Things To Know

Validation of a Model

- Validate that the values the model predicts quantify the relationship between independent and dependent variables

Validation of a Model

- Validate that the values the model predicts quantify the relationship between independent and dependent variables
- We have already seen one way of validating a model

Validation of a Model

- Validate that the values the model predicts quantify the relationship between independent and dependent variables
- We have already seen one way of validating a model
  - GOODNESS OF FIT!

Validation of a Model

- Validate that the values the model predicts quantify the relationship between independent and dependent variables
- We have already seen one way of validating a model
    - GOODNESS OF FIT!
- But goodness of fit relies on measuring the performance of the model on the data you used to build the model in the first place

Validation by Predicting New Values

- Step 1. Generate a linear regression model on some data (sample)

Validation by Predicting New Values

- Step 1. Generate a linear regression model on some data (sample)
- Step 2. Predict the dependent values of unused data using the fit model

Validation by Predicting New Values

- But first, we split the data into training and testing sets

Validation by Predicting New Values

- But first, we split the data into training and testing sets
- Training data for fitting the model
- Testing data for validating the model performance

Validation by Predicting New Values

- But first, we first split the data into training and testing sets
- Training data for fitting the model
- Testing data for validating the model performance
- We can calculate the R-squared values of the two data sets and compare them

Cross Validation

- Say you split your data 70/30 randomly into training and testing sets

Cross Validation

- Say you split your data 70/30 randomly into training and testing sets
  - That is, you train your model (fit the best-fit line) on 70% of the data
  - And test your model's performance (predict y values and compare) on 30% of the data

Cross Validation

- Say you split your data 70/30 randomly into training and testing sets
  - That is, you train your model (fit the best-fit line) on 70% of the data
  - And test your model's performance (predict y values and compare) on 30% of the data
- In this scenario, there is a possibility that even though the data split is random, a key subset of your data ends up in the training set and is missing from the test set

Cross Validation

- Say you split your data 70/30 randomly into training and testing sets
  - That is, you train your model (fit the best-fit line) on 70% of the data
  - And test your model's performance (predict y values and compare) on 30% of the data
- In this scenario, there is a possibility that even though the data split is random, a key subset of your data ends up in the training set and is missing from the test set
  - When looking at NBA Elo data, our training set could have all play off games vs. the test set, with an 80/20 split

Cross Validation

- Say you split your data 70/30 randomly into training and testing sets
    - That is, you train your model (fit the best-fit line) on 70% of the data
    - And test your model's performance (predict y values and compare) on 30% of the data
- In this scenario, there is a possibility that even though the data split is random, a key subset of your data ends up in the training set and is missing from the test set
    - When looking at NBA Elo data, our training set could have all play off games vs. the test set, with an 80/20 split
- How could this impact our model's performance?

Cross Validation

- Say you split your data 70/30 randomly into training and testing sets
  - That is, you train your model (fit the best-fit line) on 70% of the data
  - And test your model's performance (predict y values and compare) on 30% of the data
- In this scenario, there is a possibility that even though the data split is random, a key subset of your data ends up in the training set and is missing from the test set
  - When looking at NBA Elo data, our training set could have all play off games vs. the test set, with an 80/20 split
- How could this impact our model's performance?
- What is a viable alternative? We don't want to get rid of training and testing sets!

k-Fold Cross Validation

- Instead of splitting the data into a training and testing set and validating the model once, we repeat the process, k times.

k-Fold Cross Validation

- Instead of splitting the data into a training and testing set and validating the model once, we repeat the process, k times.
- So not only are we validating the model multiple times, but with unique training and testing sets every time

k-Fold Cross Validation

1. Shuffle the data randomly
2. Split the data into k groups
3. For each unique group:
    a. Take the group as test data set
    b. Combine the remaining groups as training sets
    c. Fit the model on the training set and evaluate it on the test set
    d. Record the evaluation score and discard the model

- We can summarize the model using all evaluation scores recorded

k-fold Cross Validation

- Note that each observation (data point) in the data is assigned to an individual group and stays in that group for the duration of the procedure.

k-fold Cross Validation

- Note that each observation (data point) in the data is assigned to an individual group and stays in that group for the duration of the procedure.
- How many times is each group used in training and testing sets?

k-fold Cross Validation

- Note that each observation (data point) in the data is assigned to an individual group and stays in that group for the duration of the procedure.
- How many times is each group used in training and testing sets?
- 1 time for testing and k-1 times for training

Transformations

- One of the most important assumptions when fitting a linear regression model is that there exists a linear relationship between the independent and the dependent variables

Transformations

- One of the most important assumptions when fitting a linear regression model is that there exists a linear relationship between the independent and the dependent variables
- Often times the linear relationship is not clear

Transformations

- One of the most important assumptions when fitting a linear regression model is that there exists a linear relationship between the independent and the dependent variables
- Often times the linear relationship is not clear
- In that case, different transformations can be applied to the data to make the linear relationship clearer

Transformations

- $\hat{y}_i = b_0 + b_1x_i$ - a unit increase in x is associated with an average of b_1 units increase in y
-
- $\log(\hat{y}_i) = b_0 + b_1x_i$ - a unit increase in x is associated with an average of b_1 units increase in log(y)
- $\log(\hat{y}_i) = b_0 + b_1 \log(x_i)$ - a k-fold increase in x is associated with k^b multiplicative increase in y
    - If x doubles, y changes by a multiplicative factor of 2^b

Transformations (Standardization)

- When modeling a multiple linear regression model, we might have to deal with independent variables that correspond to different units

Transformations (Standardization)

- When modeling a multiple linear regression model, we might have to deal with independent variables that correspond to different units
    - Age column on a scale of 20-70 vs. Salary column on a scale of 30,000-70,000

Transformations (Standardization)

- When modeling a multiple linear regression model, we might have to deal with independent variables that correspond to different units
  - Age column on a scale of 20-70 vs. Salary column on a scale of 30,000-70,000
- When we want to build a model where variables have different units, so a large value in one scale does not overpower a relatively smaller value in a different scale, we standardize our data

Transformations (Standardization)

- When modeling a multiple linear regression model, we might have to deal with independent variables that correspond to different units
    - Age column on a scale of 20-70 vs. Salary column on a scale of 30,000-70,000
- When we want to build a model where variables have different units, so a large value in one scale does not overpower a relatively smaller value in a different scale, we standardize our data
- Performed by converting data into Z-scores
    - mean = 0, sd = 1

Transformations (Standardization)

- $Z \sim N(0,1)$
- $z = (x - \text{mean}) / sd$
-
- Done separately for each attribute