Time Series Prediction Task

TOP DATA SCIENCE

CONFIDENTIAL

September, 2018

Abstract

This document is confidential and contains a predictive analysis task. This task is specifically created *only* for recruitment purposes. For questions, please contact `oguzhan.gencoglu@topdatascience.com`.

1 Data

The data includes weekly occurrence counts of 11 keywords related to a disease in an online discussion forum. The goal is to predict the official weekly disease counts. The name of the disease and keywords are not relevant, thus are not given.

The data is divided into training and test sets.

Training set has 282 rows (weeks), 13 columns. First column is the date of the start of that particular week, next 11 columns are the keyword counts during that week and the last column is the target which is the official weekly count of number of people having that disease in that week.

The test set (52 rows) structure is the same, except that it does not have the last column, i.e., target values.

2 Task

The task is to create a machine learning model, trained on the training set, to predict the weekly disease counts on the test set. There are several approaches to this task, feel free to be creative.

3 Deliverables

- A jupyter notebook (preferably Python 3) combining the script (code) and your reporting together

The predicted values should be assigned to a variable (1D numpy array or python list of length 52), *predictions*, in your implementation.