accern

Thank you for your interest in the **Data Science Intern** position at Accern.
We find your profile a great match to our requirements, and would like to go ahead with the our standard 2-stages screening process -- *1) completion of an assignment,* and *2) final interview with our team upon successful completion of the assignment.*

Assignment
Accern is a data science startup that monitors over 300 million websites, extracts content and derives actionable analytics that can be used by investment banks, traders and hedge funds to make better investment decisions.

You are given a data file with about 8 months of Accern data (Jan - August, 2016).

You task is as following:

1) Develop a model for **anomaly detection**. It could be a stochastic or completely deterministic/rule based model. You goal here is to identify anomalies in the data distribution, if any. The model should answer questions such as -- Is there a field/attribute that behaves inconsistently over time? Are there timeframes when certain attributes had issues like too many missing values, too much variance in the value distribution etc.? You must clearly state your hypothesis, assumption and definition of how your model describes anomaly.

2) Write a script (Python, R or IPython Notebook) that can be used to pass any data file (similar format), passes it through the model (you developed), and the model detects and returns any anomalies. It is an open-ended task, so you have the power to decide how the user should use the script. Should he just pass the data file, or also call different functions for each attribute individually. Your goal is to make this process as efficient as possible.

3) Write a 2-3 pages report of your findings from the distribution of the provided data sample. (*Did you find something interesting/shocking/surprising? What approaches did you follow? What were your hypotheses? What was the rationale behind those hypotheses? Did they hold or fail? Did you learn something?*)

Data - https://dl.dropboxusercontent.com/u/428478238/research/accern_data_sample.csv.gz
User Guide - http://docs.accern.com/

Time : You must complete and submit this assignment in **5 days.**

How to submit the assignment?

1. **Codebase:** Push your code to a public repository on Github and provide us the link.
2. **Documentation:** Please provide a clear documentation (in the Readme file of your repo) of how to run the code, how the code is organized, any assumptions made and the list of libraries used.

Contact: You can connect with Anshul (anshul@accern.com) for any related queries.