Predicting the likelihood of a posting being filled

To build a model that can be used to predict the likelihood of a posted job being filled, I made sure that the data samples used represent the different classes of status in the data as shown below and then followed by data treatment and resampling;

```
data_new = data_new.replace({'completed': '1', 'filled': '2', 'withdrawn': '3', 'expired': 'na', 'system_invalidated':'na'})
```

I then dropped rows that have na as status. These were postings that expired or system_invalidated

```
data = data_new.dropna()
```

The data was split into train and validate sets. A snippet of the code training a model on the data is shown below;

```
[1484]  train-mlogloss:0.108553 eval-mlogloss:0.183060

[1485]  train-mlogloss:0.108507 eval-mlogloss:0.183018

[1486]  train-mlogloss:0.108475 eval-mlogloss:0.182999

[1487]  train-mlogloss:0.108394 eval-mlogloss:0.182992

[1488]  train-mlogloss:0.108382 eval-mlogloss:0.182996

[1489]  train-mlogloss:0.108358 eval-mlogloss:0.182992

[1490]  train-mlogloss:0.108340 eval-mlogloss:0.183002

Stopping. Best iteration:

[1390]  train-mlogloss:0.111835 eval-mlogloss:0.182929
```

The classification performance of the model was evaluated using a classification report as shown below;

	precision	recall	f1-score	support
1.0	0.96	0.99	0.98	345
2.0	0.94	0.97	0.96	345
3.0	0.97	0.89	0.92	290
avg / total	0.96	0.96	0.95	980

Precision measures the percentage of classified filled jobs (class 2) that are actually filled jobs while recall measures the percentage of filled jobs identified from all possible filled jobs. The precision and recall for filled jobs are 94% and 97% respectively.

Why use xgb and not any other algorithm? Well, It performed better than the other algorithms I researched, the classification of tp, tn, fp and fn is the best. Xgb also offers better computational speed than the others.

Some recommendations on how to improve the likelihood of a session being filled

The analysis of the features that were assigned the most importance in constructing the model shows that Network Locum could improve the likelihood of a job being filled by;

1. Improving staff planning with the different CCGs or practice. My model gave the highest importance to the date a job was posted ('date_posted': 19106) and start date of the job ('date_started': 10722)

2. Where the job is available was also assigned a high importance ('practice_id': 8735). Therefore, Network Locum may prioritize their resources on achieving the first recommendation with certain practices while improving the collection of relevant data on Locums to better understand why they prefer certain practices