Data:

Integrating Predictive Artificial Intelligence and Machine Learning with Public Data

Data:
Integrating Predictive Artificial Intelligence and Machine Learning with Public Data

Problem:

CTAC has been supporting a large health-focused agency in preparing for disasters. Through this support, CTAC established an interactive map consolidating data from multiple sources to assist state, local, and federal agencies in reacting to catastrophic events and deploying power-dependent medical devices. However, CTAC realized this data was largely reactive to events that may cause power outages and sought to leverage advanced technologies to predict these catastrophes and provide more time for planning the deployment of resources. This would springboard planning from reactive to proactive.

Solution:

We combined cutting-edge technology, agile methodologies, public big-data sets, and years of experience to build a novel proof-of-concept tool – the Climate-based Outage Analysis and Tracking Intelligence, or COATI. COATI was born to illustrate what can be done when you analyze and visualize data while leveraging machine learning and predictive artificial intelligence.

Data Prep:

- Data comes from multiple sources in different formats. The COATI core data sets are a mix of JSON, CSV, parquet, and some proprietary text-based formats. The variety of formats, compositions, and partitions necessitated preprocessing and transforming data into a common format before processing with machine learning. Python + Jupyter notebooks made this easy with a combination of pyspark, pandas, numpy, and fastparquet. Datasets were processed and merged in a multi-stage approach to arrive at a common data set that combines DOE outage data, US County map data, NOAA weather data, and third-party power outage data.

Model Training:

- We experimented with as many inputs as we could get for training the model. Utilizing tools from the popular scikit-learn python package, it is possible to graph the relative importance of each parameter (feature) to the overall classification of data (see below). Interestingly, the county of residence was by far the most important factor in determining electrical outage probability. Upon reflection, this makes some sense as outages are probably strongly correlated with factors like geography, local tree failure, above vs. below ground wiring, etc.

- The training was performed on a random subset of historical data using tools from scikit-learn for sampling. Multiple classifiers were analyzed to find the one that best fits our data including DecisionTree, LogisticRegression, RandomForest, AdaBoost, and KNeighbors. In the end, the RandomForest classifier gave us the best accuracy.

Predictions:

- We developed our tools to get seven-day forecasts for each country in the US and export them as a large CSV table. Our saved classifier model is loaded from a pickled state and the forecast data is fed in. The result is a large data frame that includes outage probabilities for each county on each day of the forecast.

Calculating Vulnerability:

- Simply predicting when and where outages would occur was only a part of this project – the real goal was to use this information to illustrate differences in health equity. Not every area of the country has the same access to health services with factors like geographic proximity to hospitals, bed counts, and population density have potentially significant impacts on the availability and quality of care available.

- The vulnerability score is calculated with a variety of static factors (population, land area, hospital & bed count, and the number of local electrically dependent people) and dynamic factors from the forecast (extreme temperatures, excessive precipitation, high sustained winds & large gusts). These factors are scored individually and then combined to form an aggregate Vulnerability Score for a county. A high vulnerability combined with a high outage probability becomes an area of higher concern as the potential repercussions of a regional power outage are more severe (think heat stroke, hypothermia, flooding, downed trees, hospitals over capacity, etc.).

Presentation:

- The result of all this computation culminates into a large table of numbers, however, the CTAC team incorporated kepler.gl maps, an open-source framework for working with geospatial map data, to make the data more useful and compelling. It provides us with tools for filtering, coloring, and displaying data in both two and three dimensions to add visual appeal and readability.

Outcome:

COATI features three main reporting views: 1) Probability of a Power Outage, 2) Vulnerability of the County's Population, and 3) Regional Vulnerability (Heat map). The view which shows the Probability of a Power Outage view is titled **'High Outage Chance'**. Based on an upcoming seven-day forecast, it shows the entire US and each county's predicted probability percentage of losing power. By hovering over a selected county, the user can see several variables that contribute to the probability and vulnerability scores.

The second **'High Vulnerability'** view provides COATI users with a filtered view of high vulnerability areas (See 'Calculating Vulnerability' below). The view is filtered to only show counties with an outage probability of over 40%. This highlights the inequity that different populations may experience during a power outage and can help federal, state, and local authorities focus attention on those vulnerable populations.

The third **'Heat Map'** view highlights clusters of vulnerable regions of the country, defined by the proximity of vulnerable counties during a predicted outage. For example, the test period of June 6-8, 2023, showed clusters of vulnerable counties in the Carolinas and Alabama. This means that the entire southeastern region of the US could be impacted, as affected residents from neighboring counties cross county and state lines in search of services. COATI map can support governments in planning and preparing the necessary emergency services.