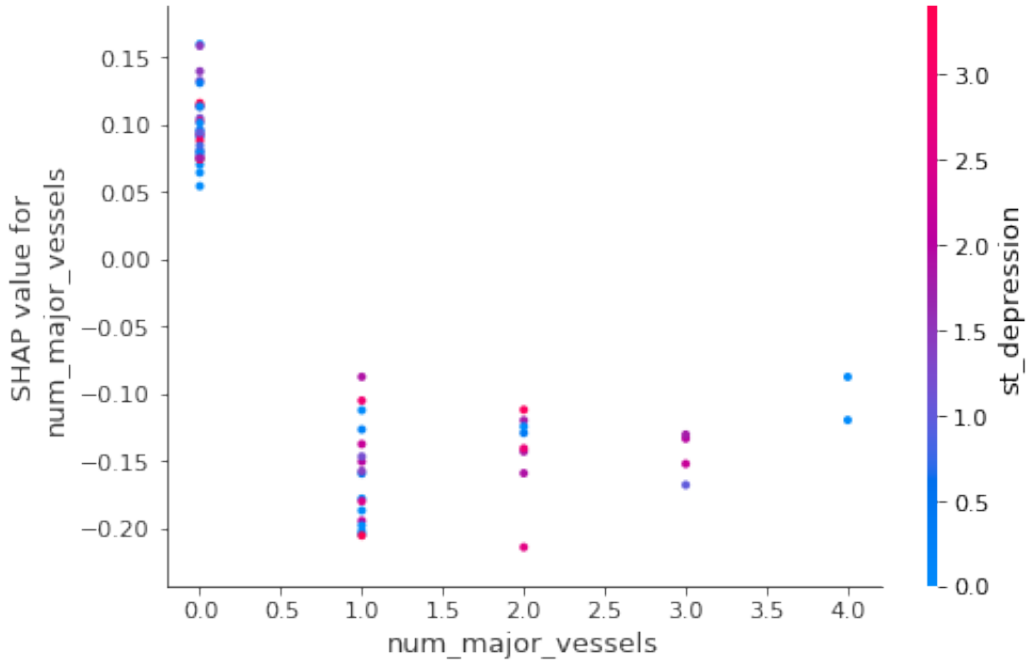
Let's focus on number of major vessels. It tells us that lower (blue) it is, the worse (1=heart disease). The blue dots are on the right. We can also plot violin figure below:



Moreover, we can see how the feature interact with each other...not a lot of insight though. So let's look at just two



Below is between number of major vessels and st_depression.



Conclusion

Althought this is a rather small dataset, we can actually create a simple model and use different machine learning explainability tools to see what is inside the black box. Especially SHAP (based on game theory) could even be applied to deep learning neuron networks. Local Interpretable Model-agnostic Explanations (*LIME*) is also used to explain the black box of computer vision. I feel very excited about explainability for machine learning and deep learning methods, which means that we could apply machine learning to more different areas such as medical, financial, and even juridical fields.

Overall, the course project give me opportunities to work on both regression (house price prediction) and classification (heart disease prediction). I focused more on applying different machine learning methods for house price prediction, and focused more on data visualization and explainability for heart disease prediction. I think these two projects give me comprehensive understanding about the classic machine learning methods. For next semester, I would appreciate if I could still audit on your CS677 deep learning class. Thanks again!