

**CMPE 255: Data Mining Project**

**Deep Solar: Solar Deployment Analysis**

**Project Group - 15**

**Instructor: Professor Gheorghi Guzun**

**Submission Date: May 5, 2020**

**Submitted by:**

**Fanus Arefaine – SID: 014535859**

**Sen**

**Steven**

**Section 1: Introduction**

* **Motivation**
* **Objective**
* **Literature/Market Review**

**Analysis 1: The State of California Subset Analysis**

**Section 2: System Design and Implementation Details**

In this analysis, the State of California socio-economic subset is extracted from the DeepSolar dataset for analysis.

* **System Design/Architecture/Data Flow**





























For the State of California subset, I decided to look into the impact of socioeconomic and environmental parameters on the solar system count in each tract area. Hence, I split the DeepSolar dataset into four subsets of parameters. These subsets belong to different aspects of socioeconomic status of the State of California. Each subset is examined carefully for missing values and irregularities within each subset. After each subset is cleaned, they are merged and fed into classifier.

The tract areas are then classified into areas with solar systems and areas without solar system using the classification algorithm. Finally, the tract areas’ subset is fed into regression algorithm for a prediction. The regression algorithm gives the predictions of solar system count per thousand households across the tract areas within the State of California.

* **Architecture Related Decisions**

After I decided to perform socioeconomic data analysis for the State of California, I chose to split the dataset into four subsets of socioeconomic parameters to handle them with their relevant parameters. These sets of parameters are cleaned and fed into the classifier to prune the tract areas without solar systems. Hence, the regression algorithm will be able to only deals with tract areas with solar systems which makes the data smaller, more convenient for computation and more accurate.

* **Algorithms Considered/Used**

I have made use of major algorithms in the classification and regression sections of the implementation. Although I have tried couple of algorithms for comparison and regression, some of the algorithms shine over others.

I compared three major classification algorithms for classification of the dataset. Logistic Regression, Support Vector Machine, and Ensemble Methods are the algorithms I compared for my dataset. These classification algorithms are perfect for smaller dataset and they have shown good result.

I have also compared and tuned couple of regression algorithms. Due the non-linearity of the data, I did not use linear regression. Algorithms that support non-linear relationship that I used are Decision Tree Regressor, Random Forest Regressor, and Support Vector Machines. These algorithms have shown fairly good result on the dataset.

* **Technologies and Tools Used**

In this analysis, I have used python libraries such as numpy, pandas, scikit-learn, matplotlib, seaborn and several other supporting components. I used these libraries because they are very powerful for data computation and visualization.

* **Use Cases**

Solar panel retailers, solar installers, and regulatory agencies can use this predictive model to make essential decisions. The model can help these sectors to understand and predict the distribution of solar system in relation to socioeconomic and environmental parameters across the State of California.

**Section 3: Experiments / Proof of Concept Evaluation**

* **Dataset Used and Preprocessing Performed**

The DeepSolar dataset provided by Stanford University is used for this analysis. The dataset has 72537 data records of tract areas across the United States of America and 169 features. The data is composed of numerical and categorical features.

For this particular analysis, Analysis 1, a subset of the dataset that represents only the State of California is used. The subset is divided into four sets of parameters as described above. Various preprocessing techniques are applied on the subsets as per their requirements.

These subsets are first examined for missing values which could be inconvenient for computation and otherwise misleading to the predictive model. Moreover, the parameter values are normalized to eliminate in the range of values while keeping the distribution and information embedded in the data. Mapping of categorical values to representative numerical values is performed as part of cleaning.

Moreover, feature reduction is applied on the dataset. Since features with constant values doesn’t not contribute much to the classification algorithm, they are dropped before the data is fed into the classification algorithm. Feature creation is also used to make a meaningful label for the regression algorithm.

* **Methodology Followed**

For optimal training and evaluating the classification and regression algorithms, the dataset is split into training, validation and testing subsets. The split proportion used is 60% training data, 20% validation data, 20% testing data. Moreover, I used *GridSearchCV* and *RandomizedSearchCV* for cross validation to optimally train algorithms with different parameters. I used 5-fold and 10-fold combinations to get best results from the algorithms.

* **Graphs of parameters/algorithms evaluated in a comparative manner**
* **Analysis Result**

**Section 4: Discussion & Conclusions**

* **Decisions made**

After I decided to build predictive model reflecting the State of California, I decided to use the socioeconomic and environmental parameters. I hand-picked those parameters to support my analysis

* **Difficulties Faced**

Some of the difficulties I faced include picking the right socioeconomic and environmental parameters that correctly support my analysis. Moreover, cleaning the data was major task in the project as it is very crucial for the performance of the algorithms.

* **Things That Worked**

Before feeding the dataset into the classification algorithm, feature reduction is performed on the dataset which really made great difference. In addition, normalization of the dataset before feeding to algorithms helped the algorithms to learn and perform better.

* **Things That Did Not Work Well**
* **Conclusion**

**Analysis 2:**

**Section 2: System Design & Implementation details**

**•Algorithm(s) considered/selected**

I choose to use Random Forest Regression as the machine learning algorithm because it has the features I needed. Random Forest is one of the most accurate machine learning models available. It runs efficiently on large databases. It gives estimates of what variables that are important in the classification. Moreover, Random forest will handle the missing values and maintain the accuracy of a large proportion of data.

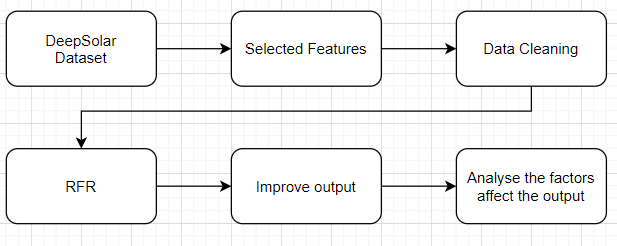
**•Technologies & Tools used**

For this project I use python libraries such as matplotlib, numpy, sklearn and seaborn. By using these libraries I can easily apply machine learning models to the project, plot diagrams to visualize the data and help me to understand the data.

**•Architecture - related decisions**

I will use Random Forest Regression to predict the panel area, then identify the correlation socioeconomic factors with residential solar deployment density. In order to achieve this purpose, I first selected the features that I think are most important. Then I generate a new feature called solar density and compare it with other important socioeconomic factors to see the relationship between them.

**•System design/architecture/data flow**

****

**•Use cases**

Since my goal is to predict the area of solar panels according to the data markers, it can be used by power companies. By predicting the area, power companies can estimate how much power is being generated and fed back to the power grid. Also, it will so the potential market for solar panel retails.

**Section3: Experiments / Proof of concept evaluation**

**•Dataset(s) used**

The dataset we use is DeepSolar from the DeepSolar Project by Stanford University. The dataset has 72537 data records of tract areas across the United States of America and 169 features. The data is composed of numerical and categorical features.

For my analysis, I choose the features that relate to demographic, socioeconomic, and environmental parameters as my subset.Then clean the data by deleting the missing value. Then I create a new feature to help me analyse.

**•Methodology followed (e.g. n - fold-cross validation, number of folds, size of training/test set etc.)**

In this analysis, I used Random Forest Regression. I split the data into 90% training and 10% testing, and I used 10 fold to get the result. Then I use the locally weighted linear regression to analyse some important features with solar density.

**•Graphs showing different parameters/algorithms evaluated in a comparative manner,**

**along with some supportive text.**

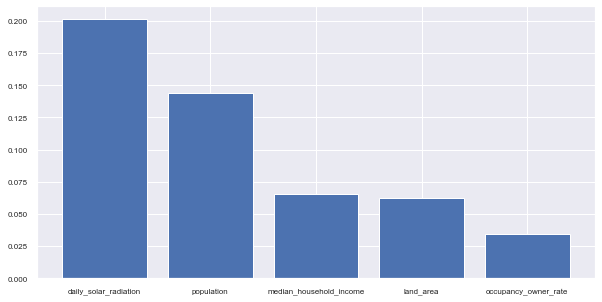


Figure1: relative importance of factors for panel deployment prediction

In Figure1 shows the 5 most important factors from RF, and they are daily solar radiation, population density, the median household income, the land area, and the ratio of houses occupied by their owners. I will use the correlations between the density of solar systems deployments and some important features to give us an insight into the factors that increase or decrease the likelihood of solar panel deployment rate.

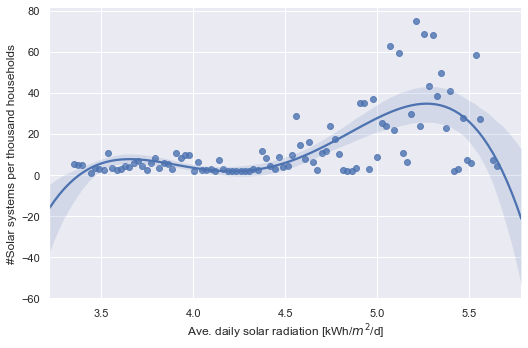


Figure2:solar panel deployment density vs. the average daily solar radiation

It is very obvious that the Sun is the most important factor. The solar panel installation peaks at a radiation level of about 5.25 kWh/m2 per day.

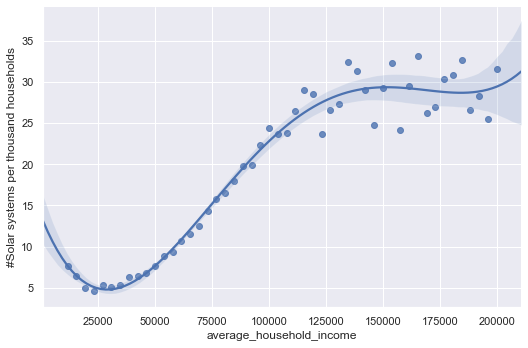


Figure3: solar panel deployment density vs. population density

Figure 3 shows that the solar panel installation peaks when the population density is about 10,000 people per square mile.

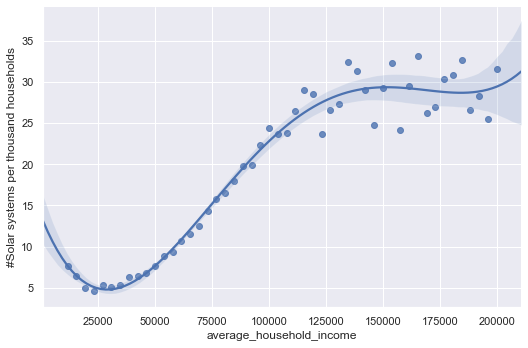


Figure4: solar panel deployment density vs. average household annual income

Figure shows the solar panel installation increase as the household annual income increases, and it peaks when household annual income reaches 150K.

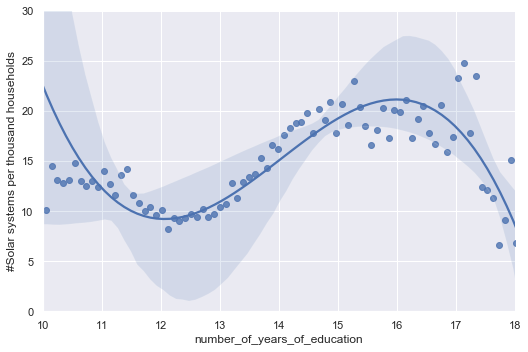


Figure5: solar panel deployment density vs. number of years of education

Figure 5 shows the solar panel installation peaks when the number of years of education reaches 16, but starts to decrease when it’s over 16 years.

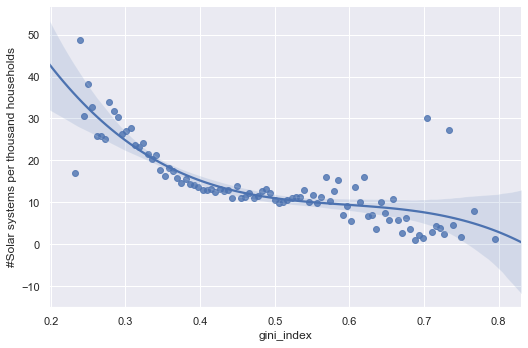


Figure6: solar panel deployment density vs. gini index

Gini index is a measurement of equality where 0 means “total equality” and 1 means “total inequality”The figure shows the high inequality reduces solar panel installations.

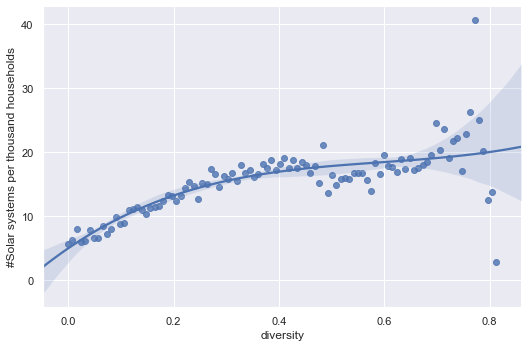


Figure7: solar panel deployment density vs. Simpson’s Diversity Index

Simpson’s Diversity Index is a measurement of racial diversity, and the figure shows the solar panels installations decrease as a region becomes more racially diverse.

**•Analysis of results**

The plots showed what are the optimal points in terms of maximizing solar deployment density. Some of the features are for factors we can’t control such as solar radiation but it’s still good to know their impact. However, other ones we can control. We see that the more equal the society, the more educated, the greater the likelihood there will be more solar panel installations.

**Section 4: Discussion & Conclusions**

**•Decisions made**

I decided to predict the solar panel area. I choose the socioeconomic and environmental parameters for my machine model. Then I use the important parameters to do the correlation analysis.

**•Difficulties faced**

One of the biggest difficulties is choosing the parameters to support my analysis. There are too many features in this dataset, a total of 169 features. I had to look into each feature and then decide to use it or not. I also need to clean the missing value in the data.

**•Things that worked**

I chose part of the dataset as my sub dataset and that reduced a lot of the redundant data which improved my result. I have cleaned the data, deleted all the missing values. The Random Forest Regression shows an acceptable result. I also try to improve the output of RFR by adding some demographic factors, and the result indeed improved (R2 value is 0.668).

**•Things that didn’t work well**

The Random Forest Regression gives an acceptable result, but not great, the model needs to be tuned. In my analysis, I also try to use the number of years of education to predict the income. In theory, when you have a higher education, they usually get a better job and make more money. However, the result isn’t good, R2 is 0.55.

**•Conclusion**

In this analysis, I have predicted the solar panel area and analysed some important socioeconomic, environmental and demographic factors that correlate with the solar density. Further work can apply to this analysis, such as tuning the ML model. Also, the dataset has 169 features, there are a lot of features I haven't used. I can go deeper into the data looking for more related features or looking at it from a different angle and do some different analysis. The analysis can help to increase solar panel deployment rate which will maximize the solar power generation potential and reduce carbon emissions.

**Analysis 3:**