Data Science Internship - *Month1*

January 2018

Task1: Python for DS

Improve your python skills concentrating on libraries used in data science projects: numpy, pandas, scikit-learn, matplotlib.

Task2: Classification in DS

Use case scenario:
On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. Translated 32% survival rate.
One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew.
Latest discovery have shown that there were actually 2234 people people on Titanic, with relevant information about additional 10 passengers such are name, ticket class, age, sex, but without information whether they survived the sinking.
Create the system which would predict probability of their survival.

Subtask1: Analyze given dataset (performing basic stats, missing values analysis, detecting outliers, etc.). Document your discoveries.

Subtask2: Using expert knowledge (if you watched the movie, you already have it, otherwise gain some by research or from other Titanic (history or movie) experts) to perform data dimensionality reduction and choose optimal/suboptimal attributes for further processing.

Subtask3: Perform preprocessing (fill missing values, data normalization, deal with categorical attributes, etc.)

Subtask4: Research on best classification models/algorithms for your use case and implement three best ones (take into account 'probability measure' from top task). Compare those three approaches using appropriate metrics and splitting dataset in train and test data. Document your decisions and conclusions.

Subtask5: Test your system with 10 new discovered passengers whose information would be provided after successfully completion of first four subtasks.

Task3: Anomaly detection

Your task is to develop system which can be used for **smart production monitoring** and **fault detection** in manufacturing processes.
More accurately, you will perform **anomaly detection on measurement data** collected during semi-conductor manufacturing process which can be found at http://archive.ics.uci.edu/ml/datasets/secom.
Data contains labels for final state of the production entity (faulty or not). However, your job is to perform anomaly detection using unsupervised (and/or semi-supervised) learning, using these labels only for testing purposes. Additionally, try to detect which part of measurement line is most important for fault detection.

Subtask1: Download data and perform data analysis. Document basic statistics.

Subtask2: Research on methodology used in anomaly and fault detection.

Subtask3: Develop anomaly detection system.

Subtask4: Perform postprocessing to detect measurement point(s) which is(are) most relevant for fault detection.

Subtask5: Document your discoveries, together with short proposition on how would you implement and use your system in the real world production processes.