

Final Project

04.14..2020

Team: The Credibles

Menghong Han, Peihan Tian, Yanghe Liu, Laurence Finch



Part 1. Introduction

Introduction – Renewable energy sources in California

The state of California has the goal of 100% “zero-carbon” sourced power by 2045. A preliminary target is for 60% of its power to be from renewable sources by 2030. Our data set contains hourly data on the amount of energy produced from various sources of renewable energy including solar.

The business relevance of renewable energy sources is obvious. Given California’s, and other states’/countries’ emissions reduction targets there is growing demand for renewables and there are opportunities for high growth.

Climate change in general, driven by greenhouse gas emissions partly from fossil fuel powered energy companies is a very pressing issue for businesses as well as society more broadly. The implications of climate change are far-reaching and will affect all industries from agriculture to energy as environmental conditions worsen, consumer awareness and preferences change and environmental disasters become more frequent. These challenges will pose a threat to some businesses, and offer opportunities to others.

Guiding Questions and Avenues of Inquiry

1. Seasonality in energy production

We were curious about how different energy sources production level could be affected by the time of day and time of year, to see if there was any seasonality in the production that could potentially be used for better forecasting or to understand production variation. Fortunately, our data contained a timestamp variable containing the date and time of the observation. After transforming our data to include time and month indicators, we used these new variables to try to improve our predictive modelling of the small hydro energy source as an example of the value of including temporal variables in predictions.

2. Complementary / non-complementary energy sources

We were also interested in how different energy sources production levels might covary and to see which energy sources are complementary and which are non-complementary: which energy sources production levels increase at the same time and which tend to decrease when the other is increasing. We did not have a clear idea what sources would be complementary, so started down this avenue with little idea of where it would lead. Again, our predictive modelling described below provided an opportunity to walk down this avenue.

3. Valuable and practical insights for energy planners and investors

We believed that to be useful to policy makers and investors, our analysis and results had to be accessible and understandable. To be able to predict the general level of energy output of an energy source based on the output level of other energy sources and temporal indicators would be valuable in directing energy plant production by building plants in locations that are likely to produce high energy levels. Our classification model, described below, offered the perfect opportunity to do this. We used the small hydro energy source, classifying it into low, medium and high output classes. This provided an accessible output, that is easily understood by a non-expert audience.

Data source

The data set is from Kaggle

<https://www.kaggle.com/cheedcheed/california-renewable-production-20102018>

The original source is California ISO (CISO): an independent non-profit, public-benefit corporation power grid operator, who supply 80% of California's power lines. The nature of the CISO and their transparent and open approach to their data makes us believe that this is a highly reliable source of data.

Our initial inspection found that there are some missing values, especially in the solar variable. CISO states that the data is unverified raw data and is not intended to be used as the basis for operational or financial decisions, so we knew from the start that it would be necessary to wrangle it.

Data Preparation

Initially the data contained 10 variables: TIMESTAMP, BIOGAS, GEOTHERMAL, Hour, SMALL HYDRO, SOLAR, SOLAR PV, SOLAR THERMAL and WIND TOTAL.

As we went through the data, the first step we did was to check the extreme values of each variable. Since the dataset contains hourly reports on the power production from various sources, all numerical variables should have contained positive numbers. However, we found out there's one negative number in the variable WIND TOTAL, so we replace it with 0.

Then, as we found out variable SOLAR missed around 70 percent of data, we chose to delete the whole column. Also, we discovered that there's a pattern for the missing values in SOLAR PV and SOLAR THERMAL. Data of SOLAR PV and SOLAR THERMAL are missing at the same time, and all missing values combine to be in a whole missing day. So, the missing data of solar energy seems only due to randomness. Therefore, we just deleted all missing rows of SOLAR PV and SOLAR THERMAL.

As a result, since all missing values are missing completely at random, we believe our data is relatively clean.

Finally, to take advantage of potential seasonality in energy production that we suspected, we converted the variable “TIMESTAMP” into two categorical variables “Hour” and “Month”.

Post-preparation Variable Description:

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, 24 values in total	Categorical
	Month	12 values in total	Categorical
Target Variable	SMALL.HYDRO ³	Small hydro production in MW	Integer

Figure 1. 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 Statistics

First of all, we choose the relevant and important variables for our analysis based on the distributions and boxplot included below. They are: BIOGAS, BIOMASS, GEOTHERMAL, SMALL.HYDRO, SOLAR.PV, SOLAR.THERMAL, WIND.TOTAL, representing power production from various power sources (measured in megawatts) and also Hour and Month.

We found that all energy variables fitted approximately well with a normal distribution, with the exception of SOLAR.PV and WIND.TOTAL.

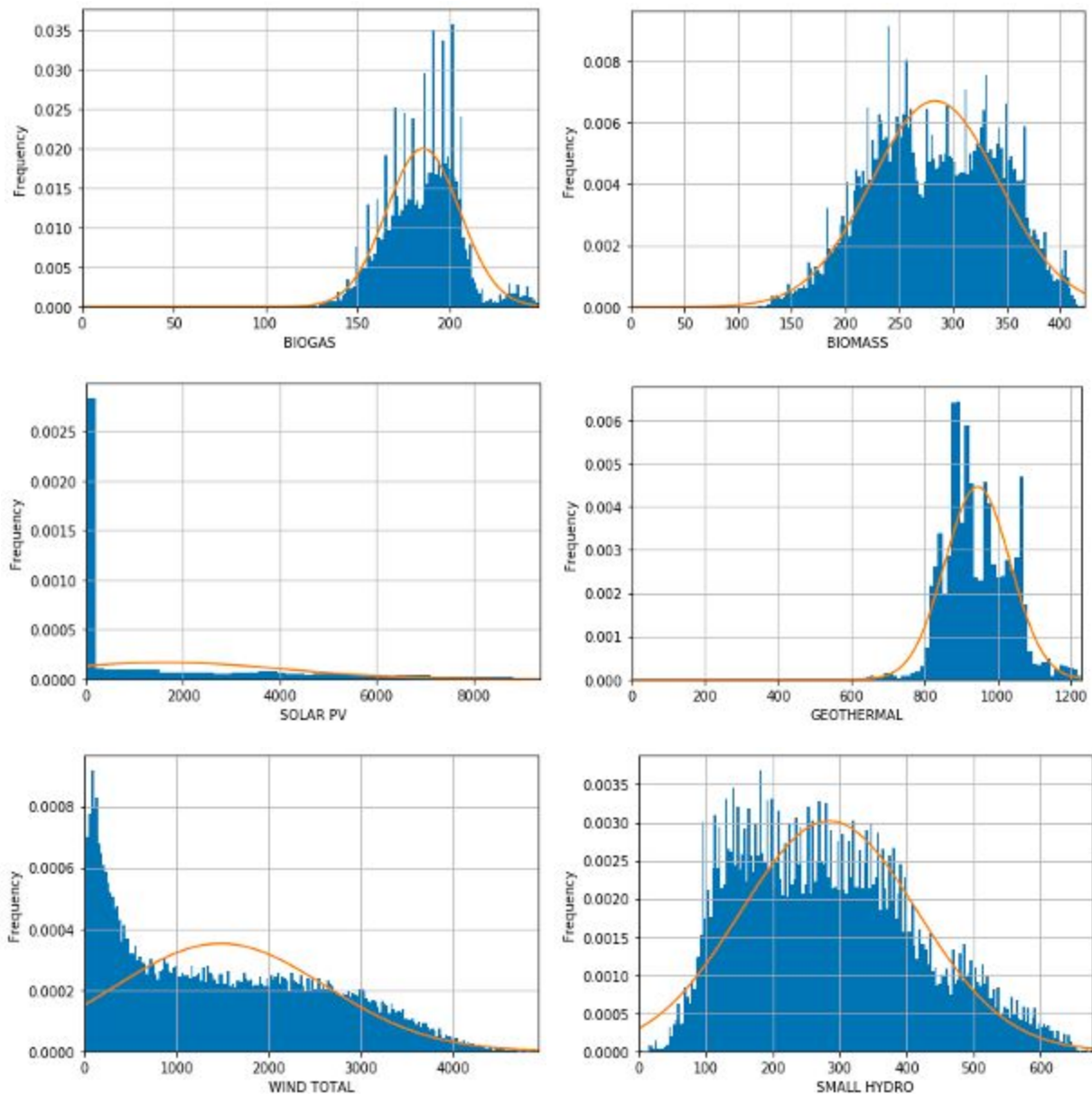


Figure 2. Variable Distributions

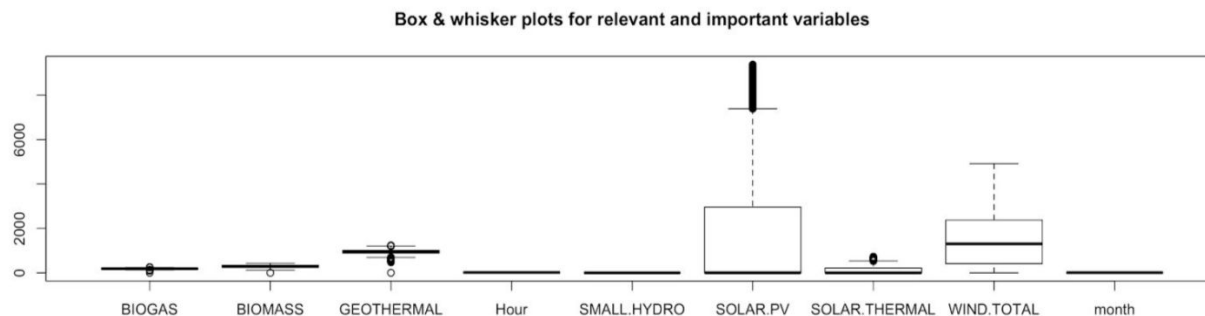


Figure 3. Variable Box-and-whisker plot

From the distribution charts, we can see the distribution of the first 4 variables, "BIOGAS", "BIOMASS", "GEOTHERMAL", "SMALL.HYDRO" are relatively close to normal distribution compared to the distribution of the last two variables "SOLAR.PV" and "WIND.TOTAL". Let's look at the first 4 variables in detail.

From the statistic summary, the distribution of "BIOGAS" presents a slightly negative skewness and a kurtosis slightly larger than 3 (skewness = -0.03, kurtosis= 0.81), "BIOMASS" shows negative skewness and a kurtosis slightly less than 3 (skewness = -0.06, kurtosis= -0.78), "SMALL.HYDRO" shows a large positive skewness and a kurtosis slightly less than 3 (skewness = 0.45, kurtosis= -0.52), "GEOTHERMAL" shows a slightly positive skewness and a kurtosis larger than 3 (skewness = 0.14, kurtosis= 0.92). The parameters show the distributions of the above variables are rather close to normal distribution. Furthermore, based on the box plots of those 4 variables below we can see fewer outliers therefore we conclude that the data quality of the first 4 variables seems good in general.

From the distribution of "SOLAR.PV", we found that data highly concentrated on the frequency of 0 and shows positive skewness. However, this makes sense since "SOLAR.PV" is Solar Photovoltaic which works only when the sun rises. From the box chat of "SOLAR.PV" we can see the outliers show a large right deviation, that is to say 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".

The distribution of "WIND.TOTAL" is similar. From evening to the following morning and during noon, there would not be so much wind which explains the positive skewness of the distribution. From the box chat of "WIND.TOTAL", we can also see right deviation of outliers which illustrate the comparatively high intensity of wind power during a short period of time. Therefore, the data quality of "WIND.TOTAL" seems good.

From the box-and-whiskers 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 regression modelling.

Then we calculated 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
SMALL.HYDRO	0	678	284.3	272	134	132.1	17454.3
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

Figure 4. Variable key statistics

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

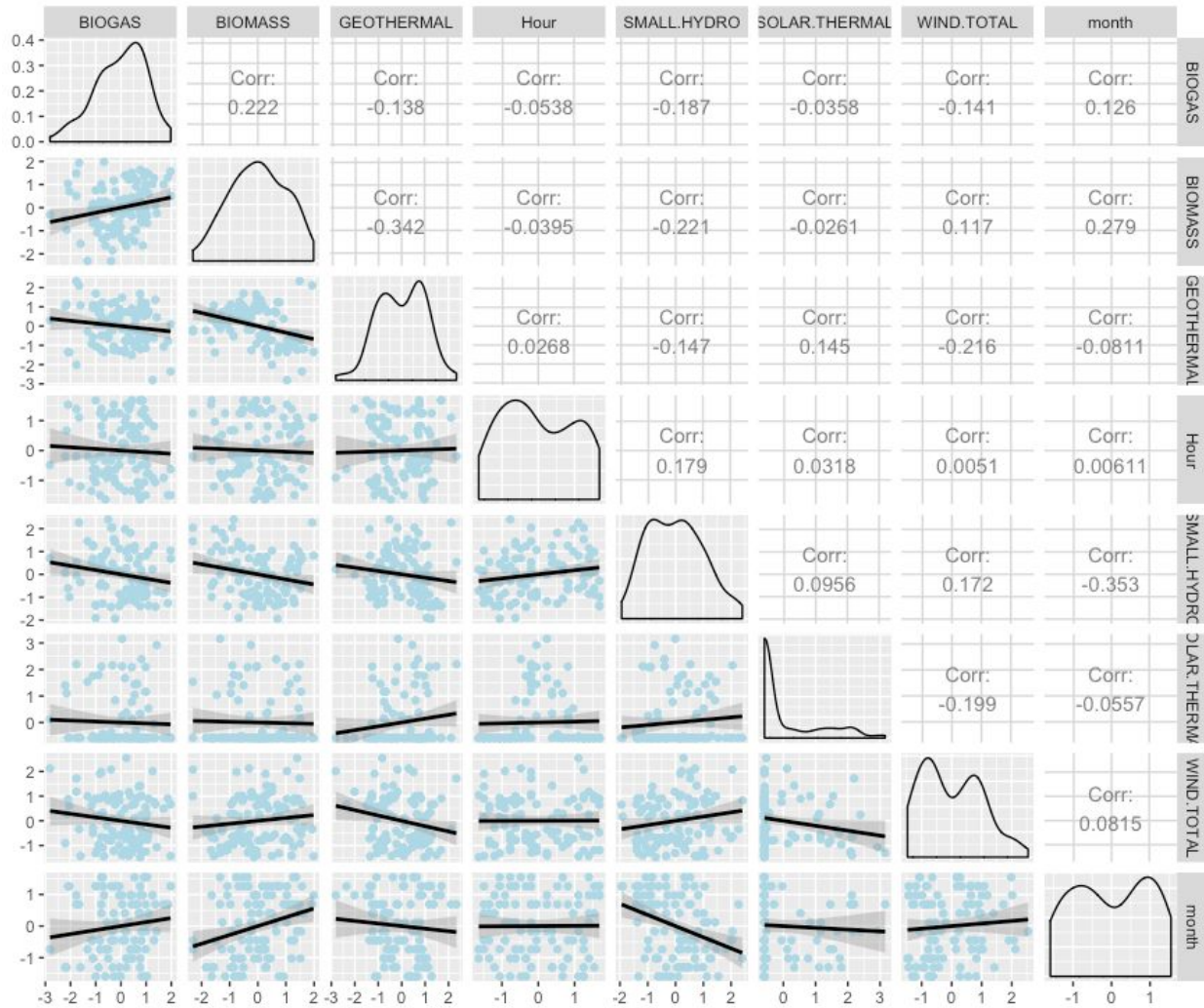



Figure 5. Correlation plots after sampling

Target Variable Justification

We chose small hydro as our target variable, after statistical, policy and economic considerations.

From a statistical perspective, from the descriptive statistics above 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.

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.

Part 2. Model Comparison

Hand-Crafted Models

Data leakage

Data leakage is when information from outside the training dataset is used to create the model. Our datasource is from Kaggle, and most data leakage cases in Kaggle are from deliberately using information outside the training dataset. In other words, the data itself is less likely to have this kind of problem. What is more, one important sign of data leakage is an overly optimistic model, which is not really in our case because our model only has an accuracy of nearly 90%, acceptable but not exaggerative.

Outliers management

To remove outliers, we set 5% and 95% confidence intervals and only consider data within this range. However, we must take special cases into consideration such as the feature 'solar.pv', most of whose values are zero but we can tell by common knowledge that they are not outliers since Solar Photovoltaic works only when the sun rises. Moreover, we will further remove model sensitive outliers accordingly.

Numeric model:

Multi-linear Regression

The target variable 'small-hydro' is numeric; we are interested in how the amount of small hydro energy production can be predicted using the other type of renewable energy and some time and month indicator variable. We then fit a linear regression model, regressing the target variable, SMALL.HYDRO on all other variables. We used a forward stepwise fit to train our model. We also tried the backward and exhaustive regression, and the backward method has the highest R squared.

As a result, we chose backward regression for further prediction. Then, we make predictions for the validation set.

Looking at specific parameter estimates, we see that there is a strong positive relationship between biogas production and small-hydro production (coefficient of 47.48 on BIOGAS, meaning an increase in 1 unit of biogas production predicts an increase of 47.48 units of small hydro production, not considering the small polynomial terms), although this effect gets slightly weaker at higher levels of biogas production, seen by the negative coefficient on the $BIOGAS^2$ coefficient.

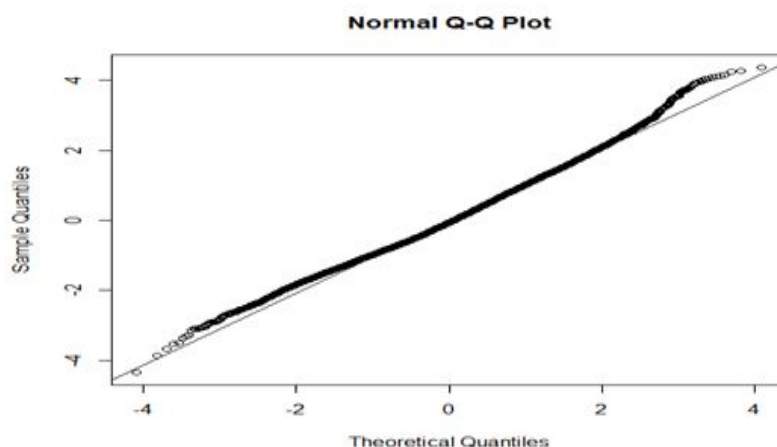
SMALL.HYDRO is also predicted increasing with BIOMASS energy production (coefficient of 1.91, meaning an increase of one unit of biomass energy predicts a 1.91 unit increase in

small hydro production, not including the small polynomial terms) and decreasing in GEOTHERMAL (coefficient of -5.53, meaning an increase of one unit of geothermal energy predicts a decrease of 5.53 units of small hydro, not considering the small polynomial terms). The coefficients on SOLAR.PV, SOLAR.THERMAL and WIND.TOTAL are small, so are not very correlated with small hydro power production. (SOLAR.PV and SOLAR.THERMAL coefficients are not statistically significant, possibly because of their high collinearity)

The coefficients on the polynomial terms are generally negative, meaning that small hydro power is predicted to be increasing slightly less at higher levels of alternative renewable energy production, given the generally positive relationships between small hydro and alternative power sources. The coefficient on BIOGAS_2 is -0.2616 meaning that for an increase of one unit of biogas production, small hydro production is predicted to be 0.2616 x biogas energy production level, not including the standalone effect of biogas production. Similarly, the coefficient on BIOMAS_2 is -0.04421, meaning that a one unit increase in biomass energy production predicts a 0.04421 x biomass energy production level in small hydro, not considering the standalone effect of biomass energy production on small hydro. These findings are in agreement with our intuition that higher energy production from other sources could take focus away from small hydro, even if environmental conditions or energy demand mean that small hydro production is higher when alternative energy production is higher.

Perhaps the most interesting and most interpretable is the coefficients on the time and month indicator variables. For time, we use midnight (hour 24) as the omitted reference hour. In the early hours, before 7am, small.hydro energy production is below the midnight reference (negative coefficients: coefficient on hour 2 of -29.14 meaning small hydro energy production is 29.14 units per hour less at 2am than at midnight for example). After 7am production is above that of midnight (positive coefficients: the coefficient of 116.4 on hour 12, being the peak production hour, meaning at 12pm hourly small hydro production is 116.4 units higher than at midnight), with the highest production between 10am and 3pm.

For months, December (month 12) is the omitted reference month. Small hydro energy production is predicted lower than December in the months of September, October and November, and higher in all other months.



	RMSE (Validation set)	Adjusted R squared
Backward	89.948	0.5087
Exhaustive	89.572	0.5085
Forward	87.644	0.5082

As the result shows, residuals appear normally distributed and the linearity assumption holds. We can therefore be confident using the model for prediction.

The RMSE for the validation set is 89.948 in the backward regression, which means that our prediction results are a little far away from the real values. The adjusted R squared is 0.5087 and it explains almost 50% of the variability of the response data around its mean.

Overfitting problems

In our regression model, overfitting is less likely to happen because we have 68 features but more than 40 thousand observations. Before adding interaction terms and polynomial features, the R squared is only 0.16, an underfitting problem. After the improvement, the R squared is as large as 0.51. In conclusion, we are more likely to face underfitting rather than overfitting in the regression model.

Categorical model:

Random Forest

Now the target variable 'small-hydro' is categorical and it has 3 classes: low, medium and high energy. We are interested in how our model can label each variable based on other features, such as hour and solar power. To fulfill our goal, we can build the random forest model because we have lots of observations and a reasonable number of features, based on which we can make our model more complex by increasing ntree and mtry without overfitting.

	Training Set	Testing Set
Accuracy	99.42%	84.21%

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 accuracy.

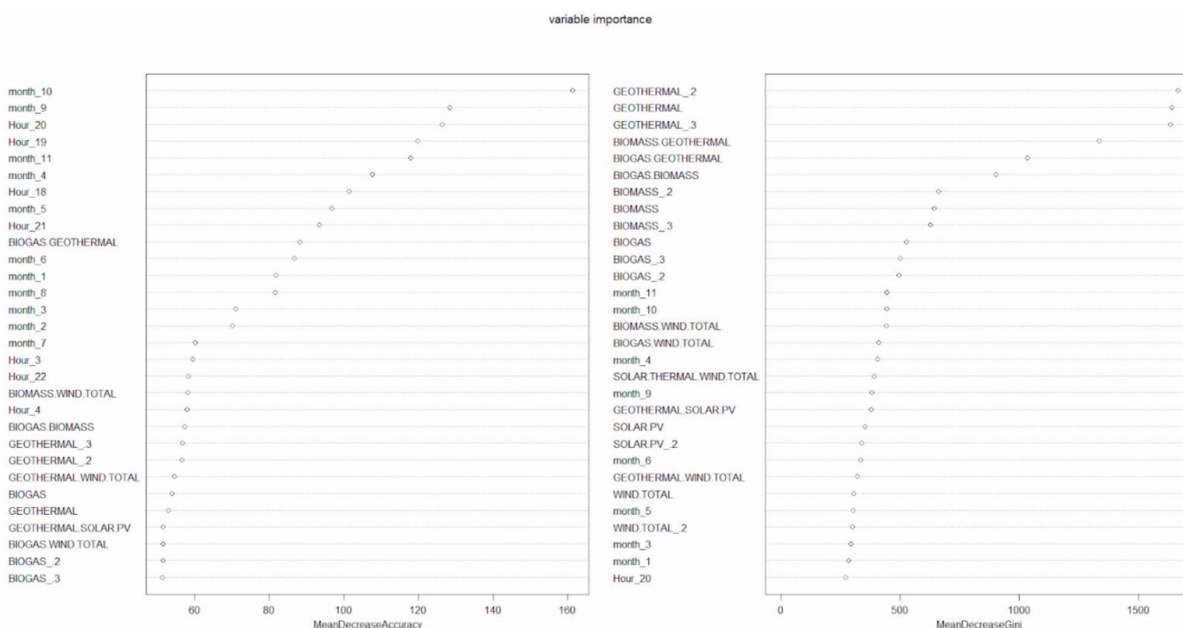


Figure 6. Variable Importance

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.

Overfitting problems

Overfitting is a serious problem for decision trees and hard to avoid. To solve it, we can use the grid search on `tree_depth` and `max_features`. For the number of trees, it will not cause overfitting as its number increases. When tree-depth is 5 and `max_feature` is 20 we can get the highest accuracy in the testing set, 84.21%. Even though the training set has an accuracy of 99.42%, we can still conclude that we solve this problem because they do not have much difference.

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.

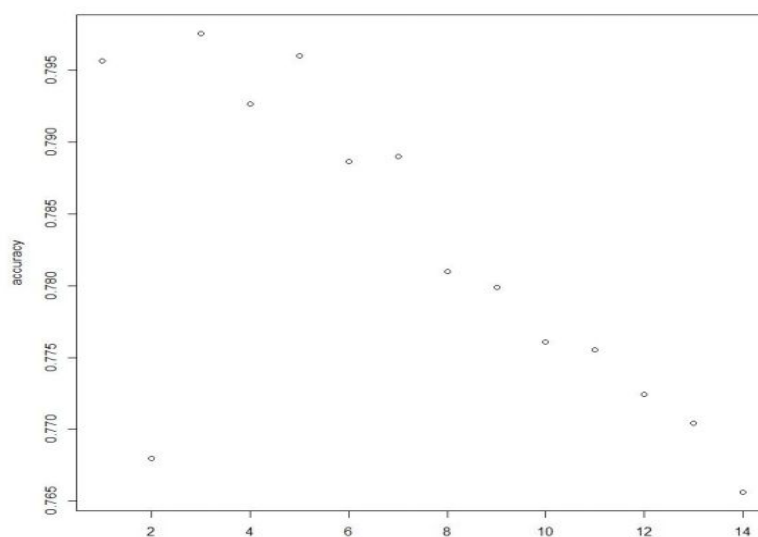


Figure 7. Accuracy with different K

After determining the best K, we test the accuracy for both the 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%

	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

Conclusion

In conclusion, for the classification problem, random forest and KNN has the highest accuracy of 84.21% and 79.76%. For the numeric model, the backward regression has the highest R square 0.5087 and RMSE 89.948.

DataRobot Models

The Hand-Crafted Models we chose based on accuracy are classification models including Random Forest, KNN and multi-linear regression, therefore, we also ran classification and regression models in DataRobot. For classification, the dataset we uploaded is the same as the one we conducted our own models which contains 66 numerical variables including the square, cubic and interactive terms and one 3-level categorical target variable (1- high,

2-low, 3-medium). For regression, the dataset we uploaded is the same as the one we conducted our own models which contains 67 numerical variables including the target. We let DataRobot do further feature engineering.

Outliers, overfitting and Target Leakage

We did remove outliers in former assessments, so the uploaded data is quite clear and had already got rid of the outliers issues. To be consistent with our own models, we divided the data into training (80%) and holdout dataset (20%). We noticed that DataRobot already considered the target leakage problem by providing the store away validation dataset as holdout. From the picture we can see that the three most accurate have holdout accuracy between 80%-90%, which seems not overly optimistic. In addition, there are nearly no gaps between the accuracy of training and holdout datasets of all the models. Therefore, we are confident that the DataRobot models have no overfitting and target leakage problems.

Regression Model

We used the 34 variables DataRobot recommended which are the most important features based on feature impact scores from a particular model, and the best regression model is **Light Gradient Boosted Trees Regressor with Early Stopping**, with the RMSE of 39, R square 0.91.

Model Name & Description	Feature List & Sample Size	Validation	Cross Validation	Holdout
Light Gradient Boosted Trees Regressor with Early Stopping <small>Tree-based Algorithm Preprocessing v1</small> DM TK	DR Reduced Features M17 80.0 %	40.6047 *	40.3470 *	39.0375
M73 BP74				

Figure 8. Top 1 accurate regression model provided by DataRobot

In this model, the regressor uses the LightGBM implementation of Gradient Boosted Trees. It uses least squares loss by default, but can also use poisson loss for count problems, tweedie loss for zero-inflated count problems, and gamma loss for right skewed, positive distributions. It is a cutting-edge algorithm for fitting extremely accurate predictive models. Early stopping is used to determine the best number of trees where overfitting begins. In this manner GBMs are usually capable of squeezing every last bit of information out of the training set and producing the model with the highest possible accuracy. The advantages of LightGBM include faster training speed and higher efficiency; lower memory usage; better accuracy; parallel learning supported; capable of handling large-scale data.

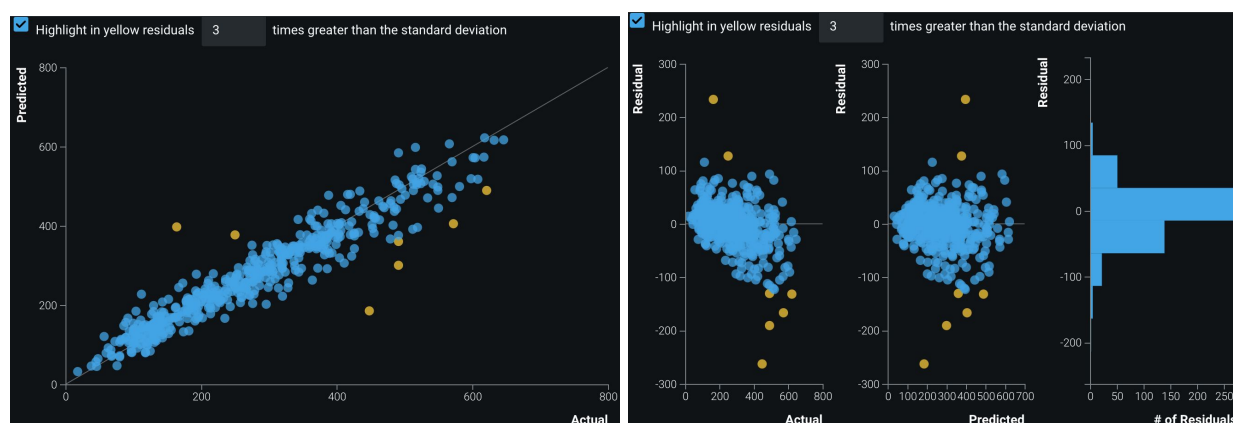


Figure 9. Residual Plot for LightGBM Tree

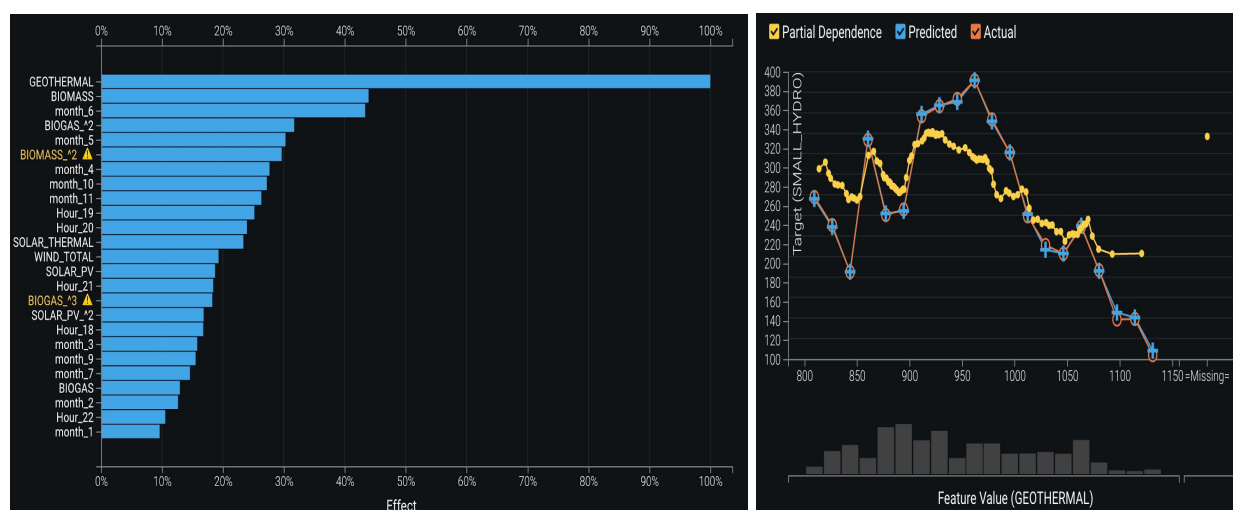


Figure 10. Feature Impact & Feature Effect - Partial Dependence Plot - LightGBM Tree

The residual plot proves we have little outliers, and from the feature impact chart, we found geothermal, biomass and month_6 are the most impactful to the model. Also, from the Partial Dependence Plot of geothermal, we can see how changes in the values of geothermal affects the model's predictions.

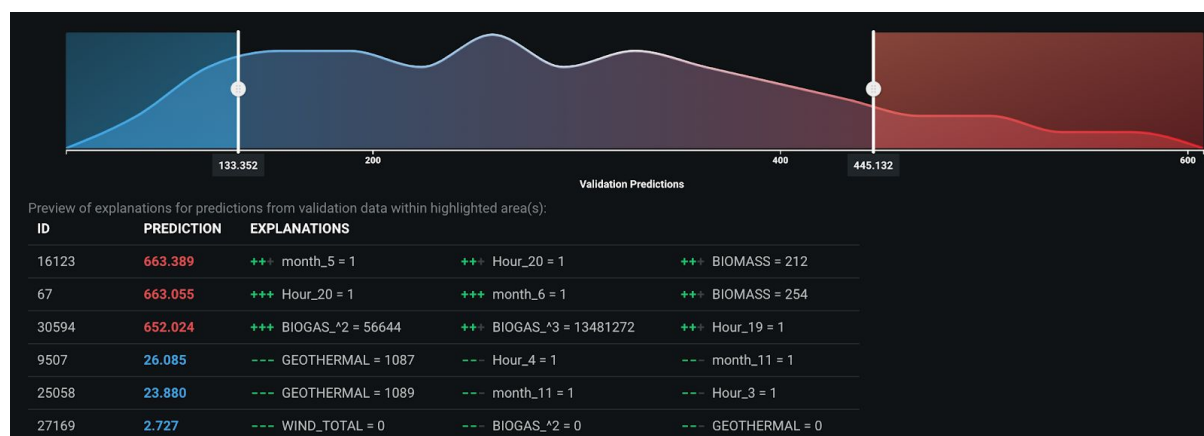


Figure 11. Prediction Explanations - LightGBM Tree

From the prediction explanations chart, we can see the top 3 explanations of high or low prediction, and for high predictions, month_5, month_6, hour_20, BIOGAS and BIOMASS are the largest contributors, while GEOTHERMAL, month_11, hour_4, and hour_3 are largest contributors for low predictions.

Classification Model

We used the 65 variables DataRobot provides which removes variables that may cause data leakage, and the best two classification models based on accuracy are **Light Gradient Boosted Trees Classifier with Early Stopping (88.9%)**, **RandomForest Classifier (85.9%)**.

Model Name & Description	Feature List & Sample Size	Validation	Cross Validation	Holdout
Light Gradient Boosted Trees Classifier with Early Stopping (SoftMax Loss) (16 leaves) Tree-based Algorithm Preprocessing v1 M50 BP24	Informative Features - Leakage Removed 80.0 %	0.8831 *	0.8829 *	0.8887
RandomForest Classifier (Gini) Tree-based Algorithm Preprocessing v1 M2228 BP22	Informative Features - Leakage Removed 80.0 %	0.8536 *	0.8509 *	0.8594

Figure 12. Top 2 accurate models provided by DataRobot

Light Gradient Boosted Trees Classifier is the LightGBM implementation of Gradient Boosted Trees which can be used both on classification and regression.

Random Forest Classifier is an ensemble method where hundreds (or thousands) of individual decision trees are fit to bootstrap re-samples of the original dataset, with each

tree being allowed to use a random selection of N variables, where N is the major configurable parameter of this algorithm. Ensembling many re-sampled decision trees serves to reduce their variance, producing more stable estimators that generalize well out-of-sample. Random forests are extremely hard to over-fit, are very accurate, generalize well, and require little tuning, all of which are desirable properties in a predictive algorithm. Random forests have recently been overshadowed by Gradient Boosting Machines, but enjoy a major advantage in that they are embarrassingly parallel and therefore scale much better to larger datasets. Actually the accuracy gap between the two models on our dataset is really small, only 4%.

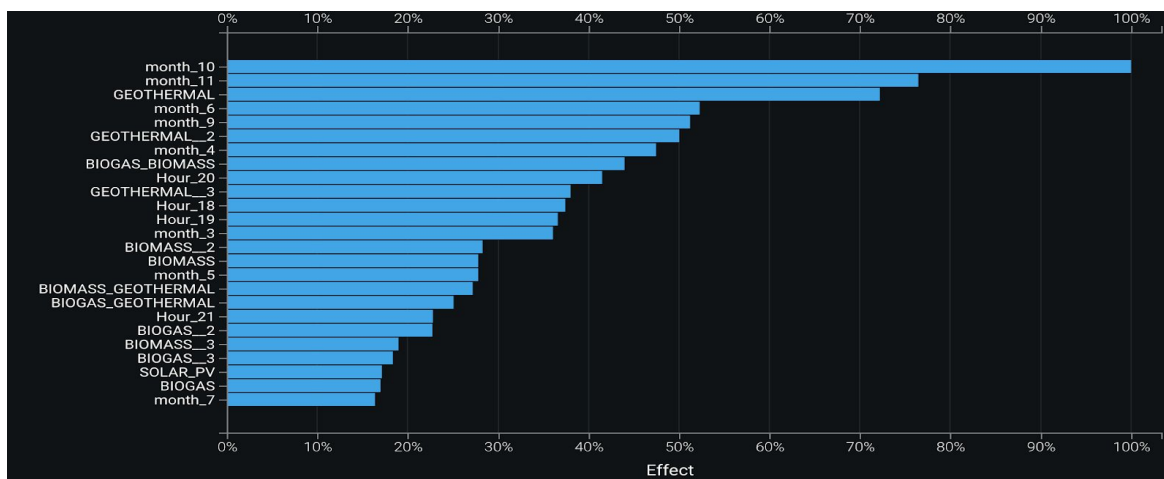


Figure 13. Feature impact for aggregate classes - LightGBM Tree

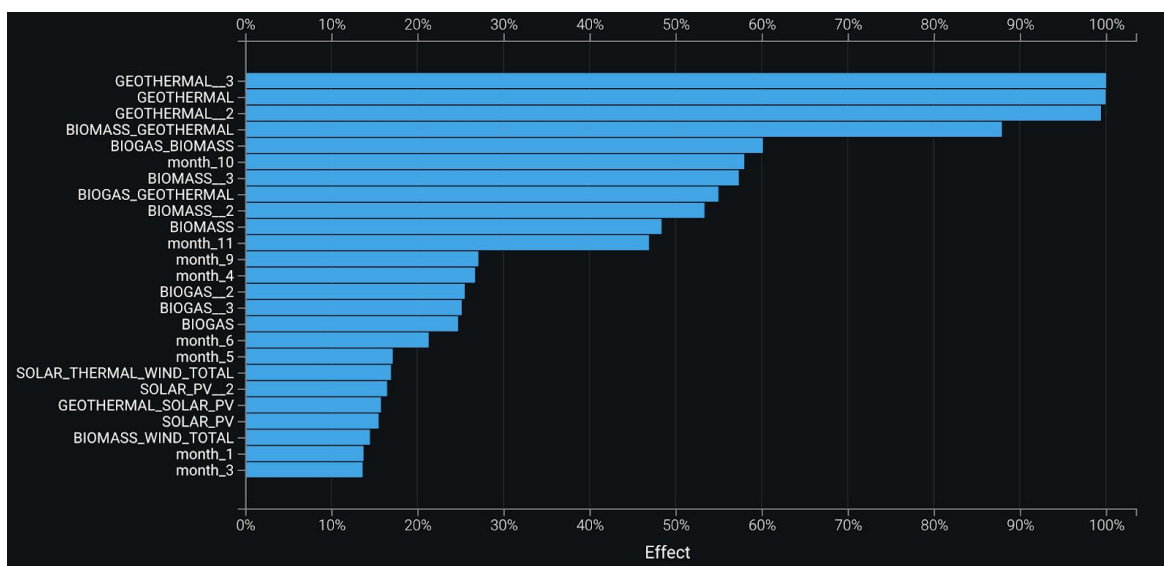


Figure 14. Feature impact for aggregate classes - Random Forest

From the feature impact chart of both models, we see month_10, month_11 and GEOTHERMAL (including square and cubic forms) are the top influencers. They are reasonable because the climate in California is Mediterranean, so in winter especially in October and November the precipitation is relatively high. In addition, the correlation between small hydro and geothermal is quite high and we are not sure whether this is a coincidence.

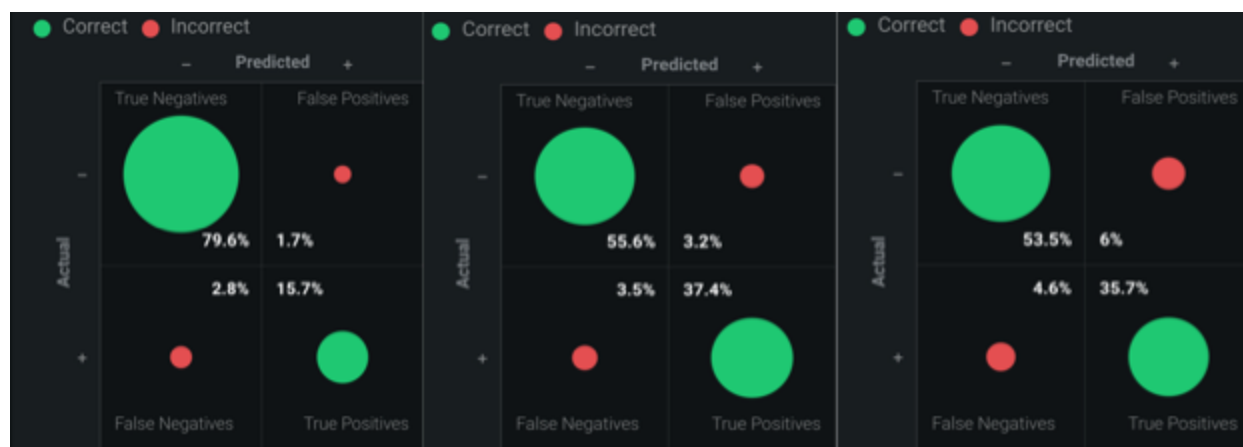


Figure x. Confusion Matrix from level 1-3 - LightGBM Tree



Figure 15. Confusion Matrix from level 1-3 - Random Forest

	LightGBM Tree			Random Forest		
Class	1	2	3	1	2	3
F1 Score	87%	92%	87%	82%	90%	84%
Recall	85%	91%	88%	79%	90%	85%
Precision	90%	92%	86%	86%	90%	82%

Figure 16. Confusion Matrix Scores

From the confusion matrix and the related scores we can see the performance of both models are good, and their predictability of the 2 (low) level is slightly better.

Comparison


The best Hand-Crafted Models we chose are the Multi-linear Regression model (Regression), Random Forest and KNN (Classification). The DataRobot best models are Light Gradient Boosted Trees Regressor with Early Stopping (Regression), Light Gradient Boosted Trees Classifier with Early Stopping and RandomForest Classifier (Classification).

	Classification		Regression
Hand-Crafted	Random Forest	KNN	Multi-linear Regression
Accuracy	84.21%	79.76%	R square 0.51 RMSE 89.9
DataRobot	Random Forest	LightGBM Tree	LightGBM Regressor
Accuracy	85.90%	88.90%	R square 0.91 RMSE 39.0

Figure 17. Performance of Hand-Crafted & DataRobot Models

DataRobot models have higher accuracy compared to our models, but the gap is not large except for the regression, and multi-linear regression may not be the appropriate model for this task. The feature impact of our model is highly related to DataRobot models, which is the month and hour as well as biogas and biomass are main permutation impactors. We did a pretty good job.

The best DataRobot model is Light Gradient Boosted Trees for both regression and classification methods, and we learnt that this model uses least squares loss by default, but can also use poisson loss for count problems, tweedie loss for zero-inflated count problems, and gamma loss for right skewed, positive distributions. It is a cutting-edge algorithm for fitting extremely accurate predictive models. Early stopping is used to determine the best number of trees where overfitting begins. In this manner GBMs are



usually capable of squeezing every last bit of information out of the training set and producing the model with the highest possible accuracy. We will certainly consider this powerful model in other tasks.

Part 3. Reflection:

Roadblocks

Along the way our group went through this final report, we encountered many difficulties. From wrangling data to comparing our models with DataRobot, we overcame a great number of burdens. This project is the first time we handled real life raw data, so when we went through data wrangling, the number of missing values in the data set is beyond our expectations. Especially for the variable Solar, which contains two thirds of missing values. Therefore, we decided to remove the whole variable out of the data set and make predictions based on the other two variables Solar PV and Solar Thermal. Then, we deleted some data that is not reasonable in a real situation. For example, we found out some data in variable Wind Total is negative, so we remove them from the data set.

Although our project was processing well, we did have several roadblocks. In A3: Clustering, when clustering the variables, we did not remove the numerical form of our target variable which has the risk of collinearity. We solve this problem by deleting the target variable and running the model again.

Model Summary

Going through our project, we conducted mainly two kinds of models, regression and classification models on our california-renewable-production data. For regression, firstly we did linear regression with 9 variables but the R square was quite low, therefore, we generated square, cubic and interactive terms and used the 66 variables to conduct the multi-linear regression with backward method, R squared improved. For classification, we did tree class including decision tree, deep tree and random forest, also KNN and logistic regression. The random forest had the best performance. However, the accuracy with logistic regression was quite low and we did regularization with Ridge and LASSO and also included the two clustering variables, and the accuracy increased a lot to 66%.

DataRobot


The experience using DataRobot was awesome, and we were impressed by its convenience and accuracy in several ways. First of all, it's more convenient to do data preparation. Once you uploaded the data, the system automatically calculated the descriptive statistical information for you and aware you if the data might have target leakage problems, and you can also change the type of the variable with a simple click. What's more, DataRobot helps a lot in terms of feature engineering by creating several feature lists which includes the most important variables, the variables that have relatively higher correlation with the target, and the variables removed from the leakage problem. In addition, it enables you to choose and run multiple regression or classification models fast and far more conveniently. The models DataRobot provided are famous and professional with high accuracy, and you do not necessarily need to set the parameters by yourself. The definition of the models are introduced clearly by the form of blueprint. Moreover, you can split the data into train, validation, and holdout dataset in one step. The time efficiency is also a big advantage. Most importantly, DataRobot provides a comprehensive and extreme clear explanation of the result. You can see the confusion matrix, the feature impact and the insights of the selected model. Model comparison can easily be done with DataRobot, and you can always rerun the model after some adjustment.

The best model DataRobot provided in both regression and classification questions is Light Gradient Boosted Trees, and we learnt that this model uses least squares loss by default, but can also use poisson loss for count problems, tweedie loss for zero-inflated count problems, and gamma loss for right skewed, positive distributions, which is a cutting-edge algorithm for fitting extremely accurate predictive models. We will definitely consider this method when we encounter similar problems.

Part 4. Conclusions & Insights

In recent years, Congress unanimously approved changes to simplify federal permitting requirements for small hydropower and provided financial support to lower the cost. Therefore, small hydro development is expected to accelerate in response to the convenience government offers. Our project builds classification models to explore the relationship among output of renewable energy and seasonality, providing in-depth study in the small hydro industry.

The ability to predict small hydro energy production using alternative energy sources and time and month categories is useful for all who care about specific sources of energy



production for whatever reason. For policymakers and energy planners, it is useful to be able to estimate a time/year dependence of small hydro energy production, to be able to plan for energy demand or supply shocks. For example, in the case of a particularly cloudy summer where solar energy supply drops, policy makers could look to our model to estimate knock-on or indirect effects to small hydro energy production, that might not be immediately obvious otherwise.

Small hydro energy production is very environmentally friendly, producing very little if any carbon emissions. We believe that our model is most useful for obtaining a better understanding of the renewable energy production landscape, for identifying complementary and non-complementary energy sources to small hydro and finally for identifying and forecasting time and month specific 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>

<https://info.ornl.gov/sites/publications/files/Pub56556.pdf>