Published in Towards Data Science

Hennie de Harder Follow
Jan 4, 2020 · 7 min read ★ · Listen

Save

Interpretable Machine Learning Models

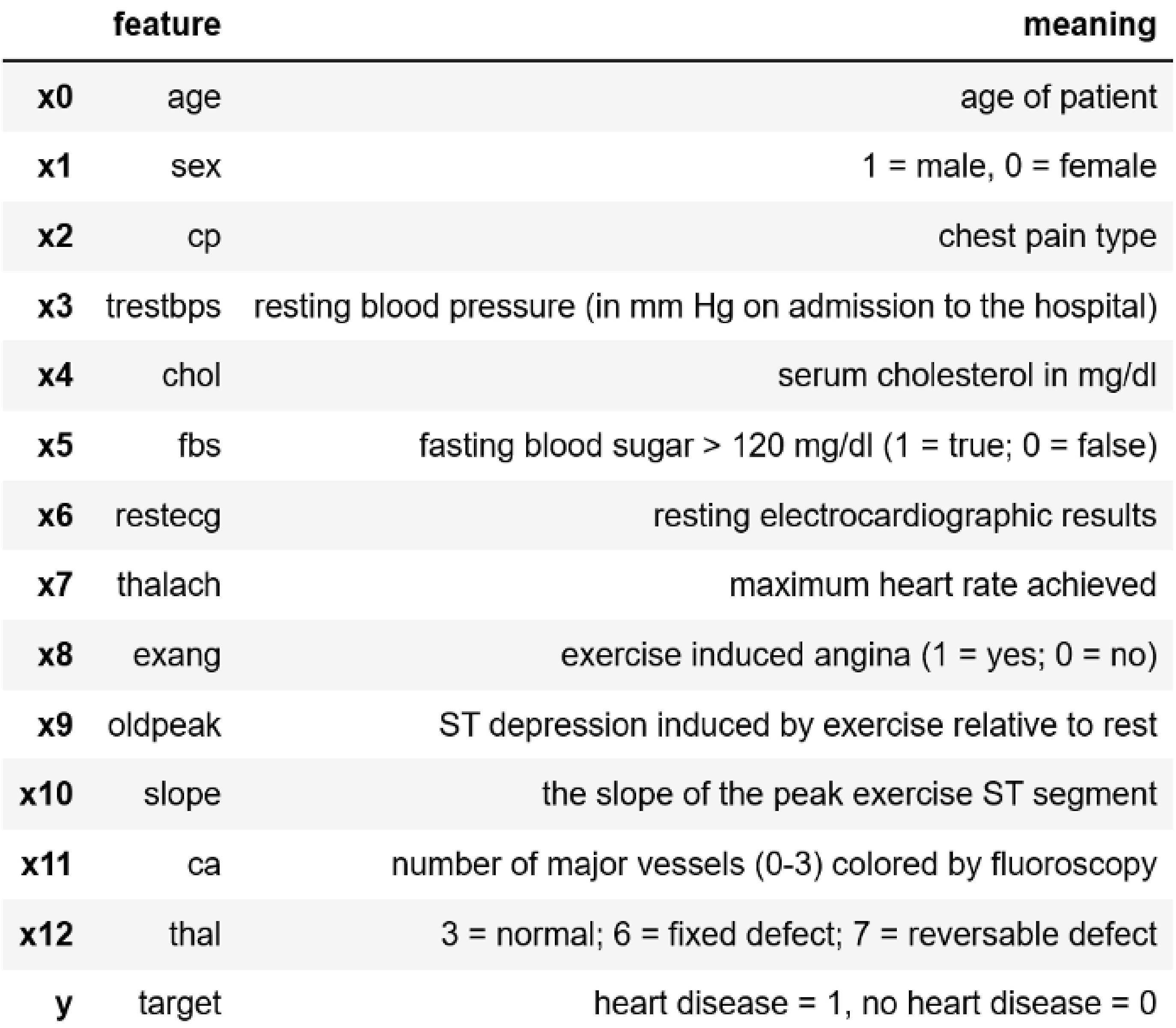
A thorough introduction to the interpretation of logistic regression models and decision trees.

A machine learning model from Amazon selected only males from a pile of resumes[1]. Another model fired teachers who were underperforming, according to the model[2]. Such models are discriminatory and can be bad for society. They can make wrong decisions and affect people's lives in a negative way. To solve this problem, you can start interpreting your models. Interpretability means that a human can understand the cause of the decision.

A lot of research has been done on the interpretability of machine learning models. There are different ways to interpret your machine learning models. The easiest split is between interpretable models and model-agnostic methods. Interpretable models are models who explain themselves, for instance from a decision tree you can easily extract decision rules. Model-agnostic methods are methods you can use for any machine learning model, from support vector machines to neural nets. In this article, the focus will be on interpretable models, like linear regression, logistic regression and decision trees. Here's a different article about model-agnostic methods.

Dataset

A field where you can use model interpretability is health care. To find out how a model decides whether or not a person has heart disease, we use a dataset from the Cleveland database with the following features:



	feature	meaning
x0	age	age of patient
x1	sex	1 = male, 0 = female
x2	cp	chest pain type
x3	trestbps	resting blood pressure (in mm Hg on admission to the hospital)
x4	chol	serum cholesterol in mg/dl
x5	fbs	fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
x6	restecg	resting electrocardiographic results
x7	thalach	maximum heart rate achieved
x8	exang	exercise induced angina (1 = yes; 0 = no)
x9	oldpeak	ST depression induced by exercise relative to rest
x10	slope	the slope of the peak exercise ST segment
x11	ca	number of major vessels (0-3) colored by fluoroscopy
x12	thal	3 = normal; 6 = fixed defect; 7 = reversable defect
y	target	heart disease = 1, no heart disease = 0

We will try to predict the target with logistic regression and a decision tree and interpret the models we have built. You can find the Heart Disease UCI dataset on Kaggle.

Code

The code for this article about interpretable models (and for the article about model-agnostic methods) can be found on GitHub.

Models you can interpret are, among others, linear regression models, logistic regression models and decision trees. So there is no need for a model-agnostic method in these cases, although you could use them too for more insights. We will build a logistic regression model and decision tree on the dataset and see how we can interpret the results, after a short note about linear regression.

Linear Regression

In linear regression you can use the weights (or coefficients) to find out which features are the most important, these are the features with the highest weight. If, for one record, you add one to a feature value while keeping the others the same the prediction will increase by the weight of the feature.

Logistic Regression

For logistic regression, the way to interpret is a bit different because a coefficient is the natural logarithm of the odds ratio. To interpret the coefficients, you need to do the transformation from the coefficient to the odds ratio. We can use the functions below:

$$\ln(a) = b$$

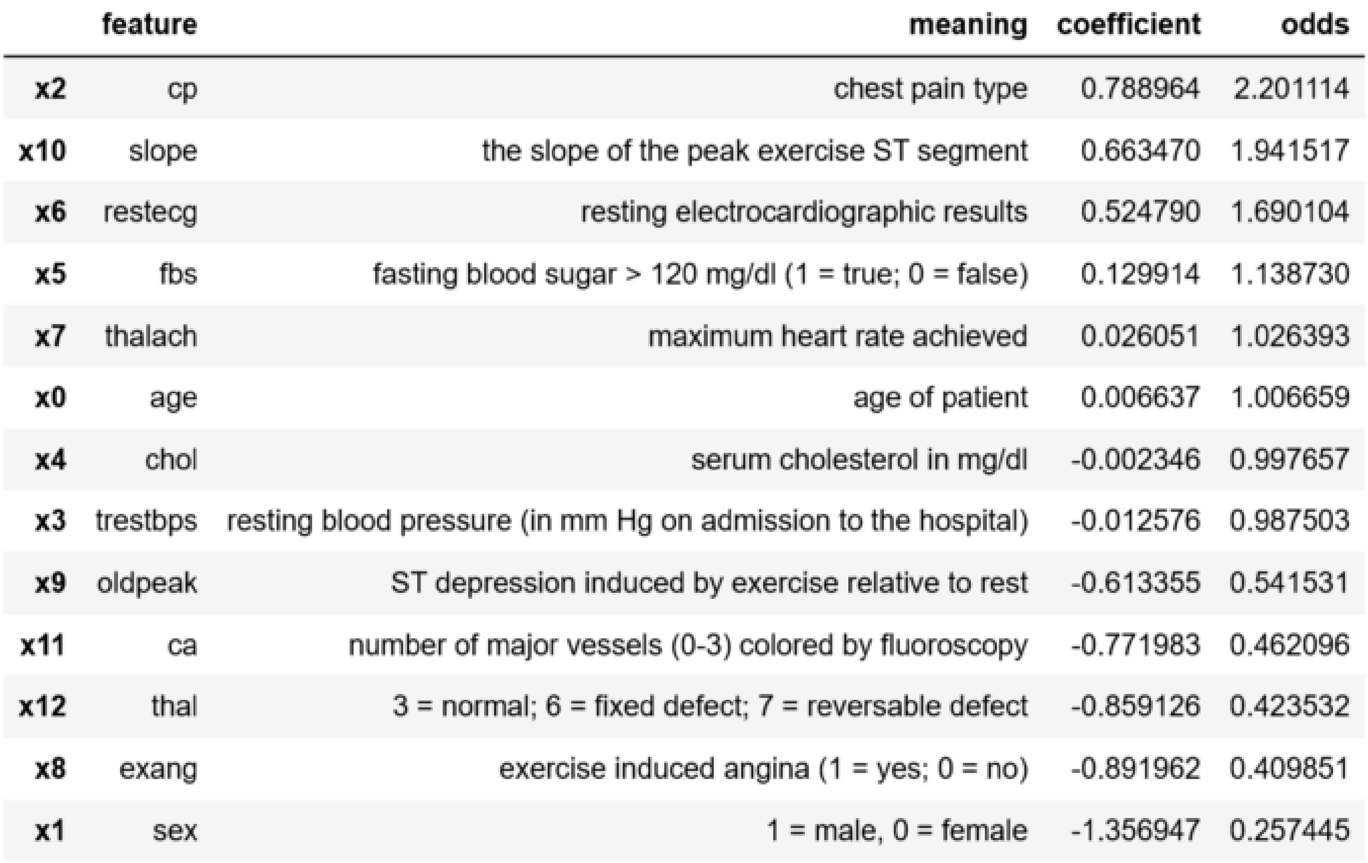
$$e^{b} = a$$

$$e^{\ln(a)} = a$$

According to the first two equations, we can replace b with $\ln(a)$, to get a. So if we raise e to the power of the natural logarithm of the odds ratio (equal to the coefficient we get from a logistic regression model), we get the odds ratio.

Great! After this transformation, it's a bit the same as linear regression, but instead of increasing it when you change one feature value it will change by a factor. An odds ratio with a value around 1 has the least impact, while a higher or lower one has more influence[3].

So let's build a logistic regression model on the heart disease dataset to understand what happens. We got the following coefficients for each feature, and we can transform them (*e* raised to the power of the coefficient) to receive the odds ratio. You can see the values in the table below, the values are sorted in descending order:



	feature	meaning	coefficient	odds
x2	cp	chest pain type	0.788964	2.201114
x10	slope	the slope of the peak exercise ST segment	0.663470	1.941517
x6	restecg	resting electrocardiographic results	0.524790	1.690104
x5	fbs	fasting blood sugar > 120 mg/dl (1 = true; 0 = false)	0.129914	1.138730
x7	thalach	maximum heart rate achieved	0.026051	1.026393
x0	age	age of patient	0.006637	1.006659
x4	chol	serum cholesterol in mg/dl	-0.002346	0.997657
x3	trestbps	resting blood pressure (in mm Hg on admission to the hospital)	-0.012576	0.987503
x9	oldpeak	ST depression induced by exercise relative to rest	-0.613355	0.541531
x11	ca	number of major vessels (0-3) colored by fluoroscopy	-0.771983	0.462096
x12	thal	3 = normal; 6 = fixed defect; 7 = reversable defect	-0.859126	0.423532
x8	exang	exercise induced angina (1 = yes; 0 = no)	-0.891962	0.409851
x1	sex	1 = male, 0 = female	-1.356947	0.257445

Now we can start interpreting the model:

- Chest pain type has a high impact. When this value increases by one, the probability having a heart disease increases by 120.1 percent.

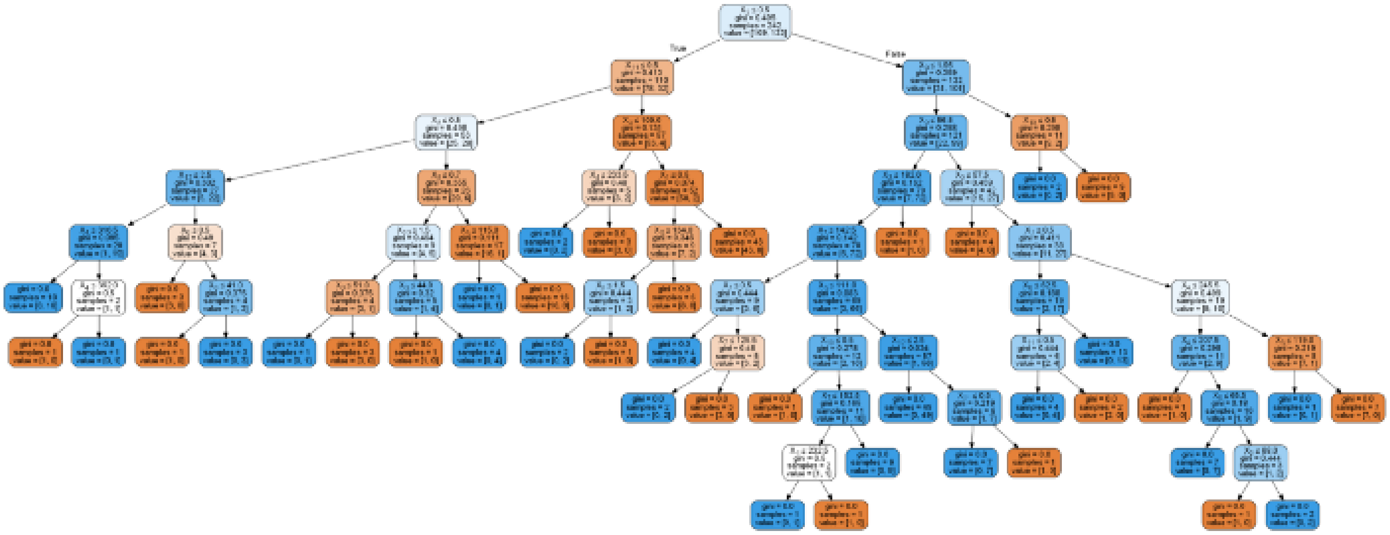
- On the bottom we see the sex variable. This odds ratio is below 1, around 1/4. This means that having a heart disease is almost four times more likely when you're female!

When results like this seems strange, the best thing to do is investigate your data. You might be wondering why age and cholesterol level have a really low impact. This model tells us that a higher cholesterol level is a bit better than a lower one. Is this the truth? The big difference between males and females also raises questions! To investigate this we could use a target plot, you can find one for the gender feature in the notebook on GitHub.

Decision Trees

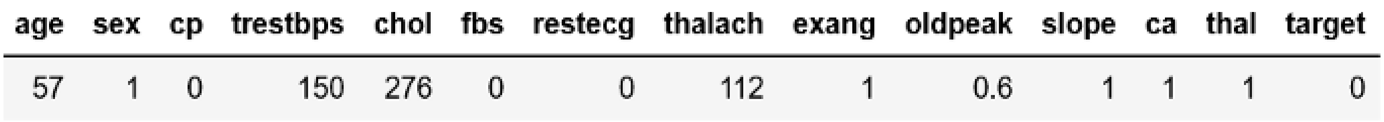
Let's see what a decision tree shows us! Decision trees are like a rule system, you start in the root node and then you follow the path for a record to a leaf node where you can see the prediction. It's harder to interpret when your tree has a higher depth, but still doable.

Below you can see an image of a decision tree, build on the heart disease dataset.



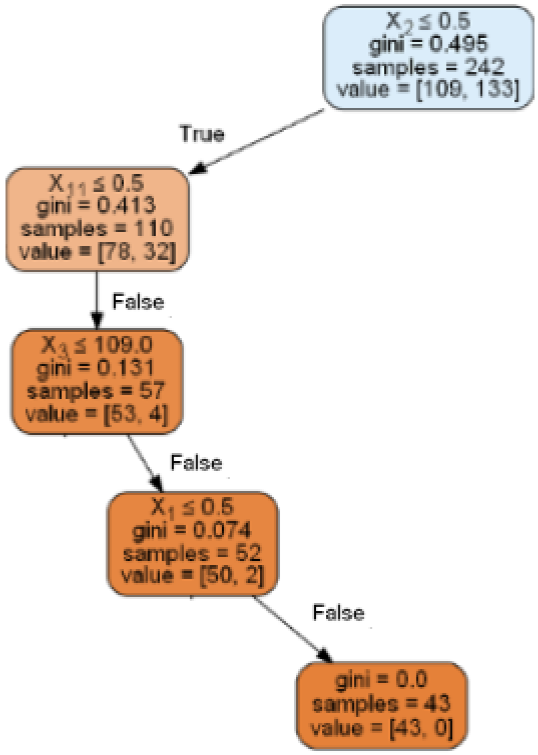
On every decision node you see four lines with text. The first line represents the decision rule. If the condition is true for a record, you follow the left branche, if it's false, you follow the right one. Gini, on the second line, is the value of the Gini Impurity, the lower this value, the better the split[4]. The third and fourth line show how many samples from

So let's take a record from the test set and follow the nodes of the tree to see what the final prediction will be. This is the new record:

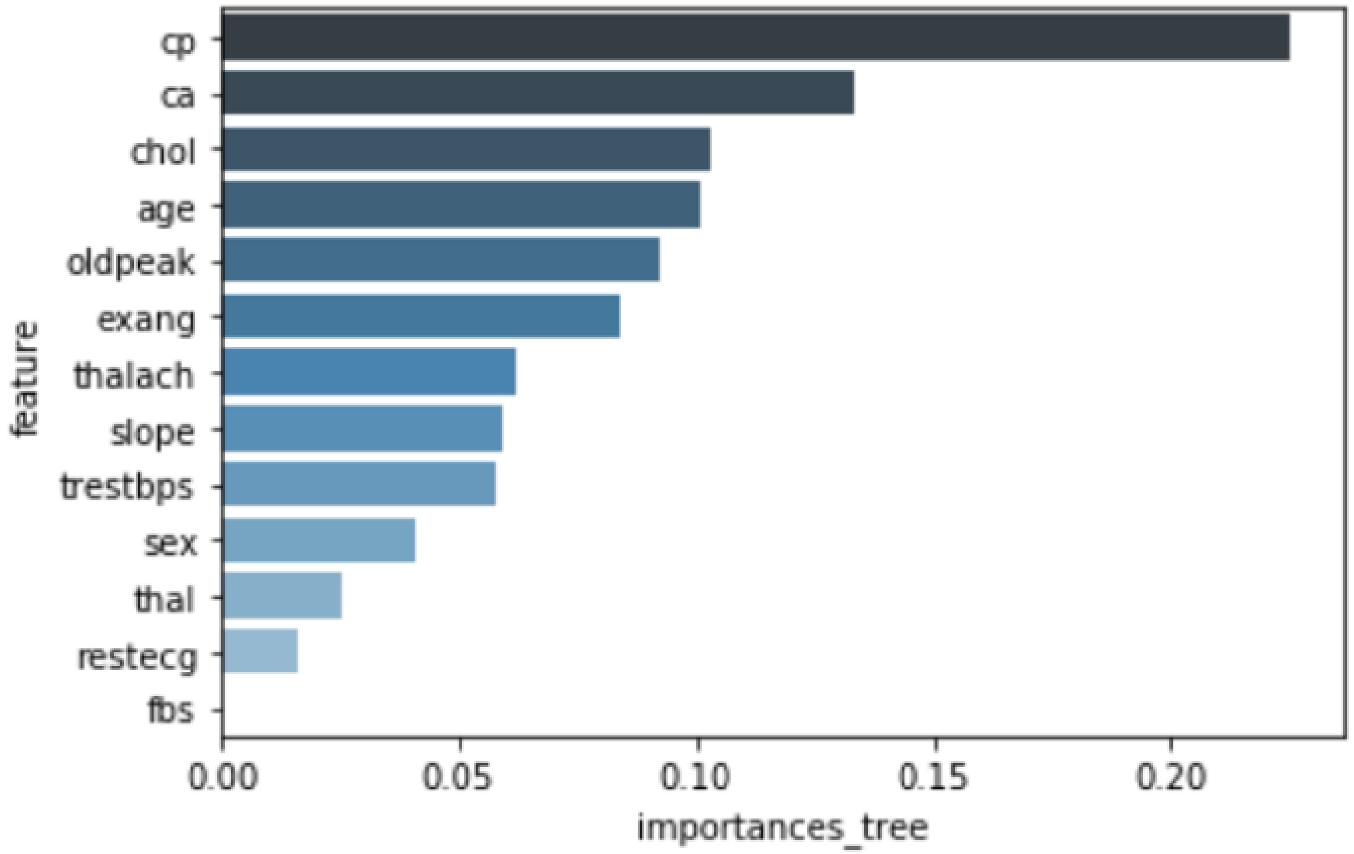


age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
57	1	0	150	276	0	0	112	1	0.6	1	1	1	0

Let's start in the root node of the decision tree. The decision rule is: feature X_2 (cp, chest pain type) smaller than or equal to 0.5. For our new record, the value is 0, so smaller than 0.5. This statement is true and we follow the left branche. The following feature is X_{11} (ca, number of major vessels colored by fluoroscopy), this value is greater than 0.5, so the statement is false and we continue to see if X_3 (trestbps, resting blood pressure) is smaller or equal to 109. This is not the case, since it is equal to 150. So false again. The next node checks if X_1 (sex) is smaller than 0.5. No, it's 1 for this record, he is male. Now we found the leaf node for this record, the prediction is 0. This prediction is correct! So for this record, the model uses only four features, cp, ca, trestbps and sex.



feature importances of trees in different ways. The easiest way is to use the feature importances of scikit-learn. The importance of a feature is computed as the (normalized) total reduction of the criterion brought by that feature. It is also known as the Gini importance.



Should you trust these feature importances? You could say that features like cp, ca, chol and age are really important in predicting a heart disease. It's possible they do, but first test your model! Keep in mind that those feature importances are based only on the training set and the current tree. The way feature importances are calculated in scikit-learn isn't the best one. It is better to use one of the model-agnostic methods, like permutation feature importance. Curious? Read about it in the next article about model-agnostic methods!

[2] C. O'Neil, Weapons Of Math Destruction (2016), Crown New York

[3] How do I interpret odds ratios in logistic regression?, UCLA: Statistical Consulting Group