**Proposal for Capstone Project 2**

First idea: Primary image Identification

With the latest development of E-Shop platform, it is important to develop suitable computer vision algorithm and data science method to automate the image classification process. Until now there is a large amount of human involvement for identifying primary image out of all images associated with one product. The definition of a primary image that can be directly use on an E-shop platform is: The primary image stands first for that Item. Technically Primary is defined as one of the Product Image style. A Primary image can be either RAW or OUT OF PACKAGING style.

Without developing this automatic image identifier, human is required to make judgement call of selecting a primary image. There are two major pitfalls for this method: 1. Lack of a consistent standard of selecting primary images. Different people will make different choices of primary images, and different types of products will have different preferences for primary images. 2. The large amount of images need to be scanned for primary images will make it impossible for human classification.

The idea of using data science model for primary image identification is to utilize the deep neural network for image classification from separating images as primary and secondary.

Second idea: Predictive maintenance

Predictive maintenance is a popular application of predictive analytics that can help businesses in several industries achieve high asset utilization and savings in operational costs. The tricky part of the predictive maintenance model is to improve its confusion matrix performance, and achieve a balanced false positive and false negative performance. In this project, an industry-specific business scenario will be studied with Machine learning model. The main content of this project is on the data science process - including the steps of data preparation, feature engineering, model creation, and model operationalization. Two questions will be answered from this exercise: 1. How to design suitable target for predictive maintenance so it can achieve advanced warning, and provide values for business owner. 2. How to use suitable DS algorithm to improve its accuracy, in terms of confusion matrix. For this type of problem, accuracy is not a good measure because the dataset is extremely imbalanced. Confusion matrix is a more accurate way of measuring its performance.

Third idea: Well logging prediction

Well logs are interpreted/processed to estimate the in-situ petrophysical and geomechanical properties, which is essential for subsurface characterization. Various types of logs exist, and each provides distinct information about subsurface properties. Certain well logs, like gamma ray (GR), resistivity, density, and neutron logs, are considered as “easy-to-acquire” conventional well logs that are run in most of the wells. Other well logs, like nuclear magnetic resonance, dielectric dispersion, elemental spectroscopy, and sometimes sonic logs, are only run in limited number of wells.

Sonic travel-time logs contain critical geomechanical information for subsurface characterization around the wellbore. Often, sonic logs are required to complete the well-seismic tie workflow or geomechanical properties prediction. When sonic logs are absent in a well or an interval, a common practice is to synthesize them based on its neighboring wells that have sonic logs. This is referred to as sonic log synthesis or pseudo sonic log generation.

Compressional travel-time (DTC) and shear travel-time (DTS) logs are not acquired in all the wells drilled in a field due to financial or operational constraints. Under such circumstances, machine learning techniques can be used to predict DTC and DTS logs to improve subsurface characterization. The goal of the “SPWLA’s 1st Petrophysical Data-Driven Analytics Contest” is to develop data-driven models by processing “easy-to-acquire” conventional logs from Well #1, and use the data-driven models to generate synthetic compressional and shear travel-time logs (DTC and DTS, respectively) in Well #2. A robust data-driven model for the desired sonic-log synthesis will result in low prediction errors, which can be quantified in terms of Root Mean Squared Error by comparing the synthesized and the original DTC and DTS logs.