

**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**

Incentive variables

Tract areas with solar system count = 0

Demographic variables

Classifier Algorithm

Merge

DeepSolar Dataset

Tract areas with solar system count >= 1

Environmental variables

Regression Algorithm

Political View variables

Output

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 phyton 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:**

**Analysis 3:**