Yinan Guo

guoy221@wfu.edu

2021/11/07

# Executive Summary

## Problem

This task is to help online job posting service detect fake job postings and prevent identity theft. We have previous job postings with labels of whether they are fake or not. I built a random forest model and an xgboost model to learn from past postings and to predict new postings.

## Key Findings

1. There are 157 out of 1979 (7.93%) of postings with salary range which is fraud while 460 out of 10537 (4.37%) of postings without salary range which is fraud. Thus, although whether or not the company provide a salary range is not most important factor to help decide whether the job posting is fake since it is not among the top 20 important factors for either random forest model or xgboost model, it shows that postings does provide salary range will be about 1 time more likely to be fraud.
2. The presence of a logo is an important indicator for whether or not it is a fake posting since it has been ranked 1, 8 in random forest and xgboost. 209 out of 9939 (2.10%) of postings with company logo which is fraud while 408 out of 2577 (15.8%) of postings without company logo which is fraud, which is 6 times higher than those with logos.
3. There are 184 out of 6168 (2.98%) of postings has questions which is fraud while 433 out of 6348 (6.82%) of postings without questions which is fraud. Thus, although whether or not the company has a question is not most important factor to help decide whether the job posting is fake since it is not among the top 20 important factors for either random forest model or xgboost model, it shows that postings does not have questions will be about 1 time more likely to be fraud.

## Model Performance Summary & Interpretation

1. My final model is a random forest model with 127 trees, min\_n of 9 and tree depth of 5 and gives out a final AUC of 0.9757632. Base on AUC score, we can say our model can identify right from wrong 97.6% of the time.
2. Company profile, whether or not has company logo, description, industry and job function are top 5 important factors to the final model.

|  |  |  |
| --- | --- | --- |
|  | Random Forest | XGBoost |
| AUC on test set | 0.9757632. | 0.9295655 |

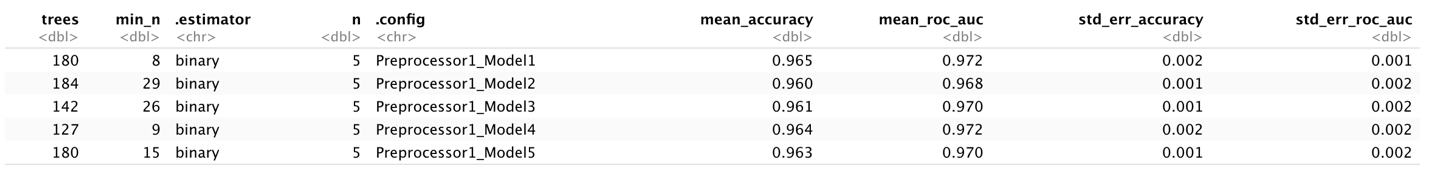
## Recommendations

1. Online job posting service websites can require real companies to upload their logo to be shown on their profile page to ensure the authenticity of the information posted.
2. Since company profile is an important indicator to tell fake job postings, the online job posting service websites can require real companies to upload detailed company description to ensure the authenticity of the information posted. What’s more, the websites can step up the review of the company's description.
3. Since job description is an important indicator to tell fake job postings, the online job posting service websites can increase requirements for description and information release, like to restrict filling form for job descriptions, etc.

# Detailed Analysis

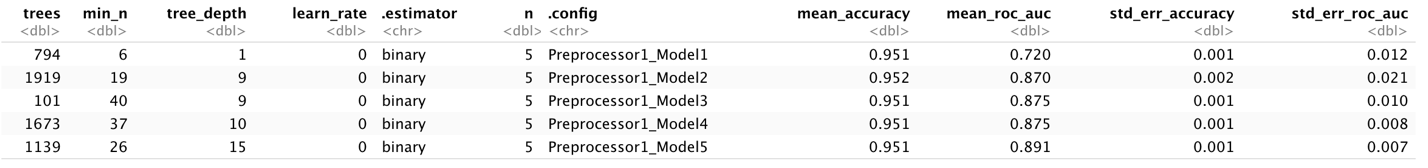
## Comparing tuning performance of RF and XGB models

* Random Forest



Tune grid randomly selected tree numbers between 100 and 200, min\_n between 5 and 30 and generated 5 models with different tunning parameters. The best random forest model is model 4, which gives out a mean\_accuracy of 0.964 and a mean\_roc\_auc of 0.972.

* XGBoost



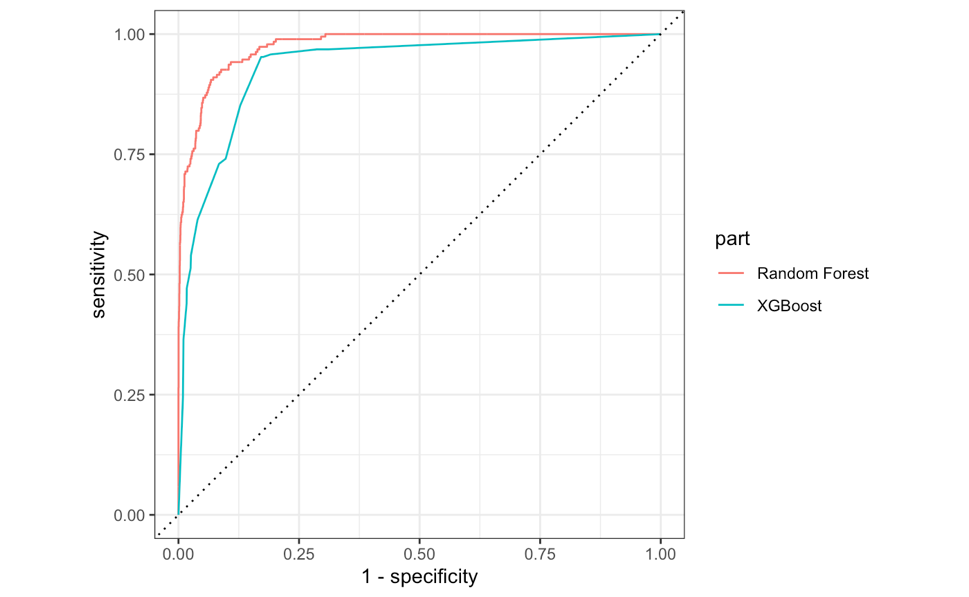
Tune grid randomly selected tree numbers between 100 and 2000, min\_n between 5 and 50, tree\_depth between 1 and 20, learn\_rate between 0 and 1, and generated 5 models with different tunning parameters. The best XGBoost model is model 5, which gives out a mean\_accuracy of 0.951 and a mean\_roc\_auc of 0.891.

## Comparing the final fit metrics of RF and XGB model on train and test

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Random forest | | XGBoost | |
| train | test | train | test |
| Accuracy | 0.99623331 | 0.9685752 | 0.9525168 | 0.9496671 |
| AUC | 1.00000000 | 0.9757632 | 0.9164312 | 0.9295655 |

The random forest model gives out higher accuracy and AUC for both train and test dataset. And the difference in accuracy and auc between train and test set are acceptable, which doesn’t show serious overfitting problem.

## Comparing the ROC curve of RF and XGB model on Test set.



According to the ROC, random forest has much better performance than XGBoost.