A3: Classification

03.11.2020

Team: The Credibles
Menghong Han, Peihan Tian, Yanghe Liu, Laurence Finch

Part 1. Descriptive Statistics

Variable Description

First of all, we choose the relevant and important variables based on the distribution and boxplot from our last report as well as the economical and environmental significance, they are: "BIOGAS", "BIOMASS", "GEOTHERMAL", "SMALL.HYDRO", "SOLAR.PV", "SOLAR.THERMAL", "WIND.TOTAL", representing power production from various power sources (measured in megawatts). In addition, because seasonality has a large influence on energy generation, we convert the variable “TIMESTAMP” into two categorical variables “Hour” and “Month”.

In order to carry out the classification analysis, we transform our former target variable "SMALL.HYDRO" which is continuous into categorical variable. In order to get appropriate segment thresholds, we used unsupervised K-means clustering method to find the reasonable k and labeled "SMALL.HYDRO" data points accordingly.

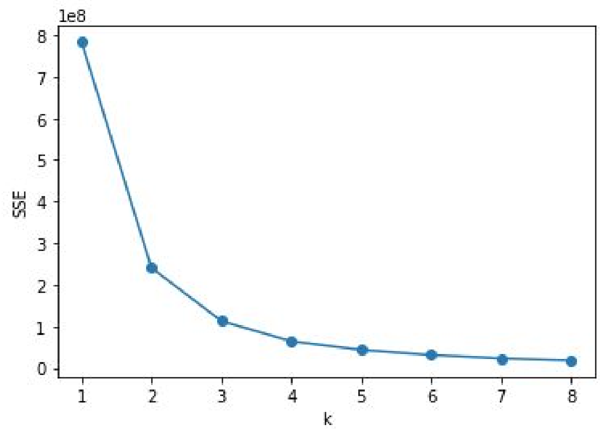

Figure 1. SSE for different k

0	1	2
18482	17872	8550

Figure 2. Number of datapoints inside each segmentation

From the SSE plot, we can see k=3 is the most reasonable choice and based on the labels we segmented the target variable into 3 levels (0,1,2) with the threshold of 236 and 402 after running the clustering several times. Then we rechecked the number of datapoints inside each segmentation and their distributions and we confirmed our labeling result is justifiable.

Variable Type	Variable Name	Description	Type
Important Variables	BIOGAS	Biogas production in MW	Integer
	BIOMASS	Biomass production in MW	Integer
	GEOTHERMAL	Geothermal production in MW	Integer
	SOLAR.PV[1]	Solar Photovoltaic production in MW	Integer
	SOLAR.THERMAL[2]	Solar thermal production in MW	Integer
	WIND.TOTAL	Wind power production in MW	Integer
	Hour	00:00 as 1, totally 24 values	Categorical
	Month	Totally 12 values	Categorical
Target Variable	SMALL.HYDRO3	Small hydro production in MW	Categorical 0 [0,235] 1 [236,401] 2 [402,678]

Figure 3. Variable Description

[1] Solar Photovoltaic (PV) is a technology that converts sunlight (solar radiation) into direct current electricity by using semiconductors. When the sun hits the semiconductor within the PV cell, electrons are freed and form an electric current.

[2] Solar thermal technologies capture the heat energy from the sun and use it for heating and/or the production of electricity[1]. This is different from photovoltaic solar

[3] Small hydro energy production in California is defined as energy production from water related sources, at a facility with a capacity of 30MW or less. More information about small hydro can be found [here](https://www.hydro.org/policy/technology/small-hydro/).

Descriptive Analysis

And then we draw box-and-whisker plots for relevant and important variables.

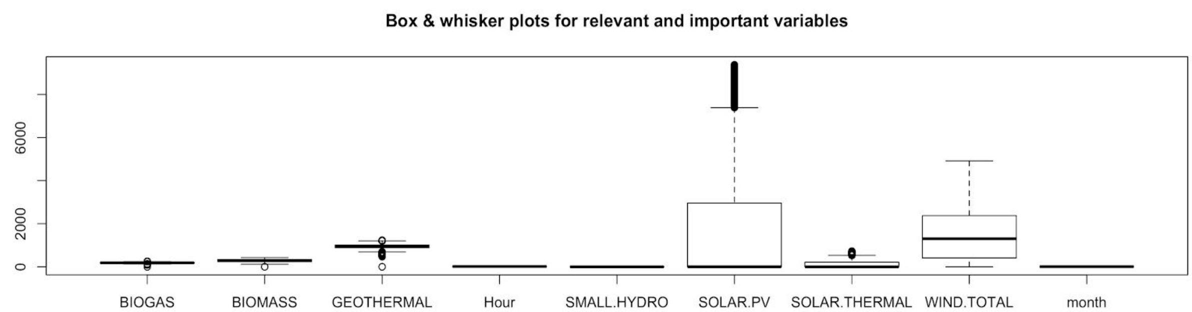

Figure 4. Box-and-whisker plots for predictor variables

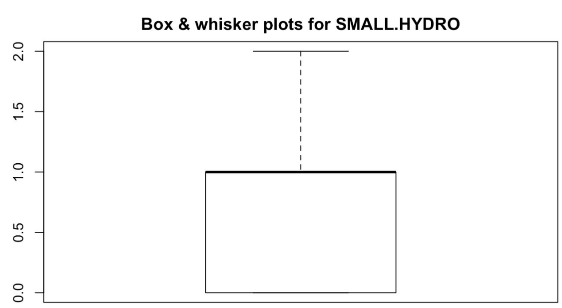

Figure 4. Box-and-whisker plots for target variable

From the plot, we can see that all variables have few outliers except for "SOLAR.PV", which shows a very large right deviation. We found that data highly concentrated on the frequency of 0 and shows positive skewness. However, this situation is reasonable since "SOLAR.PV" is Solar Photovoltaic which works only when the sun rises. Then, from the box chat of "SOLAR.PV" we can see the outliers show a large right deviation. Although Solar Photovoltaic works relatively fewer hours a day, it generates much more energy than others in a short period of time. Therefore, the large number of outliers would not destroy the data quality of "SOLAR.PV" and we should deal with the outliers later under the context of the classification model.

Then we calculate the minimum, maximum, and average (mean, median, mode) and standard deviation and variance of important variables.

VARIABLES	MIN	MAX	MEAN	MEDIAN	MODE	STD	VARIANCE
BIOGAS	0	248	185.7	187	199	19.9	397.9
BIOMASS	0	423	283.6	283	232	59.6	3552
GEOTHERMAL	0	1230	945.3	928	921	89.5	8011.5
SOLAR.PV	0	5558	1491	3	0	2031.2	4125921
SOLAR.THERMAL	0	725	117.3	0	0	118.7	35621.6
WIND.TOTAL	0	4914	1478.7	1301	129	1135	1288768
Month	1	12	6.6	7	12	3.5	12.4
Hour	1	24	12.5	12.5	1	6.9	47.9
SMALL.HYDRO	0	2	0.8	1	0	0.7	0.6

Figure 5. Descriptive statistics of important variables

To figure out potentially linear or curvilinear relationships among variables, we create scatter plots as follows. Since there are too many data points, the scatterplots are hard to read, therefore, we randomly selected 100 variables to find clear patterns.

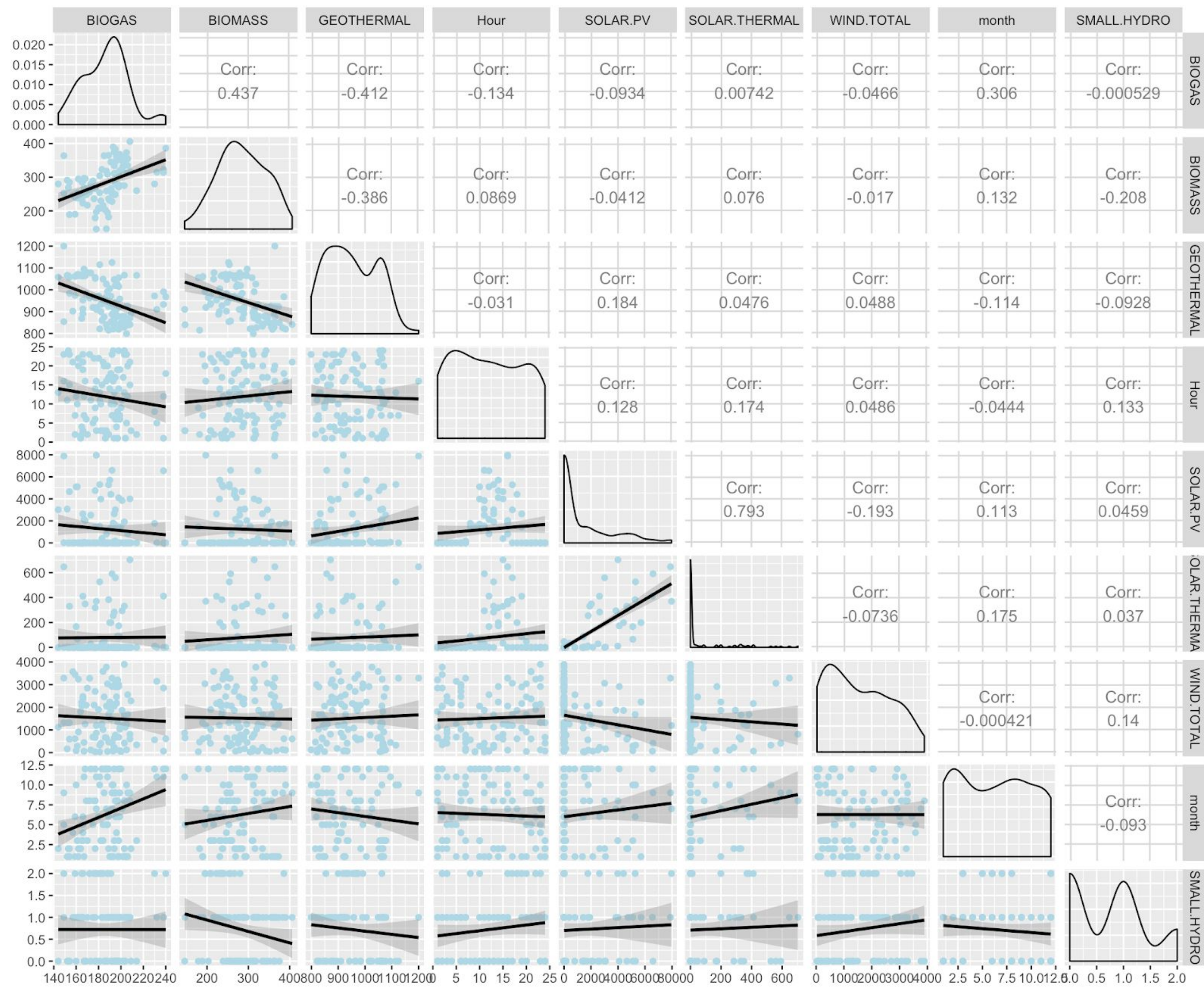

Figure 6. Correlation plots after sampling

Justify target variable

From a statistical perspective, we can see the distribution of "SMALL.HYDRO" is close to normal distribution and the absolute value of correlation coefficients between "SMALL.HYDRO" and other variables are larger relatively, which we can see clearly from the slopes although all linear relationships are not that obvious.

From an policy perspective, as a leader in renewable energy, California has pledged to use only clean sources for electricity, including wind and solar power by 2045, however, one hurdle is energy storage, while small hydro may help the state reach its goal of zero emissions by providing the solution "pumped storage," which uses water in reservoirs at different elevations to smooth the fluctuations of intermittent power from the wind and sun, and makes electricity available when it is needed. Moreover, spinning a turbine using water offers many benefits beyond simply producing electricity. It also offers a tremendous amount of operational flexibility and rapid start/shutdown capabilities.Therefore, it's meaningful to figure out how other energys interact with the production of small hydro which can be a great measurement to evaluate the overall effectiveness of renewable energy.

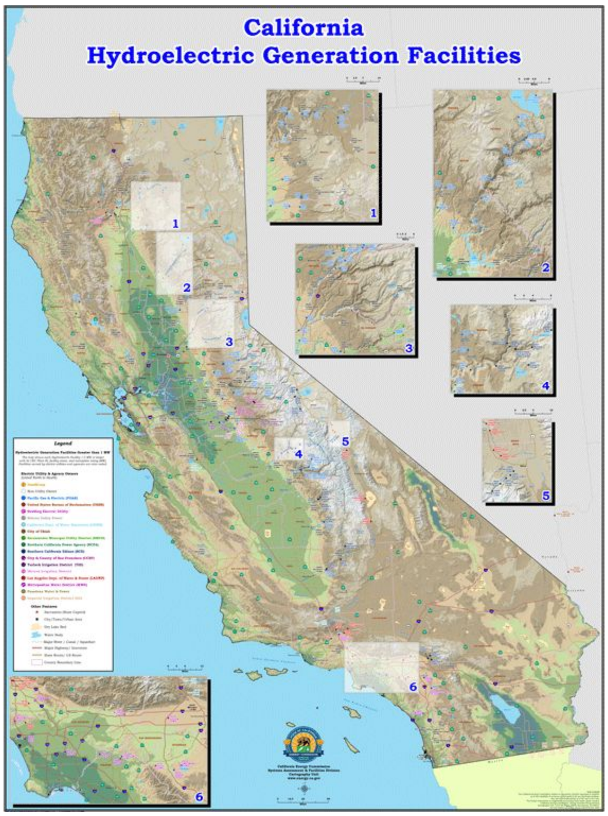

Figure 7. Hydroelectric Generation Facilities greater than 1 MW

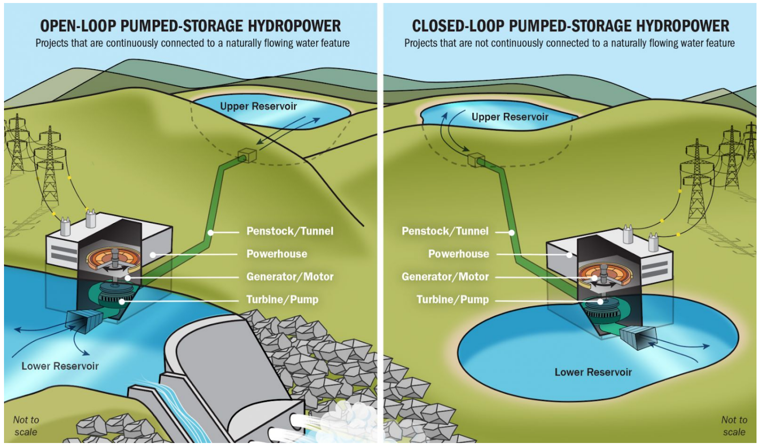

Figure 8. Pumped-storage hydropower (PSH)

From an economic and environmental perspective, due to the advantage illustrated above, California planned to build more hydro plants, however, many professionals questioned the efficiency of hydroelectric especially small hydro, at the meanwhile, some people are worrying about the environmental disruption caused by building new plants, climate advocates say, this would reduce the need to build new solar and wind farms between now and 2030 and as a result, more gas plants would continue to operate, spewing planet-warming pollution into the atmosphere. Therefore, it's urgent to evaluate the effectiveness of small hydro itself to compare with the other renewable energy which would be extremely helpful for economic and environmental decisions.

In general, we chose small hydro as our target variable, after statistical, policy, economic and environmental considerations.

Part 2. Classification

Next step, we will use our cleaned data to build 3 prediction models against the level of small hydrogen (low, medium, high).

Model 1-1: Classification Tree

At first, we built a simple classification tree to label the small hydrogen with 3 levels.

	Training Set	Testing Set
Accuracy	68.21%	67.89%

Figure 9. Accuracy of Classification Tree

From the table, we can see that the difference between training set and testing set is insignificant, which means the model is not overfitting. However, its accuracy rate is a little low, only 67.89%, which is an underfitting situation. We can make our model more complex to avoid this situation.

	low	medium	high
Sensitivity	0.8366	0.6154	0.4884
Specificity	0.757	0.770	0.947
Pos Pred Value	0.6994	0.6663	0.665
Prevalence	0.4033	0.4198	0.1769
Detection Rate	0.3374	0.2584	0.0864
Detection Prevalence	0.4824	0.3877	0.1299
Balanced Accuracy	0.7968	0.6962	0.7177
Neg Pred Value	0.8727	0.7363	0.8960

Figure 10. Confusion Matrix of Classification Tree

Through Table 2 we can see that our model are more likely to have a higher sensitivity in the low 'small hydrogen' class and a higher specificity in the high 'small hydrogen' class. Both of the factors(sensitivity, specificity) are not very good.

Model 1-2: Deeper Classification Tree

Second, we build a more deeper classification tree by adding minsplit, the minimum number of observations in a node for a split to be attempted. We also used a 5-fold cross-validation to make our model more complex.

	Training Set	Testing Set
Accuracy	99.8913%	78.0314%

Figure 11. Accuracy of Deeper Classification Tree

From this table, we can see that the accuracy rate in the training set is 99.89% and is 78.03% in the testing set, which means that there is a potential overfitting problem and we should fix it. After trying different parameters, we still cannot improve the accuracy rate. Thus, we may exchange our model to random forest.

	low	medium	high
Sensitivity	0.8487	0.7390	0.7225
Specificity	0.8858	0.8300	0.9357
Pos Pred Value	0.8340	0.7587	0.7073
Prevalence	0.4033	0.4198	0.1769
Detection Rate	0.3423	0.3102	0.1278
Detection Prevalence	0.3423	0.4089	0.1807
Balanced Accuracy	0.8673	0.7845	0.8291
Neg Pred Value	0.8965	0.8147	0.9401

Figure 12.. Confusion Matrix of Deeper Classification Tree

We can see that the sensitivity and specificity are relatively high compared with the classification tree.

Model 1-3: Random Forest

Before using the random forest, we added more variables, such as the interaction term and dummy variables to improve the complexity of our model. Then we change the parameters of the model to ntree as 500 and mtry as 20. Because we have overall 70 variables, the number of trees and tries are reasonable in the model.

	Training Set	Testing Set
Accuracy	99.42%	84.21%

Figure 13.. Accuracy of Random Forest

We can see that the accuracy in the training set is 99.42% and 84.21% in the testing set. We tried different parameters and the accuracy in the testing set does not change much. We can conclude that the capacity for the random forest is nearly 84.21%. The model will determine the lower limit of the accuracy and the data itself will determine the higher limit of the accuracy.

	low	medium	high
Sensitivity	0.8934	0.8149	0.7747
Specificity	0.9173	0.8670	0.9586
Pos Pred Value	0.8796	0.8159	0.8009
Prevalence	0.4033	0.4198	0.1769
Detection Rate	0.3603	0.3421	0.1371
Detection Prevalence	0.4096	0.4192	0.1711
Balanced Accuracy	0.9054	0.8409	0.8667
Neg Pred Value	0.9272	0.8662	0.9519

Figure 14.. Confusion Matrix of Random Fore

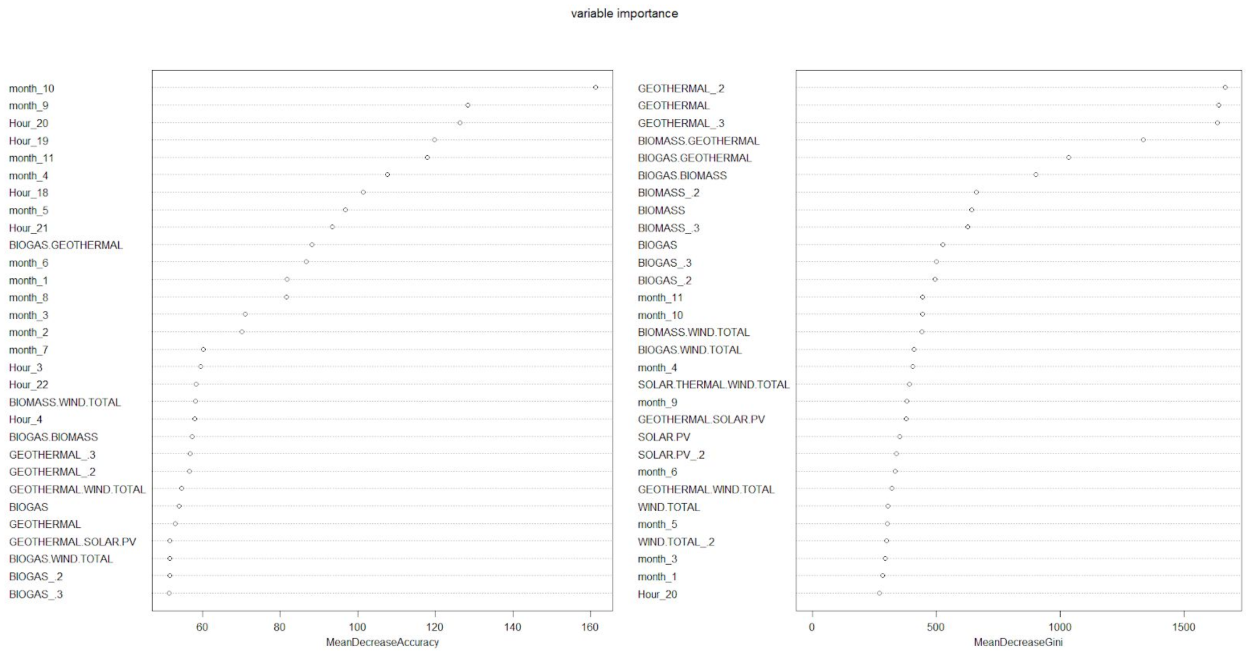

Figure 15.. The Importance of The Variables in Random Forest

From figure 9, we can see that the most important variables are month and hour, which are reasonable because natural energy, especially hydrogen is highly related to weather, temperature and sunlight.

Other variables, such as the geothermal, can also significantly affect the accuracy rate due to its high correlation with hydrogen.

At this time we have already reached the higher limit of the model through changing parameters. If we want to further improve the accuracy, we can only improve our data by creating more variables. Based on the fact that we have included interaction terms, dummy variables and higher order variables, perhaps we have reached the limit of this kind of model.

Model 2: K-Nearest Neighbors

For the K-NN method, we want to see which K gives the best performance. So we run a loop which contains K from 1 to 14. The accuracy plot shows that when K equals 3, this classification performs the best.

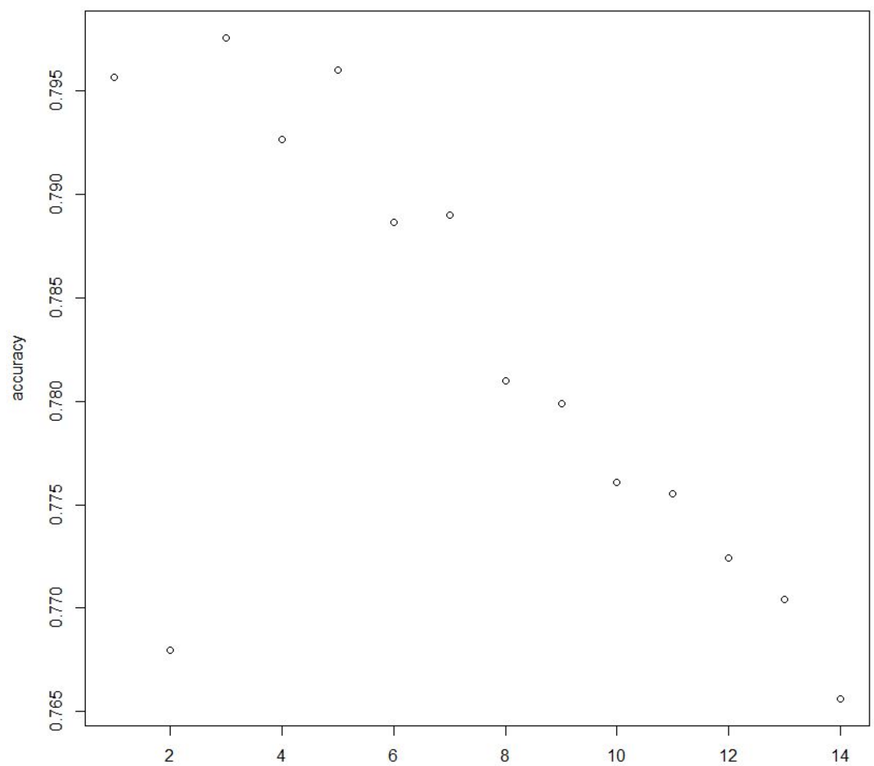

Figure 16.. K-NN accuracy for different K

After determining the best K, we test the accuracy for both training set and testing set. The training set has an accuracy of 89.39% and the testing set has an accuracy of 79.76%. The difference in accuracy for in-sample and out-of-sample data is reasonable. Also, From the confusion matrix of KNN, we can see that the sensitivity, specificity, and F1 scores are all relatively high. Therefore, we can conclude that there's no overfitting here.

	Training Set	Testing Set
Accuracy	89.39%	79.76%

Figure 17.. Accuracy of K-NN

	low	medium	high
Sensitivity	0.8155	0.7822	0.7885
Specificity	0.8876	0.8359	0.9526
Pos Pred Value	0.8364	0.7608	0.7923
Prevalence	0.4133	0.4002	0.1864
Detection Rate	0.3370	0.3131	0.1469
Detection Prevalence	0.4029	0.4115	0.1855
Balanced Accuracy	0.8515	0.8090	0.8705
F1	0.8258	0.7713	0.7904

Figure 18.. Confusion Matrix of K-NN

Model 3: Logistic Regression

We use the clean data from the last step to build our logistic model, since linear and logistic models are similar to each other to some extent.

Given that our target is ordinal categorical variable and glm function cannot deal with it, we built our model with multinom function in the ‘nnet’ package.

When we build the logistic model we have to set one of the levels of the dependant variables as baseline. In this example, we set the ‘medium’ as the baseline and we achieved this by the ‘revel’ function.

Furthermore, usually we get one set of estimates from the model but here we clearly see two sets in two rows. What’s going on? In logistic regression, one level of the dependent variable is taken as reference and separate model coefficients are estimated on the remaining levels. In our case Target Variable has 3 levels (low, medium and high), and by default the medium level is taken as a reference. Therefore, for the remaining two levels low and high we get model coefficients. That’s why in the output above you see the rows for Coefficients are marked 2 and 3.

We first built a logistic regression with all the variables and then checked the significance of each variable.

```
                  BIOGAS       BIOMASS    GEOTHERMAL      SOLAR.PV SOLAR.THERMAL    WIND.TOTAL     BIOGAS_.2     BIOGAS_.3    BIOMASS_.2    BIOMASS_.3
Coefficient 6.083578e-05 4.339504e-04 1.134173e-03 -1.364397e-03 -3.896222e-03 -1.151755e-03 -7.069143e-04 3.823395e-07 5.869538e-04 -8.531825e-07
Std. Errors 9.701598e-13 1.768130e-12 4.826419e-12  5.424245e-11  1.811962e-11  3.787928e-11  1.006762e-10 6.180962e-09 3.438568e-10  1.017199e-09
z stat      6.270697e+07 2.454290e+08 2.349927e+08 -2.515367e+07 -2.150278e+08 -3.040593e+07 -7.021663e+06 6.185761e+01 1.706972e+06 -8.387569e+02
p value     0.000000e+00 0.000000e+00 0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00 0.000000e+00 0.000000e+00  0.000000e+00
            GEOTHERMAL_.2 GEOTHERMAL_.3   SOLAR.PV_.2 SOLAR.THERMAL_.2 SOLAR.THERMAL_.3 WIND.TOTAL_.2       Hour_1       Hour_2       Hour_3
Coefficient -3.840753e-05  2.062058e-08 -1.387825e-07     3.922247e-05    -3.771718e-08 -9.695734e-08 1.111311e-04 1.531812e-04 1.797098e-04
Std. Errors  2.366625e-09  4.246408e-11  4.151732e-09     4.310504e-09     7.023618e-10  1.114485e-08 1.545137e-15 1.335572e-15 1.662585e-15
z stat      -1.622882e+04  4.856006e+02 -3.342761e+01     9.099277e+03    -5.370050e+01 -8.699741e+00 7.192312e+10 1.146933e+11 1.080906e+11
p value      0.000000e+00  0.000000e+00  0.000000e+00     0.000000e+00     0.000000e+00  0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
                  Hour_4       Hour_5       Hour_6        Hour_7        Hour_8        Hour_9       Hour_10      Hour_11      Hour_12      Hour_13
Coefficient 1.838315e-04 1.524619e-04 9.254797e-05 -1.948035e-05 -1.278890e-04 -5.878087e-05 -1.326985e-06 2.321518e-05 3.240433e-05 3.142773e-05
Std. Errors 2.025290e-15 2.309761e-15 2.213805e-15  3.910173e-15  8.591221e-15  3.420534e-15  2.116789e-15 4.729309e-15 7.077480e-15 7.604760e-15
z stat      9.076800e+10 6.600766e+10 4.180494e+10 -4.981967e+09 -1.488601e+10 -1.718471e+10 -6.268858e+08 4.908789e+09 4.578513e+09 4.132639e+09
p value     0.000000e+00 0.000000e+00 0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
                 Hour_14       Hour_15       Hour_16       Hour_17       Hour_18       Hour_19       Hour_20       Hour_21       Hour_22      Hour_23
Coefficient 1.575653e-05 -9.866641e-06 -3.725540e-05 -6.809934e-05 -1.199542e-04 -1.415630e-04 -2.131151e-04 -1.620831e-04 -8.237022e-05 6.520909e-06
Std. Errors 7.398653e-15  6.788040e-15  4.487210e-15  3.097966e-15  7.676151e-15  6.798159e-15  3.040084e-15  2.916419e-15  2.459486e-15 1.846194e-15
z stat      2.129649e+09 -1.453533e+09 -8.302575e+09 -2.198196e+10 -1.562687e+10 -2.082373e+10 -7.010173e+10 -5.557605e+10 -3.349083e+10 3.532082e+09
p value     0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00 0.000000e+00
                  month_1       month_2       month_3       month_4       month_5       month_6       month_7       month_8      month_9     month_10
Coefficient -1.488639e-04 -1.040277e-04 -1.239736e-04 -1.281544e-04 -1.487511e-04 -1.673043e-04 -1.051067e-04 -8.257236e-05 2.936160e-04 3.951367e-04
Std. Errors  1.382408e-14  5.767141e-15  7.070688e-15  4.151302e-15  5.469161e-15  8.998545e-15  5.545913e-15  9.028137e-15 5.954139e-15 1.202969e-14
z stat      -1.076844e+10 -1.803801e+10 -1.753346e+10 -3.087089e+10 -2.719815e+10 -1.859237e+10 -1.895210e+10 -9.146114e+09 4.931293e+10 3.284679e+10
p value      0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00 0.000000e+00 0.000000e+00
                month_11 BIOGAS.BIOMASS BIOGAS.GEOTHERMAL BIOGAS.SOLAR.PV BIOGAS.SOLAR.THERMAL BIOGAS.WIND.TOTAL BIOMASS.GEOTHERMAL BIOMASS.SOLAR.PV
Coefficient 4.661608e-04  -1.230001e-04      2.477630e-04    4.658392e-06        -1.408844e-05      7.736171e-06      -9.891973e-05     2.205048e-06
Std. Errors 6.622447e-15   1.703708e-10      5.481915e-10    1.910663e-08         3.569912e-09      7.147417e-09       1.129468e-09     7.256159e-08
z stat      7.039101e+10  -7.219551e+05      4.519644e+05    2.438102e+02        -3.946437e+03      1.082373e+03      -8.758078e+04     3.038864e+01
p value     0.000000e+00   0.000000e+00      0.000000e+00    0.000000e+00         0.000000e+00      0.000000e+00       0.000000e+00     0.000000e+00
            BIOMASS.SOLAR.THERMAL BIOMASS.WIND.TOTAL GEOTHERMAL.SOLAR.PV GEOTHERMAL.SOLAR.THERMAL GEOTHERMAL.WIND.TOTAL SOLAR.PV.SOLAR.THERMAL
Coefficient         -3.425420e-05      -1.566840e-06        6.064032e-07            -9.607342e-07          3.045691e-07           1.454626e-06
Std. Errors          7.616522e-09       1.462658e-08        3.093545e-08             1.565030e-08          3.781043e-08           6.461531e-08
z stat              -4.497355e+03      -1.071228e+02        1.960221e+01            -6.138757e+01          8.055162e+00           2.251210e+01
p value              0.000000e+00       0.000000e+00        0.000000e+00             0.000000e+00          8.881784e-16           0.000000e+00
            SOLAR.PV.WIND.TOTAL SOLAR.THERMAL.WIND.TOTAL
Coefficient        2.351369e-08            -2.975932e-07
Std. Errors        7.774949e-09             5.374907e-08
z stat             3.024288e+00            -5.536713e+00
p value            2.492189e-03             3.082016e-08
```

Figure 19. Significance of Each Variable in The Logistic Regression Target Variable Low

	BIOGAS	BIOMASS	GEOTHERMAL	SOLAR.PV	SOLAR.THERMAL	WIND.TOTAL	BIOGAS_.2	BIOGAS_.3	BIOMASS_.2	BIOMASS_.3
Coefficient	2.354282e-04	9.635942e-04	1.235354e-04	6.030353e-04	-8.224626e-04	-1.562303e-03	-2.859346e-06	4.942246e-07	-4.772929e-04	5.970938e-07
Std. Errors	8.510766e-13	6.681935e-13	4.004468e-12	2.336214e-11	1.349823e-11	4.933707e-11	9.821348e-11	6.034962e-09	9.920279e-11	1.297535e-09
z stat	2.766240e+08	1.442089e+09	3.084938e+07	2.581251e+07	-6.093116e+07	-3.166591e+07	-2.911358e+04	8.189358e+01	-4.811285e+06	4.601754e+02
p value	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00

	GEOTHERMAL_.2	GEOTHERMAL_.3	SOLAR.PV_.2	SOLAR.THERMAL_.2	SOLAR.THERMAL_.3	WIND.TOTAL_.2	Hour_1	Hour_2	Hour_3
Coefficient	2.938164e-05	-2.076974e-08	3.836976e-08	-2.860003e-06	2.707397e-09	4.168952e-08	-9.525842e-05	-1.192811e-04	-1.215926e-04
Std. Errors	2.106194e-09	6.132369e-11	4.016521e-09	3.190439e-09	7.550578e-10	1.282329e-08	1.185477e-15	1.444826e-15	1.282179e-15
z stat	1.395011e+04	-3.386904e+02	9.552984e+00	-8.964295e+02	3.585681e+00	3.251078e+00	-8.035449e+10	-8.255742e+10	-9.483277e+10
p value	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	3.361992e-04	1.149682e-03	0.000000e+00	0.000000e+00	0.000000e+00

	Hour_4	Hour_5	Hour_6	Hour_7	Hour_8	Hour_9	Hour_10	Hour_11	Hour_12	Hour_13
Coefficient	-1.127808e-04	-9.540501e-05	-2.000627e-05	5.311110e-05	8.741869e-05	1.563668e-05	-5.009570e-05	-8.508988e-05	-9.663137e-05	-9.520788e-05
Std. Errors	1.146719e-15	1.088381e-15	9.976582e-16	3.204073e-15	7.644624e-15	2.720415e-15	2.365888e-15	1.515253e-15	2.842618e-15	3.641440e-15
z stat	-9.835080e+10	-8.765771e+10	-2.005324e+10	1.657612e+10	1.143532e+10	5.747904e+09	-2.117416e+10	-5.615556e+10	-3.399379e+10	-2.614567e+10
p value	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00

	Hour_14	Hour_15	Hour_16	Hour_17	Hour_18	Hour_19	Hour_20	Hour_21	Hour_22	Hour_23
Coefficient	-8.508106e-05	-5.187557e-05	2.228006e-05	9.731723e-05	1.874495e-04	1.831534e-04	1.861604e-04	1.365047e-04	7.257516e-05	9.827755e-06
Std. Errors	3.286736e-15	2.850916e-15	1.901234e-15	4.981475e-15	1.239281e-14	1.187943e-14	1.783888e-15	1.944727e-15	2.001124e-15	1.715041e-15
z stat	-2.588619e+10	-1.819610e+10	1.171873e+10	1.953583e+10	1.512567e+10	1.541769e+10	1.043566e+11	7.019219e+10	3.626720e+10	5.730332e+09
p value	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00

	month_1	month_2	month_3	month_4	month_5	month_6	month_7	month_8	month_9	month_10
Coefficient	-9.491031e-05	-1.311509e-04	2.129143e-04	5.563338e-05	1.249570e-04	3.027798e-04	1.139313e-04	9.243438e-05	-2.084392e-04	-2.328980e-04
Std. Errors	7.972749e-15	3.569261e-15	5.102798e-15	5.887404e-15	8.229720e-15	1.550687e-14	9.593205e-15	6.589448e-15	1.026810e-14	9.930426e-15
z stat	-1.190434e+10	-3.674455e+10	4.172500e+10	9.449561e+09	1.518363e+10	1.952553e+10	1.187625e+10	1.402764e+10	-2.029968e+10	-2.345297e+10
p value	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00

	month_11	BIOGAS.BIOMASS	BIOGAS.GEOTHERMAL	BIOGAS.SOLAR.PV	BIOGAS.SOLAR.THERMAL	BIOGAS.WIND.TOTAL	BIOMASS.GEOTHERMAL	BIOMASS.SOLAR.PV
Coefficient	-1.637487e-04	2.542655e-04	-1.246521e-04	-3.896717e-06	5.331836e-06	2.229428e-06	7.308742e-05	-1.190528e-06
Std. Errors	7.174472e-15	8.422301e-11	5.786860e-10	1.050207e-08	2.646254e-09	9.805646e-09	4.932975e-10	5.021974e-08
z stat	-2.282380e+10	3.018955e+06	-2.154054e+05	-3.710427e+02	2.014862e+03	2.273617e+02	1.481609e+05	-2.370637e+01
p value	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00

	BIOMASS.SOLAR.THERMAL	BIOMASS.WIND.TOTAL	GEOTHERMAL.SOLAR.PV	GEOTHERMAL.SOLAR.THERMAL	GEOTHERMAL.WIND.TOTAL	SOLAR.PV.SOLAR.THERMAL
Coefficient	1.002287e-05	-3.771869e-06	3.410615e-07	-4.852500e-06	2.336295e-06	2.893640e-07
Std. Errors	4.708483e-09	1.831667e-08	2.633313e-08	1.269127e-08	4.738713e-08	6.422305e-08
z stat	2.128684e+03	-2.059255e+02	1.295180e+01	-3.823493e+02	4.930232e+01	4.505611e+00
p value	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	6.618233e-06

	SOLAR.PV.WIND.TOTAL	SOLAR.THERMAL.WIND.TOTAL
Coefficient	-2.689318e-08	5.888416e-07
Std. Errors	7.527735e-09	8.351796e-08
z stat	-3.572546e+00	7.050479e+00
p value	3.535269e-04	1.783018e-12

Figure 20. Significance of Each Variable in The Logistic Regression Target Variable High

From the above plots, we get the coefficients, p-value of target variable low and high. We can see that all the variables are significant at the 5% level.

For the coefficients, we can see that month-2, month-1, hour-1 to hour-6 all are negative correlated with the small hydrogen's high level.

Correspondingly, the coefficients of month-2,month-1, hour-1 to hour-6 all are positively correlated with the small hydrogen's low level, which is reasonable and makes sense. When temperature is low and weather is cold, the production of small hydros is more likely to be in the low class and when temperature is high and weather is warm, it is more likely to be in the high class.

We do not include the intercept because it might cause multicollinearity.

Identify outliers

We have already removed obvious outliers in the regression process. Compared the prediction and the ground truth we found out the most obvious outliers. After removing 12 data points, which is 0.2% data points from the dataset, the accuracy increased also the some of the leverage points left in since extreme leverage points are removed.

Interpretation of odd ratio

Although we can calculate change of probability when one of variables changes holding others constant, we usually do not do that, for it is not intuitive, straightforward and hard to interpret. However, we do have a measurement that is easy to interpret—odd ratio.

This ratio of the probability of choosing low over the baseline that is medium is referred to as relative risk (often described as odds). However, the output of the model is the log of odds. To get the relative risk IE odds ratio, we need to exponentiate the coefficients.

```
> exp(coef(multinom.fit))
       BIOGAS  BIOMASS GEOTHERMAL  SOLAR.PV SOLAR.THERMAL WIND.TOTAL BIOGAS_.2 BIOGAS_.3 BIOMASS_.2 BIOMASS_.3 GEOTHERMAL_.2 GEOTHERMAL_.3 SOLAR.PV_.2
low  1.000061 1.000434   1.001135 0.9986365     0.9961114  0.9988489 0.9992933         1  1.0005871  0.9999991     0.9999616             1   0.9999999
high 1.000235 1.000964   1.000124 1.0006032     0.9991779  0.9984389 0.9999971         1  0.9995228  1.0000006     1.0000294             1   1.0000000
     SOLAR.THERMAL_.2 SOLAR.THERMAL_.3 WIND.TOTAL_.2    Hour_1    Hour_2    Hour_3    Hour_4    Hour_5   Hour_6    Hour_7    Hour_8    Hour_9   Hour_10
low         1.0000392                1     0.9999999 1.0001111 1.0001532 1.0001797 1.0001838 1.0001525 1.000093 0.9999805 0.9998721 0.9999412 0.9999987
high        0.9999971                1     1.0000000 0.9999047 0.9998807 0.9998784 0.9998872 0.9999046 0.999980 1.0000531 1.0000874 1.0000156 0.9999499
       Hour_11   Hour_12   Hour_13   Hour_14   Hour_15   Hour_16   Hour_17   Hour_18   Hour_19   Hour_20   Hour_21   Hour_22  Hour_23   month_1
low  1.0000232 1.0000324 1.0000314 1.0000158 0.9999901 0.9999627 0.9999319 0.9998801 0.9998584 0.9997869 0.9998379 0.9999176 1.000007 0.9998511
high 0.9999149 0.9999034 0.9999048 0.9999149 0.9999481 1.0000223 1.0000973 1.0001875 1.0001832 1.0001862 1.0001365 1.0000726 1.000010 0.9999051
       month_2  month_3   month_4   month_5   month_6   month_7   month_8   month_9  month_10  month_11 BIOGAS.BIOMASS BIOGAS.GEOTHERMAL
low  0.9998960 0.999876 0.9998719 0.9998513 0.9998327 0.9998949 0.9999174 1.0002937 1.0003952 1.0004663       0.999877         1.0002478
high 0.9998689 1.000213 1.0000556 1.0001250 1.0003028 1.0001139 1.0000924 0.9997916 0.9997671 0.9998363       1.000254         0.9998754
     BIOGAS.SOLAR.PV BIOGAS.SOLAR.THERMAL BIOGAS.WIND.TOTAL BIOMASS.GEOTHERMAL BIOMASS.SOLAR.PV BIOMASS.SOLAR.THERMAL BIOMASS.WIND.TOTAL
low        1.0000047            0.9999859          1.000008          0.9999011        1.0000022             0.9999657          0.9999984
high       0.9999961            1.0000053          1.000002          1.0000731        0.9999988             1.0000100          0.9999962
     GEOTHERMAL.SOLAR.PV GEOTHERMAL.SOLAR.THERMAL GEOTHERMAL.WIND.TOTAL SOLAR.PV.SOLAR.THERMAL SOLAR.PV.WIND.TOTAL SOLAR.THERMAL.WIND.TOTAL
low             1.000001                0.9999990              1.000000               1.000001                   1                0.9999997
high            1.000000                0.9999951              1.000002               1.000000                   1                1.0000006
```

Figure 21. Log Ratio of The Target Variable Low and High

Comparing two logit models, coefficients of each predictor are same, but intercepts are different. For each predictor, compare the odd ratio against 1.0. Less than 1.0 means a decrease of 1.0-odds ratio percent. Greater than 1.0 means an increase of odd ratio-1.0 percent vs. the baseline category.

The relative risk ratio for a one-unit increase in the variable BIOGAS is1.00061 for being in class low vs being in class medium and 1.000235 for being in class high vs being in class medium.

For another instance, the relative risk ratio for a one-unit increase in the variable BIOMASS is1.000434 for being in class low vs being in class medium and 1.000964 for being in class high vs being in class medium.

The other benefits of interpreting odds instead of total percentage is that statements such as those above are true for any values of X_1. However, the change in probability, p, for a unit increase in a particular predictor is not a constant—it depends on the specific values of the predictor variables.

Accuracy of the model:

	Training Set	Testing Set
Accuracy	59.55%	32.48%

Figure 22. Accuracy of Logistic Regression

We can see that the logistic model performs the worst among the 3 models(Random Forest, KNN, Logistic) . It only has an accuracy rate of 59.55% in the training set and 32.48% in the testing set.

Part3. Conclusion

Comparison of different classification methods

	Random Forest	KNN	Logistics Regression
Accuracy for testing	84%	80%	32%

Figure 23. Accuracy of 3 Best Models

From the accuracy of the three best models from their category, we found Random Forest and KNN have relatively high accuracy which shows the reinforcement of each other, while the result of Logistics Regression can be contradict. The difference can be reasonable since different models have different algorithms to identify boundaries.

Finally, we can conclude that the Random Forest performs the best among the three, which has an accuracy rate of nearly 85% in the testing set.

Reflections - Insights for Business and Policy Leaders

The accurate classification of high, medium and low instances of small hydro power generation is a valuable result for businesses and policy leaders alike.

For both business leaders and local governments involved in energy planning and infrastructure, who are invested in small hydro energy producing areas, understanding when an area is likely to be a high production area can be used to attract further investments into both small hydro facility production and related businesses like energy storage and distribution. Classification into three clear categories, “low”, “medium” and “high”, provides a simple interface with complex classification methods behind, that is ready to go for investors and politicians.

The classification is also useful for investigating further idiosyncratic reasons for low or high energy production, that is otherwise unseen in our data. This could be old generators in some plants, or worse functioning generators at certain times or places. Similarly, investigating high producing instances could lead to insight about what factors make small hydro production greater than other areas - insight that could be used to increase the efficiency of small hydro production elsewhere.

Furthermore, and perhaps most valuably, identifying which areas are high energy producing areas, compared to low and medium producing areas is very useful for energy and infrastructure planners, to enable them to focus new plant production in high production areas. This would save resources and be more energy efficient, delivering financial and environmental benefits. The classification of low and medium energy producing areas, might encourage energy planners to evaluate the value of maintaining existing plants, compared to expanding the high producing areas, or developing facilities in areas classified as high production.

Reference:

https://ww2.energy.ca.gov/maps/powerplants/hydroelectric_facilities.html

https://www.voanews.com/usa/power-plants-create-giant-water-battery

https://www.energy.gov/eere/water/pumped-storage-hydropower

https://www.intechopen.com/books/renewable-hydropower-technologies/prospects-of-small-hydropower-technology