Capstone Project Ideas

Springboard Data Science Career Track

I. **Predicting Mobile User Demographics from Phone Usage**
`https://www.kaggle.com/c/talkingdata-mobile-user-demographics/data`

I would utilize the data logged from the mobile apps to predict age/gender of mobile users (regression/classification, respectively). I would likely spend a fair amount of time with exploratory data analysis and feature engineering prior to making predictions. Once I get to the prediction stage, I plan to compare/contrast a variety of different methods to evaluate which performs best on this dataset.

II. **Building a Topic Model for Predicting Jeopardy Question Categories**
(long URL, click here)

I would use different types of topic models (tf-idf, LDA, PLSA, BTM, etc.) to classify Jeopardy questions/answers into categories. The dataset contains categories for each given question that would allow me to assess the success rate of the model. This project would require a bit of preprocessing (tokenize/vectorize questions), demonstrate general knowledge with best practices in NLP, and show some familiarity with my choices of (probably linear) classifiers.

Alternative Dataset (collection of blog posts, $\sim$ 300 MB):
`http://u.cs.biu.ac.il/%7Ekoppel/BlogCorpus.htm`

III. **Anomaly Detection for Time Series Data**

I would use this project to explore the many different methods currently used for anomaly detection in time series data [1]. I would likely tackle this problem in the form of a blog post that outlines the various techniques available for time series forecasting, focusing in on the realm of applicability for each technique (i.e., which techniques are ideally suited for particular patterns in time series data) with either real-world examples or general discussion, then have a nice piece at the end about detecting anomalies. I have found a number of datasets that could be used for this project, which I have listed below.

Datasets:
Cryptocurrency Market Data: `https://www.kaggle.com/jessevent/all-crypto-currencies/data`
Yahoo Web Traffic Data: (long URL, click here)
Cell Phone Sensor Data: (long URL, click here)

[1] `https://blog.statsbot.co/time-series-anomaly-detection-algorithms-1cef5519aef2`