Project 3 – is it Fake?

# Fake Job Postings

You now work for an online job posting service like Indeed, Monster or LinkedIn and the service has been hit with a number of fake job postings. There are a number of reasons why fake job postings exist. Scammers post fake jobs to commit some form of identity theft. Companies have a number of reasons to post fake positions one is to gauge the current talent pool another is to build up a list of potential candidates for another position. It’s annoying and since we are deep diving into machine learning we aren’t going to take it. The firm believes that the presence of a logo, a salary range and if they have questions is an indicator of a legit job posting. This week your job is to identify fake job postings using what we’ve learned. Note this will be a challenging task (though actually easy to model) but you’ll need to break it up into steps and be diligent in your coding, and like always there is a bar to meet on Kaggle.

You have been provided two datasets

* Job\_training.csv – use this one to train your model
* Job\_holdout.csv – use this one to score your hold out for Kaggle.

## Tasks

Unlike our previous projects I want you to focus on your executive summary this week, your executive summary is worth 35 points and I’m asking for a simple model summary (worth 10) includes tuning, final performance, and a roc chart – see below.

### Recipe Feature Processing

* Use text stack to process job description and any other variables you think can benefit
  1. What does min\_times do?

step\_tokenize(description) %>%

step\_stopwords(description) %>%

step\_tokenfilter(description, min\_times = 20) %>%

step\_tf(description)

or use step\_tfidf(description)

#### *Free Advice*

*Prep your recipe and check to see what it is doing by Baking it on a small sample of data – you want to do this to make sure that your recipe is producing what is expected, check the number of columns the recipe produces – is it creating thousands of features does it even work?*

### Create two models and tune them.

To meet expectations, you need to create two models, one Random Forest and one XGboost. Compare the performance (standard classification stuff), you should tune them using k-fold cross validation and at least 2 parameters.

## Executive Summary (35 points)

Your challenge is to concisely present your findings and results without writing a book. A good executive summary will leave the reader with **actionable** takeaways that they can remember and regurgitate at the next meeting. Your executive summary should be just that a ***summary***. What problem are you challenged with? What were 3 or 4 key findings (things you found interesting that influenced the model). What was result of your model, and 2-3 recommendations that you’d make.

* State the problem
* 3-4 Key findings
* Model Performance & Interpretation of it.
* 2-3 actionable Recommendations

Helpful hint: do not attempt to draft an executive summary until after you’ve done the analysis and written the modeling report. The executive summary should be the last thing you should do.

## Detailed Analysis (10 points)

Here all we want is the following.

1. Two Tables comparing tuning performance of your RF and XGB models
2. Single Table comparing the final fit metrics of your RF and XGB model on train and test
3. Single Graph comparing the ROC curve of your RF and XGB model on Test set.

**Grading:**

This project is worth a total of 50 points

* Executive Summary is worth 35 points
* Detailed Analysis is worth 10 points
* Kaggle AUC up to 5 points