MITx: 15.071x The Analytics Edge

Help

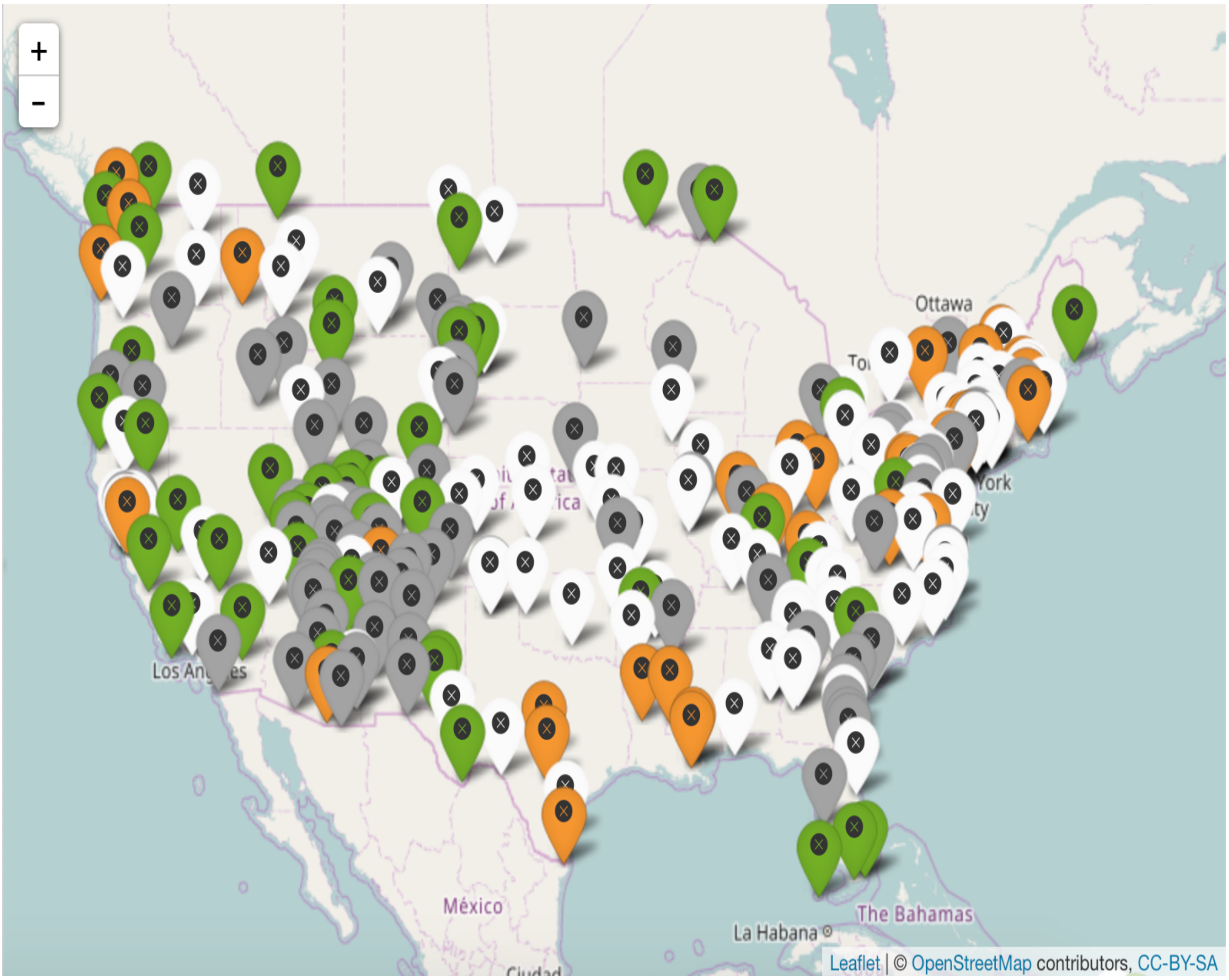
Course > Final Exam > Final Exam > FORECASTING NATIONAL PARKS VISITS

FORECASTING NATIONAL PARKS VISITS

Bookmark this page

The U.S. National Parks System includes 417 areas including national parks, monuments, battlefields, military parks, historical parks, historical sites, lakeshores, seashores, recreation areas, scenic rivers and trails, and the White House (see map in Figure 1). Every year, hundreds of millions of recreational visitors come to the parks. What do we know about the parks that can affect the visitor counts? Can we forecast the monthly visits to a given park accurately? To derive insights and answer these questions, we take a look at the historical visits data and the parks information released by the National Parks Service (NPS).

Figure 1: A map of the U.S. National Parks System areas. Green: National Parks; Grey: National Memorial/National Monument; Orange: national Historical Park/Site; White: others. Made with *leaflet* package in R with NPS data.



For this problem, we obtained monthly visits data between 2010 and 2016 (source: https://irma.nps.gov/Stats/Reports/National). We also got park-specific data via the NPS API (https://developer.nps.gov/api/index.htm). The aggregated dataset park_visits.csv results in a total of 12 variables and 25587 observations. Each observation contains one record per park per month. Here's a detailed description of the variables:

- **ParkName**: The full name of the park.

- **ParkType**: The type of the park. For this study we restrict ourselves to the following more frequently visited types: National Battlefield, National Historic Site, National Historical Park, National Memorial, National Monument, National Park, National Recreation Area, and National Seashore.

- **Region**: The region of the park, including Alaska, Intermountain, Midwest, National Capital, Northeast, Pacific West, and Southeast.

- **State**: The abbreviation of the state where the park resides.

- **Year**, **Month**: the year and the month for the visits.

- **lat**, **long**: Latitude and longitude of the park.

- **Cost**: a simple extraction of the park's entrance fee. Some parks may have multiple levels of entrance fees (differ by transportation methods, age, military status, etc.); for this problem, we only extracted the first available cost information.

- **logVisits**: Natural logarithm of the recreational visits (with one added to the visits to avoid taking logs of zero) to the park in the given year and month.

- **laglogVisits**: the logVisits from last month.

-**laglogVisitsYear**: the logVisits from last year.

Problem 1 - Number of National Parks in Jan 2016

2/2 points (graded)

Load park_visits.csv into a data frame called visits.

Let's first look at the visits in July 2016. Subset the observations to this year and month, name it visits2016jul. Work with this data subset for the next three problems.

Which park type has the most number of parks?

● National Historic Site ✓

○ National Historical Park

○ National Monument

○ National Park

Which specific park has the most number of visitors?

○ Yellowstone NP

○ Golden Gate NRA

● Great Smoky Mountains NP ✓

○ Cape Cod NS

Explanation

Use the table command to tabulate the counts by park types, and which.max to find the one with maximum number of log visits.

Submit You have used 2 of 2 attempts

Answers are displayed within the problem

Problem 2 - Relationship Between Region and Visits

3/3 points (graded)

Which region has the highest average log visits in July 2016?

○ Intermountain

○ National Capital

● Pacific West ✓

○ Southeast

What is the average log visits for the region in July 2016 with:

1. the highest average log visits?

10.767849 ✓ **Answer:** 10.767849

10.767849

2. the lowest average log visits?

9.374157 ✓**Answer:** 9.374157

9.374157

Explanation
You can answer this question by using the tapply function on the visits by region using mean.

Submit You have used 1 of 3 attempts

ℹ Answers are displayed within the problem

Problem 3 - Relationship Between Cost and Visits

2/2 points (graded)
What is the correlation between entrance fee (the variable cost) and the log visits in July 2016?

0.4010611 ✓**Answer:** 0.4010611

0.4010611

Choose the most reasonable possible answer from the following statements:

○ Higher entrance fees are associated with lower log visits, likely because visitors are cost sensitive

● Higher entrance fees are associated with higher log visits, likely because more expensive parks are often more popular due to other features of the parks ✓

○ There is no association between entrance fees and the log visits

Explanation

Use the cor function to solve this question.

Submit You have used 1 of 2 attempts

Answers are displayed within the problem

Problem 4 - Time Series Plot of Visits

1/1 point (graded)

Let's now look at the time dimension of the data. Subset the original data (visits) to "Yellowstone NP" only and save as ys. Use the following code to plot the logVisits through the months between 2010 and 2016:

```
ys_ts=ts(ys$logVisits,start=c(2010,1),freq=12)
```

```
plot(ys_ts)
```

What observations do you make?

☑ Between the years, the shapes are largely similar. ✓

☑ The log visits are highly cyclical, with the peaks in the summer time. ✓

☐ There is a trend of substantial increase in log visits over recent years.

✓

Explanation

Use the provided code and make the observations.

Submit You have used 1 of 2 attempts

Answers are displayed within the problem

Problem 5 - Missing Values

2/2 points (graded)

Note that there are some NA's in the data - you can run colSums(is.na(visits)) to see the summary.

Why do we have NA's in the laglogVisits and laglogVisitsYear? These variables were created by lagging the log visits by a month or by a year.

○ The dataset inevitably have missing data due to human entry negligence.

● These are lagged variables and the earlier data is not available for the first months. ✓

○ The values were outliers and therefore removed.

To deal with the missing values, we will simply remove the observations with the missing values first (there are more sophisticated ways to work with missing values, but for this purpose removing the observations is fine). Run the following:

```
visits = visits[rowSums(is.na(visits)) == 0, ]
```

How many observations are there in visits now?

21855 ✓ **Answer:** 21855

21855

Explanation
Use nrow after running the command.

Submit You have used 1 of 2 attempts

Answers are displayed within the problem

Problem 6 - Predicting Visits

3/3 points (graded)
We are interested in predicting the log visits. Before doing the split, let's also make Month a factor variable by including the following:

visits$Month = as.factor(visits$Month)

Subset our dataset into a training and a testing set by splitting based on the year: training would contain 2010-2014 years of data, and testing would be 2015-2016 data.

Let's build now a simple linear regression model "mod" using the training set to predict the log visits. As a first step, we only use the laglogVisits variable (log visits from last month).

What's the coefficient of the laglogVisits variable?

0.927945 ✓ **Answer:** 0.927945

0.927945

What's the out-of-sample R2 in the testing set for this simple model?

0.8973278 ✓ **Answer:** 0.8975923

0.8973278

Explanation
Run the linear regression with lm and look at the summary.
Then calculate the out-of-sample R2 using the test data.

Submit You have used 1 of 3 attempts

ℹ Answers are displayed within the problem

Problem 7 - Add New Variables

0/2 points (graded)
We see that the model achieves good predictive power already simply using the previous month's visits. To see if the other knowledge we have about the parks can improve the model, let's add these variables in a new model.

The new model would have the following variables:

laglogVisits, laglogVisitsYear, Year, Month, Region, ParkType, and cost

Looking at the model summary, which of the following statements are correct (significance at 0.05 level)?

☑ Both the log visits from last month and last year are significant and are positively associated with the current log visits. ✓

☑ None of the regions are significant from the baseline region (Alaska).

□ None of the park types are significant from the baseline park type (National Battlefield). ✓

□ The cost is no longer significant.

✗

Explanation

Run lm with new set of variables, summary on the model, and look at the significant variables.

Submit You have used 2 of 2 attempts

Answers are displayed within the problem

Problem 8 - Out-of-Sample R2

2/2 points (graded)

In the new model, what's the out-of-sample R2 in the testing set?

0.9370909 ✓ **Answer:** 0.937253

0.9370909

Explanation

Calculate the out-of-sample R2 using the testing data with new model.

Submit You have used 1 of 2 attempts

Answers are displayed within the problem

Problem 9 - Regression Trees

3/3 points (graded)

In addition to the logistic regression model, we can also train a regression tree. Use the same set of variables as the previous problem (laglogVisits, laglogVisitsYear, Year, Month, Region, ParkType, and cost), train a regression tree with cp = 0.05.

Looking at the plot of the tree, how many different predicted values are there?

4 ✓ **Answer:** 4

4

What is the out-of-sample R2 on the testing set?

0.7858791 ✓ **Answer:** 0.7864307

0.7858791

Explanation

Run with rpart function and look at the prp plot. Calculate out-of-sample R2 based on this new model.

Submit You have used 1 of 3 attempts

Answers are displayed within the problem

Problem 10 - Regression Trees with CV

3/3 points (graded)

The out-of-sample R2 does not appear to be very good under regression trees, compared to a linear regression model. We could potentially improve it via cross validation.

Set seed to 201, run a 10-fold cross-validated cart model, with cp ranging from 0.0001 to 0.005 in increments of 0.0001. What is optimal cp value on this grid?

0.0001 ✓**Answer:** 0.0001

0.0001

Looking at the validation R2 versus the cp value, we can further refine the cp range. In what direction should it change?

● smaller values of cp ✓

○ larger values of cp

Explanation
Use train function from caret package with tuneGrid=expand.grid(cp=seq(0.0001, 0.005,0.0001).
By looking at the plot, the R2 seems to get worse starting from 0.0001, therefore smaller values of cp may give better validation results.

Submit You have used 1 of 3 attempts

ℹ Answers are displayed within the problem

Problem 11 - Final Regression Tree

2/2 points (graded)
Rerun the regression tree on the training data, now using the cp value equal to the one selected in the previous problem (under the original range). Note: do not get the tree from the cross-validation directly.

What is the out-of-sample R2 in the testing set?

0.9369506 ✓**Answer:** 0.937113

0.9369506

Explanation
Use cp = 0.0001. Calculate the out-of-sample R2 using the test data.

Submit You have used 1 of 2 attempts

Answers are displayed within the problem

Problem 12 - Random Forest

2/2 points (graded)

We can potentially further improve the models by using a random forest. Set seed to 201 again. Train a random forest model with the same set of covariates, and using just default parameters (no need to specify). This may take a few minutes.

What is the R2 on the testing set for the random forest model?

0.9472945 ✓ **Answer:** 0.947239

0.9472945

Explanation

Use the randomForest package and function to train the model. Calculate the out-of-sample R2 using the testing data.

Submit You have used 1 of 2 attempts

Answers are displayed within the problem

English ▼