Rensselaer

why not change the world?®

Regression: Least Squares Method, Decision Trees

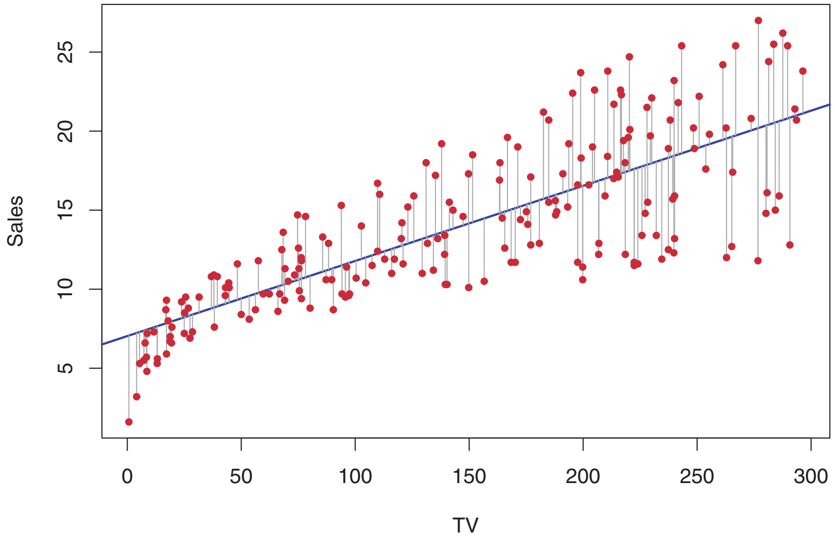
Ahmed Eleish

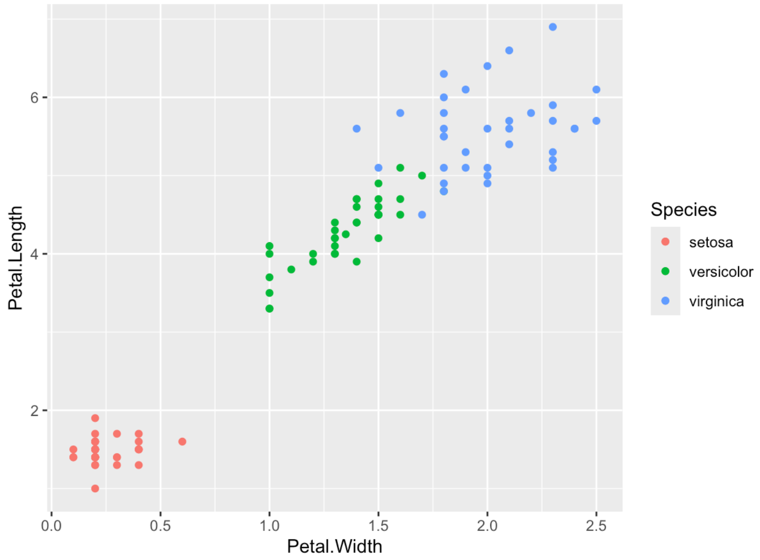
Data Analytics ITWS-4600/ITWS-6600/MATP-4450/CSCI-4960

October 8th 2024

Tetherless World Constellation
Rensselaer Polytechnic Institute

Quick Review





x-axis: independent numeric variable
y-axis: dependent numeric variable

Look for:

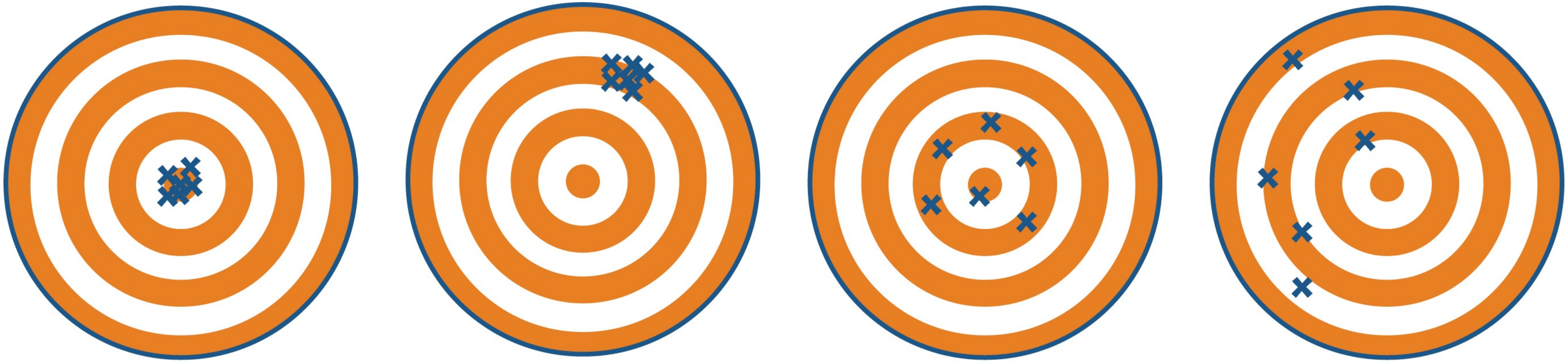
- trend? direction?
- are points tightly grouped?

x-axis: numeric variable
y-axis: numeric variable

Look for:

- structure: groups? group separation?

Accurate vs. Precise



High Accuracy
High Precision

Low Accuracy
High Precision

High Accuracy
Low Precision

Low Accuracy
Low Precision

http://climatica.org.uk/climate-science-information/uncertainty

Linear Regression

Regression

Linear Regression: In regression, fitting covariate and response data to a line is referred to as linear regression.
Covariate: A variable that is possibly predictive of the outcome under study control variable, ***explanatory variable, independent variable, predictor***
Response: dependent variable
Intercept: The expected value of the response variable when the value of the predictor variable is 0.
Slope: the average increase in Y associated with a one-unit increase in X

Reference/Resources:
The Elements of Statistical Learning. Hastie • Tibshirani • Friedman, 2nd Edition.
Introduction to Probability and Statistics, 4th Edition by Beaver.
Introduction to Statistical Learning with R, 7th Edition (ISLR).

Simple Linea Regression

- Let's take a look at the Least Squares Method for a single covariate (single regression).

- Utilizing the statistical notion of estimating parameters from data points, we find the estimates (coefficients) using the least squares method.

- We will look at evaluating linear models.

Least Squares Method

Equation of line: $\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x$

Let *n* be a positive integer. For a given data *(x_1,y_1), ..., (x_n,y_n)* $\in \mathbb{R} \times \mathbb{R}$,

- we obtain the intercept β_0 and slope β_1 using the least squares method.
- Residual Sum of Squares (RSS), the *i*th residual $e_i = y_i - \hat{y}_i$

$$\mathrm{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2$$

Or

$$\mathrm{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

More precisely, we minimize RSS

$$\text{RSS} = \sum_{i=1}^{n}(y_i - \widehat{\beta_0} - \widehat{\beta_1}x_i)^2$$

Sum of squared distances between (x_i, y_i) and $(x_i, \widehat{\beta_0} + \widehat{\beta_1}x_i)$ over i = 1,...,n

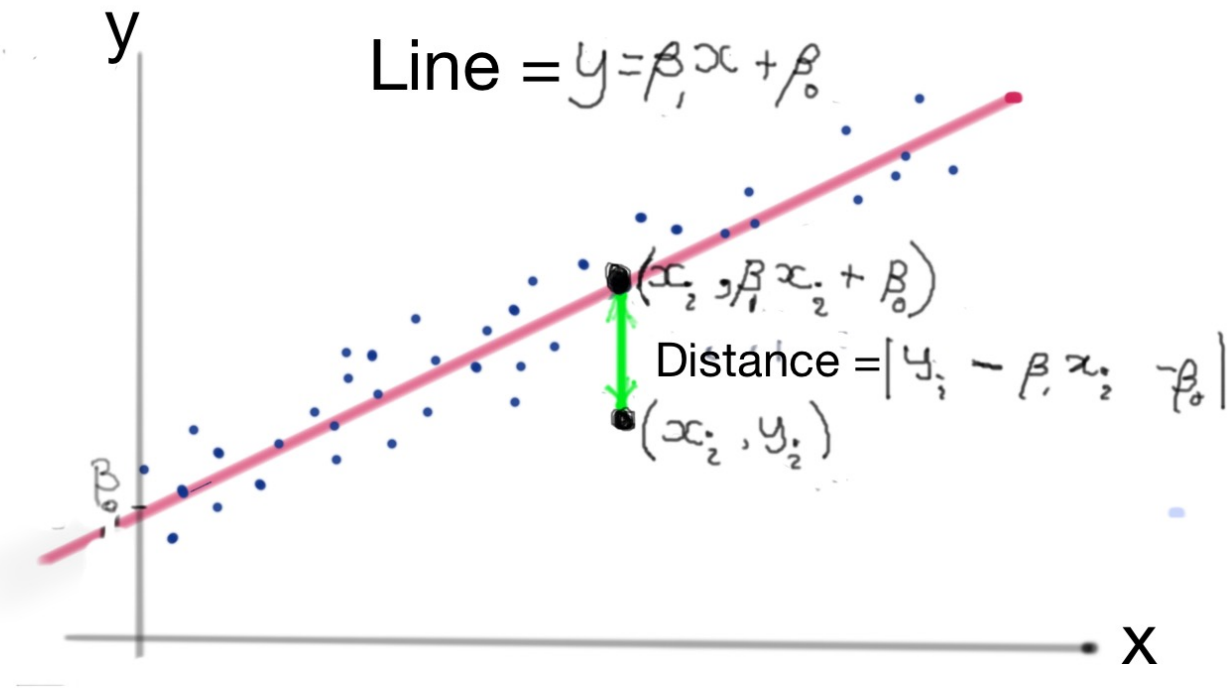


Figure: obtain $\widehat{\beta_0}$ and $\widehat{\beta_1}$ that minimize $\sum_{i=1}^{n}(y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i)$ via least squares method

- We partially differentiate L by β_0 and β_1 and let them be equal to zero, we obtain the following equations:

$$\frac{\partial L}{\partial \widehat{\beta_0}} = -2\left(\sum_{i=1}^{n}\left(y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i\right)\right) = 0 \qquad \text{Eq(1)}$$

$$\frac{\partial L}{\partial \widehat{\beta_1}} = -2\left(\sum_{i=1}^{n} x_i\left(y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i\right)\right) = 0 \qquad \text{Eq(2)}$$

Where the partial derivative is calculated by differentiating each variable and regarding the other variables as constants. In this case, β_0 and β_1 are regarded as constants when differentiating L by β_0 and β_1 respectively.

- By solving Eq (1) and Eq (2) when:

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 \neq 0 \qquad \text{Eq(3)}$$

i.e., $x_1 = x_2 = \cdots = x_N$ is not true.

Where:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{n} x_i \qquad \bar{y} = \frac{1}{N}\sum_{i=1}^{n} y_i$$

- We can obtain:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})} \qquad \text{Eq(4)}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_0\bar{x} \qquad \text{Eq(5)}$$

Assessing the Coefficient Estimates

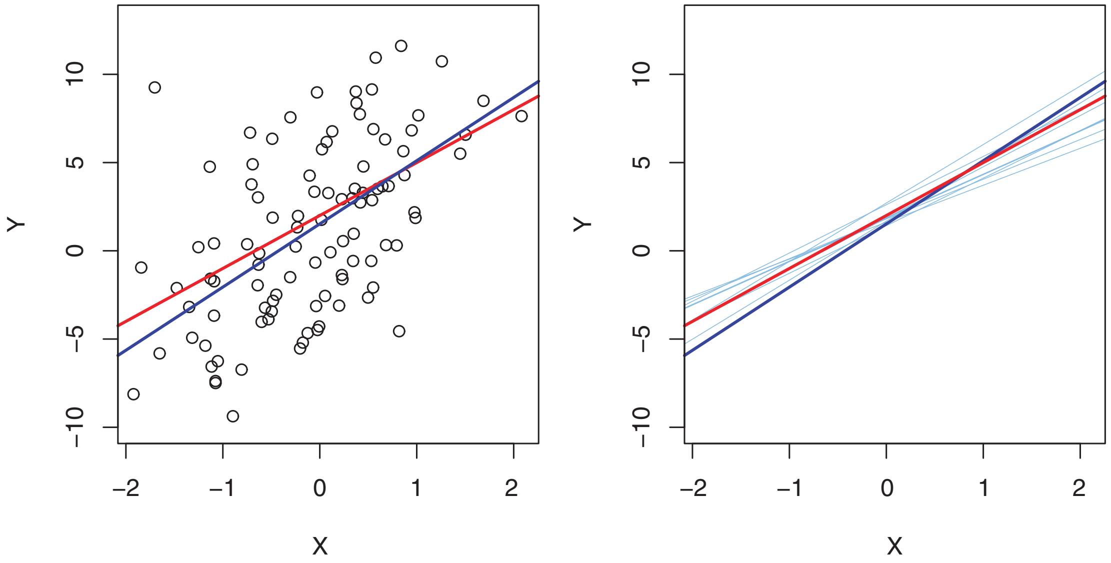
True relationship between X and Y: $$Y = \beta_0 + \beta_1 X + \epsilon.$$

- Where ϵ is a mean-zero random error

Red: true relationship

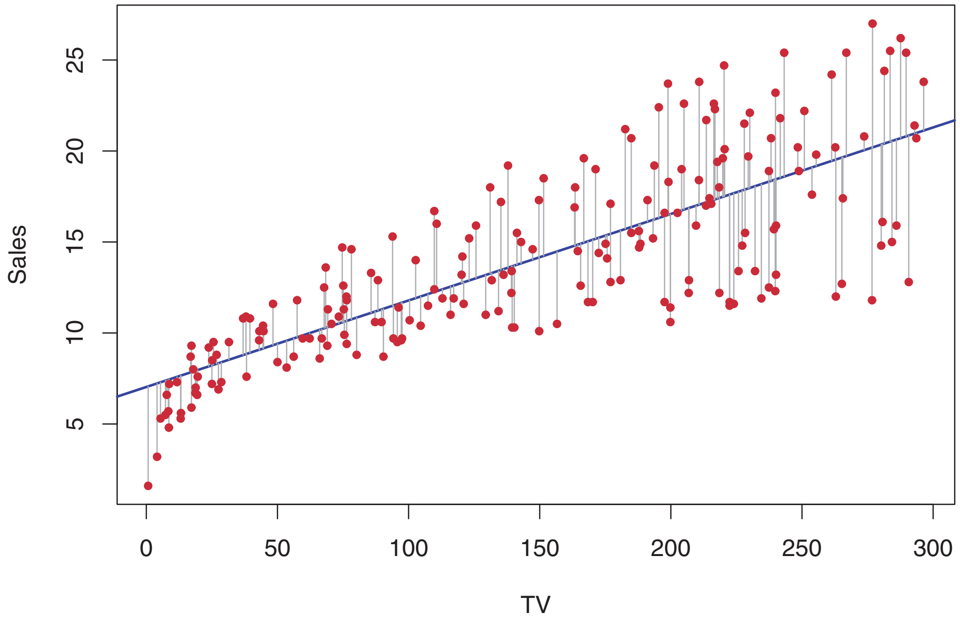
Dark Blue: least squares regression line

Light Blue: least squares regression lines for multiple random subsets



Evaluating Linear Models

- Sales vs. TV ad spending
- Sales in 1000s of units
- TV ad spending in 1000s of $



Evaluating Linear Models

Values of coefficients >> their Std. errors

High t-statistic

Very low p-value

	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

Hypothesis (more TV ads → more sales)

H0 : There is no relationship between X and Y

Ha : There is some relationship between X and Y

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)},$$

Reject the null hypothesis!

Residual Standard Error

- Mean sales $\approx$ 14,000 units

RSE = 3.26 = 3,260 units
good/bad?

R^2
- measures the proportion of the variability in *Y* that can be explained using *X*
- has a value between 0,1

Quantity	Value
Residual standard error	3.26
R^2	0.612
F-statistic	312.1

$$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

$$\mathrm{TSS} = \sum (y_i - \bar{y})^2$$

Random Forest(s)

Random Forest

- Random Forest is based on decision trees.
- In Random Forest we build large number of trees, where each tree is based on a bootstrap sample.
- Then, what we do is we average those predictions together in order to get the predictive probabilities of each class across all the different trees.

Random Forest

Cons:

- Speed (it can be slow; it has to build large numbers of trees)

- Interpretability (it can be hard to interpret in the sense that you have large number of trees that are averaged together and those trees represent the bootstrap samples and are complicated to understand)

Random Forest

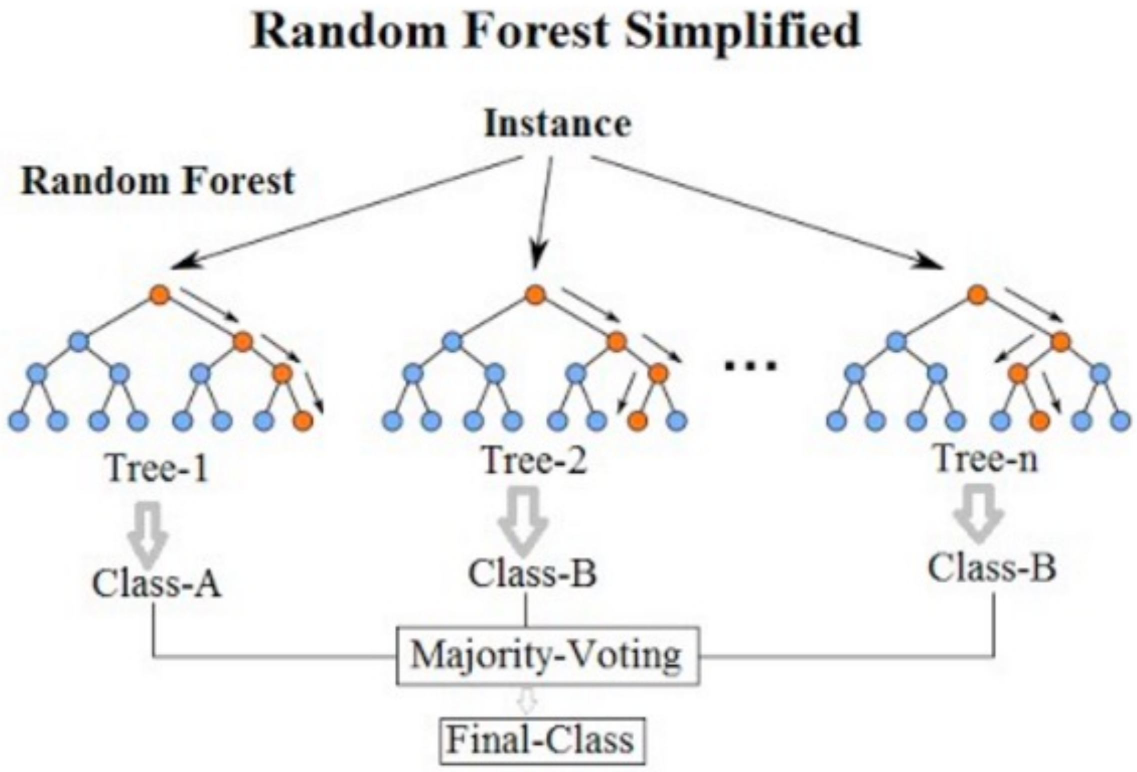


Image Resource: https://commons.wikimedia.org/wiki/File:Random_forest_diagram_complete.png

Random Forest

The original algorithm was created in 1995 by Tin Kam Ho.

An extension of the algorithm was developed by Leo Breiman and Adele Cutler, who registered "Random Forests" as a trademark in 2006

- http://www.stat.berkeley.edu/~breiman/RandomForests/

Random Forest Algorithm

- Let N_{trees} be the number of trees to build

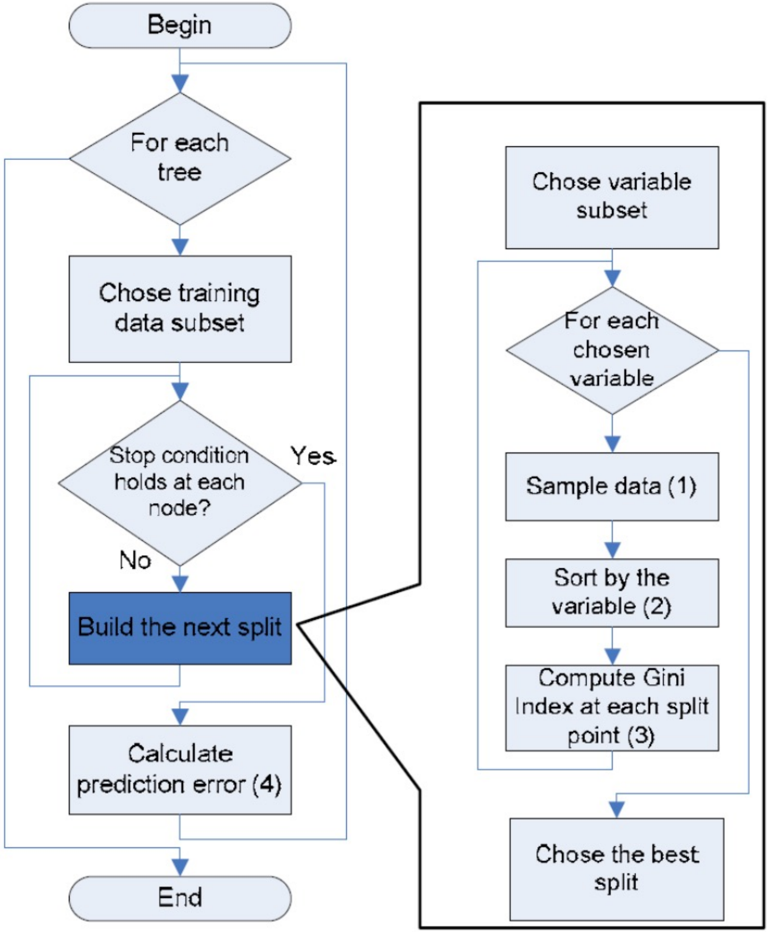
for each of N_{trees} iterations:

1. Select a new bootstrap sample from training set
2. Grow an un-pruned tree on this bootstrap.
3. At each internal node, randomly select m_{try} predictors and determine the best split using only these predictors.
4. Do not perform cost complexity pruning. Save tree as is, along side those built thus far.

Output overall prediction as the average response (regression) or majority vote (classification) from all individually trained trees

Ref: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=5f31bcc21ab2155c084527648d436b036126b30d

Random Forest



Image/ Photo Credit: Albert A. Montillo

Random Forest exercise

Code: https://rpi.box.com/s/bhdyyq3otux7kurbn7jnf6jrestrmle3

Thanks!