Intro to Data Analytics and Visualizations

Lecture 5
Fall 2014, September3

Outline

1. The tasks in a Data Science Project
2. Most common data science modeling tasks
3. A first modeling example
4. In-class coding exercise

Data Scientist's Tasks in a Data Science Project

- Define goal
- Collect and manage data
- Build model (including visualization of data)
- Evaluate model
- Present results
- Implement (deliver/deploy) model

Most Common Modeling Tasks in Data Science

- Classification
- Scoring
- Ranking
- Clustering
- Finding Relations
- Characterizations

Data Science Problem

- **Progressive, the insurance company, would like to have a quick way to quote the premium for an insurance policy on a car.**
- **The insurance agent only has 5 minutes to spend on the phone with a potential new customer.**
- **The only information the agent gets is the caller's age and the caller's vehicle age.**
- **How can a data scientist help with this problem?**

Data Science Problem

- Suppose the company has data from the insurance policies written in the past. If saved in a data frame, we would have three variables: Premium, Driver Age and Vehicle Age. How can we fit a model so that we can predict the Premium for a future, only knowing the Driver Age and Vehicle Age?

Data Science Problem

- Historical Data

	Driver Age																							
Vehicle Age	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	
0	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	3	3	3	3	3	3	3	
1	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	3	3	3	3	3	3	3	
2	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	3	3	3	3	3	3	3	
3	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	3	3	3	3	3	3	3	
4	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	3	3	3	3	3	3	3	
5	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	5	5	5	5	5	5	5	5	
6	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	5	5	5	5	5	5	5	5	
7	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	5	5	5	5	5	5	5	5	
8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	5	5	5	5	5	5	5	5	
9	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	5	5	5	5	5	5	5	5	
10	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	5	5	5	5	5	5	5	5	

Decision Trees (Predictive task)

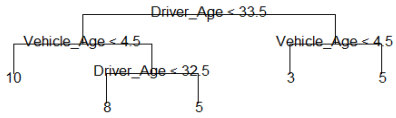
- We can build a decision tree. Decision trees are a class of techniques used to characterize the relationship between a response and a collection of covariates. In R, you can fit a decision tree, and then plot it to have a visualization of the tree.

- In R:

- ***library(tree)***
- ***insurance_tree <- tree(Premium ~ Driver_age + Vehicle_age, data = cars)***
- ***plot(insurance_tree)***

Decision Trees (Predictive task)

	Driver Age																						
Vehicle Age	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
0	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	3	3	3	3	3	3	3
1	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	3	3	3	3	3	3	3
2	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	3	3	3	3	3	3	3
3	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	3	3	3	3	3	3	3
4	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	3	3	3	3	3	3	3
5	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	5	5	5	5	5	5	5	5
6	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	5	5	5	5	5	5	5	5
7	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	5	5	5	5	5	5	5	5
8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	5	5	5	5	5	5	5	5
9	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	5	5	5	5	5	5	5	5
10	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	5	5	5	5	5	5	5	5



The Data Science Solution

- **Insurance Agent: My caller wants to get a quote. He is 20 and has a 3 year old car. What should I say?**
- **What the Data Scientist wanted to say: After getting data, coding, crunching, summarizing, visualizing, and building this model…and considering… and accounting for … a good guess might be ..10 plus or minus…**
- **What the Agent wanted to hear: 10.**

IN-CLASS Coding Exercise

Complete and submit to Dropbox as R script “Name_inclass1.r”

For In-class Exercise: R Coding Conventions and Esthetics

Structuring your code:

-make it understandable for your future self and others;

-good names for objects and functions;

-keep lines short;

-have good comments to explain code;

-make it concise and efficient.

Exercise

1. Create new R script “Inclass1” in your folder CMDA that is a Git repository.
2. Set working directory to your folder.
3. Import “cars1” dataset. Download it first from Scholar into your CMDA folder.
3. What is the dimension of your R data frame, i.e, how many rows and columns? Use a command to get the answer.
4. Assign the value of the cell [2,3] to a new variable var1.
5. What are the variable names in this data frame?
6. Output the content of the first and second columns separately.
7. Assign the values of variable “speed” from the data frame to a new variable “SPEED”. Print the new variable.
8. Output the value of row 15 from the data frame.
9. Save and close the R file. Commit changes locally to your CMDA folder. “Sync” to see the updates on your GitHub web account.