MITx: 15.071x The Analytics Edge

Bookmarks

Unit 1: An Introduction to Analytics

Entrance Survey

Unit 2: Linear Regression

Unit 3: Logistic Regression

Unit 4: Trees

Unit 5: Text Analytics

Unit 6: Clustering

Kaggle Competition

Unit 7: Visualization

Unit 8: Linear Optimization

Exit Survey

Unit 9: Integer Optimization

Final Exam > Final Exam > PATTERNS IN RENEWABLE ENERGY GENERATION

Bookmark

The use of coal in the United States peaked in 2005, and since then has decreased by 25%, being replaced by renewable energy sources and more efficient use (Lovins, 2014). As the United States pursues a portfolio of more diverse, sustainable and secure energy sources, there are many questions to consider. What are effective factors in incentivizing states to adopt more environmentally friendly energy generation methods? How do these factors vary by state? How can we direct resources to different places in the country and ensure that they effectively drive renewable energy sources adoption? To derive insights and answer these questions, we take a combination of generation, usage, and greenhouse emission data by state and combine it with macro-economic and political information.

For this problem, we gathered data from various sources to include the following information for each state within the U.S. for the years spanning year 2000 to year 2013. The aggregated dataset energy.csv results in a total of 27 variables and 699 observations. Each observation contains one record per state per year. Here's a detailed description of the variables:

- GenTotal: Annual generation of energy using all types of energy sources (coal, nuclear, hydroelectric, solar, etc.) normalized by the state population at a given year.

- GenTotalRenewable: Annual generation of energy using all renewable energy sources normalized by the state population at a given year.

- GenHydro, GenSolar: Annual generation of energy using each type of energy source as a percent of the total energy generation.

- GenTotalRenewableBinary, GenSolarBinary: 1 if generation from solar or other renewable energy sources increased between a year n and a year n+1. 0 if it did not increase.

- AllSourcesCO2, AllSourcesSO2 and AllSourcesNOx: Annual emissions per state in metric tons, normalized by the respective state population at a given

▾ Final Exam

Final Exam
Final Exam due Jul 05, 2016 at 00:00 UTC

year and caused by all energy generation sources.

- EPriceResidential, EPriceCommercial, EPriceIndustrial, EPriceTransportation, EPriceTotal: Average electricity price per state, per sector (residential, industrial, commercial, etc.)

- ESalesResidential, ESalesCommercial, ESalesIndustrial, ESalesTransportation, ESalesTotal: Annual normalized sales of electricity per state, per sector.

- CumlRegulatory, CumlFinancial: Number of energy-related financial incentives and regulations created by a state per year.

- Demographic data such as annual wages per capita and presidential results (0 if a state voted republican, 1 is democrat).

Problem 1 - Total Renewable Energy Generation

(2 points possible)
Load energy.csv into a data frame called energy.

Renewable energy sources are considered to include geothermal, hydroelectric, biomass, solar and wind.

Which state in the United States seems to have the highest total generation of energy from renewable sources (use the variable GenTotalRenewable)?

○ Arizona (AZ)

○ California (CA)

○ Idaho (ID)

○ Massachusetts (MA)

?

Which year did the above state produce the highest energy generation from renewable resources?

○ 2000

○ 2002

○ 2007

○ 2011

?

You have used 0 of 2 submissions

Problem 2 - Relationship Between Politics and Greenhouse Emissions

(3 points possible)

What is the average CO2 emissions from all sources of energy for:

- states during years in which they voted republican?

?

- states during years in which they voted democrat?

?

Note: Please use na.rm = TRUE in your calculations!

States that voted democrat have on average higher NOx emissions than states that voted republican across all years. Is this statement true or false?

□ True

□ False

?

You have used 0 of 3 submissions

Problem 3 - Relationship Between Greenhouse Emissions and Energy Sales

(2 points possible)

What is the correlation between overall CO2 emissions and energy sales made to industrial facilities? Note that the variables AllSourcesCO2 and EsalesIndustrial contain NAs. Use the parameter: use="complete" to handle NAs in this question.

?

Choose the correct answers from the following statements:

□ Overall SO2 emissions are likely higher with increased industrial energy sales

□ Overall NOx emissions are likely higher with increased residential energy sales

□ Overall CO2 emissions are likely higher with increased commercial energy sales

?

You have used 0 of 1 submissions

Problem 4 - Boxplot of Energy Prices per State

(3 points possible)

Create a boxplot of the total energy price (EPriceTotal) by State across the data, and a table summarizing the mean of EPriceTotal by State.

What observations do you make?

□ The boxplot shows a clear outlier, the state of California, with much higher energy price compared to the rest of the U.S.

□ The boxplot shows a clear outlier, the state of Hawaii, with much higher energy price compared to the rest of the U.S.

□ There are no clear outliers in the boxplot and prices seem to be equal among all states within the U.S.

□ When looking at the average energy prices, there seems to be three price tiers ($5-$9, $10-$14, and $20+)

?

Which state has the lowest average energy price of all? You might want to make a table to answer this question.

○ Alabama (AL)

○ Texas (TX)

○ California (CA)

○ Wyoming (WY)

?

Is this state associated with the highest mean total energy generation (GenTotal)?

□ True

□ False

?

You have used 0 of 2 submissions

Problem 5 - Prediction Model for Solar Generation

(2 points possible)

We are interested in predicting whether states are going to increase their solar energy generation over the next year. Let's subset our dataset into a training and a testing set by using the following commands:

```
set.seed(144)

spl = sample(1:nrow(energy), size = 0.7*nrow(energy))

train = energy[spl,]

test = energy[-spl,]
```

Let's build now a logistic regression model "mod" using the train set to predict the binary variable GenSolarBinary. To do so, we consider the following as potential predictive variables: GenHydro, GenSolar, CumlFinancial, CumlRegulatory, Total.salary, Import.

Which variable is most predictive in the model?

○ GenHydro

○ GenSolar

○ Total.salary

○ CumlRegulatory

○ CumlFinancial

○ Import

?

You have used 0 of 2 submissions

Problem 6 - Performance on the Test Set

(3 points possible)

Compute the predictions on the test set. Using a threshold of 0.5, what is the accuracy of our model on the test set?

?

What is the accuracy for states voting republican?

?

What is the accuracy for states voting democrat?

?

You have used 0 of 3 submissions

Problem 7 - Clustering of the Observations

(3 points possible)
We can perhaps improve our accuracy if we implement a cluster-the-predict approach. We are interested in clustering the observations based on information about the regulatory and financial incentives, the elections outcome and the population wealth in each state across the years, in addition to whether the state was an energy importer or not.

Let us create a train.limited and test.limited datasets, where we only keep the variables CumlRegulatory, CumlFinancial, presidential.results, Total.salary, and Import.

Using the "preProcess" function on the train.limited set, we can compute the train.norm and test.norm.

Why didn't we include the dependent variable GenSolarBinary in this clustering phase?

□ Leaving the dependent variable might lead to unbalanced clusters

□ Needing to know the dependent variable value to assign an observation to a cluster defeats the purpose of the cluster-then-predict methodology

□ Removing the dependent variable decreases the computational effort needed for the clustering algorithm

?

Let's use kmeans clustering for this problem with a seed of 144, k=2 and keep the maximum number of iterations at 1,000.

Using the flexclust package, identify the clusters and call train1 the subset of train corresponding to the first cluster, and train2 the subset of train corresponding to the second cluster.

Select the correct statement(s) below:

□ On average, train1 contains more republican states than train2

□ On average, train1 contains states that have instituted considerably more financial and regulatory policies than train2

□ On average, train1 contains states that have recorded more CO2, SO2 and NOx emissions than train2

?

You have used 0 of 1 submissions

Problem 8 - Creating the Model on the First Cluster

(1 point possible)
Using the variable GenHydro, GenSolar, CumlFinancial, CumlRegulatory, Total.salary and Import, create mod1 using a logistic regression on train1.

What variable is most predictive?

○ GenHydro

○ GenSolar

○ Total.salary

○ CumlRegulatory

○ CumlFinancial

○ Import

?

You have used 0 of 1 submissions

Problem 9 - Evaluating the Model Obtained Using the First Cluster

(2 points possible)

What is the accuracy on test1, the subset of test corresponding to the first cluster?

?

We would like to know if mod1 gives us an edge over mod on the dataset test1. Using mod, predict GenSolarBinary for the observation in test1 and report the accuracy below:

?

You have used 0 of 2 submissions

Problem 10 - Creating the Model on the Second Cluster

(1 point possible)

Using the variables GenHydro, GenSolar, CumlFinancial, CumlRegulatory, Total.salary and Import, create mod2 using a logistic regression on train2.

Select the correct statement(s) below?

□ Unlike mod1, the number of regulatory policies is more predictive than the number of financial incentives in mod2

□ Unlike mod1, the number of regulatory policies is less predictive than the number of financial incentives in mod2

□ Similarly to mod1, the number of regulatory policies is more predictive than the number of financial incentives in mod2

□ Similarly to mod1, the number of regulatory policies is less predictive than the number of financial incentives in mod2

?

You have used 0 of 1 submissions

Problem 11 - Evaluating the Model Obtained Using the Second Cluster

(2 points possible)

Using the threshold of 0.5, what is the accuracy on test2, the subset of test corresponding to the second cluster?

?

We would like to know if mod2 gives us an edge over mod on the dataset test2. Using mod, predict GenSolarBinary for the observation in test2 and report the accuracy below:

?

You have used 0 of 2 submissions

Problem 12 - Evaluating the Performance of the Cluster-the-Predict Algorithm

(1 point possible)

To compute the overall test-set accuracy of the cluster-the-predict approach, we can combine all the test-set predictions into a single vector "AllPredictions" and all the true outcomes into a single vector "AllOutcomes".

What is the overall accuracy on the test set, using the cluster-then-predict approach, again using a threshold of 0.5?

?

You have used 0 of 2 submissions