CS-4331: Data Mining

Project 1

Pritish Ayer

Nicholas Lovera

Jiayu yan

**Contribution:**

Problem Understanding Phase: Pritish Ayer, Nicholas Lovera

Data Preparation Phase: Major Contribution: Nicholas Lovera

Minor Contribution: Pritish Ayer, Jiayu Yan

Exploratory Data Analysis Phase: Major Contribution: Jiayu Yun

Minor Contribution: Pritish Ayer, Nicholas Lovera

Report Created by : Pritish Ayer

For this project, we chose dataset which contained the information of the peoples from the state of New York who switched to the Solar power. The dataset was taken from the official site of the New York state.

Link to the dataset: <https://data.ny.gov/Energy-Environment/Solar-Electric-Programs-Reported-by-NYSERDA-Beginn/3x8r-34rs>

**• Problem Understanding Phase**

Goal:

This dataset has a ton of information which can be used to generate numerous conclusions. However our main goal during this project will be but not limited to:

1) Compare solar energy (‘mean\_value\_KWh\_Annual\_Production’) in New York to average electric power in New York

2) Compare the cost and production of the solar energy and calculate the breakeven point.

3) Compare Residential and Non-residential solar energy production and identify which establishment is better benefitted from the solar energy.

4) Cluster different variables to determine if any groupings occur such as Residential/ less- KWh-Annual-Production or Non-Residential/more-KWh-Annual-Projection.

**• Data Preparation Phase**

Since the selected dataset had an abundant data, we had to prepare it well and only keep the data that we anticipate on using for this project. This dataset originally comprised the data for more than 150,000 projects with 30 variables. A lot of variables were insignificant for this project hence we got rid of many numbers of columns/variables to make the data more manageable with the help of just one line of code:

solarInfo=solarInfo [, -c(1,4,5,6,8,9,13,15,16,27,28,30)]

After that we gave our imported and more manageable dataset an index. Since we are mostly interested in the amount of energy produced by the solar energy system, we changed the Annual production (‘Expected.KWh.Annual.Production’) to numeric because of thousand separator. Now, since we had no outliers in the data, so we did not deal with any of that. However, since the data had much variance, binning the data was a tough task. Finally, after discussing with the team members we concluded to set the labels as "Under 10000","10000 to 100000","Over 100000". After doing some plots and contingency tables we saved the new dataset with new information.

• **Exploratory Data Analysis Phase**

The first step of this phase is reading in the file obtained after the Data preparation phase. After reading the file we move towards Partitioning the data. The function ‘runif()’ is used which randomly draws the numbers between 0 and 1 and assign it to the dataset each with equal probability.

train\_ind<-runif(n) < 0.75

Here ‘n’ generates numbers and 0.75 indicates that 75% of the observations will be true. Now, we divide the dataset into training and test data. After identifying the predictor variable as Project Cost and Incentive and our target variable as Expected KWh Annual, we rewrite the value for all the variables as numeric. After that we plot the graph for project vs expected KWh Annual Production and x-Incentive vs Expected Kwh Annual Production. After obtaining some satisfying results form the graph, we move towards binning the Project-Cost and Incentive variable. Lastly, we write all the new data in a new csv file named as “solarInfoAd2.csv” to apply suitable models in future.