Data Science Case Study Options

Please select and complete *one* of the following case studies. We are looking for you to show off your machine learning and coding skills using Python or R.

You are not required to use AWS in your solution, but you are welcome to use a $10 AWS credit to spin up an EC2 instance if you would like to. If you use any AWS services please remember to terminate them after you complete the exercise.

Please send a document displaying your code and thought process to aidatascienceinterviewers@amazon.com and CC your recruiter at least 24 hours prior to your interview.

Option 1: Sentiment Identification

BACKGROUND

A large multinational corporation is seeking to automatically identify the sentiment that their customer base talks about on social media. They would like to expand this capability into multiple languages. Many 3rd party tools exist for sentiment analysis, however, they need help with under-resourced languages.

GOAL

Train a sentiment classifier (Positive, Negative, Neutral) on a corpus of the provided documents. Your goal is to maximize accuracy. There is special interest in being able to accurately detect negative sentiment. The training data includes documents from a wide variety of sources, not merely social media, and some of it may be inconsistently labeled. Please describe the business outcomes in your work sample including how data limitations impact your results and how these limitations could be addressed in a larger project.

DATA

Link to data: **http://archive.ics.uci.edu/ml/datasets/Roman+Urdu+Data+Set**

Option 2: Geological Image Similarity

BACKGROUND

A geology research company wants to create a tool for identifying interesting patterns in their imagery data. This tool will possess a search capability whereby an analyst provides an image of interest and is presented with other images which are similar to it.

GOAL

Your task is to create the machine learning component for this image similarity application. The machine learning model should return the top K images that are most similar to this image based on a single image input.

DATA

Download link: http://aws-proserve-data-science.s3.amazonaws.com/geological_similarity.zip

Option 3: Maintenance cost reduction through predictive techniques

BACKGROUND

A company has a fleet of devices transmitting daily telemetry readings. They would like to create a predictive maintenance solution to proactively identify when maintenance should be performed. This approach promises cost savings over routine or time-based preventive maintenance, because tasks are performed only when warranted.

GOAL

You are tasked with building a predictive model using machine learning to predict the probability of a device failure. When building this model, be sure to minimize false positives and false negatives. The column you are trying to predict is called failure with binary value 0 for non-failure and 1 for failure.

DATA

Download link: http://aws-proserve-data-science.s3.amazonaws.com/predictive_maintenance.csv