St Olaf Energy and Environmental Values

From MSCS

Contents

St. Olaf Energy and Environmental Values Team Project

For this team project, please write up your code and written responses in R markdown yourself. Make sure your code is concise, make sure your file knits, and that the overall appearance is professional and readable for your TA's.

The datasets that we will be using for this team project have all been generated on the St. Olaf campus! One dataset comes from a survey about environmental values of incoming first year St. Olaf students in 2011 and 2012. This survey was used to gauge student responses to sustainability initiatives, and led to the development of programs like SustainAbilities (http://sustainabilities.stolaf.edu/) and Environmental Conversations (http://wp.stolaf.edu/admissions/academics/environmental-conversations/) for incoming first years. This survey is one example where 'classification methods' can be used for predicting environmental values. Effective use of random forests, boosted trees, and other classification methods goes beyond simply being able to write the code to execute a model. This project will introduce some the basics of how to interpret and analyze the results of classification models, as well as how to modify models to possibly improve performance.

In the above-mentioned survey, over 94% of incoming St. Olaf students responded that “at a place like St. Olaf, I expect to learn how to live sustainably in the residence halls.” An important step in developing the skills to live more sustainably on campus is gaining a basic level of understanding as to how and where our electricity is used. The second dataset we will be analyzing is a long-term electricity usage for all of the buildings on St. Olaf campus. We will be generating a map to visualize the data, then use classification models to predict the efficiency of some buildings on campus.

Project A: St. Olaf Environmental Values

Note, this should have been completed in your Track homework! You do not need to do it again or include it in the R markdown document if you've already completed it.

Part 1

The original Environmental values dataset has been split into training data and testing data for you. Each column corresponds to a question asked of the incoming student. Each column names consist of a few keywords that are representative of the question being asked. To find the entire question, you will need to consult the PDF of the full survey (http://www.stolaf.edu/people/olaf/cs125/SurveyQuestions.pdf) .

Load the training and testing data into R as follows:

```
train <- readRDS(gzcon(url('http://www.stolaf.edu/people/olaf/cs125/train.rds')))
test <- readRDS(gzcon(url('http://www.stolaf.edu/people/olaf/cs125/test.rds')))
```

If it works as expected, you should now have two dataframes called "train" and "test" in your workspace.

Part 2

Using the training set, grow a random forest model of 500 trees to classify answers to the question, “Private enterprise is more likely than government to find solutions to environmental problems”, using all of the other variables in the set. Apply your forest to the testing set, and make a table of the model predictions versus the students’ actual answers in the testing set. Did this model perform well? Explain.

Part 3

What variable contributes most to the classification process in your model? Should this variable be used in the model at all? Are there any other variables that you think should definitely not be a part of the model? Create new training and testing sets, removing any variables you think necessary before moving on to the next part.

Part 4

Many of the questions in this survey data have four answers: Strongly Agree, Agree, Disagree, and Strongly Disagree. Answers such as these have an inherent order to them, which adds important information to the data. Are factor levels of the relevant variables in the datasets ordered? If not, order only the relevant variables (placing No Answer responses in between Agree and Disagree), confirm that they are ordered, and repeat the analysis in Part 2. Do the results change?

Part 5

Look again at the table you generated of model predictions versus actual values. You will notice that the model seems to predict a disproportionately large number of Agree answers, and hardly any of the other factor levels. Why do you think this is? A possible solution to this problem is to simplify the four level factor variable to two levels: Strongly Agree and Agree become one category, and Strongly Disagree and Disagree become another. Create a new training and testing dataset, and replace the original variable with a new simplified version. Leave the No Answer factor level as it is. Rerun the model on this new variable, and repeat the analysis in Part 2. How does your new model perform? Did the simplification help to remedy the problem mentioned above? Are there any potential disadvantages to this kind of variable transformation?

Part 6

When fitting any model, parsimony is generally considered good practice. In other words, if a simpler model can be fit that is just as effective, it is generally the better choice. Currently, the model you have fit considers all variables in the dataset when growing the random forest. Can some of these variables be removed while preserving the effectiveness of the model? Build what you think is the most parsimonious model for predicting this variable.

Part 7

What do you think of the effectiveness of random forest models for these data? Given the source of these data, and how it was collected, are you surprised by the results? Are there any other survey questions whose responses might be easier to predict?

Project B: Visualizing and Predicting St. Olaf Energy Usage Data

Part 1

Load the following csv file:

```
read.csv(url('http://www.stolaf.edu/people/olaf/cs125/residence/st_olaf_electricity.csv'),stringsAsFactors=F
```

If the file has been loaded correctly, you should have a data frame with 7 columns that specify the St. Olaf building name, square footage, and electricity usage per year in kWh from 2007-2012. Why is there an 'X' attached to the front of the column names that specify the yearly usage?

Part 2

You need to remove several buildings from your dataset. Note that there is both a "SKOGLUND", a "TOSTRUD", and a "SKOGLUND-TOSTRUD". For later analysis, we will only need the "SKOGLUND-TOSTRUD" category. Create a new data frame which excludes the following buildings and their electricity data: "ADMINISTRATION", "CHILLER PLANT", "MANITOU FIELDHOUSE","SKOGLUND","TOSTRUD".

Part 3

First convert kWh to kBTUs (kilo-British Thermal Units), then convert to kBTUs per square foot (you will have to look up the conversion unit from kWh to kBTUs). Apply this to the yearly usage columns in your data frame.

Part 4

Add a column to your data frame called Average.Annual.Usage that calculates the average of the yearly energy usage columns from 2007-2012. You will most likely need to use a member of the apply() family.

Part 5

Add a column to your data frame called 'Efficiency' which labels the electricity usage of each building as 'poor','medium', or 'good.' To do this, classify the efficiency of the building as 'poor' if average annual kBTUs/sqft usage is greater than 25, 'medium' if it's between 15 and 25, and 'good' if usage is less than 15.

Part 6

Copy and paste the following commands into R:

This command loads the draw() function that you used in HW15 (You can look at the code by typing draw without parentheses--it uses geom_polygon in ggplot2!).

```
load(url('http://www.stolaf.edu/people/olaf/cs125/hwk15.RData'))
```

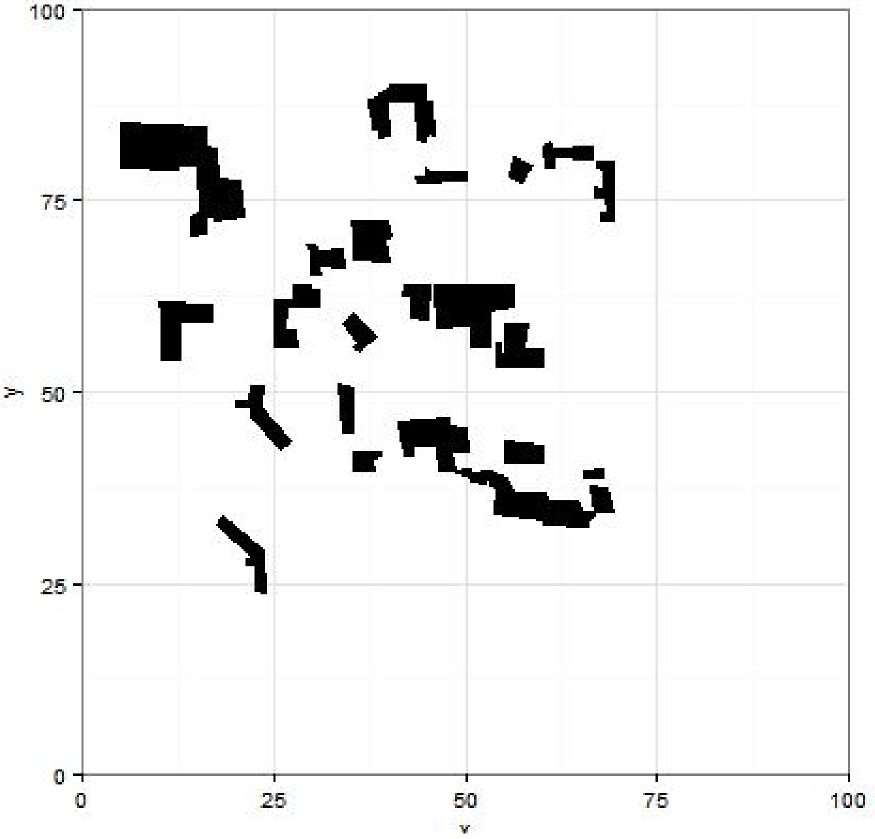
Next, read in the data:

```
file<-(read.table('http://www.stolaf.edu/people/olaf/cs125/residence/map_of_campus.txt',header=T,stringsAsFac
```

Now draw the data:

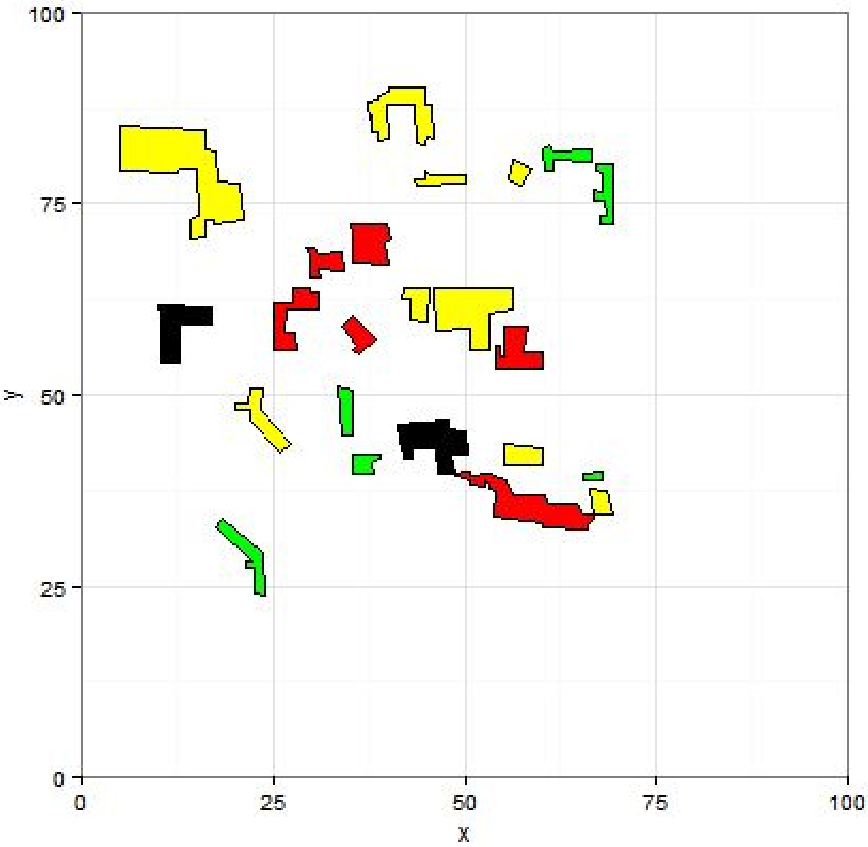
```
draw(file)
```

View the result when you draw map_of_campus.txt. If it has been loaded correctly it should appear to look something like this:



Part 7

Color each building in the map based on the efficiency ratings you calculated in Part 5. This is similar to the Calvin and Hobbes exercises we did earlier, where we colored Calvin's hair blue. Buildings that have a 'poor' rating should be colored red, those with a 'medium' rating should be colored yellow, and those with a 'good' rating should be colored green. To do this you will need to reference the data frame you worked with in Parts 1-5. Note that the energy data for Ytterboe and Tomson is difficult to access and will not be included in the final map. If you have done it correctly the result will look like this:



Part 8

Now we will add several variables to the data you were working with in Part 1-5. These variables include: Year.const.or.renovation, Era, Function, 40.year.life.cycle, Years.overdue, and Hours.open.weekly. We will use these variables to make predictions about the energy efficiency of Tomson Hall and Ytterboe Hall.

Load the testing and training sets into R as follows:

```
train<-read.csv(url('http://www.cs.stolaf.edu/wiki/images/b/b0/Electrain.csv'), stringsAsFactors=TRUE)
test<-read.csv(url('http://www.cs.stolaf.edu/wiki/images/3/34/Electest.csv'), stringsAsFactors=TRUE)
```

Part 9

Use a randomForest model or some other classification method of your choosing to predict the efficiency of Ytterboe and Tomson Hall. The final result should be a prediction of either 'poor', 'medium', or 'good' electricity usage for Tomson hall and Ytterboe hall.

Note that you may receive the following error:

```
New factor levels not present in the training data
```

This error is because Tomson and Ytterboe have different square footage than the buildings in the training set, which introduces new factor levels. To fix this, you will need to relevel the 'sqft' variable with intervals of your choosing such that there are no new factor levels in the test and training sets. Depending on time allowance, you may also choose to leave out variables such as 'sqft' to build a parsimonius model. Also, if you are including 'building' as a variable in your model it will give you the same error, which makes sense because every building has a different name. Do not include 'building' as a variable :)

You're done!

Your team will turn in an R markdown file. Make sure your code is concise, make sure your file knits, and that the overall appearance is professional and readable for your TA's.

Some questions to consider:

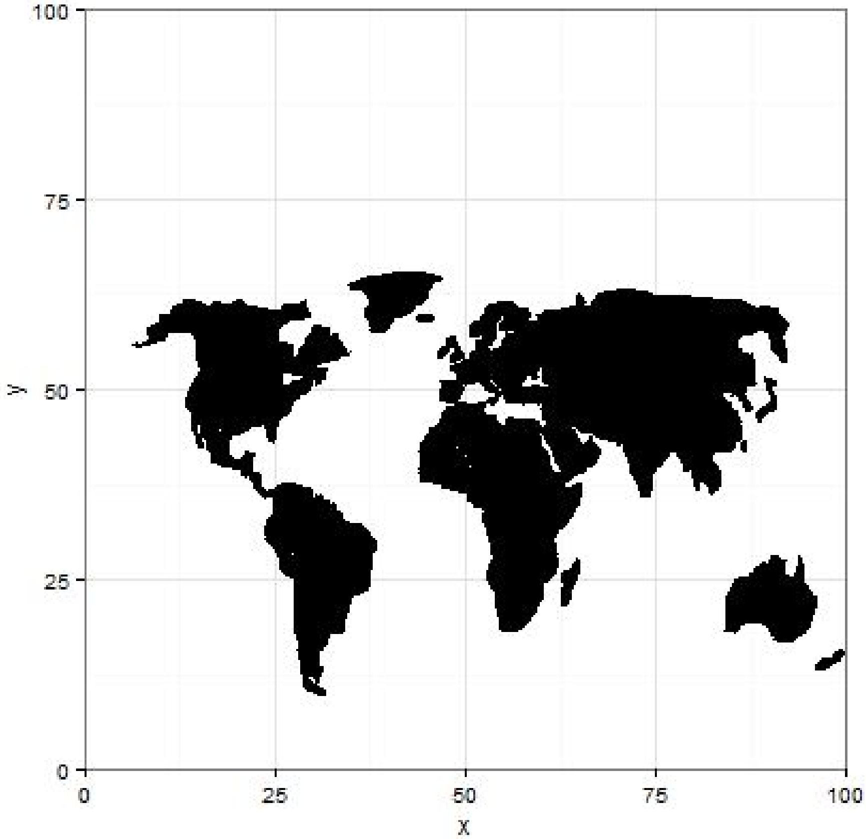
1. What surprised you about the St. Olaf data? Why do you think Regents hall has such high electricity consumption, even though it is considered to be one of the 'greenest' buildings (http://www.stolaf.edu/news/index.cfm?fuseaction=NewsDetails&id=4767) on campus?

2. Do residence halls comprise a small or large portion of the total electricity usage on campus? Do the results differ from what you expected?

3. Were you surprised by the predictions for the electricity usage of Ytterboe and Tomson? Are your predictions reasonable? St. Olaf alumni Elizabeth Turner completed her Masters of Architecture from the University of Minnesota and her thesis was on envisioning a carbon neutrality plan for St. Olaf. Her thesis is found here (http://www.aashe.org/files/resources/student-research/2009/envisioning_the_carbon_neutral_campus-final2.pdf) , and if you go to page 25 you will see a map similar to yours. Do you predictions for Tomson and Ytterboe match Turner's calculations?

Optional

Follow this link (http://www.cs.stolaf.edu/wiki/index.php/Global_Energy_Visualization) to create a visual for energy usage on a global scale!



Retrieved from "http://www.cs.stolaf.edu/wiki/index.php/St_Olaf_Energy_and_Environmental_Values"

- This page was last modified on 7 May 2015, at 08:54.
- Content is available under Attribution-Share Alike 3.0 .