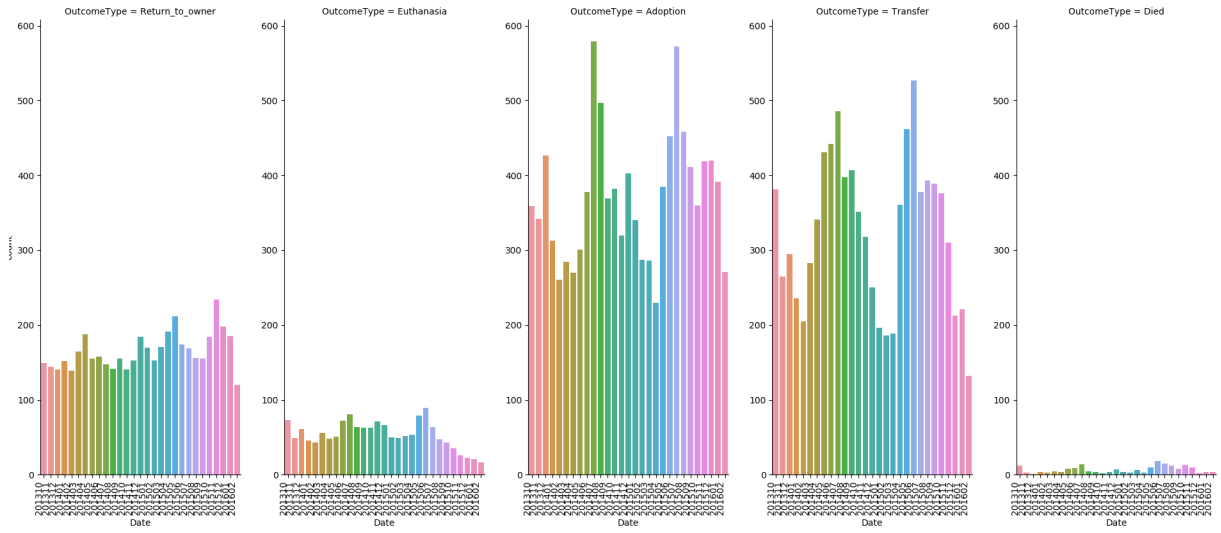
Best practice for encoding datetime in machine learning

stats.stackexchange.com/questions/311494/best-practice-for-encoding-datetime-in-machine-learning

asked Nov 2 '17 at 16:19 cck3 18 1 3

2

I'm working on a Kaggle problem (the problem closed some time ago, but doing it for self-study/practice) where the output is clearly affected by both the year and the month.



The original datetime data provides year/month/day/hour information and I felt that year and month were probably the only necessary data. So I've currently modified the feature such that the data is represented only by year and month (ex) March of 2016 would be 201603) and graphed each outcome with respect to the modified time variable consisting of year/month pair.

As you can see here, the 1st outcome has some minor seasonal fluctuations whereas the 3rd and 4th outcomes have clear seasonal trends. On the other hand, the 2nd outcome drastically decreases after May of 2015 (201505).

For my model prediction, I'd like to somehow incorporate time as a variable in a way that makes sense. What would be the best approach here? Can I just assume the earliest time period in the data to equal to 1 and increment by 1 for every month and treat the variable as a nominal category variable? Or something else?

Thanks

5

You want to preserve the cyclical nature of your inputs. One approach is to cut the datetime variable into four variables: year, month, day, and hour. Then, decompose each of these (*except* for year) variables in two.

You create a sine and a cosine facet of each of these three variables (i.e., month, day, hour), which will retain the fact that hour 24 is closer to hour 0 than to hour 21, and that month 12 is closer to month 1 than to month 10.

A quick Google search got me a few links on how to do it:

Your Answer

Post as a guest

Required, but never shown