COMP 562 Final Project Report

Edward Baker, Ronit Joshi, Aayush Singh, Neil Vakharia
University of North Carolina at Chapel Hill

December 11, 2022

1 Context and motivation

The rise of sustainable energy is revolutionizing both industrial and domestic consumption. Because of this, education and awareness about the impact of certain kinds of sustainable energy sources is necessary. Solar energy, in particular, is relatively easy for everyday homeowners to capitalize on, but is also difficult to calculate the overall cost and benefit that such a choice can provide. One way to understand the potential benefit is to examine the carbon offset caused by switching to solar energy. A carbon offset is a reduction or removal of emissions of carbon dioxide or other greenhouse gases made in order to compensate for emissions made elsewhere. Our goal was to create machine learning models to predict this carbon offset, which would hopefully provide insight into the effectiveness of solar energy for a given area. Predicting the impact on global warming and environmental health that our decisions will have is crucial to sustainable development, which is what this project aims to do.

2 Project Sunroof dataset

Our project tests a variety of machine learning models using the popular scikit-learn Python library. We sourced the data from Google's Project Sunroof dataset on Kaggle (1). This is census-tract level data for over 10000 regions that is retrieved from Google's extensive satellite mapping, 3D modeling, image analysis, and weather/climate data. This dataset was inspired by the need to determine the best potential coverage of sunlight, and the regions that would benefit the most from switching to solar energy.

3 Exploring our data

In order to create models for our data, we first had to understand what we were working with. The first thing we did was create a visualization of the values that were missing. Then the data was cleaned by removing the rows with missing values, and was normalized with min-max normalization formula:

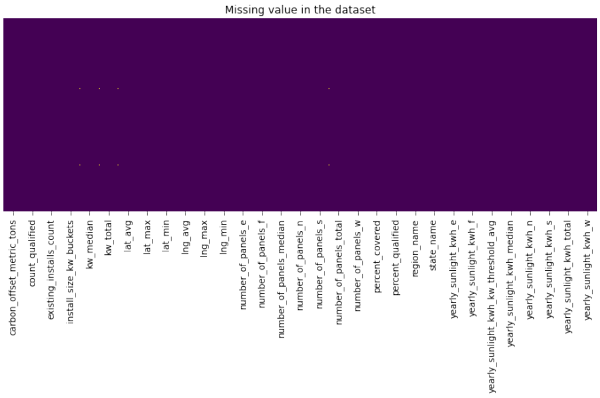

Figure 1: Exploring our data

$$x_{scaled} = \frac{x - min(x)}{max(x) - min(x)}$$

4 Model selection and training

We tried two different approaches to predict carbon offset. For our first approach, we used regression models to predict the exact carbon offset number. We did this by selecting certain features highly correlated with carbon offset and then training a standard linear regression model as well as a ridge regression model. The predictors and their descriptions are shown below. For our second approach we reformatted our data so that the problem could be recast as a classification problem. This involved creating a 'carbon score' associated with each data entry that represented how effective solar energy would be.

Carbon scores were calculated by classifying the data into quantiles. Equal interval classification was not used due to our data distribution being non-uniform. Below is an example of the frequency distribution of data (after it was normalized) for one of our key descriptors, kw_total :

As shown above, this data is not uniform, instead following closely to a logarithmic curve. Therefore, we distributed our data into equal quantiles, and any data points that lay outside of our data set were clamped to the respective minimums and maximums. Carbon scoring for an element of value x is assigned according to the following formula:

$$Carbon\ Score = \left\{ \begin{array}{ll} 1 & if\ x_{scaled} \leq 0 \\ \lceil x_{scaled} * b \rceil & if\ x_{scaled} \in (0, 1) \\ b & if\ x_{scaled} \geq 1 \end{array} \right\} \tag{1}$$

where b is the number of bins we set, (the highest carbon score possible). Note that the smallest carbon score is 1, and the highest is b.

After the carbon scores were created, we tested out a few different classification algorithms to see which would perform best in correctly predicting carbon scores. For classification algorithms, we tried a decision tree model, an SVM model, and a random forest model. The train/test split was set to 25/75 for every model.

Predictor	Description
yearly_sunlight_kwh_kw_threshold_avg	75% of the optimimum sunlight in the county
yearly_sunlight_kwh_f	Total solar energy generation potential for flat roofs
yearly_sunlight_kwh_n	Total solar energy generation potential for north-facing roofs
yearly_sunlight_kwh_s	Total solar energy generation potential for south-facing roofs
yearly_sunlight_kwh_e	Total solar energy generation potential for east-facing roofs
yearly_sunlight_kwh_w	Total solar energy generation potential for west-facing roofs
number_of_panels_f	Solar panel potential for flat roof space
number_of_panels_n	Solar panel potential for north-facing roof space
number_of_panels_s	Solar panel potential for south-facing roof space
number_of_panels_e	Solar panel potential for east-facing roof space
number_of_panels_w	Solar panel potential for west-facing roof space
lat_avg	Average regional latitude
lng_avg	Average regional longitude
kw_total	Solar potential for all roofs (kW)

4.1 Linear regression model

Linear regression had a R^2 value of 1.0, which almost perfectly fit the test data (the mean squared error between predicted and actual carbon offset was 2.196×10^{-33} metric tons).

4.2 Ridge regression model

The regularization parameter α was set to 0.01 which gave us an R^2 which was 5×10^{-9} away from 1.0. This corresponded with a mean squared error of 2.455×10^{-3}.

4.3 Decision tree model

Our decision tree had an entropy split criterion using the scikit function, and had a maximum depth of 5. The accuracy of the score was 94.476%.

4.4 SVM model

The support vector machine model was set to have a gamma value of $\frac{1}{p}$, where p is the number of features we selected. This accuracy was at 42.569%. This was the worst performing model out of all tested.

4.5 Random forest model

Our random forest model consisted of 20 trees. Its accuracy was 97.937%.

5 Accuracy and analysis

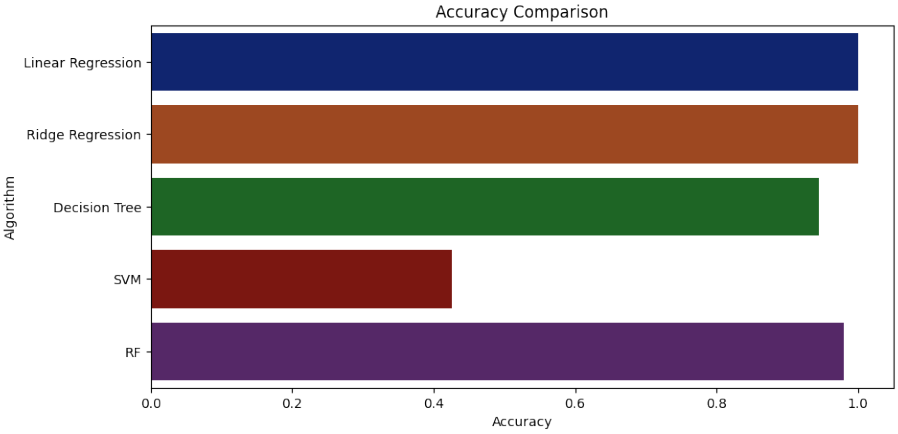

Figure 2: Accuracy comparison of 5 trained models

Given that the data had low noise and didn't have categorical features independent of others, linear regression outperforms the decision tree model, as well as random forests. The variables had a very fixed range, which is why the features didn't have outliers. This made the SVM model less reliable than most of the others.

6 Related works

Methods of predicting solar potential using LIDAR data and roof geometry (2) have been used before. This is similar to other papers about predicting solar potential using image segmentation (3). Both areas of study suggest that a building's roof structure alone isn't sufficient to characterize its solar input. Additionally, studies of analyzing data for specific regions to determine the accuracy of Project Sunroof is also prevalent (4), and point to the conclusion that the data is accurate.

Our project combines image data with numerical, accurate data of the region to predict solar potential in regions that don't have data supported by Project Sunroof. Using the models provided in this paper, we can robustly determine the financial savings and environmental impact.

References

Boysen, J. (2017, September 11). Google project sunroof. Kaggle. Retrieved December 10, 2022, from https://www.kaggle.com/datasets/jboysen/google-project-sunroof

De Barros Soares, D., Andrieux, F.,; Hell, B. (n.d.). Predicting the Solar Potential of Rooftops using Image Segmentation and Structured Data. Retrieved December 11, 2022, from https://arxiv.org/pdf/2106.15268v1.pdf

Lee, S., Iyengar, S., Feng, M., Shenoy, P.; Maji, S. (2019). DeepRoof. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery; Data Mining. https://doi.org/10.1145/3292500.3330741

Mhlanga, T. S.; Ercoskun, O. Y. (n.d.). Mapping, modeling and measuring photovoltaic potential in urban ... Retrieved December 11, 2022.