Day 14 7:21 am 25th December

Dataset source : kaggle

1990 – 2021

Model → Random forest

Input → ML model → Output

ibis → interactive and beautiful

↓

variable

↓

X.head() → details in a few rows and columns

X.info() → more details about dataset

Nulls ↪ nothing is there

Switch from ibis ↪ still a new fxn to pandas

matches = matches.execute()

Why change the away team code into a numeric?

Rand forest Classifier => can pick up non-linearities

n-estimators => the higher the number, the longer it's going to take and better accuracy

Building the model

We're now ready to create our model, train it and get some predictions.

Random forest is a type of ML model that can pick up non-linearities in the data for example for our away team code, doesn't necessarily have a linear relationship, so an away team could ne number 20 but that doesn't imply that the team is better or worse than those with a number higher or lower. They are just values for different teams. A RF model can pick that up whereas a linear model can't

n_estimators, is the number of individual decision trees we want to train. A random forest is a series of decision trees but each decision tree has slightky different parameters. The higher this number is the longer it will take for the algorithm to run but potentially the more accurate it will be.

min_sample_split, is the number of samples we want to have in a decision tree before splitting to a different node. The higher this is the less likely we are to overfit but the lower the accuracy could potentially be.

random_state, just ensures that when you are using the same data you get the same result back.

predictors are parameters that will influence the final decision

Question

- Why numerical values for args. learn? What will be the problem if we still used the int?
- Apart from random forest classifier what other ML model could have been used for the problem.?
- Is there an ideal range for the n-estimator

Barbara Aboagye 45 seconds ago (edited)

Hi Marlene,

Beautiful video and analysis. I loved it and thanks. I am new to machine learning so this was refreshing to see and also from a WOC. I have a few questions

1. Why did you decide to change the type of the away team to a numberic value? You did give a reason but I want to understand the concept behind it to allow me to make future decisions. We could have still used the string type and it will work right?

1b. I believe changing the type influenced the type of ML model you used that is the random forrest classifier, if it was still a string would you still use the same and why? if not what other ML models would have been suitable

2. What other ML model could be use for this type of problem?

3. Is there an ideal range for the choice of n - estimator?

Thanks and you bagged another subscriber ! 🙂

Marlene Mhangami 5 hours ago 399 subscribers

Hi Barbara,

1. For models built using scikit learn they usually and in some case only work with numerical values. For some models outside of scikit learn this is not the case, but I think for best practices it's best to try to convert to numbers wherever you can. This will usually also have a positive effect on the performance of your model.

2. I've seen other people using Logistic Regression for this but the process for this looked longer in terms of isolating the data etc.

3. One way I've read of doing this is to use something like GridSearchCV (read on this here https://www.datasciencelearner.com/how-to-choose-n_estimators-in-random-forest/). In general I think 50 could be a good place to start and then if you have time experiment with reducing or increasing it. ML is largely experimental, and things that work for one problem might not work for another.

Hope that helps! Thank you for subscribing and for the thoughtful questions! Good luck with your journey <3