

**CMPE 255: Data Mining Project**

**Deep Solar: Solar Deployment Analysis**

**Project Group - 15**

**Instructor: Professor Gheorghi Guzun**

**Submission Date: May 10th, 2020**

**Submitted by:**

**Fanus Arefaine – SID: 014535859**

**Sen Zheng - SID: 010611029**

**Steven Louie - SID: 009853961**

**Section 1: Introduction**

* **Motivation**

In recent years, solar systems have been getting better attention due to the increase in the price of electric energy and decreasing cost of solar panels across the world. Moreover, governments have been supporting the public with incentives to install solar systems in their homes. The use of solar systems in homes plays a crucial part in protecting the universe by reducing the amount of harmful gasses released to generate electrical energy. Due to these environmental and economic benefits, the deployment of solar energy facilities has recently been growing across many parts of the world. However, in this project, we will analyze and illustrate solar systems deployment in the United State of America. To better understand the distribution and relationships of solar systems deployment using the data, we will numerically and visually present the relationships between different parameters of solar system deployment across the United States of America.

* **Objective**

Focusing on analyzing and formulating a predictive model for solar power deployment across the United States of America, our project will be a great help to solar panel retailers, regulatory agencies, power generating companies. Our project will use the SolarDeep dataset to analyze the distribution of solar power systems installed in relation with the socioeconomic and environmental parameters. These analyses will help solar retailers and regulatory agencies and power system companies to comprehend the relationships and hence, expand their business. Moreover, we have built a predictive model that would help these companies and agencies in taking business decisions based on the projections of the data at different times and different locations. Hence, our project will be a great help in the above-mentioned areas to power system companies, solar panel retailers, solar installers and regulatory agencies.

* **Market Review**

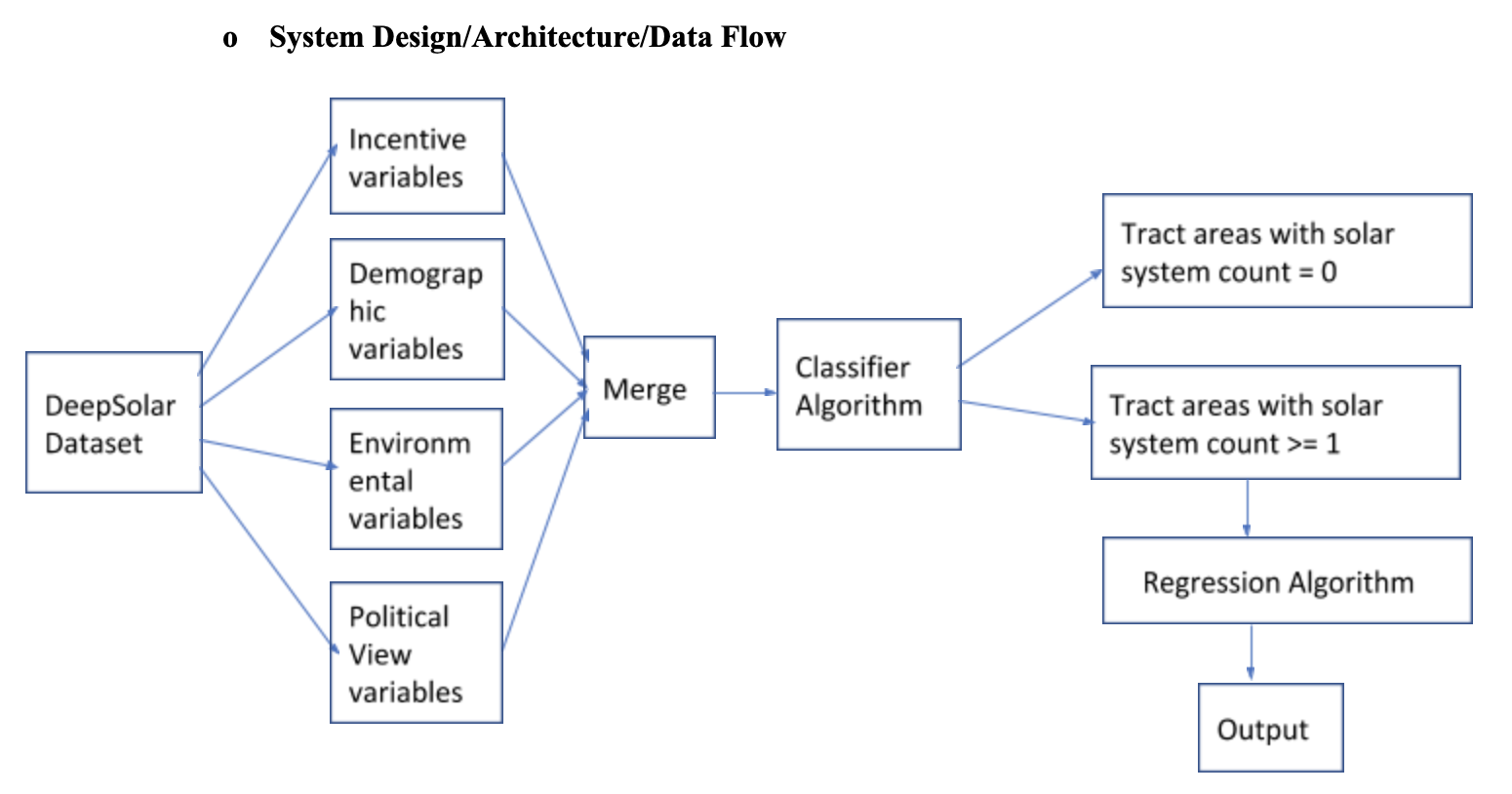
Deployment of solar systems has been growing in the United States of America. The solar power industry was valued at $52.5 billion and projected to reach $223.3 billion by 2026 and showed a 23% increase from 2018. In 2019 alone, the United States of America has installed 13.3 gigawatts (GW) of solar PV capacity, enough to support 14.5 million American homes. More than any other energy generation sources in 2019, solar power has accounted for 40% of total electric energy generation in the United States of America. This increase and highest proportion of solar power deployment is recorded as the highest share in the power generation industry. Although the solar power industry is estimated to grow at a CAGR 20.5% annually until 2026, the decreasing price rate of solar panels can make the progress much faster.

**Analysis 1: The State of California Solar Systems 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. Each subset is examined carefully for missing values and irregularities within each subset. After each subset is cleaned, they are merged and fed into a classifier. The tract areas are then classified into areas with solar systems and areas without solar systems using the classification algorithm. Finally, the tract areas’ subset is fed into a 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 and environmental data analysis for the State of California, I chose to split the dataset into four subsets to handle them with their respective 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 deal with census 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 tried a couple of algorithms for comparison and regression, some of the algorithms shine over others.

I compared two major classification and regression algorithms for my dataset. Support Vector Machine and Random Forest Classifier are the algorithms I compared for my dataset. Moreover, algorithms that support non-linear relationships that I used are Random Forest Regressor and Support Vector Machines. These algorithms have shown fairly good results 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 convenient and 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 the 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**

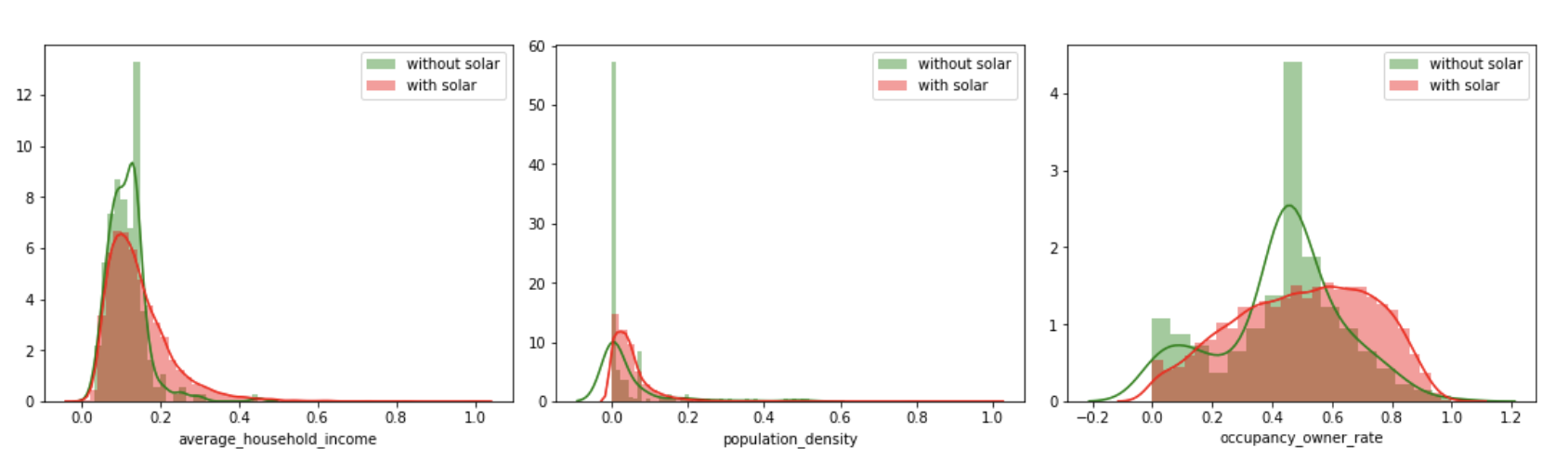
These subsets are first examined for missing values which could be inconvenient for computation and otherwise misleading to the predictive model. Moreover, the parameter values are normalized to eliminate in the range of values while keeping the distribution and information embedded in the data. The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

Moreover, feature selection is applied to 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. I also utilized a feature creation technique where I found out that the ratio of air temperature and wind speed is more correlated with the target value than the constituent parameters.

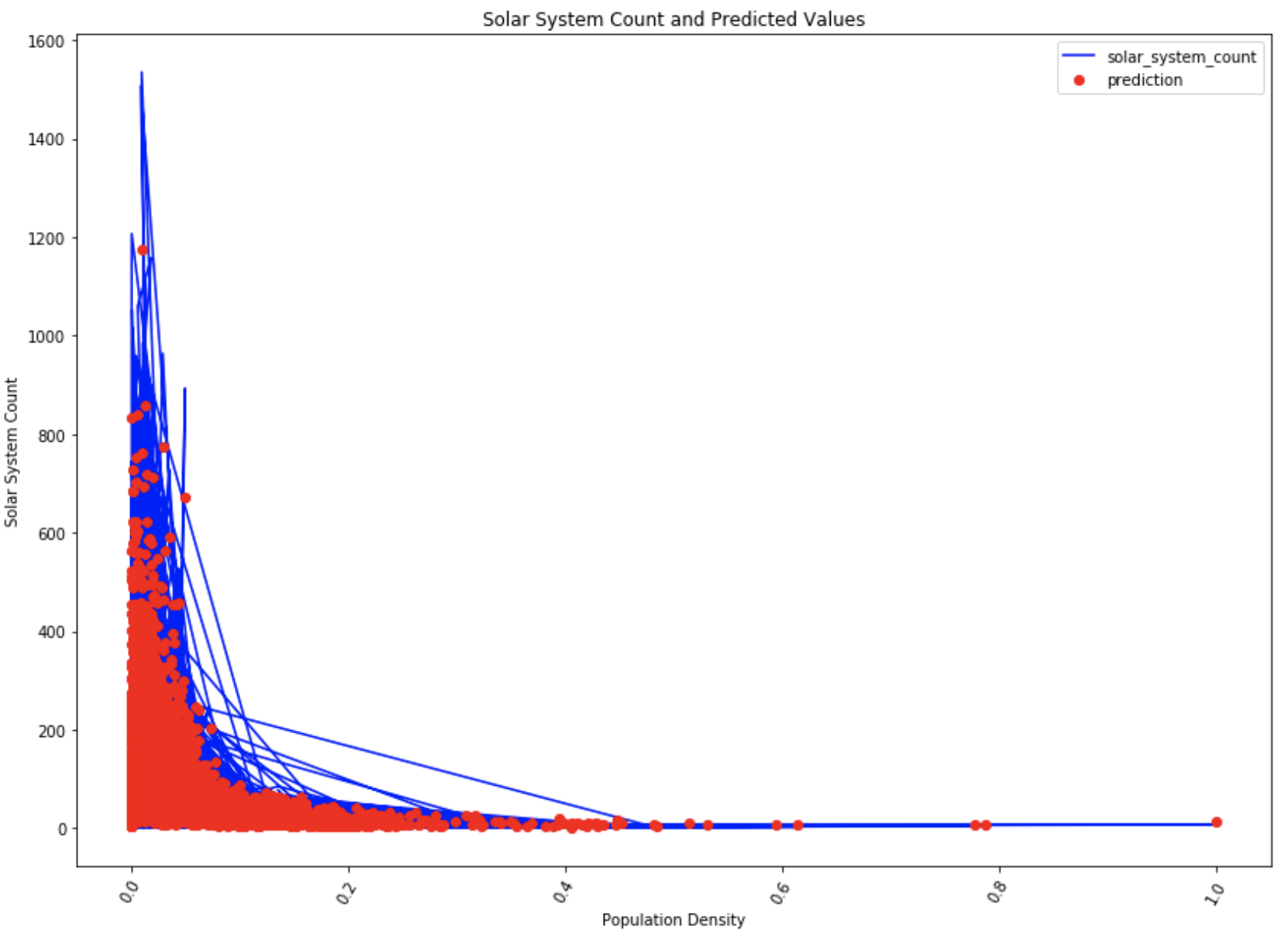
* **Methodology Followed**

For optimal training and evaluating the classification and regression algorithms, the dataset is split into training and validation subsets. The split proportion used is 80% training data, 20% validation data. Moreover, I used *GridSearchCV* for cross-validation to optimally train algorithms with different combinations of parameters of classification as well as regression algorithms. I used 5-fold and 10-fold combinations to get the best results from the algorithms.

* **Graphs of parameters/algorithms evaluated in a comparative manner and Analysis Report**
* **Classification Performance Report - Regression Performance Report**
* Random Forest Classifier - Random Forest Regressor
* Training Accuracy : 97.5% - R2 score: 0.587
* Validation Accuracy: 97.6%
* **Here is a visual presentation of the classification report on selected parameters.**



* **Regression visual performance report is also provided below:**



**Section 4: Discussion & Conclusions**

* **Decisions made**

After I decided to build a 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 a 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 a 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**

One of the techniques I firstly used to study the performance of the algorithms better was splitting the data into training, validation and testing subsets in a 60:20:20 ratio. However, the performance of the algorithms was not satisfactory even when I tried different tuning techniques. The performance was then enhanced when I split the data into only training and validation subsets. I believe the model was not performing good due to the small amount of training data.

* **Conclusion**

Considering the amount of time given for the project, considerable insights is laid out regarding the solar system count analysis in the State of California. California based solar system retailers and installers, and policy regulators can definitely benefit from the analysis and insights pointed out in this project in taking one step further into more detailed analysis with aggregate and similar datasets to make a decisive analysis.

**Analysis 2:**

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

**•Algorithm(s) considered/selected**

I choose to use Random Forest Regression as the machine learning algorithm since it is the most popular and accurate model we have right now. It has many advanced features, such as handling missing values, good accuracy for large data, and high efficiency on large size data.

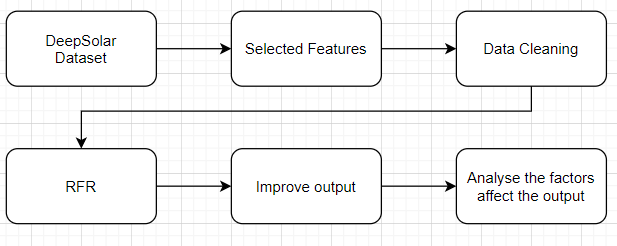
**•Technologies & Tools used**

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 correlation 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**

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% for training and 10% for testing, and I used 10 fold to get the result. Then I use the locally weighted linear regression to analyze some important features with solar density.

**•Graphs showing different parameters/algorithms evaluated.**

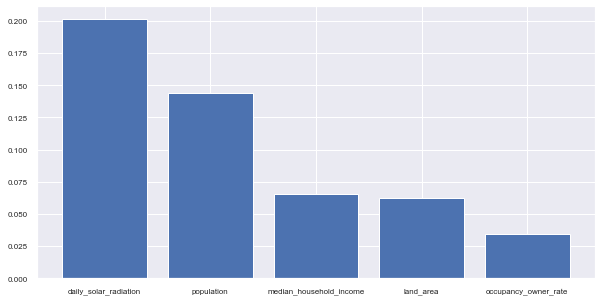


Figure1: relative importance of factors for panel deployment prediction

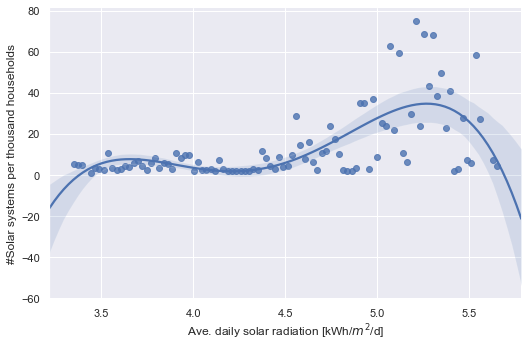
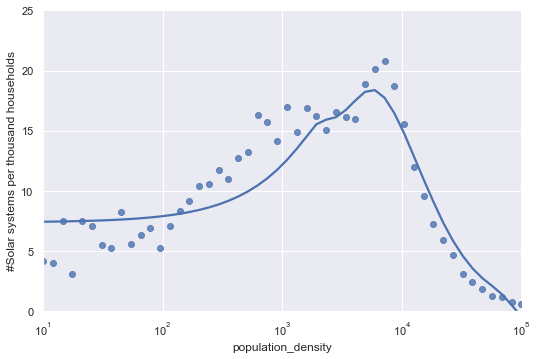
 

Figure2:solar density vs. average daily solar radiation Figure3: solar density vs. population density

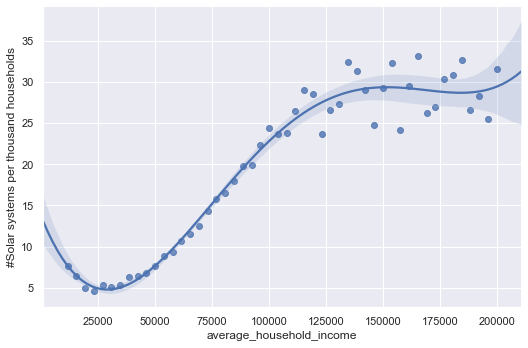
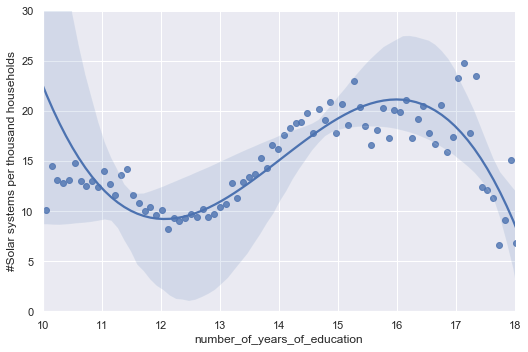
 

Figure4: solar density vs. average annual income Figure5: solar density vs. years of education

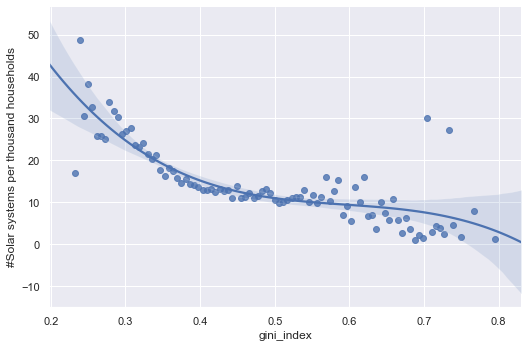
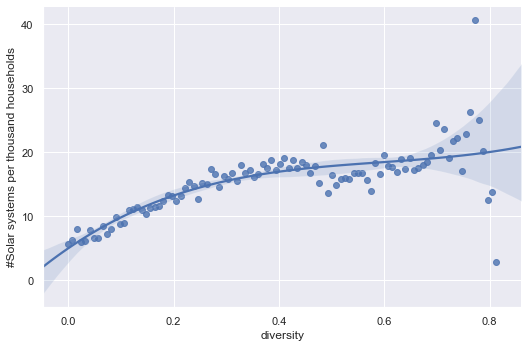
 

Figure6: solar density vs. gini index Figure7: solar density vs. Simpson’s Diversity Index

**•Analysis of results**

The plots above showed the correlation between solar density and other factors and when the solar panel deployment density reached the peak. We can see that as solar radiation increases, income increases, solar deployment increases. Moreover, the solar panel deployment reaches a peak when population density reaches 10000 people per square mile. Also, more education and more equality will also increase solar panel deployment density.

**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**

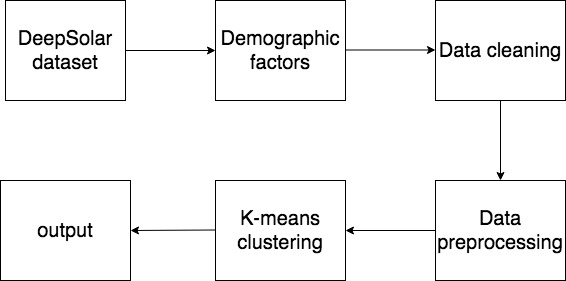
In this analysis, I have predicted the solar panel area and analysed some important socioeconomic, environmental and demographic factors that correlate with the solar density. Further work can apply to this analysis, such as tuning the ML model. Also, the dataset has 169 features, there are a lot of 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:**

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

In this analysis of relationships and k-means clustering, I focused on demographic factors such as average household income, racial ethnicity as well as solar panel count in the top five states of solar panel growth which are California, Arizona, Florida, New York, and New Jersey. Before working on the dataset, I will be cleaning up the sub dataset by dropping any rows that contain missing values. Missing values can cause problems in our dataset because it can skew calculations and projections, which will result in inaccurate projections and analysis. For each of the five states, I will be grouping the dataset into their respective state counties and perform k-means clustering to observe as to which cluster a data point belongs. As a result, it will give us a good visualization between the original groupings of the sub dataset vs the predicted groupings of the dataset using the k-means clustering algorithm.

**System design/data flow**



**Algorithms used**

In my implementation of clustering data points, I used the k-means clustering algorithm offered by scikit-learn. I made the choice of using k-means because the algorithm does a good job of finding groups within a dataset by minimizing the distances of data point to centroid in order to find the most optimal solution on a given dataset.

**Technologies used**

In my analysis and implementation, I've used libraries such as NumPy, pandas, Matplotlib, and scikit-learn.

**Use Cases**

My goal is to show the correlation between demographic factors such as average household income, racial ethnicity and solar panel system counts. This information can be used by solar panel suppliers, so suppliers will know which group of population to appease to in terms of average income and racial ethnicity.

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

Since I am only focusing on a few of the attributes out of 169 total features provided in the DeepSolar dataset. I made sure to clean up any missing values that were present in features such as solar panel count, average household income, and racial ethnic attributes; otherwise, results and projections would be misleading and inaccurate. In order to compare the results of the top five states in terms of solar panel growth, preprocessing was required because I was dealing with a sub dataset that contains multiple labels and label encoding, a preprocessing technique, was able to convert the labels (different states in the sub dataset used) into numeric form.

**methodology followed**

K-means clustering was performed on five different states of data that were categorized into two separate predictors, tile count and average household income. The number of clusters to form was chosen to be five because five different states of information were used. As a result, new labels were formed from k-means clustering and fed into a scatter plot graph along with the predictors to get a good visualization of the newly formed groupings compared with the graph of the original dataset.

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

**along with some supportive text.**

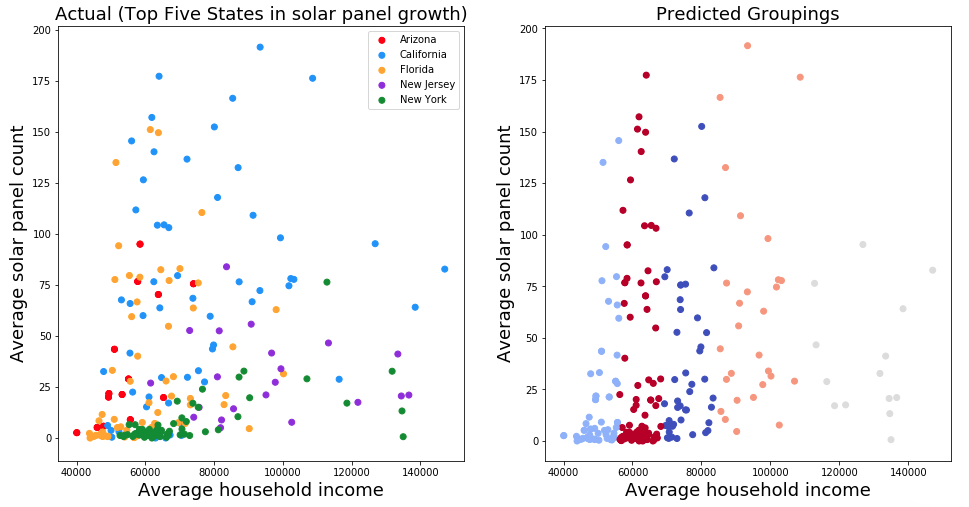
****

Figure 1. Visualization between original groupings vs predicted groupings of sub dataset

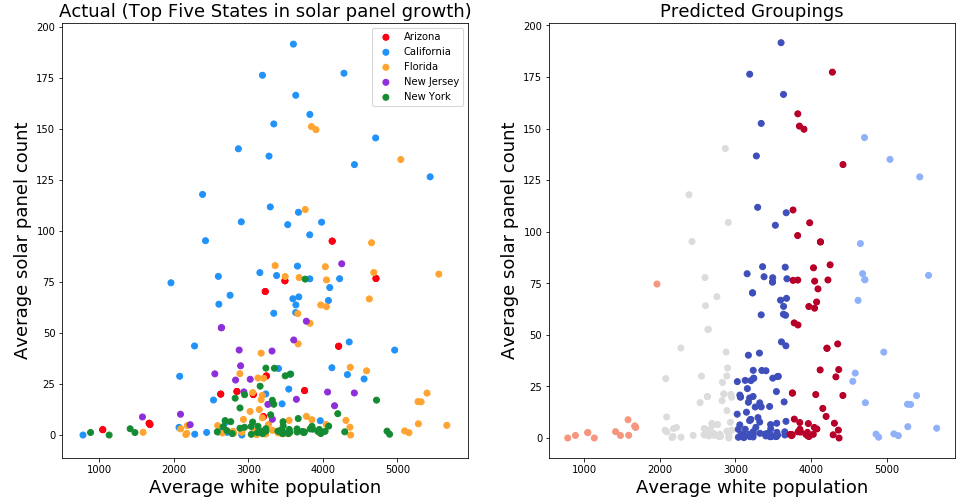


Figure 2. Visualization between solar panel counts and population of white people within the five states

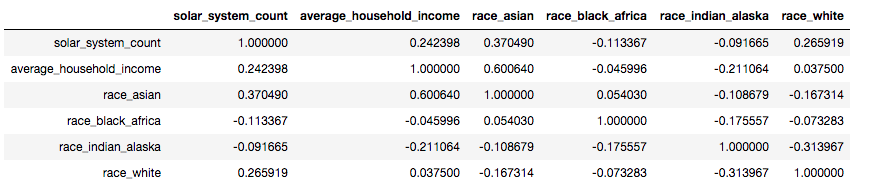


Figure 3. Correlation between demographic factors and solar panel tile counts within top 5 states in solar panel growth (California, Arizona, Florida, New York, New Jersey)

**•Analysis of results**

The first figure of this analysis shows the groupings between the original data and predicted groupings formed by k-means. The visualization shows that on average people with lower average household income are more willing to use solar panels as opposed to people with higher household income. The second figure is a visualization of the average white population within the state’s counties vs average tile counts. The results show that as the average white population increases, so does average solar panel counts.

**Section 4: Discussion & Conclusions**

**•Decisions made**

I made the decision of using only the selected demographic features offered in the dataset for my implementation.

**•Difficulties faced**

Understanding which features to select and those that would fit my model the best was a challenge because the dataset consists of a total of 169 attributes.

**•Things that worked**

Cleaning the dataset was very important because missing values can cause problems such as misleading, inaccurate projections and analysis.

**•Conclusion**

After performing the tasks and analyzing the results, I was able to predict groupings within the given sub dataset and output the correlation between the selected demographic factors and solar panel tile counts. As a result, this analysis can be valuable to solar panel suppliers that want to understand which group of people they should target when pitching their product.

**Section 5: Project Plan / Task Distribution**

**•Project Plan**

In our group, we decided to focus on different aspects of the DeepSolar dataset. The analysis and contribution of each member of our is as follows:

**•Task Distribution**

Fanus was assigned to analyze the DeepSolar dataset solely focusing on the State of California. Fanus performed this task by using different classification and regression algorithms.

Sen’s analysis was to predict the solar panel area using Random Forest Regression and to be able to identify the correlation socioeconomic factors with residential solar deployment density.

Steven’s analysis was to identify the correlation between solar panel count and demographic factors within the top five states of solar panel growth using k-means clustering.

**FULL IMPLEMENTATION OF OUR PROJECT CAN BE FOUND AT A PUBLIC GITHUB REPOSITORY:** [**https://github.com/zs19930830/CMPE255**](https://github.com/zs19930830/CMPE255)