Data Science Internship Assignment

This is a short project for you to show off your analysis and ML skills using a slice of data similar to what we work with at Glassdoor. We hope you have fun with it. Address the questions in section 3 below, aiming to be efficient, clear, and a little creative. To be considered for our Data Science, ML internship in Mill Valley, CA you must **submit your work to alan.wilson@glassdoor.com with the subject "ML Intern Summer 2018" by Monday, April 9th at 9am PDT**.

Useful shortcut for some: If you applied to multiple internships at Glassdoor you may have already submitted work to Ozan Koyluoglu on the topic of "Prediction of Apply Rate". If so, simply share your work on that project with the above email and subject line (no new work is necessary).

1) Introduction

Employers can create a Free Employer Account (FEA) on Glassdoor that allows them to update basic information about the company and respond to reviews. These employers can become Glassdoor's customer by buying job ads and/or investing in their branding (by paying for enhanced features of their employer profile as well as additional analytics).

There is a large number of potential customers for Glassdoor among those employers with an FEA, and our sales team needs help identifying the most promising candidates to reach out to. In this project you are trying to predict the likelihood that an employer with an FEA would become a client.

2) Data

There are two sets of data that you will use. They will be found in this same directory.

1. **`data_challenge_traindata.csv`** – This contains employer features at the time they may become a customer. This is from at time at least 3 months earlier than data_challenge_rundata.csv. This is for training and testing a model.
2. **`data_challenge_rundata.csv`** – This contains employer features from the previous complete month, i.e. if right now the month is February, this reflects the month of January. This is for running the model on.

Data explanations:

Column	Description
employerId	Uniquely identifiable key for each employer, anonymized and randomized for legal purposes.
employeesTotalNum	Number of employees of the employer
monthOfSignupPageViews	Number of pageviews that the employer profile received
clicks	Number of clicks the employer profile received
industry	Industry of the employer
ATS__c	The name of the applicant tracking system (ATS) used by the employer, or ‘UNKNOWN’ if they do not have one
timeOnSite	Number of days the employer has had an FEA
hasInterviews	Whether or not the employer has interviews on their profile
starRating	The employer’s star rating
numJobs	Number of open jobs the employer has
subsidiary	Whether the employer is a subsidiary of a parent company
awards	Number of awards the employer has won
contentCount	Number of reviews and interviews an employer has
freshContent	Number of reviews and interviews in the past 3 months
followers	Number of followers the employer has
photos	Number of photos the employer has
intDifficulty	Average difficulty of the employer’s interviews (ranges from 1-5)
intDuration	Average number of days the employer’s interviews take
basePayAmount	Average salary based on reviews
respondedContacts	Number of contacts linked to the account that responded
isWon	Indicates whether an account has converted to a paying client. Only in traindata.

3) Questions

a) First evaluate the data - what initial observations can you make from this data? Are there any concerns about the data?

b) Imagine you are presenting your results to a non-technical audience. Summarize your findings - What are the top 3 insights that you can draw from this data set? Share your plots and findings.

c) Train and test a model to predict how likely an account will be won (meaning they become a paying customer). You're welcome to just pick a few features (please describe your reasoning in picking the features). For this task, use the data set `data_challenge_traindata.csv`.

d) Briefly describe in a few words: How well does the model work? Which are the most important features? What about accuracy?

e) Run your model on the dataset `data_challenge_rundata.csv`. What insights can you draw?

f) If you run out of time/had extra time: Please explain the next steps you would take.

Please use Python or R to analyze this data. Submit a PDF document with your responses and a separate document with your code in a Jupyter notebook, R markdown, or a simple text file. We are not looking for a polished presentation or comprehensiveness; simply address each of the questions with a few thoughtful comments. We expect this exercise to take two to three hours. If you see any discrepancies or have any issues with the data set, please make your best assumptions and note them in your write up.

Good luck!