Capstone II Ideas

Below are three (3) ideas I am considering for my **Capstone Project II**. If you have time, please feel free to take a look and let me know if you have any questions or would like any additional information. Your feedback is greatly appreciated!

1. Microsoft Malware Prediction

Problem: Currently, when Windows Defender Advanced Threat Protection (Windows Defender ATP) encounters a suspicious but undetected file, it queries Microsoft's cloud protection backend. The cloud backend applies heuristics, machine learning, and automated analysis of the file to determine whether the files are malicious or clean. Predictive technologies are already effective at detecting and blocking malware **at first sight**. However, Microsoft is calling for the Data Science community to push these technologies even further—to stop malware **before it is even seen**.

Goal: Predict if a device is likely to encounter malware given the current machine state.

Data Sets

Kaggle Dataset: https://www.kaggle.com/c/microsoft-malware-prediction/data

Microsoft has provided 9.4GB of anonymized data that has been gathered from 16.8 million devices. This data is broken up into two files called test.csv and train.csv.

The train.csv file contains machine configuration information such as the country location, processor type, OS version, Windows Defender engine, whether the firewall, UAC, or Smartscreen is enabled, amount of ram, storage capacity, and more. Each row in this dataset corresponds to a machine, uniquely identified by a `MachineIdentifier`. `HasDetections` is the ground truth that can either be set to 0 or 1 to indicate that Malware was detected on the machine.

The malware industry continues to be a well-organized, well-funded market dedicated to evading traditional security measures. Once a computer is infected by malware, criminals can hurt consumers and enterprises in many ways.

The client/customer will be Microsoft but will benefit the more than one billion enterprise and consumer customers the most.

2. Sustainable Industry: Rinse Over Run - Hosted by DrivenData

Problem: Efficient cleaning of production equipment is vital in the Food & Beverage industry. Strict industry cleaning standards apply to the presence of particles, bacteria, allergens, and other potentially dangerous materials. At the same time, the execution of cleaning processes requires substantial resources in the form of time and cleaning supplies (e.g. water, caustic soda, acid, etc.).

Given these concerns, the cleaning stations inspect the turbidity—product traces or suspended solids in the effluent—remaining during the final rinse. (***Turbidity** is a measure of the degree to which the water* *loses its transparency due to the presence of suspended particulates. The more total suspended solids in the water, the murkier it seems and the higher the turbidity. Turbidity is considered a good measure of the quality of water.*) In this way, turbidity serves as an important indicator of the efficiency of a cleaning process. Depending on the expected level of turbidity, the cleaning station operator can either extend the final rinse (to eliminate remaining turbidity) or shorten it (saving time and water consumption).

Goal: Predict turbidity in the last rinsing phase in order to help minimize the use of water, energy and time, while ensuring high cleaning standards.

Data Sets
Training Set Values -- The time series for each process in the training set.
Test Set Values -- The time series for each process in the test set.
Training Labels -- The final turbidity values for all of the processes in the training set.
https://www.drivendata.org/competitions/56/predict-cleaning-time-series/data/

`train_values.csv` and **`test_values.csv`** contain metadata on the cleaning process, phase, the object as well as time series measurements, sampled every 2 seconds. The time series data pertain to the monitoring and control of different cleaning process variables in both supply and return Clean-In-Place lines as well as in cleaning material tanks during the cleaning operations.

3. DengAI: Predicting Disease Spread - Hosted by DrivenData

Problem: Dengue fever is a mosquito-borne disease that occurs in tropical and sub-tropical parts of the world. In mild cases, symptoms are similar to the flu: fever, rash, and muscle and joint pain. In severe cases, dengue fever can cause severe bleeding, low blood pressure, and even death.

Because it is carried by mosquitoes, the transmission dynamics of dengue are related to climate variables such as temperature and precipitation. Although the relationship to climate is complex, a growing number of scientists argue that climate change is likely to produce distributional shifts that will have significant public health implications worldwide.

In recent years dengue fever has been spreading. Historically, the disease has been most prevalent in Southeast Asia and the Pacific islands. These days many of the nearly half-billion cases per year are occurring in Latin America.

An understanding of the relationship between climate and dengue dynamics can improve research initiatives and resource allocation to help fight life-threatening pandemics.

Goal: Predict the number of dengue cases (total_cases) for each (city, year, week of the year) based on environmental variables describing changes in temperature, precipitation, vegetation, etc., for two cities, San Juan and Iquitos, spanning 5 and 3 years respectively.

Data Sets

Training Data Features -- The features for the training dataset.
Training Data Labels -- The number of dengue cases for each row in the training dataset.
Test Data Features -- The features for the testing dataset
https://www.drivendata.org/competitions/44/dengai-predicting-disease-spread/data/

The data comes from multiple sources aimed at supporting the Predict the Next Pandemic Initiative. Dengue surveillance data is provided by the U.S. Centers for Disease Control and prevention, as well as the Department of Defense's Naval Medical Research Unit 6 and the Armed Forces Health Surveillance Center, in collaboration with the Peruvian government and U.S. universities. Environmental and climate data is provided by the National Oceanic and Atmospheric Administration (NOAA), an agency of the U.S. Department of Commerce.

According to DrivenData:
"*This is a complicated and messy problem, to be sure. But real data is often complicated and messy. Study up using the resources below—your insights could save lives!*"

Additional Resources:

- Dengue Forecasting Homepage
- CDC Dengue Overview
- NOAA Wiki