Job Salary Prediction

Predict the salary of any UK job ad based on its contents

$6,000 · 289 teams · 5 years ago

Competition Data

Edit

- Location_Tree.csv
- mean_benchmark.csv
- random_forest_benchm...
- test.csv
- Valid_rev1.csv
- Test_rev1.7z
- Test_rev1.zip
- Train_rev1.7z
- Train_rev1.zip
- Valid_rev1.7z

Location_Tree.csv 1.6 MB Download

Data Description

Code to create the benchmarks is available on Github

Job Data

The main dataset consists of a large number of rows representing individual job ads, and a series of fields about each job ad. These fields are as follows:

Id - A unique identifier for each job ad

Title - A freetext field supplied to us by the job advertiser as the Title of the job ad. Normally this is a summary of the job title or role.

FullDescription - The full text of the job ad as provided by the job advertiser. Where you see ***s, we have stripped values from the description in order to ensure that no salary information appears within the descriptions. There may be some collateral damage here where we have also removed other numerics.

LocationRaw - The freetext location as provided by the job advertiser.

LocationNormalized - Adzuna's normalised location from within our own location tree, interpreted by us based on the raw location. Our normaliser is not perfect!

ContractType - full_time or part_time, interpreted by Adzuna from description or a specific additional field we received from the advertiser.

ContractTime - permanent or contract, interpreted by Adzuna from description or a specific additional field we received from the advertiser.

Company - the name of the employer as supplied to us by the job advertiser.

Category - which of 30 standard job categories this ad fits into, inferred in a very messy way based on the source the ad came from. We know there is a lot of noise and error in this field.

SalaryRaw - the freetext salary field we received in the job advert from the advertiser.

SalaryNormalised - the annualised salary interpreted by Adzuna from the raw salary. Note that this is always a single value based on the midpoint of any range found in the raw salary. This is the value we are trying to predict.

SourceName - the name of the website or advertiser from whom we received the job advert.

All of the data is real, live data used in job ads so is clearly subject to lots of real world noise, including but not limited to: ads that are not UK based, salaries that are incorrectly stated, fields that are incorrectly normalised and duplicate adverts.

Location Tree

This is a supplemental data set that describes the hierarchical relationship between the different Normalised Locations shown in the job data. It it is likely that there are meaningful relationships between the salaries of jobs in a similar geographical area, for example average salaries in London and the South East are higher than in the rest of the UK.

© 2017 Kaggle Inc