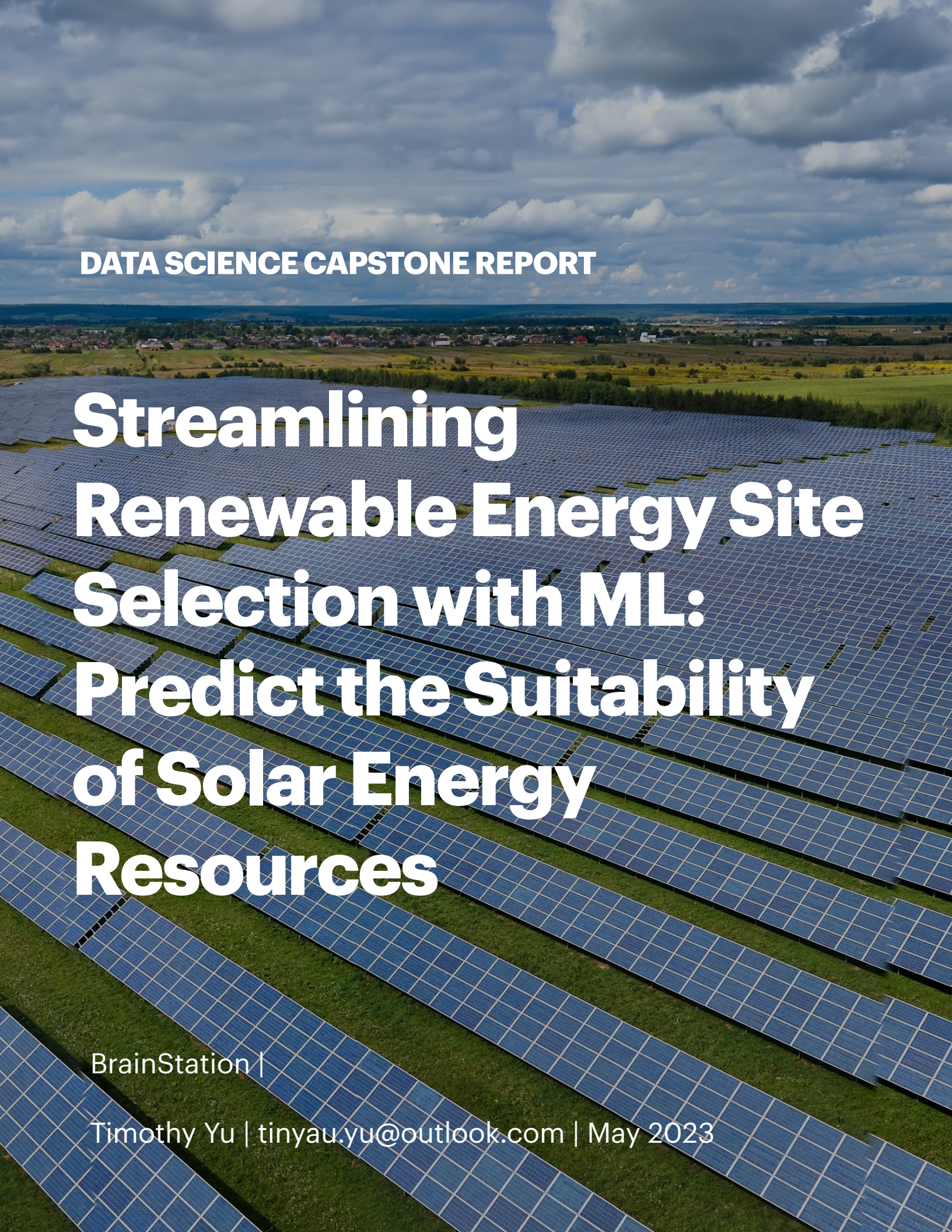

DATA SCIENCE CAPSTONE REPORT

Streamlining Renewable Energy Site Selection with ML: Predict the Suitability of Solar Energy Resources

BrainStation |

Timothy Yu | tinyau.yu@outlook.com | May 2023

PROBLEM STATEMENT

The goal of this project is to apply machine learning techniques to classify whether solar is suitable as the primary energy source for a location.

This can help **energy decision makers** and **energy companies** to simplify the complex site selection process, thereby accelerating the transition to clean and sustainable energy sources, ultimately contributing to the fight against **climate change** and **energy crisis**.

BACKGROUND

Energy decision makers and the energy sector are facing challenges in meeting increasing energy demand and stabilizing the energy price from volatility of fossil fuels, while facing trying to reduce greenhouse gas emissions to combat global warming [1]. Solar power is a popular and cost-effective clean energy source that being considered as the biggest source of electricity by 2050. However, finding the best site for solar power plants is challenging and requires detailed assessment of various geospatial factors and socio-economic factors. With the advancement of technology and the availability of GIS data, machine learning models can be trained to identify suitable sites for renewable energy sources.

DATA ACQUISITION

The core data for this project is a census of U.S. operable power plants as of September 2022. It is freely available for download as CSV file on the U.S. Energy Atlas website [2]. The data set has 11,569 unique power plants and 34 columns which represent the geophysical, power generation, energy source, and other general characteristics of power plants. The Longitude and Latitude columns are used to extract geospatial data from raster maps.

Raster is a digital aerial imagery from satellite. It consists a matrix of pixels where each cell contains a value representing geospatial information. There are seven rasters used in this project, representing seven climate variables: Wind speed[3], Wind power density[3], Solar radiation[4], Air temperature[4], Cloud cover index[5], Slope[6], and Elevation[6]. The value in each cell is an aggregated average annually over decades.

DATA PREPARATION

In this project, we focus on renewable energy sources including solar, wind, hydroelectric, geothermal, and biomass. We filtered out non-renewable power plants based on the column *PrimSource*, then filtered further to only include power plants that exclusively use their primary energy source. After merging geospatial data and removing irrelevant columns and duplicates, we filled missing and extreme values with the median. Finally, we set *PrimSource* as the target variable and assigned *solar* a value of 1 and *non-solar* a value of 0 to ensure a balanced dataset. The resulting clean binary classification dataset contains 7,784 renewable power plants and 11 columns.

FEATURE ENGINEERING

We performed OneHotEncoding to create dummy variable for the *State* column, and then inspected the multicollinearity between the features. We ended up dropping all the dummy variables of *State*, *Latitude, CloudCoverIndex, AirTemperature,* and *WindPowerDensity* to achieve VIF scores less than 2 for the remaining features. The closer the VIF scores to 1, the less multicollinearity between the features. Lastly, we converted the target variable to a binary variable with class 1 (i.e. Solar), and class 0 (i.e. Wind, Geothermal, Hydroelectric, and Biomass). See **Table 1** for the final model for training.

Features	Target
Longitude	isSolar {1: solar, 0: non-solar}
WindSpeed	(i.e. 1: PrimSoure = solar)
SolarRadiation	(i.e. 0: PrimSoure = wind, geothermal, hydroelectric, biomass)
Elevation	
Slope	

Table 1: Final model for training

RESULTS

North Carolina and California are the Top 2 states with the most number of Solar Power Plants

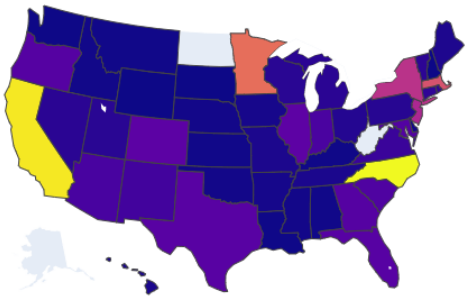

Figure 1: Choropleth of Solar Plants by States

We produced a choropleth to illustrate the number of operable solar power plants per state in U.S (see **Figure 1**). We found that the total number of solar power plants locate in North Carolina and California contributes about 30% of the total solar power plants in the U.S. We argue that the reason why these two states excel in renewable energy may not solely due to the climate and weather conditions but more to the supportive policies developed to encourage the growth of renewable energy.

Classify if solar energy is suitable with Supervised Machine Learning

Several classification models, including Logistic Regression, Decision Tree, Random Forest, and XGBoost, were utilized to categorize solar energy as the primary energy source. Each model was initially trained using a base model and then underwent hyperparameter optimization using a 5-fold cross-validation grid search. Although various evaluation metrics, including accuracy, precision, recall, f1-score, and ROC AUC score, were recorded, the primary focus was on accuracy and recall. The reason for emphasizing recall was to minimize false negatives, where a location is deemed unsuitable for solar energy when it is actually suitable. This approach ensures that we identify all potentially suitable solar energy locations and can conduct further evaluations. In contrast, we may not need to investigate any further if a location is classified as unsuitable for solar energy.

Table 2 depicts the assessment metrics for the ultimate models chosen following hyperparameter optimization. Of all the models, the Random Forest model had the highest *recall* (91.62%) for the test set.

	Test Accuracy	Test Precision	Test Recall	Test F1-Score	Test ROC-AUC
Logistic Regression	72.19	71.83	84.92	77.83	69.95
Decision Tree	77.65	76.89	87.37	81.80	75.93
Random Forest	81.89	79.84	**91.62**	85.32	80.18
XGBoost	82.59	82.10	89.16	85.48	81.43

Table 2: Evaluation metrics of the final models

RECOMMENDATIONS AND FUTURE WORKS

Develop Solar Power Plants in the California Sate

Our models indicate that the two most significant features in determining the suitability of solar energy as the primary renewable energy source are Slope and WindSpeed. The optimal Decision Tree model highlights that locations with a slope of less than 3.35m and a wind speed of less than 8.2 m/s are ideal for solar energy. California meets these requirements and also has policies in place that encourage the development of renewable energy.

Future Works

In order to enhance the accuracy and scope of our machine learning project, there are several future areas of exploration. Firstly, we aim to augment the dataset size by utilizing the global renewable power plants inventory. This would enable us to obtain more comprehensive and diverse data for training our models. Secondly, we intend to extract additional weather data such as humidity, dew, precipitation, and socio-economic data (e.g. land cover and natural disaster frequency) to investigate their potential impact on the suitability of solar energy. Lastly, we plan to experiment with neural networks to obtain higher scores, as they can potentially offer more advanced learning capabilities compared to traditional machine learning algorithms. Overall, these proposed future works will help us to gain deeper insights into the utilization of solar energy and contribute to the advancement of renewable energy development.

REFERENCE

[1] International Energy Agency. (2019). World Energy Outlook 2019. Paris: International Energy Agency.
https://www.iea.org/reports/world-energy-outlook-2019
[2] U.S. Energy Atlas
https://atlas.eia.gov/datasets/eia::power-plants/about
[3] Global Wind Atlas
https://globalwindatlas.info/
[4] Global Solar Atlas
https://globalsolaratlas.info/map
[5] Earth Env Project Cloud
https://www.earthenv.org/cloud
[6] Earth Env Project topography
https://www.earthenv.org/topography