

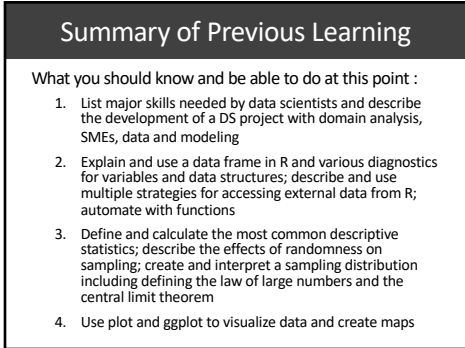
Introduction to Linear Modeling

Professor Jeff Saltz
Professor Jeff Stanton

Copyright 2022: Jeffrey Saltz and Jeffrey Stanton; please do not upload.

School of Information Studies
SYRACUSE UNIVERSITY

1

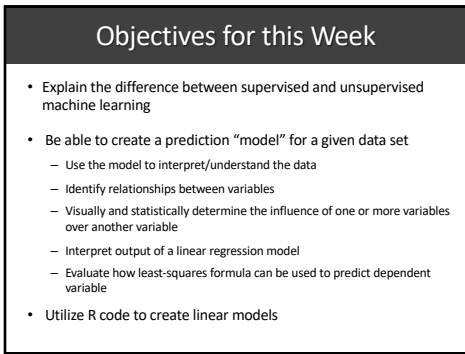


Summary of Previous Learning

What you should know and be able to do at this point :

1. List major skills needed by data scientists and describe the development of a DS project with domain analysis, SMEs, data and modeling
2. Explain and use a data frame in R and various diagnostics for variables and data structures; describe and use multiple strategies for accessing external data from R; automate with functions
3. Define and calculate the most common descriptive statistics; describe the effects of randomness on sampling; create and interpret a sampling distribution including defining the law of large numbers and the central limit theorem
4. Use plot and ggplot to visualize data and create maps

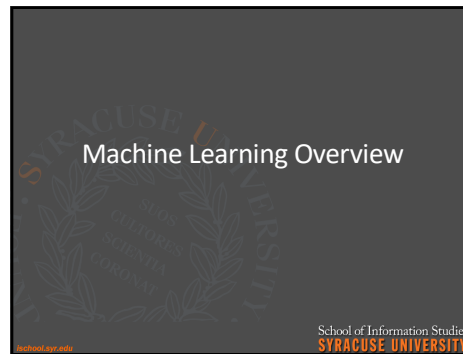
2



Objectives for this Week

- Explain the difference between supervised and unsupervised machine learning
- Be able to create a prediction "model" for a given data set
 - Use the model to interpret/understand the data
 - Identify relationships between variables
 - Visually and statistically determine the influence of one or more variables over another variable
 - Interpret output of a linear regression model
 - Evaluate how least-squares formula can be used to predict dependent variable
- Utilize R code to create linear models

3



4

Trad Statistical Analysis vs. Machine Learning / Data Science

Statistical Analysis	Data Mining / Data Science
<ul style="list-style-type: none"> Often focused on confirming a hypothesis in light of some theory Data arrives with provenance, screening & cleaning are trivial Measurement strategy planned in advance, thoughtful choice of constructs and scales Metric variables often follow the normal distribution 	<ul style="list-style-type: none"> More exploratory, looking for novel patterns and needles in haystacks Data may arrive with little provenance, may require extensive diagnosis and restructuring No planned measurement strategy: work with what you have available Mixture of textual, categorical, & metric, no presumption of distribution

5

Machine Learning in a Nutshell

- “Supervised” refers to the idea that there is a **criterion** used during an algorithm training phase
- A supervised algorithm uses a set of input variables to optimize the prediction of an outcome variable
- Supervised data mining is closely connected with machine learning techniques: What is the difference?

Supervised Learning

Unsupervised Learning

Image credit: Olivia Klose

6

Machine Learning Techniques

Unsupervised learning includes a variety of machine learning techniques that do not use a criterion or dependent variable, but rather look for patterns solely among “independent” variables.

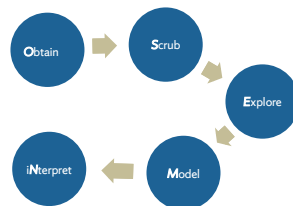


Supervised learning is parallel in concept to the predictive statistical techniques used by many social science researchers, such as linear regression, but without the restriction of only exploring linear relationships.

Another form of learning is known as “**reinforcement learning**.” Evolution of these models depends on success/failure cues from real or simulated environments.

7

The Machine Learning Process



Hilary Mason and Clark Wiggins, 2010 - <http://www.fantix.com/2010/05/05/economics-of-data-science/>

8

OSEMN Phase: Obtain

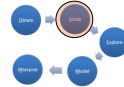
- Collect data
 - Deal with file formats and how to read data
 - Query databases or data repositories
 - Extract data from other sources
 - Generate data (e.g., surveys, sensors)



9

OSEMN Phase: Scrub

- Get data into a useable format structure
 - Filter/subset
 - Extract attributes
 - Replace/handle missing, illegal, and anomalous values
 - Transform/bin/code data attributes



Typically 50% or more of the work is in this phase

10

OSEMN Phase: Explore

- Explore patterns and trends
 - Start to “understand” the data, detect outliers
 - Visualize attributes (e.g., scatter plots, histograms)
 - Calculate/visualize descriptive statistics (e.g., distributions)
 - Feature selection (what attributes are most interesting)



11

OSEMN Phase: Model


- Build predictive models (machine learning)
 - Type of modeling:
 - Supervised (classification, regression)
 - Unsupervised (e.g., clustering)
 - Model tuning and comparing candidate models



12

OSEMN Phase: Interpret

- Understand/explain the results
 - Draw conclusions from and data models
 - Evaluate the meaning of results
 - Communicate the results
 - Ensure actionable insight



13

Question

How is the OSEMN process different from or the same as the overall data science process?

14

Supervised Data Mining



15

Question

What is the meaning of the word “supervised” in the context of Machine Learning?

16

Two types of prediction problems

Classification problems

- Predicting membership in two or more categories based on a set of predictors (model features)

• **Examples of criteria to predict:**
Medical diagnosis,
Employment outcomes (e.g., hiring),
Financial outcomes (e.g., loan default)

Regression problems

- Predicting a continuous numeric outcome based on a set of predictors

• **Examples of criteria to predict:**
sales volume,
employee engagement,
customer satisfaction

Traditional Approaches:

Classification problems
Logistic regression
Discriminant analysis

Regression problems

Ordinary least squares regression
linear models
Generalized linear models
Lasso regression

17

Supervised Learning Example

Train a machine learning algorithm to predict the weather

- Collect weather data over a period of time
 - Sunny, cloudy
 - Temperature
 - Barometer
 - Wind speed and direction
- Train a machine learning algorithm with these collected variables
- Collect more weather data and predict the weather via our trained algorithm
 - Classification would be predicting good weather or bad weather
 - Regression would be predicting the temperature
- Then validate the prediction

18

The Modeling Process

- Use a substantial number of training cases
 - The machine learning algorithm can use that data to build a model
- Use the results of this process (i.e., the model) on test data set to determine how well algorithm performed
 - Validate the model on new data
- The result is a model that can be used for prediction
 - Predict data that was not used during training
 - Predict future instances of data

*Note: The model is **not always useful** for explaining results to managers.*

Some algorithms produce results that are not easy to interpret or visualize or explain how they work; for some algorithms there is no output that is like a regression coefficient.

19

Linear Modeling Overview

School of Information Studies
SYRACUSE UNIVERSITY

20

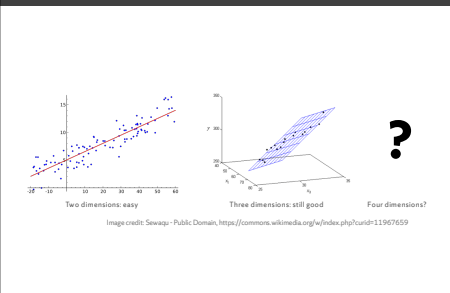
Linear Modeling



"Does X influence Y?"

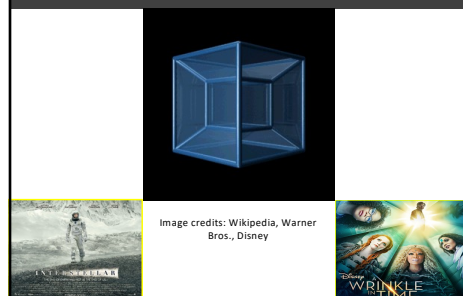
21

Linear Relationships in 2+ Dimensions



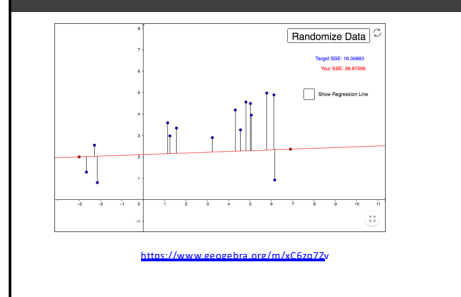
22

4D Cube: Tesseract



23

AN EXAMPLE



24

A Simple Model to Predict GPA

Predicting a student's semester GPA using three pieces of information?

Data collection:

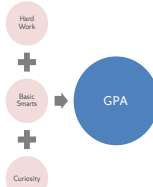
- Measure three predictors, using tests or surveys with multiple students (e.g., $n=120$)
- Then at the end of the semester we would have four pieces of information:
 - The criterion (GPA)
 - The three predictors (hard work, basic smarts, & curiosity).

Model creation:

- Using linear regression, calculate coefficients for each predictor to make an equation:

$$\text{GPA} = (B1 * \text{HardWork}) + (B2 * \text{BasicSmarts}) + (B3 * \text{Curiosity})$$

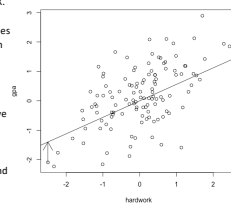
$$[Y = MX + b] \text{ or } Y = bX + e$$



25

Prediction Errors: The Least Squares Criterion

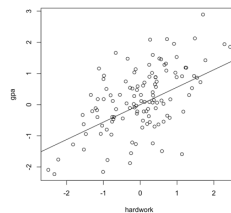
- Each point **on the line** represents our prediction of Y given a certain value of X.
- To the extent that an observed value does not fall on the line, we have a prediction error: the vertical (Y-axis) distance between the point and the line. See the arrow at the lower left.
- For the best fitting line, all of the positive errors (too high) and all of the negative errors (too low) will sum to zero.
- When we square all prediction errors and sum them, we get a measure of the overall error in the model.
- The `lm()` function chooses a slope that makes this value as small as possible.



26

Finding the Best Fitting Line

- The diagram at left visualizes a cloud of points representing scores on our "hardwork" predictor and our GPA criterion/outcome.
- The other two predictors are still in the data, we're just ignoring them for the moment.
- How can we decide the best slope and intercept for the line? What does it mean to have a good fit to these points?
- The `lm()` procedure uses the "least squares criterion" to select the one best value for the slope.



27

Regression Terminology

- Criterion/**dependent variable**: what we are trying to predict
- Predictor/**independent variable**: one of the variables we use to predict the criterion; there are usually multiple predictors
- Coefficients/weights: the strength of prediction for each predictor; sometimes also called B-weights (or the standardized version is called a beta-weight)
- Regression equation: the result of the regression analysis in the form of an algebraic equation
 - $Y = MX + B$
 - or...
 - $\hat{Y} = B_1X_1 + B_2X_2 + B_3X_3 + \dots$
 - Y-hat is the predicted Y, subscripts on Bs and Xs refer to predictor number

28

A Worked Example with R

SCHOOL OF INFORMATION STUDIES
SYRACUSE UNIVERSITY

29

GPA Model With All Three Predictors

```
regOut3 <- lm(gpa ~ hardwork + basicsmarts + curiosity, data=educdata)
summary(regOut3)
Call:
lm(formula = gpa ~ hardwork + basicsmarts + curiosity, data = educdata)
Residuals:
```

Min	1Q	Median	3Q	Max
-1.02063	-0.37301	0.00361	0.31639	1.32679

Residuals look good: median should be near zero and symmetric

30

Significance Tests on Predictors

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.08367	0.04575	1.829	0.07.
hardwork	0.56935	0.05011	11.361	<2e-16 ***
basicsmarts	0.52791	0.04928	10.712	<2e-16 ***
curiosity	0.51119	0.04363	11.715	<2e-16 ***

....

Residual standard error: 0.4978 on 116 degrees of freedom
Multiple R-squared: 0.7637; Adjusted R-squared: 0.7576
F-statistic: 125 on 3 and 116 DF, p-value: < 2.2e-16

B-weights

Significance test of the null hypothesis that $B = 0$

Notice the small penalty for having 3 predictors

F-statistic tests the null hypothesis that $R\text{-squared} = 0$

31

Interpreting the Model

- Adjusted R-squared value 0.7576
- Known as the coefficient of determination
- The proportion of the variation that is accounted for in the dependent variable by the whole set of independent variables.
- The closer to 1.0, the greater the influence the independent variable has on predicting the value of the dependent variable.
- The R-squared value of 0.7576 indicates that the three factors account for 75.76% of GPA.

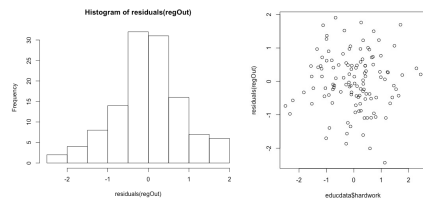
32

Evaluating Regression Results

- Examine the F-test on R-squared, **if the p-value is < 0.05, it is significant and OK to proceed**
- Examine the **adjusted R-squared value**: This translates as, "the proportion of variance in Y that is accounted for by all Xs working together."
- Look for **significant predictors** (but usually ignore the intercept) by seeing *, **, or ***
- Look in the "Estimate" column for the value of the slope on a significant predictor: One unit change in X causes that much change in Y

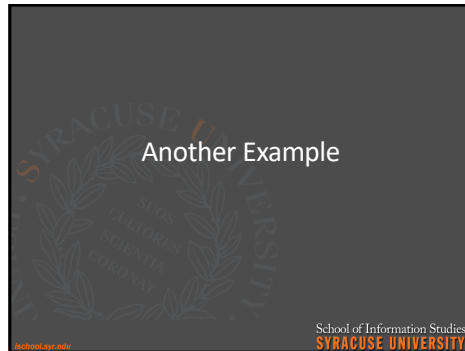
33

Residuals Are Errors of Prediction



34

Another Example



35

Car Maintenance

- We manage a "fleet" of cars
 - Cars get replaced every three years
 - Have information on:
 - Past repairs
 - Miles driven
 - # of oil changes during past three years
- How often to change the oil?
- Can we build a model to predict repair costs?

36

Question

In this new example, which are independent and which are dependent variables? Why?

	oilChanges	repairs	miles
1	3	300	100500
2	5	300	116000
3	2	500	136000
4	3	400	110500
5	2	700	150500
6	6	420	117000
7	6	100	89500
8	4	290	99500
9	3	475	100500
10	2	620	120500
11	0	600	106000
12	8	150	115000
13	7	200	104000
14	8	50	98500

37

Small Data Set, n=14

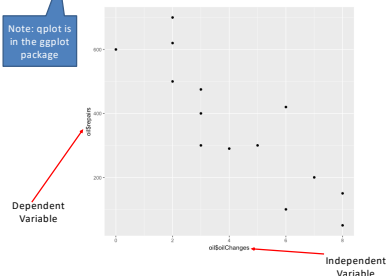
```
oilChanges <- c(3, 5, 2, 3, 2, 6, 6, 4, 3, 2, 0, 8, 7, 8)
repairs <- c(300, 300, 500, 400, 700, 420, 100, 290, 475, 620, 150, 200, 50)
miles <- c(100500, 116000, 136000, 110500, 150500, 117000, 89500, 99500, 100500, 120500, 106000, 115000, 104000, 98500)
oil <- data.frame(oilChanges, repairs, miles)
View(oil)
```

	oilChanges	repairs	miles
1	3	300	100500
2	5	300	116000
3	2	500	136000
4	3	400	110500
5	2	700	150500
6	6	420	117000
7	6	100	89500
8	4	290	99500
9	3	475	100500
10	2	620	120500
11	0	600	106000
12	8	150	115000
13	7	200	104000
14	8	50	98500

38

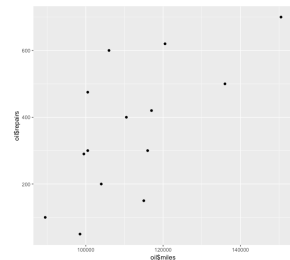
qplot(oil\$oilChanges, oil\$repairs)

Note: qplot is in the ggplot package



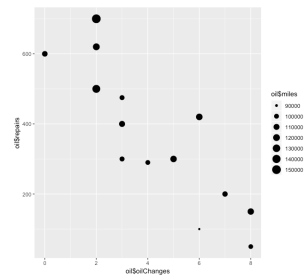
39

qplot(oil\$miles, oil\$repairs)



40

qplot(oil\$oilChanges, oil\$repairs,
size=oil\$miles)



41

```
lmOut <- lm(repairs ~ oilChanges + miles, data=oil)
summary(lmOut)
```

```
Call:
lm(formula = repairs ~ oilChanges + miles, data = oil)

Residuals:
    Min       1Q   Median       3Q      Max
-115.39  -38.76  -19.61   31.48   130.62

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  35.716647 172.343530   0.207  0.83961
oilChanges  -57.256143   8.945717  -6.400 5.08e-05 ***
miles         0.005104   0.001377   3.706  0.00346 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 74.71 on 11 degrees of freedom
Multiple R-squared:  0.8828, Adjusted R-squared:  0.8615
F-statistic: 41.44 on 2 and 11 DF, p-value: 7.56e-06
```

42

Interpreting the Model

- R-squared value 0.8615
- Known as the coefficient of determination
- The proportion of the variation that is accounted for in the dependent variable by the whole set of independent variables.
- The closer to 1.0, the greater the influence the independent variable has on predicting the value of the dependent variable.
- The R-squared value of 0.8615 indicates that the oil changes accounts for 86.15% of the cost of repairs.

43

What's an Oil Change Worth?

- The initial overall model was significant (F-test) and the oilChange predictor was significant with a slope of -57
- In the initial model, every oil change reduces the dollar amount of repairs by \$57

```
Call:
lm(formula = repairs ~ oilChanges + miles, data = oil)

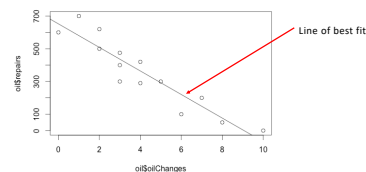
Residuals:
    Min       1Q   Median       3Q      Max
-115.39  -38.76  -19.61   31.48  130.62

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  35.716647  172.343530   0.207  0.83961
oilChanges  -57.256143   8.245717  -6.928 5.08e-05 ***
miles        0.005104    0.001377   3.706 0.00046 **
---
Signif. codes:  0.001 '***' 0.01 '**' 0.05 '*' 0.1 '.' 0.1 ' ' 1
```

44

Looking at the "abline"

abline(lmOut)



The model suggests that we should do as many oil changes as possible.
 → it predicts very low (almost 0) repairs if we do 9 or more oil changes, but about \$680 if we do no oil changes.

45

Predict New Values

```
> oilChanges <- c(1, 1, 2, 2, 3, 3)
> miles <- c(10000, 20000, 20000, 40000, 40000, 80000)
> oilPred <- data.frame(oilChanges, miles)
> predict(lmOut1, oilPred)
```

1	2	3	4	5	6
29.5	80.5	23.2	125.3	68.1	272.2

46

Questions

Is this an accurate model?

47

Working Through a Refined Example



School of Information Studies
SYRACUSE UNIVERSITY

48

Cost of Oil Change

- How “model” the cost?
- What might be some ranges of the cost?

49

Include the Cost of an Oil Change

What if oil changes cost \$350 each?

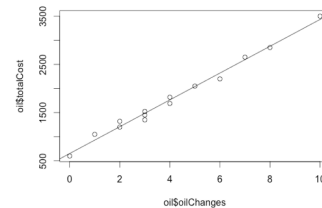
```
oil$oilChangeCost <- oil$oilChanges * 350
oil$totalCost <- oil$oilChangeCost +
  oil$repairs
```

```
m <- lm(formula=totalCost ~ oilChanges,
  data=oil)
```

50

What is oil changes cost \$350 each?

```
plot(oil$oilChanges, oil$totalCost)
abline(m)
```



51

Predict New Values

```
oilChanges <- c(1, 1, 2, 2, 3, 3)
miles <- c(10000, 20000, 20000, 40000, 40000, 80000)
oilPred <- data.frame(oilChanges, miles)
predict(m, oilPred)
```

1	2	3	4	5	6
379	430	723	825	1118	1322

52

Question

How accurate is the model?

- Did we have all the facts?
- Did we have all the data?

53

Project Information

School of Information Studies
SYRACUSE UNIVERSITY

54

Project Context

Role:

- You should act as a consultant

Goal:

- How might your client predict future energy usage?
 - Predict energy usage if the summer was 5 degrees warmer
 - Provide actionable insight into how to reduce energy costs

What data to analyze – should be:

- A function of what the team determines might be useful
- Determined by each project team
- There is *A LOT* of data

Remember this needs to be data driven –

55

Project Data

School of Information Studies
SYRACUSE UNIVERSITY

56

Static House Data

A file with basic house information for a random sample of single family houses that eSC serves.

- The file contains the list of all houses in the dataset.
- For each house, there is information describing the house.
 - The information ranges from the building id (used to access the energy data) to other house attributes that do not change (such as the size of the house).
- There are ~5,000 houses in the dataset (rows in the file)
- The file can be found at:
https://intro-datascience.s3.us-east-2.amazonaws.com/SC-data/static_house_info.parquet

Note that this file is in 'parquet' (an optimized for storage CSV file) format.

57

Energy Usage Data

- For each house, there is a file that contains energy usage data, which was collected hour-by-hour.
- There is one data file per house. Energy usage is:
 - Collected every hour
 - Collected across many sources (ex. air conditioning system, dryer
 - the 'building ID' is file name which identifies the house.
- Note that each file is in 'parquet' (an optimized for storage CSV file).
- All the data is in one folder on amazon AWS.
- For example, the following URL is for 'building_id' 102063.

<https://intro-datascience.s3.us-east-2.amazonaws.com/SC-data/2023-houseData/102063.parquet>

There are approximately 5,000 houses (i.e., different building ID's) in the directory

58

Weather Data

- Hour-by-hour weather information (one file for each geographic area)
- The weather data was collected for each county and stored based on a county code:
 - The county code for each house can be found at 'in.county' column of the house static dataset. This file is in a simple CSV format.
 - For example, the following URL provides the weather for county 'G4500010'.

<https://intro-datascience.s3.us-east-2.amazonaws.com/SC-data/weather/2023-weather-data/G4500010.csv>

There are approximately 50 counties in the directory.

59

Meta Data

- A data description file, explaining the fields used across the different housing data files.

~270 attributes

https://intro-datascience.s3.us-east-2.amazonaws.com/SC-data/data_dictionary.csv

60



61

Suggested Project Steps

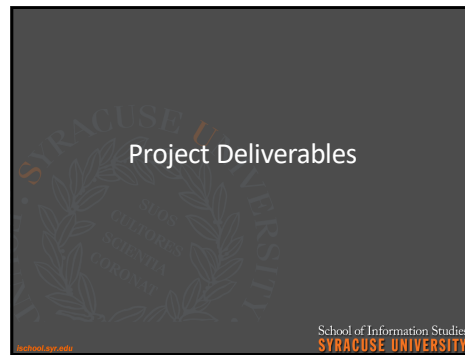
- a) Determine the best approach to read and merge the data
- b) Determine what should be the output during this 'data preparation' phase.
- b) Do exploratory analysis of the data
- c) Build a model that predicts the energy usage, for any given hour, for the month of July.
 → July was selected, as eSC thought July is typically the highest energy usage month.
 Hint: you will need to try several models and pick the best model.
- d) Understand and be able to explain your model's accuracy.

62

Suggested Project Steps

- e) Create a new weather dataset → all July temps 5 degrees warmer
- f) Use your best model to evaluate peak future energy demand
 → assume no new customers
 Note: this must be model driven, not just increasing energy usage by a percentage
- g) Show future peak energy demand in total (for an hour):
 - For different geographic regions
 - For other dimensions /attributes you think important
- h) Create a shiny application to interact with the data
- i) Identify one approach to reduce peak energy demand
- j) What would you suggest, how would you model the impact.
- k) How would you explain the impact. BE DATA DRIVEN

63



64

Project Deliverables

Word Document:

- Target audience is your manager / instructor (hint: your manager/instructor is a data science expert)
- Focus on what was accomplished
- Should describe all analysis done, even if an analysis did not generate any interesting results, it should still be included

Presentation:

- Target audience is your client (hint: the client is not a data science expert)
- Presentation length is 10 minutes (lab instructor will explain specifics)
- Be sure to include the following in your presentation:
 - Number of records in dataset evaluated
 - Key drivers identified; accuracy of results

65

Project Deliverables (continued)

Interactive Application (shiny app):

- A shiny app needs to be created and deployed on shinyapps.io
- To better understand your model's energy prediction
- To better understand the potential future energy needs and/or savings

66

Expectations

- 1) Work at a consistent pace throughout the rest of the semester
- 2) Tasks should be distributed equally across the team members
- 3) Tasks should typically not take a long time to complete – one week target, two weeks is fine, but not a month
- 4) Tasks should be at an appropriate level of effort / detail

67

Project Updates

- 1) Project Updates: Nov 2, Nov 16, Nov 30
(one per group, not per person)
- 2) For each update (including for the final submission), provide:
 - a) Work done by each person (since the last update)
 - b) Work planned to be done by each person (by the next update)
 - c) Key issues / challenges

68

Project Grading

School of Information Studies
SYRACUSE UNIVERSITY

69

Presentation (5%)

- 0.5% - Business Questions** - are they appropriate within the context
- 0.5% - Use of Descriptive statistics** - Did you provide context and a basic understanding of the data
- 1% - Use of modeling techniques** - Did you try at least 3 different models and explain why they were/were not useful
- 1% - Visualization** - Did you convey the results in an easy to understand manner
- 1% - Interpretation of the results/Actionable Insights** - Are the results actionable (as compared to just interesting)
- 1% - Know your audience** - did the presentation present findings in an easy to understand way (ex. no data science lingo, easy for others to follow the logic)

70

Shiny App (5%)

- 1% - App can load / use a data file provided by the user**
- 0.5% - Display of the first 'n' rows of the read in dataset**
- 1% - Generate predictions via a stored model**
- 0.5% - Display the Confusion Matrix**
- 2% - An explanation within the app, of how to interpret the Confusion matrix:**
- Which numbers to "look at"
 - What is a "good number"

71

Word Document (14%)

- 1% - Business Questions** - appropriate within the context?
- 1% - Data cleanse/munge/preparation** - transform/clean/munge the data appropriately? What about NAs?
- 1% - Use of Descriptive statistics** - provide context and a basic understanding of the data?
- 4% - Use of modeling techniques** - try at least 3 different models
- 3% - Visualization** - convey information in an easy to understand manner
- 4% - Interpretation of the results/Actionable Insights** - Are the results actionable (as compared to just interesting)
- 1% - Validation** - How do you know your results were correct (i.e., no errors)

72

Final Project (Word doc)

Example Table of Contents:

- Introduction (scope/context/background)
- Business Questions addressed
- Data Acquisition, Cleansing, Transformation, Munging
- Descriptive statistics & Visualizations
- Use of modeling techniques & Visualizations (noting techniques explored but not used in presentation)
- Actionable Insights / Overall interpretation of results
- Appendix – Code (can be link to the code)

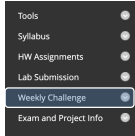
73

Weekly Quiz

In Blackboard

Time Limit:
5 Minutes

Password:
week8



74
