Handling Missing Data

- The data obtained could have certain values missing.
- There are various means to handles these situations.
- **Missing values impact training and predictions**.
  - Training Data : Contains "unknown" values.
  - Predictions : Input at prediction time contains "unknown" values.

Strategy 1 : Purification by Skipping

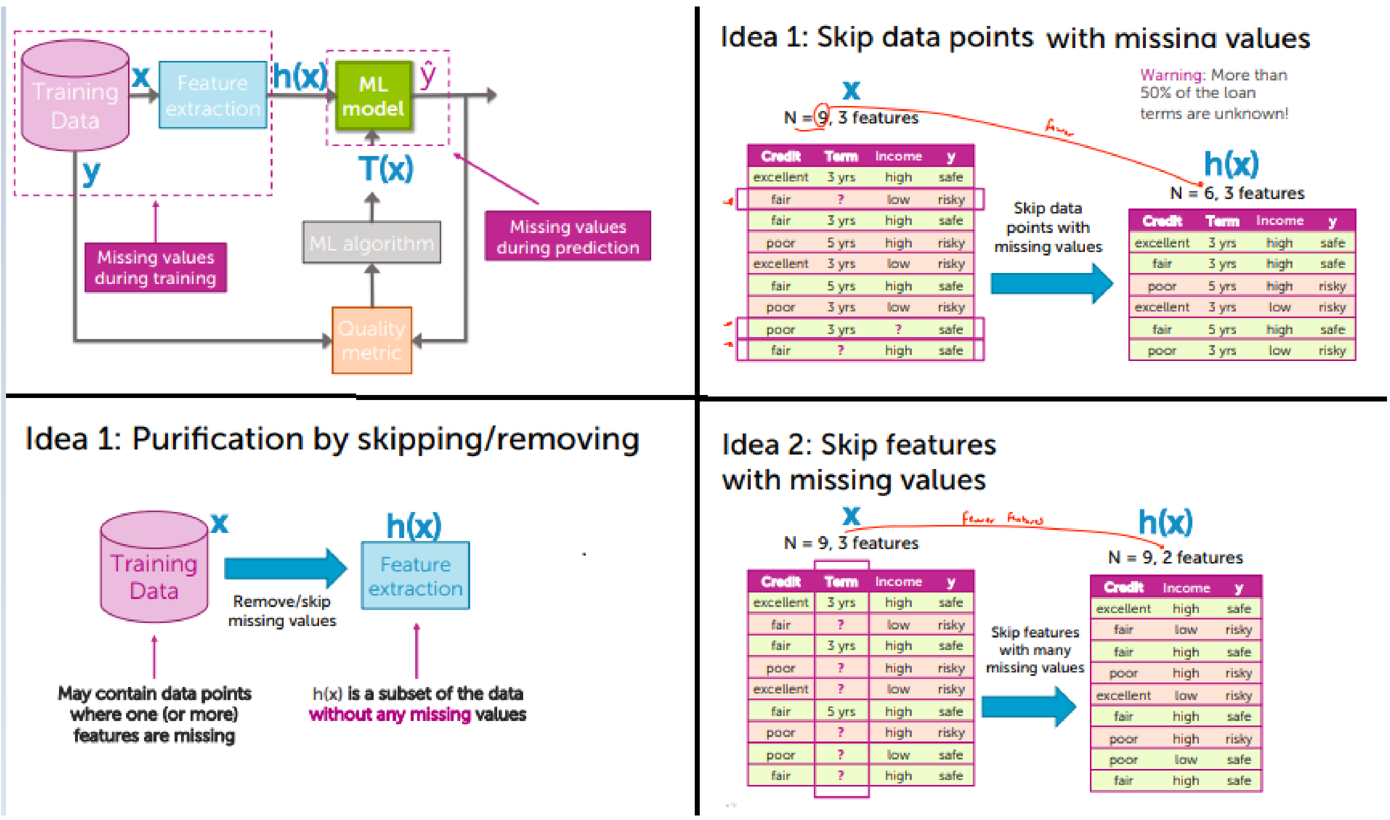
Idea 1 - Skipping datapoints missing values

```
* In case more than 50% of the data has a ceratin value missing, then removing tho
se records from the data could be problematic. It can effect the training process
 as enough scenrios will not be tested.
* Make sure only a few data points are skipped.
```

Idea 2 - Skipping features with missing values

```
* The resultant dataset will have fewer features.
* Should make sure only a few features are skipped.
```

- **Pros :**
  - Easy to understand and implement.
  - Can be applied to any model (decision trees, logistic regression, linear regression, ...)
- **Cons :**
  - Removing datapoints and features may remove important information from the data.
  - Unclear when it's better to remove data points versus features.
  - Doesn't help if data is missing at prediction time.



Strategy 2 : Purification by imputing

- Instead of throwing away the data, impute the data -> no need to reduce the datatset.

Idea - Purification by imputing

- Replace the value with the most common value.

Common rules for purification by imputation

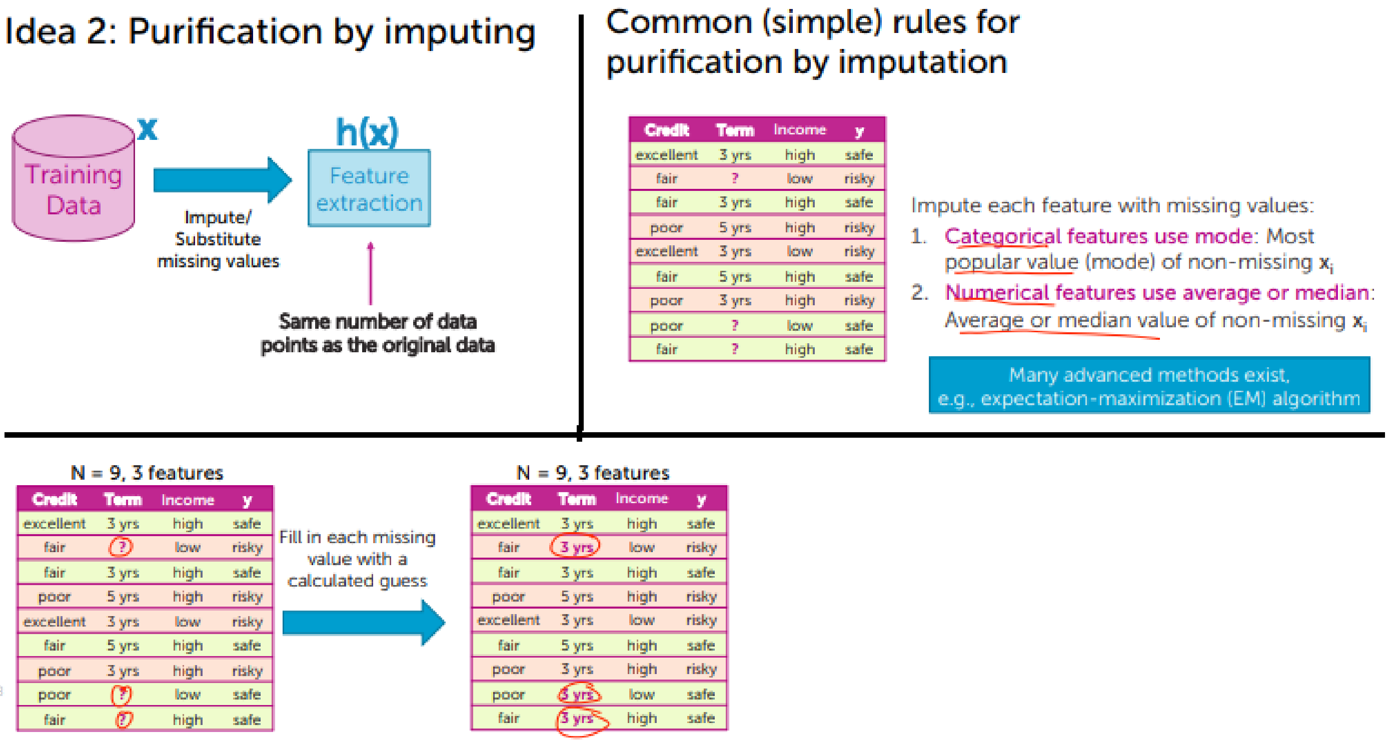
- Impute each feature with missiing values.
    1. **Categorical features use 'mode'** - Most popular value (mode) of non-missing xi.
    2. **Numerical features use average or median** - Average or median value of non-missing xi.

Pros:

- Easy to understand and implement.
- Can be applied to any model (decision trees, logistic regression, linear regression, ...)
- Can be used at prediction time - same imputation rules.

Cons:

- May result in systematic errors.
- Could be biased.
- Example - If a loan application doesn't take into consideration the age of the applicant - and assumes it to be 40, then everybody is elligible for loan.

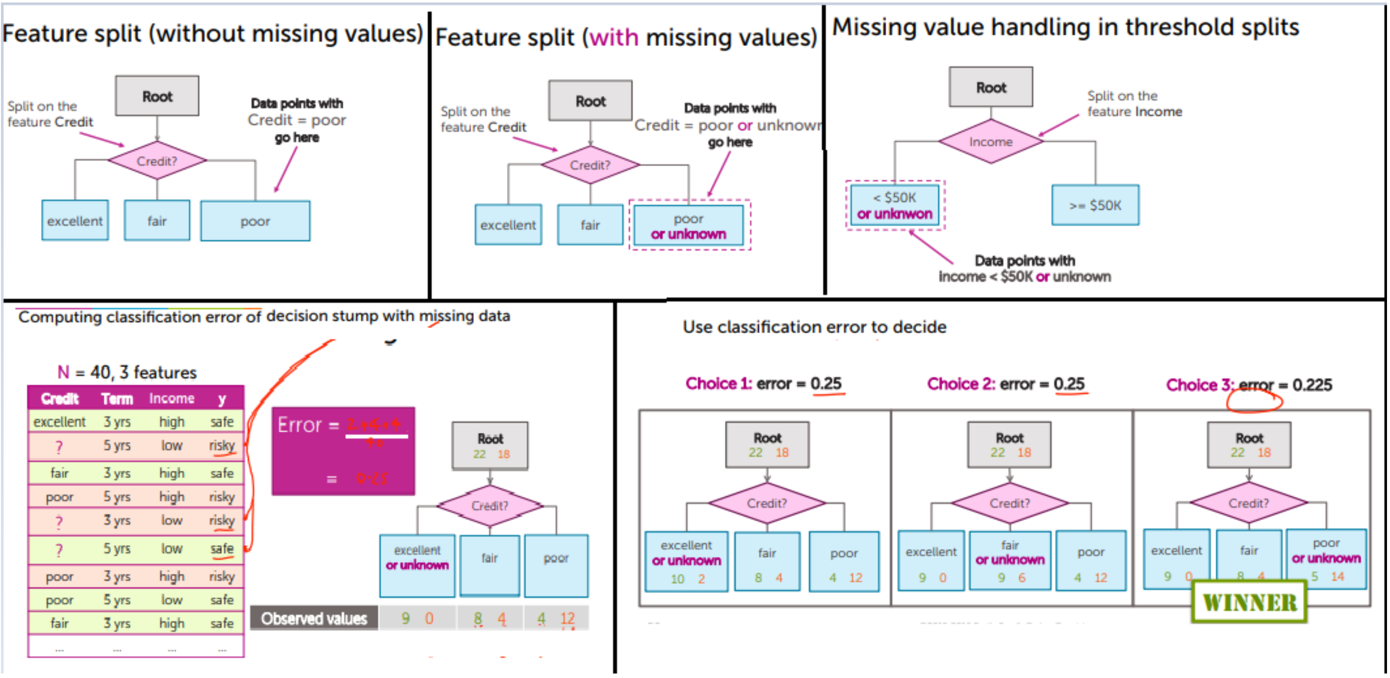


Strategy 3 : Adapt learning algorithm to be robust to missing values

- Add missing value choice to every decision node.(Default missing value node traversal decisions).
    - At credit -> **poor**

- At credit -> Term -> **3 year**
- At credit -> income -> **low**
- At credit -> income p-> term -> **5 year**

- Explicitly handle missing data by learning algorithm.
- For a given feature , while selecting the branch for the missing values, select the branch with **least classification error.**

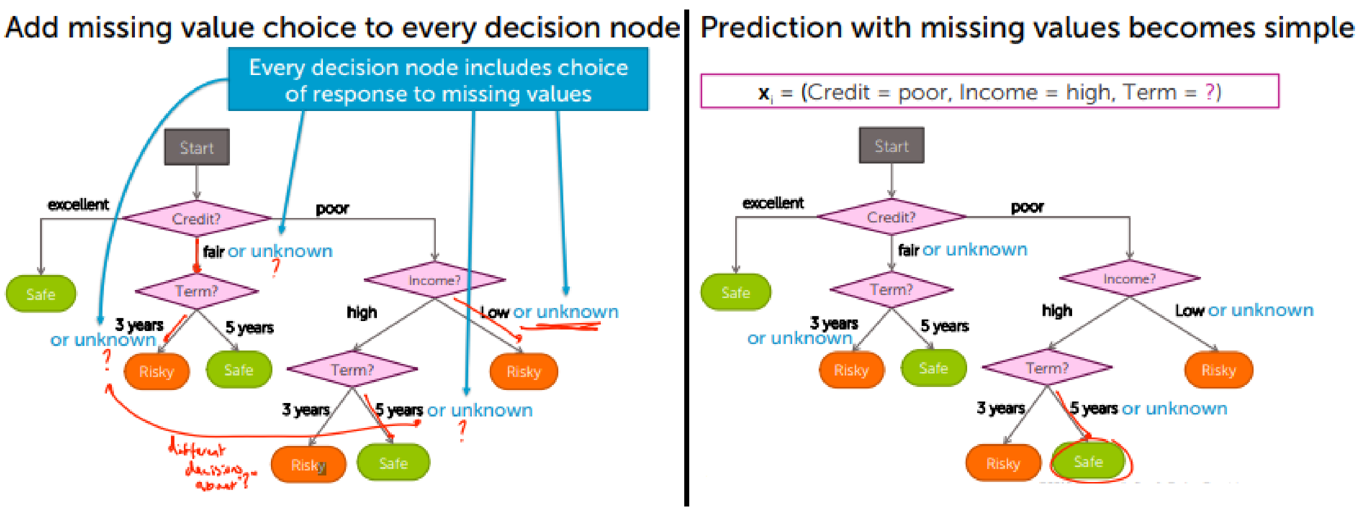


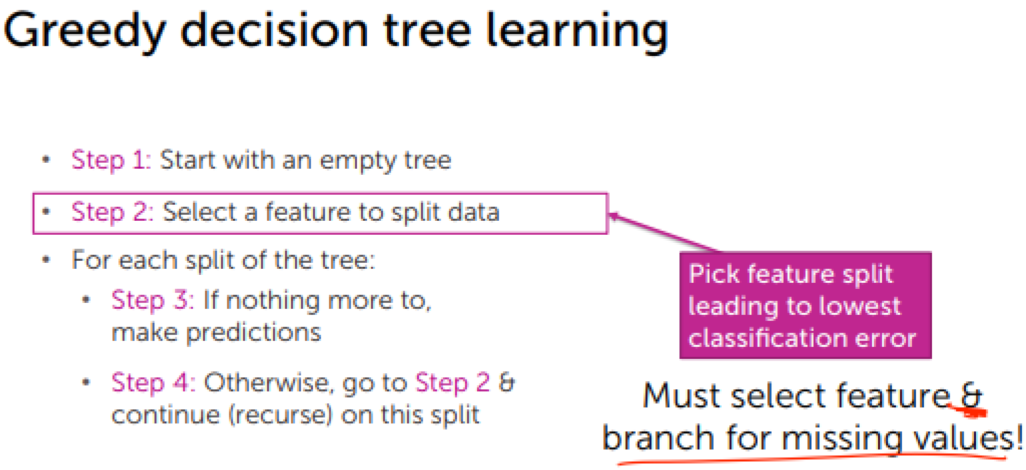
Pros :

- Addresses training and prediction time.
- More accurate predictions.

Cons :

- Requires modification of learning algorithm.
  - Very simple for decision trees.





Summary :

1. Skip all rows with any missing values.
2. Skip all features with many missing values.
3. Impute missing values using other data points.

- Modify the learning algorithm:
  - Missing values get added to one branh of split.
  - Use classification error to determine where missing values go.

Quiz

1. (True/False) Skipping data points (i.e., skipping rows of the data) that have missing features only works when the learning algorithm we are using is decision tree learning.

   - ( ) True
   - (•) False

---

2. What are potential downsides of skipping features with missing values (i.e., skipping columns of the data) to handle missing data?

   - [x] So many features are skipped that accuracy can degrade
   - [ ] The learning algorithm will have to be modified
   - [ ] You will have fewer data points (i.e., rows) in the dataset
   - [x] If an input at prediction time has a feature missing that was always present during training, this approach is not applicable.

---

3. (True/False) It's always better to remove missing data points (i.e., rows) as opposed to removing missing features (i.e., columns).

   - ( ) True
   - (•) False

---

4. Consider a dataset with N training points. After imputing missing values, the number of data points in the data set is

- ( ) 2 * N
- (x) N
- ( ) 5 * N

---

5. Consider a dataset with D features. After imputing missing values, the number of features in the data set is

- ( ) 2 * D
- (x) D
- ( ) 0.5 * D

---

6. Which of the following are always true when imputing missing data? Select all that apply.

- [x] Imputed values can be used in any classification algorithm
- [x] Imputed values can be used when there is missing data at prediction time
- [ ] Using imputed values results in higher accuracies than skipping data points or skipping features

---

7. Consider data that has binary features (i.e. the feature values are 0 or 1) with some feature values of some data points missing. When learning the best feature split at a node, how would we best modify the decision tree learning algorithm to handle data points with missing values for a feature?

- (x) We choose to assign missing values to the branch of the tree (either the one with feature value equal to 0 or with feature value equal to 1) that minimizes classification error.
- ( ) We assume missing data always has value 0.
- ( ) We ignore all data points with missing values.