ETH zürich

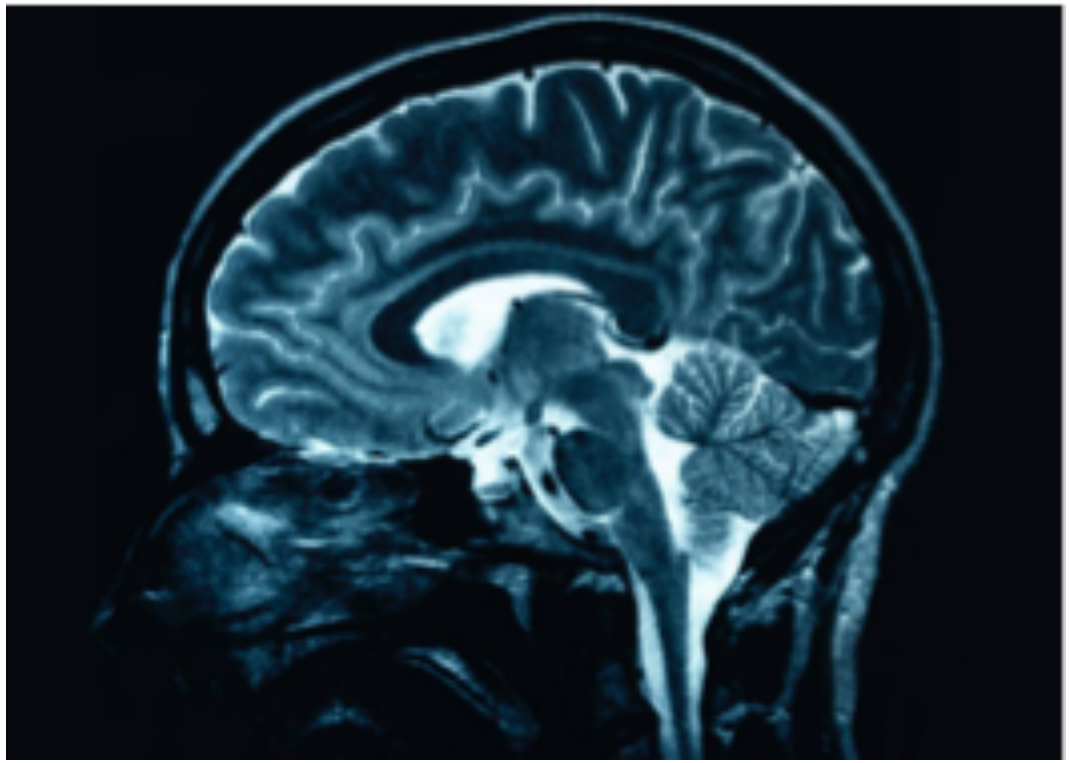
D INFK

Advance Machine Learning

Practical Exercises: **Task 1**

Laura Manduchi

Predict a person‘s age using features extracted from brain image data (MRI).



Introduction

Magnetic Resonance Imaging (MRI)

- It uses a magnetic field and radio waves to produce three dimensional detailed anatomical images.
- Non-invasive image technology and non-radiative investigation of sensitive organs.
- It produces high-resolution images.
- MRI machines are large, tube-shaped magnets.

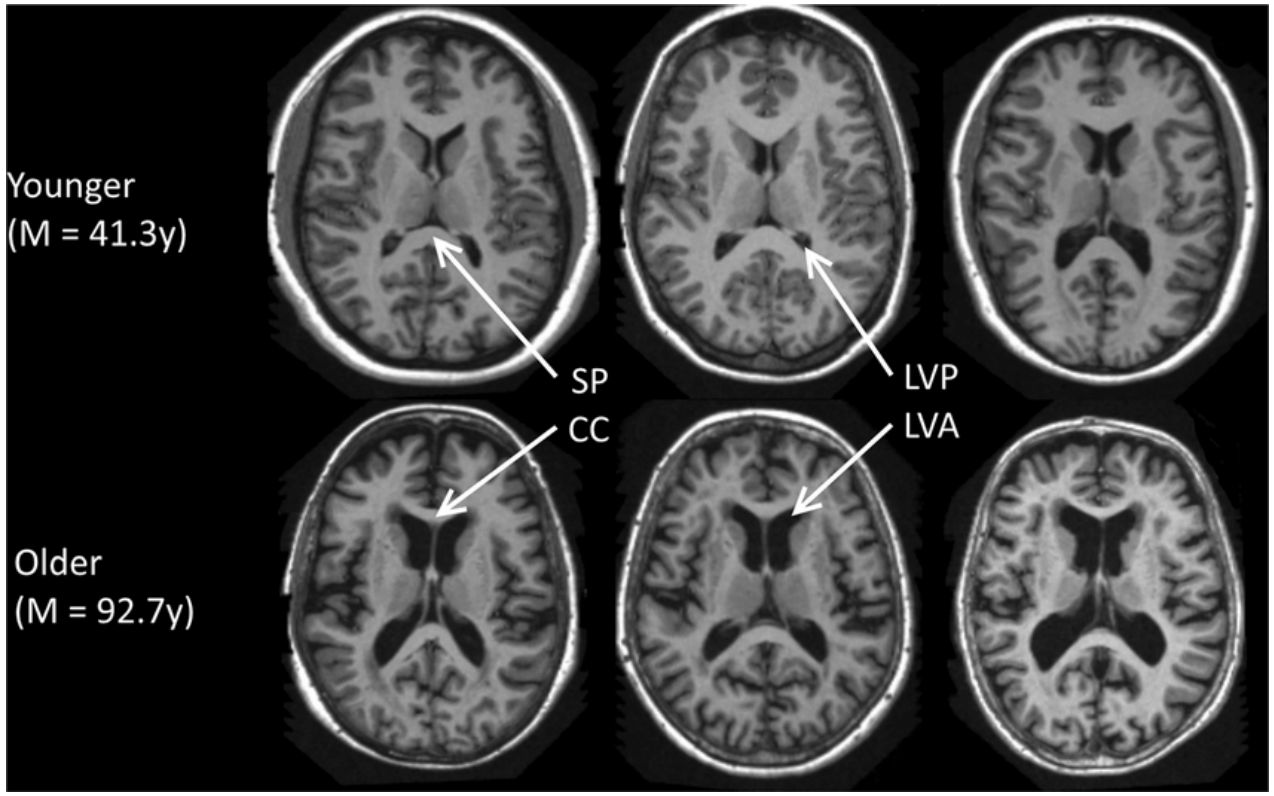
MRI in developing countries.

- Switzerland: 45.0 MRI units per 1.000.000 inhabitants (2016)

  West-Africa: 0.22 MRI units per 1.000.000 inhabitants (2018)
- Nearly two-thirds of the world's population has no access to basic diagnostic imaging!
- Budget MRI scanners: cheaper and portable versions to increase their accessibility and applicability in developing countries.
  - Wald LL et al. Low-cost and portable MRI. J Magn Reson Imaging. 2020
  - https://www.nwo.nl/en/cases/budget-mri-scanner-developing-countries

Ageing Effect on Brain MRI

- The brain undergoes profound age-related neuroanatomical changes during the aging process.
- The global grey matter volume decreased with age.
- Grey matter contains most of the brain's neuronal cell bodies



Davis, Nick. (2017). MINI-REVIEW Brain Stimulation for Cognitive Enhancement in the Older Person: State of the Art and Future Directions. Journal of Cognitive Enhancement. 1. 10.1007/s41465-017-0036-1.

Brain Age Estimation: Why is it interesting ?

- Ageing does not affect people uniformly!
- Individual rates of aging are shaped by interactions between environmental, genetic, and epigenetic factors.
- Studies based on brain MRI shows that there is a relation between accelerated aging and accelerated brain atrophy.
- It could improve early diagnosis and risk-assessments for age-associated neurodegenerative and neuropsychiatric diseases at a subject level:
  - Alzheimer, Parkinson, Huntington, etc.

Related Work

- Recent publications, have demonstrated that MRIs can be used to predict brain age with reasonably good accuracy!

- Jonsson, Benedikt A. et al. “Brain age prediction using deep learning uncovers associated sequence variants.” *Nature Communications* 10 (2019).
- Peng, Han et al. “Accurate brain age prediction with lightweight deep neural networks.” *bioRxiv*(2021).
- Jiang, Huiting et al. “Predicting Brain Age of Healthy Adults Based on Structural MRI Parcellation Using Convolutional Neural Networks.” *Frontiers in Neurology* 10 (2019).
- Cole, James H. et al. “Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker.” *NeuroImage* 163 (2017): 115-124.
- Sajedi, Hedieh and Nastaran Pardakhti. “Age Prediction Based on Brain MRI Image: A Survey.” *Journal of Medical Systems* 43 (2019): 1-30.

Description of the Dataset

MRI processing

Raw brain scans are difficult to handle.

- 3D brain scans are ~ 200x200x200 ~ 107 features/voxels (3D pixels ~ 1mm3)
- 3D structure + individual brain shapes → difficult to recognize disease patterns
- Data is scarce:

	ImageNet	MRI data set for task 1
Image size	224 x 224 x 3	~200 x 200 x 200
Data set size	1.2 million	~1200

MRI processing

We derived ~200 anatomical features for this project.

- Informative features are derived from the MRI images (with Freesurfer).
- They are extracted using **domain knowledge**.
  - e.g. cortex volume, left/right hemisphere surface area, white/gray matter volume etc.
- No need to process big images (6 GB) → csv sheet (3MB).
- No need for image analysis.
- Information loss!

MRI processing

We add additional complexity in the dataset by adding:

- Irrelevant features
- Outliers
- Perturbations (e.g. missing values, noise, etc.)

File Description

We provide the following files:

- **X_train.csv, y_train.csv:** the training set, including the features and labels
- **X_test.csv:** the test set (make predictions based on this file)
- **sample.csv:** a sample submission file in the correct format

Task Description

Task Description

- Subtask 0: Filling missing values.
- Subtask 1: Outlier detection.
- Subtask 2: Feature selection.
- Main Task: Age Prediction

Subtask 0: Filling Missing Values

Task Requirement: Fill missing values in the training and test set.

- There are missing values in the dataset that are set to NaN.
- Many methods cannot handle them automatically.
- Different strategies to impute them, could you name some ?

Imputation Recap

1. Mean/Median Values

	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	NaN
1	9	NaN	9.0	0	7.0
2	19	17.0	NaN	9	NaN

	col1	col2	col3	col4	col5
0	2.0	5.0	3.0	6.0	7.0
1	9.0	11.0	9.0	0.0	7.0
2	19.0	17.0	6.0	9.0	7.0

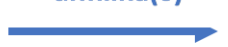
- **Pros:**
  - easy & fast.
- **Cons:**
  - Does not take into account the correlation between features.
  - Poor results on encoded categorical features.
  - Not very accurate.
  - Does not account for uncertainty.

Imputation Recap

2. Most Frequent / Constant Values

- **Pros:**
  - Easy & fast.
  - Works with categorical features.

	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	NaN
1	9	NaN	9.0	0	7.0
2	19	17.0	NaN	9	NaN



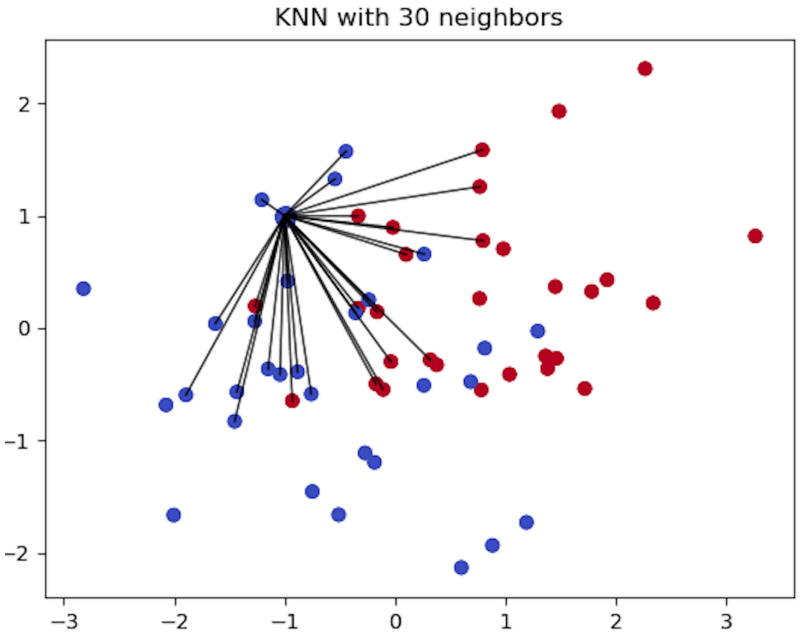
	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	0.0
1	9	0.0	9.0	0	7.0
2	19	17.0	0.0	9	0.0

- **Cons:**
  - Does not take into account the correlation between features.
  - Not very accurate.
  - Does not account for uncertainty.
  - It can also introduce bias.

Imputation Recap

3. K nearest neighbours (k-NN)

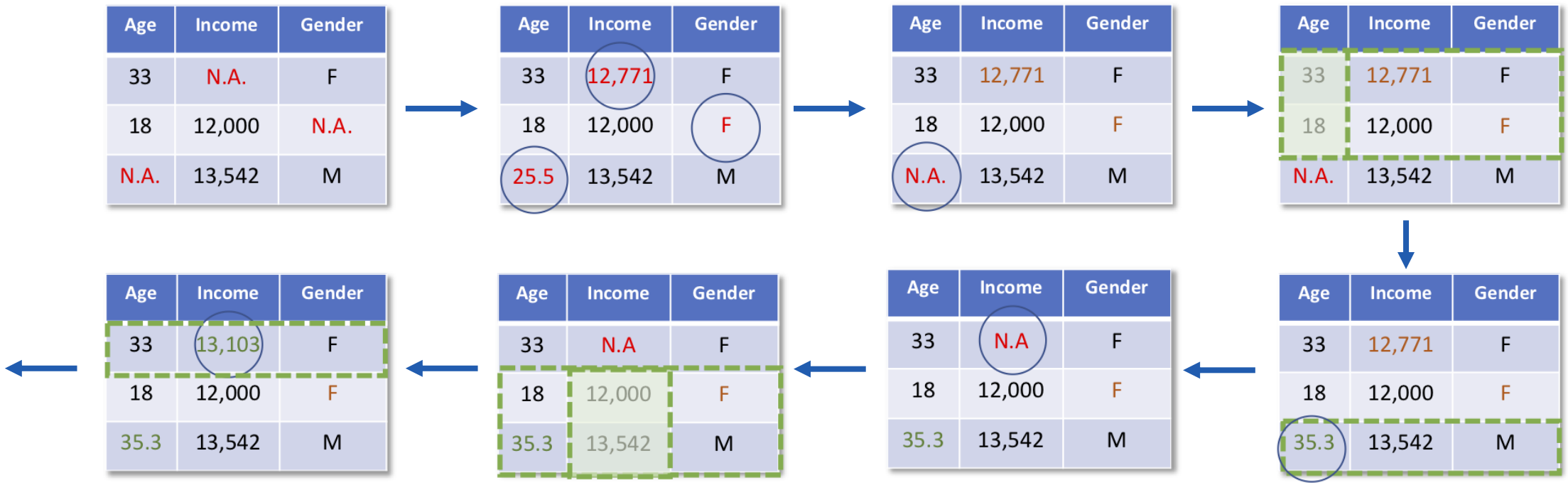
- Compute the weighted average of the features of the K-closest neighbours.
- **Pros:**
  - More accurate.
- **Cons:**
  - Computationally expensive
    (It stored the whole training dataset in memory)
  - Sensitive to outliers! (Unlike SVM)



Imputation Recap

4. Multivariate Imputation by Chained Equation (MICE)

- Iterative series of predictive models. In each iteration each specified variable in the dataset is imputed using the other variables until convergence



Imputation Recap

4. Multivariate Imputation by Chained Equation (MICE)

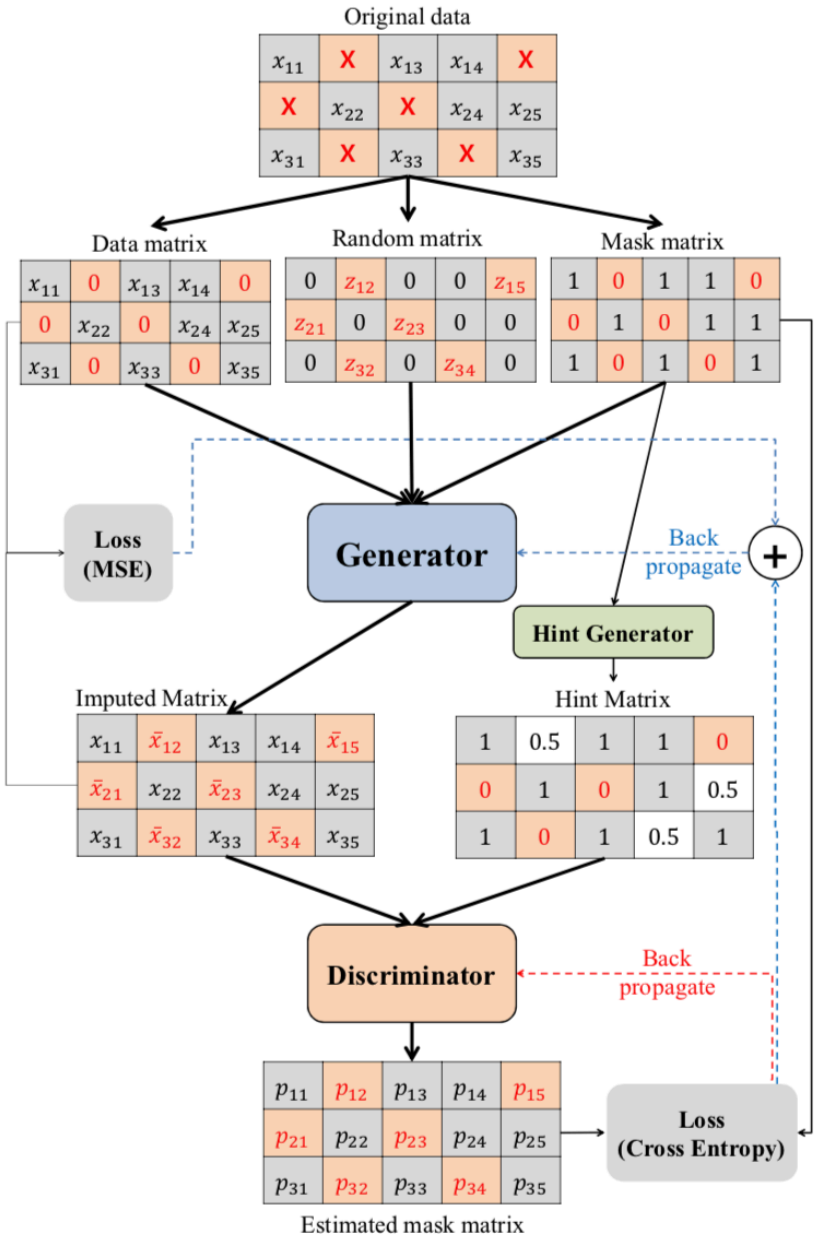
- **Pros:**
  - Handle the uncertainty of the missing values.
  - Works with different data types.
- **Cons:**
  - Can suffer from performance problems when the number of observation is large
  - Poor performance if the data have complex features, such as nonlinearities and high dimensionality.

Imputation Recap

5. Generative Adversarial Imputation Nets

(Yoon, Jinsung et al. “GAIN: Missing Data Imputation using Generative Adversarial Nets.” , ICML, 2018.)

- The Generator (G) observes some components of the real data vector and outputs a completed vector.
- The Discriminator (D) attempts to determine which components were observed and which imputed.
- To ensure that G learns the desired distribution a hint vector is given to D.



Imputation Recap

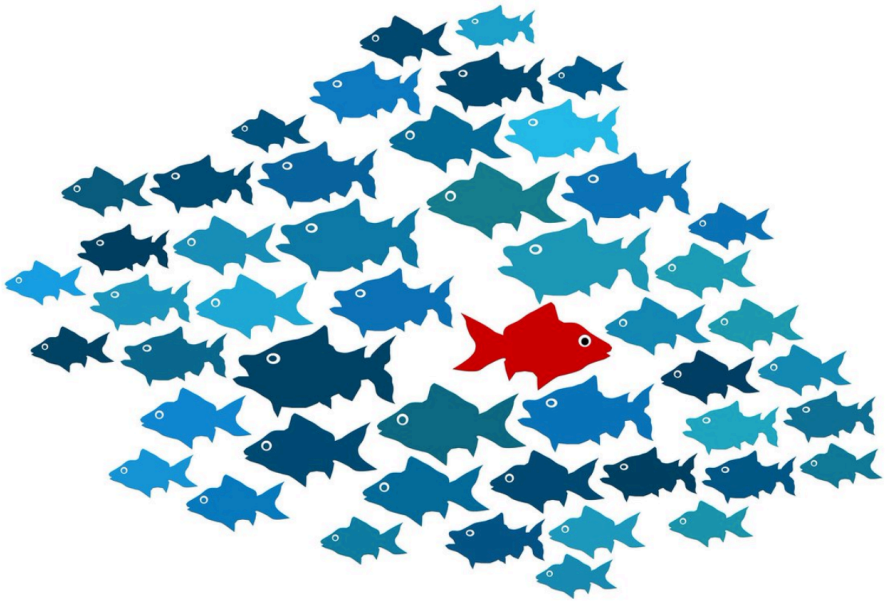
5. Many more!

- Stochastic regression imputation
- MissForest
- Hot-Deck imputation
- Matrix Completion
- Denoising Autoencoder (DAE)
- …

Subtask 1: Outlier Detection

Task Requirement: Build outlier detection model to classify samples in the training set that are outliers.

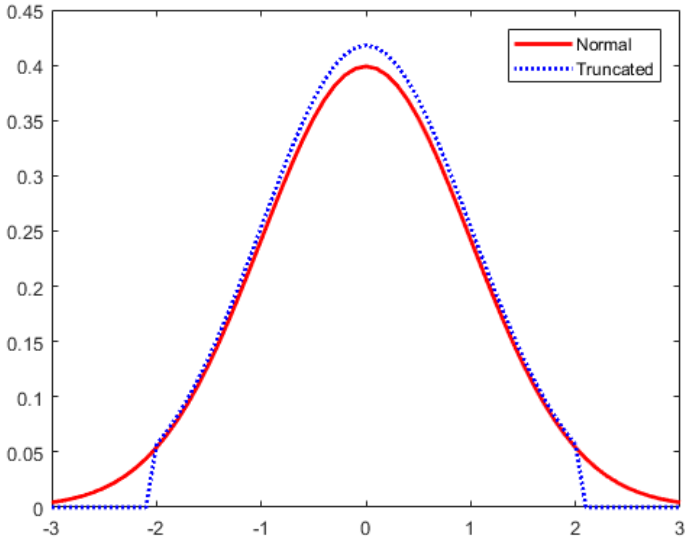
- There are outliers in the **training set** (X and y):
  - ➢ Measurement/data entry errors
- If the resulting model is not robust enough, it may be sensitive to outliers.
- Outlier deletion can be expected to lead to better results.



https://towardsdatascience.com/a-brief-overview-of-outlier-detection-techniques-1e0b2c19e561

Outlier Detection Recap

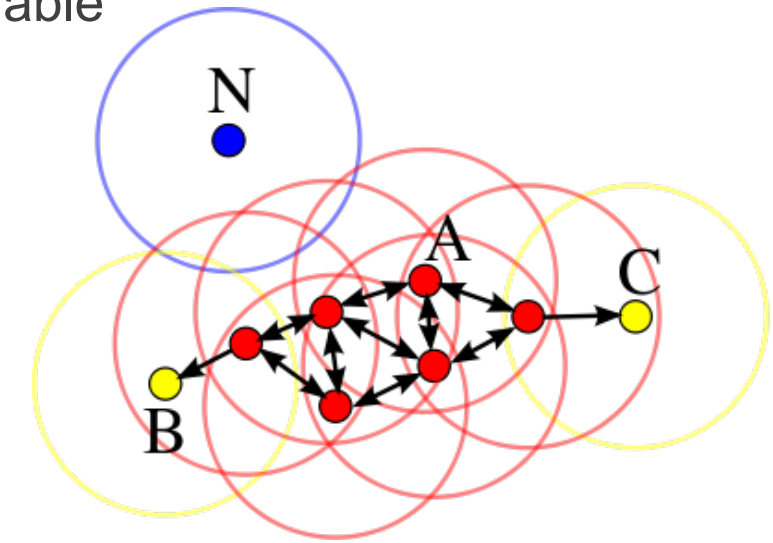
1. **Z-score:**
   - Indicates how many standard deviations a data point is from the mean assuming a gaussian distribution. Outliers lies outside a certain threshold.
   - **Pros:**
     - Effective if data (or transformations) can be described by a gaussian distribution.
   - **Cons:**
     - Requires low-dimensional feature space.
     - Does not perform well if the data distribution can't be described with a parametric model.



Outlier Detection Recap

2. Dbscan (Density Based Spatial Clustering of Applications with Noise):

- Density based clustering method, finds neighbours by density on an n-dimensional sphere with radius ε.
- **Core point: A** contains more samples than a certain value (M) in the neighbourhood.
- **Border point: C** is not a core point but it is density reachable
- **Outlier: N** lies in no cluster (not density-reachable nor density-connected to any other point).



Outlier Detection Recap

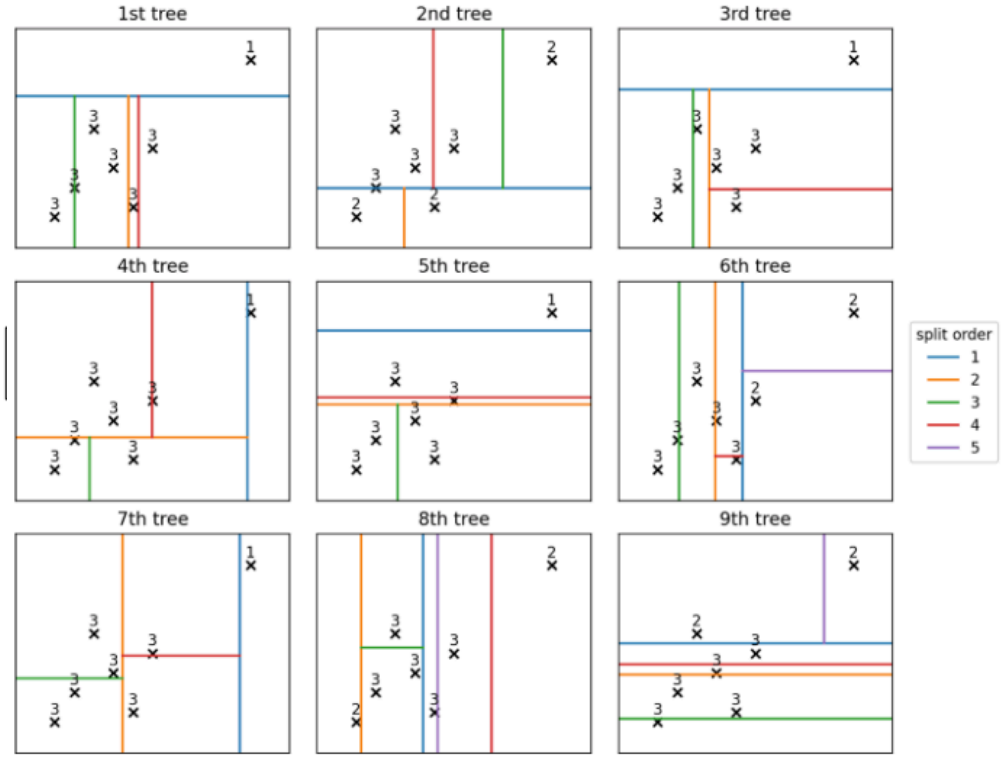
2. Dbscan (Density Based Spatial Clustering of Applications with Noise):

- **Pros:**
  - Effective method for non-parametric data distribution.
  - Method intuitive
- **Cons:**
  - Hyper-parameter tuning (M, ε)
  - Scaling matters!
  - Higher the dimensionality of the input, the less accurate it becomes.

Outlier Detection Recap

3. Isolation Forest:

- If we fit a decision tree on all observations, on average outliers should be found close to the root of the tree.
- **Pros:**
  - Works well in high-dimensional data.
  - Few parameters.
- **Cons:**
  - Could be computationally intensive.



https://towardsdatascience.com/isolation-forest-the-anomaly-detection-algorithm-any-data-scientist-should-know-1a99622eec2d

Outlier Detection Recap

4. Many more!

- Robust Random Cut Forest
- Deep Generative Models (VAE / GAN)
- Dan Hendrycks and Kevin Gimpel. “A baseline for detecting misclassified and out-of-distribution examples in neural networks.” *ICRL*, 2017.
- Kimin Lee et al. “Training confidence-calibrated classifiers for detecting out-of-distribution samples.” *ICRL*, 2018.
- Hendrycks, Dan et al. “Deep Anomaly Detection with Outlier Exposure.” *ICRL,* 2019.
- …

Outlier Detection to Discover Structural Defects in Newborns ECHOs

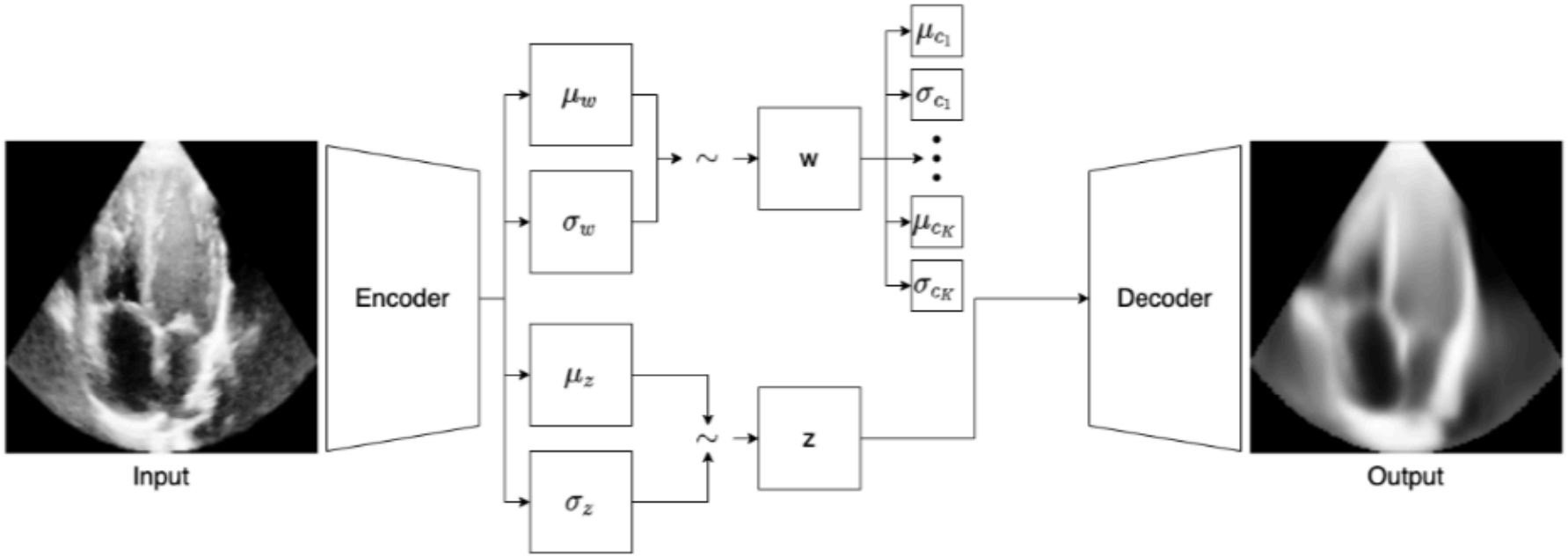
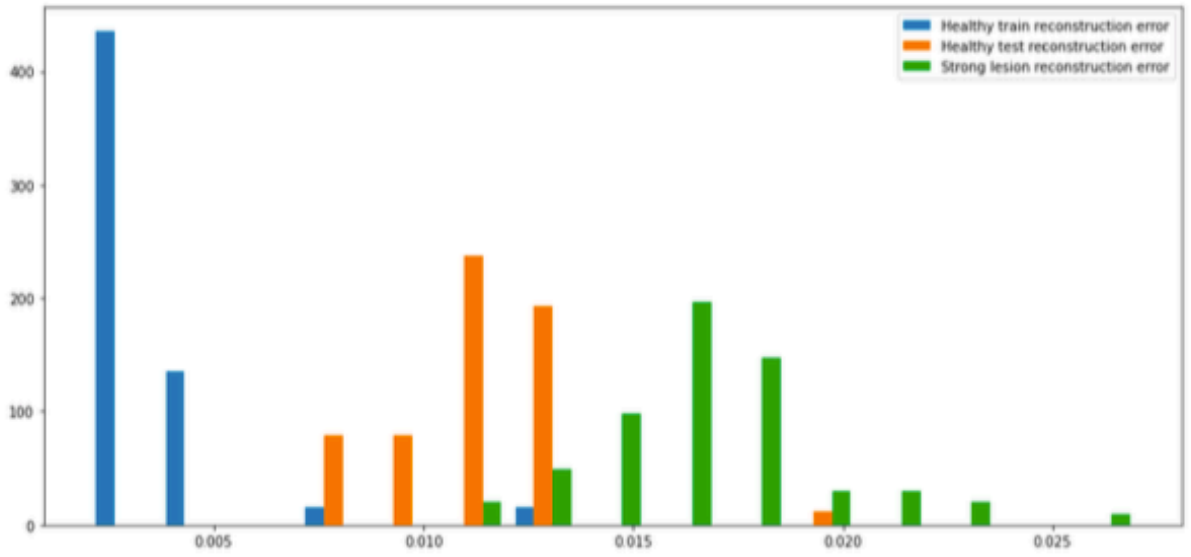
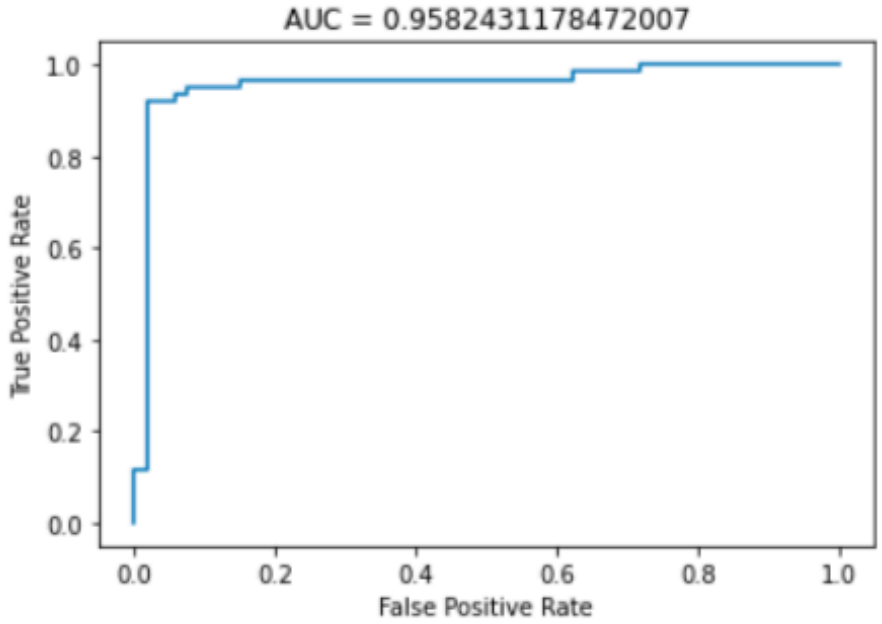


Figure 3.13: Gaussian Mixture Variational autoencoder schematic.

We fit a GMM-VAE to learn the distribution of healthy echocardiogram's samples of newborns.

Outlier Detection to Discover Structural Defects in Newborns ECHOs

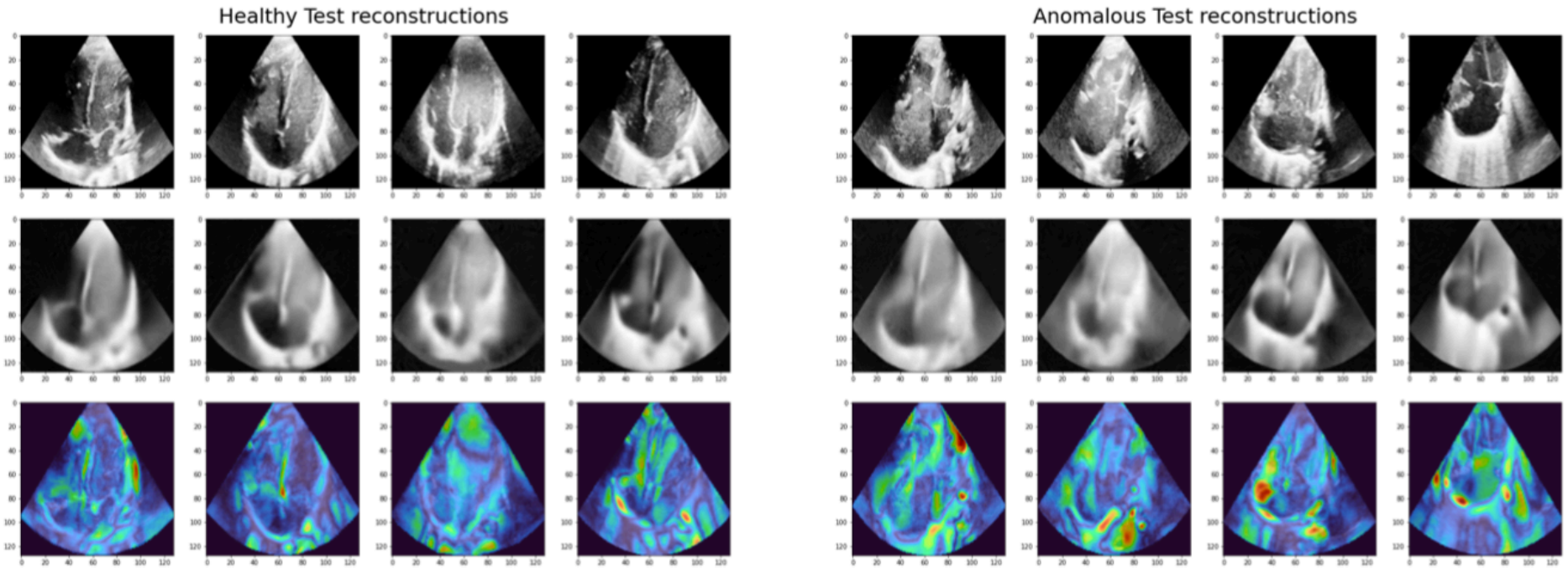




Histogram of mean squared errors of healthy and unhealthy hearts (left) and ROC curve of the prediction (right).

Outlier Detection to Discover Structural Defects in Newborns ECHOs

medical data science



When we test on anomalous samples the model outputs a healthy version of the heart!

Subtask 2: Feature Selection

Task Requirement: Delete irrelevant/redundant features.

- **Feature Selection:** select a subset of original features.

  Feature Extraction: compute new features often on a lower-dimensional space.
- We added manual features to the FreeSurfer processed dataset, which are either random or highly correlated with other features.
- Feature Selection simplifies the model (easier to interpret), leads to shorter training time and reduces overfitting.

Feature Selection Recap

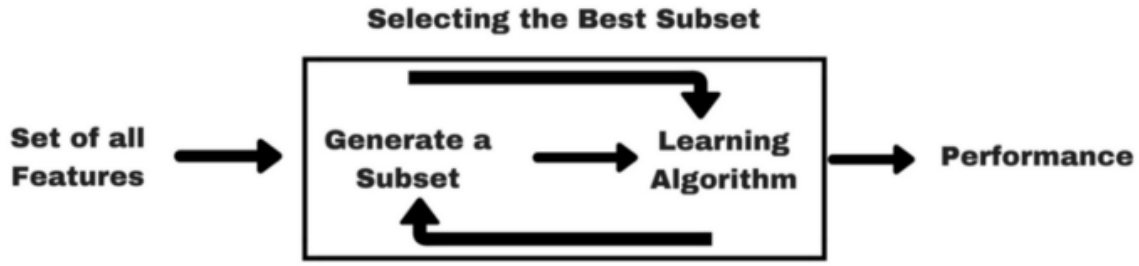
1. Filter methods:

- Pick up the intrinsic properties of the features measured via univariate statistics.
- **Relief:** sample instances and updates the relevance of each feature as following:
  - It selects two nearest points of the same and opposite class.
  - For each feature f, define its importance as p_f.
  - If a feature difference is observed in the same class neighbor then decrease p_f.
  - If a feature difference is observed in the opposite class neighbor then increase p_f.
- **Correlation Coefficient:** measure correlations of variables.
  - Pearson product-moment correlation coefficient (linear relationship)
  - If two features are correlated, one of them is redundant.

Feature Selection Recap

2. Wrapper methods:

- Greedy-search approach algorithm to select a subset of features based on the performance of the applied learning.
- **Forward Feature Selection:** start with empty set of features, then best-performing features are iteratively added.
- **Backward Feature Elimination:** start with all features, worst features removed.
- **Exhaustive Feature Selection:** brute-force evaluation of each feature subsets.
- ….



Feature Selection Recap

3. Embedded methods:

- Algorithms which simultaneously perform model fitting and feature selection
- **LASSO Regularization:** shrink some of the coefficient to zero.
- **Random Forest Importance:** most important features closer to the root of the tree.
- **Attention Networks**
- …

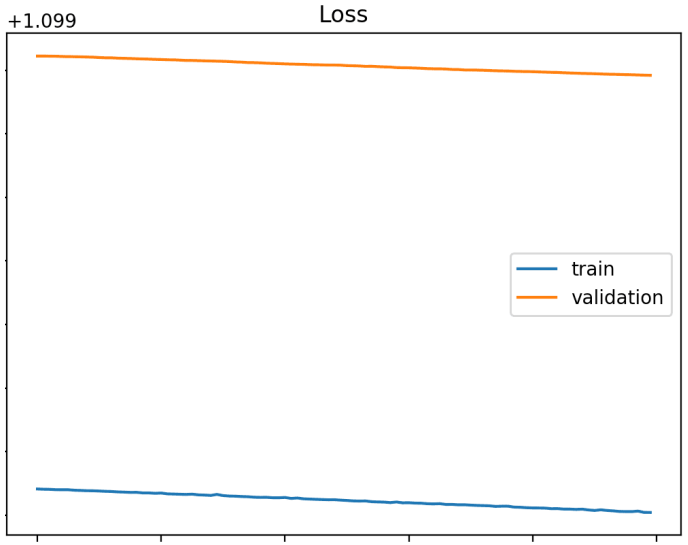
Main Task: Age Prediction

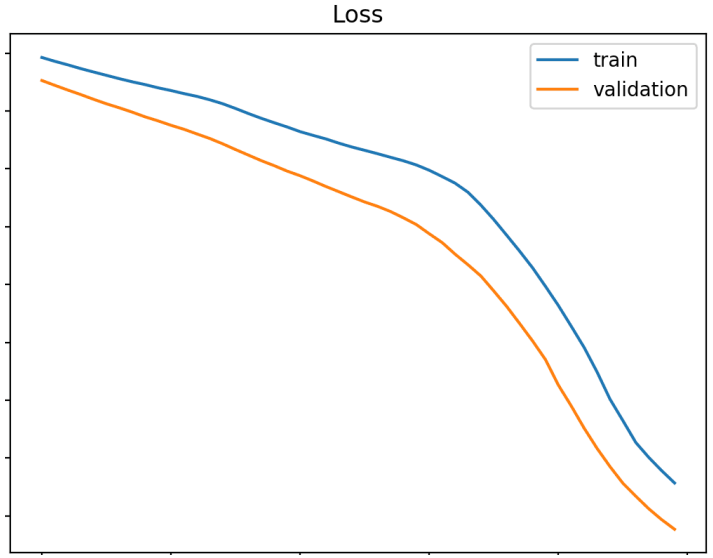
Task Requirement: Use suitable regression methods to predict the age of a person from the brain data.

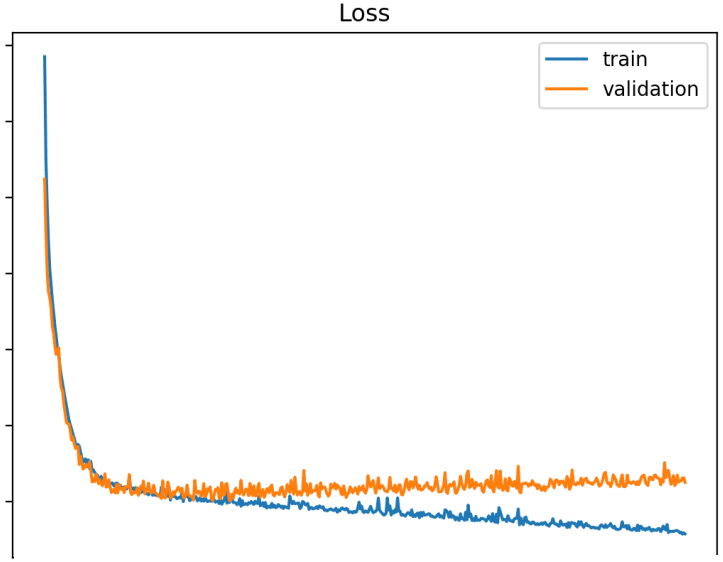
- After primary pre-processing and feature selection, the dataset is ready to perform the regression task.
- You are free to use any regression model

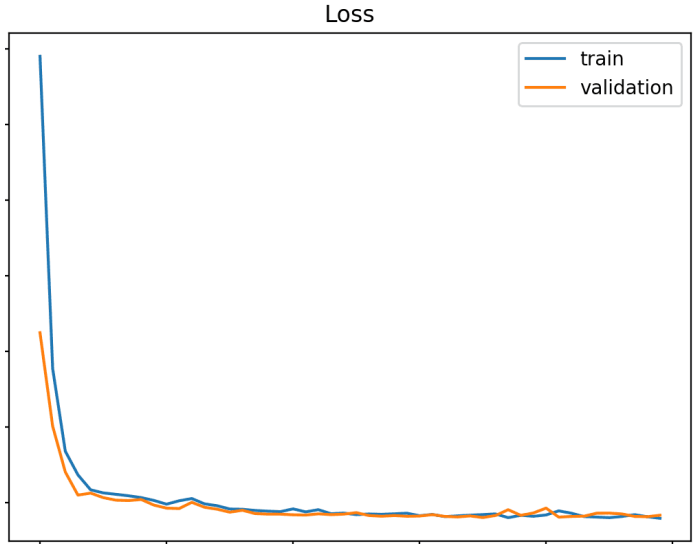
  (Lasso regression, ensemble models, MLP, TabNet etc.)
- Do not overfit!!!

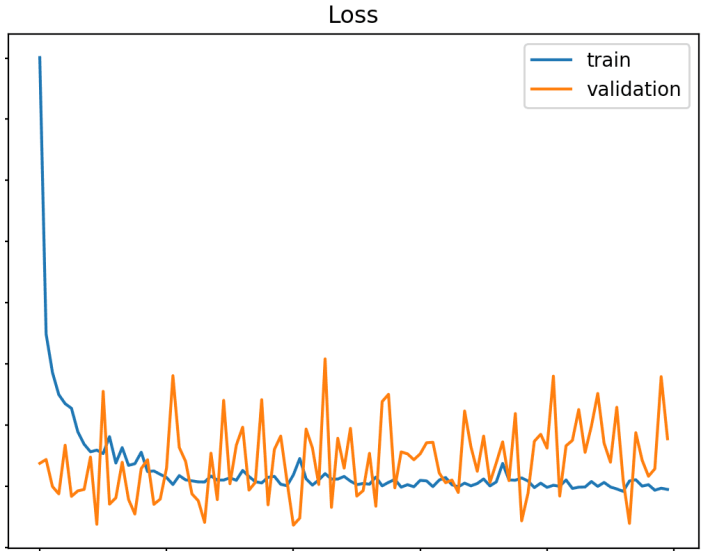
Learning Curves

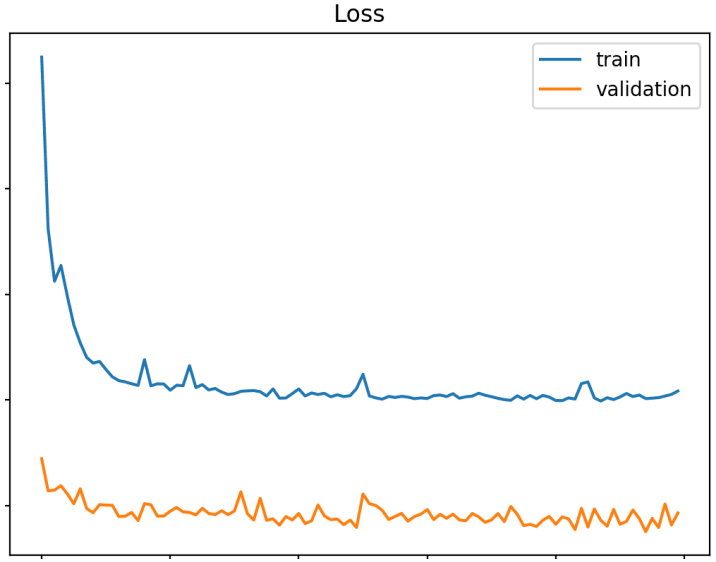






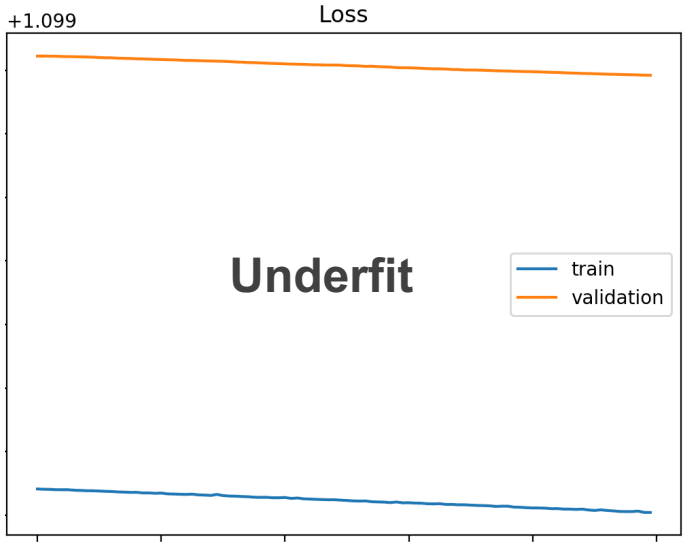


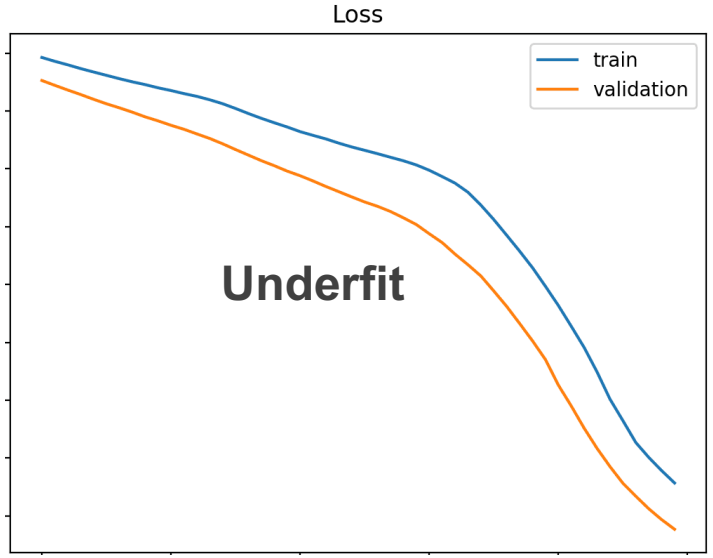


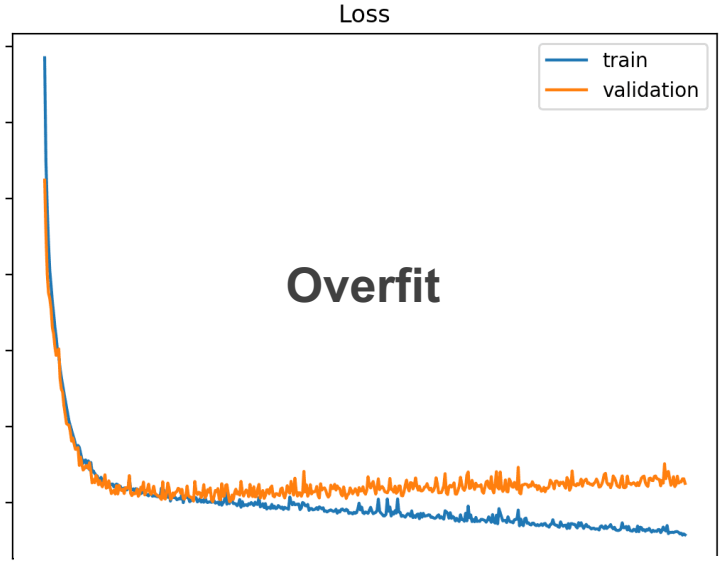


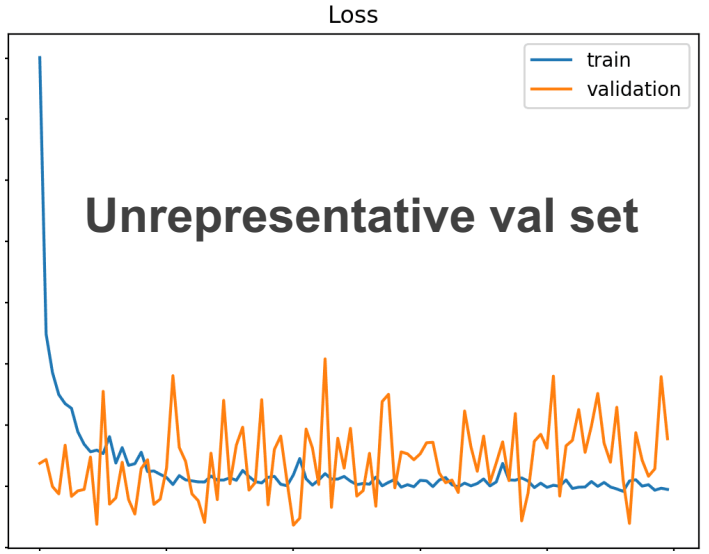
https://machinelearningmastery.com/learning-curves-for-diagnosing-machine-learning-model-performance/

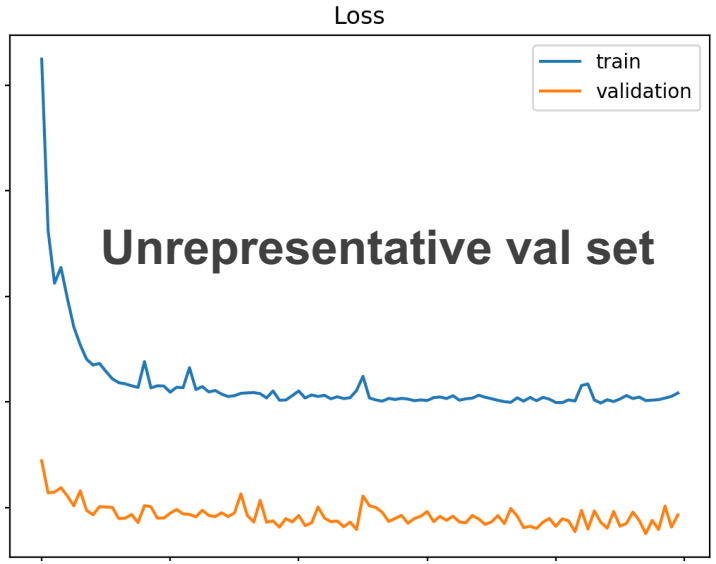
Learning Curves











https://machinelearningmastery.com/learning-curves-for-diagnosing-machine-learning-model-performance/

Evaluation Metric

Coefficient of Determination R^2:

- Is the proportion of the variation in the dependent variable (Y) that is predicted by the independent variables (X). $\bar{y}$ is the average of the true labels, f_i is the predicted label:

$$R^2 := 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}} \qquad SS_{\text{tot}} := \sum_i (y_i - \bar{y})^2 \qquad SS_{\text{res}} := \sum_i (f_i - y_i)^2$$

- How well the data fits the model.
- Varies between 1 (best) and $-\infty$ (worst)
- A model that predicts always the empirical mean of the predictior variable has R^2=0
- In Python:

```
from sklearn.metrics import r2_score
score = r2_score(y_true, y_pred)
```

Important!

- Do NOT use AutoML packages
  - This includes anything that does automatic data cleaning and model selection!
- Beware of overfitting on the public test data
- Describe what you did when you hand in the project
- Do NOT wait until the last day to submit something
  - Servers usually get overloaded and crash causing long waiting times

Q&A