University of South Wales
Prifysgol De Cymru

MS4S10 Machine Learning and Decision Making

Week1

Moizzah Asif

moizzah.asif@southwales.ac.uk

J418

1

University of South Wales
Prifysgol De Cymru

Know thy module

Week 1 – 4
Moizzah Asif

- 8 weeks intensive learning.
- All synchronous teaching on Friday - 2 hours (14 00 – 16 00)
- Pre-reading and audio visual learning content will be provided iteratively over the week
- You are required to go through the Supplementary resources (pre & post lecture) and any pre-recorded video files for weeks 1 - 4.
- It is also expected that you would undertake any required lab work before attending the synchronous live session.
- Post your queries on the SR in the discussion boards. Feel free to create new threads within each week's forum.
- The lectures are designed under the assumption that you understand the pre-lecture reading from SR for the week

- 27-11-2020 – **Basics of Machine Learning**; The ML process; Data collection & preprocessing
- 04-12-2020 - **Supervised Learning**: classification, regression, optimisation, model selection and generalisation, parametric and non-parametric learning, Decision Trees
- 11-12-2020 – **Supervised Learning**: Probabilistic learning, Bayes Learning, Naïve Bayes classifiers, Ensemble Learning, Random Forest , Support Vector Machines, Kernel functions, Hyper-parameter optimization
- 08-01-2021 – **Unsupervised learning**: Clustering and dimensionality reduction

Coursework 1 – 50% weightage

2

University of South Wales
Prifysgol De Cymru

Resources

Reading List:

- Machine Learning Tom M. Mitchell , McGraw-Hill, available at USW Treforest library
- Machine Learning, An algorithmic Perspective, Stephen Marsland, 2nd ed., CRC press, available at USW Treforest library
- Pattern Recognition and Machine Learning, Christopher M. Bishop, Springer available at USW Treforest library

Other recommended readings:

- Introduction to Machine learning with Python, available at USW Treforest library
- Hands-on Machine Learning with Scikit-Learn & Tensorflow, Aurelien Geron, O'reilly, available at USW Treforest library
- Applied predictive modelling, Max Kuhn and Kjell Johnson, Springer (eBook – access via university account while on campus)

3

University of South Wales
Prifysgol De Cymru

1. Basics of Machine Learning (ML)

What is Machine Learning? - I

Learning:

Humans learn from
Experience → *remember, adapt and generalise*
(reason, apply logic and deduce)

Machines learn from data

Machine Learning:

Concept
enables computers to modify or *adapt* their action to be more accurate while *generalising*

High Level
Field of study that gives computers the ability to learn without being explicitly programmed (Arthur Samuel - 1959)

4

University of South Wales
Prifysgol De Cymru

1. Basics of Machine Learning (ML)

What is Machine Learning? - I I

Machine Learning:

Formal

Data → Learning algorithm → Knowledge

Algorithms
- improve their performance P
- at a task T
- with experience E

(Tom Mitchell 1998)

5

University of South Wales
Prifysgol De Cymru

1. Basics of Machine Learning (ML)

Designing a learning system

Arthur Samuel's checkers-playing programme was the first self-learning programme which eventually outplayed Arthur!

Task – T : playing checkers

Performance measure – P: percent of games won against opponents

Training experience – E: playing practice games against itself

6

University of South Wales Prifysgol De Cymru

1. Basics of Machine Learning (ML)

Types of Machine Learning

Supervised Learning:
- Machine learns from a training set which has correctly assigned labels to a target variable.
- Generalises the knowledge created to predict the target variable on unseen data.

Unsupervised Learning:
- An uncategorised dataset is provided to the machine.
- It finds similarities between datapoints and categorises similar datapoints into same categories.

Reinforcement Learning:
- The machine doesn't know the outcomes, but gets awarded or penalised when it reaches the right or wrong outcome.
- Learns from experience to reach the right outcome
- Does not get told how to reach it

7 Moizzah Asif - Machine Learning and Decision Making © University of South Wales

7

University of South Wales Prifysgol De Cymru

1. Basics of Machine Learning (ML)

Types of Machine Learning

Evolutionary Learning:

- Machine doesn't know the outcomes neither is their feedback on outcome
- It's given a criterion of stop/fit/good
- Learn evolution by
  - generating initial outcomes.
  - Reproduce new outcomes based on current outcomes
  - Evaluate new outcomes
  - Form new outcomes by selecting good /better solutions from current and previous outcomes

8 Moizzah Asif - Machine Learning and Decision Making © University of South Wales

8

University of South Wales Prifysgol De Cymru

1. Basics of Machine Learning (ML)
2. The Machine Learning Process
   i. *Data Preparation*

The Machine Learning process

Data collection & preparation	Find data source, extraction methods, cleaning to suit next phase
Feature engineering	Understand the domain of the problem, create and extract, reduce redundancy, aim to aid the next phase
Algorithm selection	Type of learning, memory requirement, processing, complexity, trade offs or comparisons
Parameter & Model Selection	Type of learning, memory requirement, processing, complexity, trade offs or comparisons
Training	Training, validation and test partitions, avoiding overfitting and selection biases
Evaluation	Human experts, cost, validation metrics, result interpretation

9 Moizzah Asif - Machine Learning and Decision Making © University of South Wales

9

University of South Wales Prifysgol De Cymru

1. Basics of Machine Learning (ML)
2. The Machine Learning Process
   i. *Data Preparation*

Data Pre-processing (preparation)

Generally: addition, deletion or transformation of data before training the ML model.

But, let's take a step back – let's first agree on some definitions!

Predictor variable: independent variable

Target variable: dependent variable, which is to be predicted

Feature engineering: encoding of the predictors

10 Moizzah Asif - Machine Learning and Decision Making © University of South Wales

10

University of South Wales Prifysgol De Cymru

1. Basics of Machine Learning (ML)
2. The Machine Learning Process
   i. *Data Preparation*

Missing Values

The case of missing values

It is important to understand why the values are missing!

- Could be structurally missing
  - Ex: #births given by a man
- Informatively missing – huh?

Missing values are generally related to predictor variables than the over all dataset.
- May be concentrated in a subset of predictors rather than occurring randomly across all predictors

11 Moizzah Asif - Machine Learning and Decision Making © University of South Wales

11

University of South Wales Prifysgol De Cymru

1. Basics of Machine Learning (ML)
2. The Machine Learning Process
   i. *Data Preparation*

Missing Values

The fantastic case of missing values and what to do with them

- For large datasets, assuming the *missingness is not informative*, removal could be an option.
- If removal is not an option then **either**
  - choose predictive models which account for missing values

    Or
  - impute the missing data!

12 Moizzah Asif - Machine Learning and Decision Making © University of South Wales

12

University of South Wales Prifysgol De Cymru

1. Basics of Machine Learning (ML)
2. The Machine Learning Process
i. *Data Preparation*

Missing Values

Data Imputation

Adds another layer of modelling

where values of predictor variables are used to predict the missing predictor variable values.

If the number of predictors affected by missing values is small, an exploratory analysis of the relationships between the predictors can be conducted.

Ex: predictor A with missing values highly correlates with another predictor B which has less or no missing values. If there is strong correlation, you can either get rid of predictor A (feature selection to be discussed later) or use the correlation to impute missing values.

13 Moizzah Asif - Machine Learning and Decision Making © University of South Wales

13

University of South Wales Prifysgol De Cymru

1. Basics of Machine Learning (ML)
2. The Machine Learning Process
i. *Data Preparation*

Missing Values

Data Imputation

Another popular method is to use K-nearest neighbor model to impute missing values.

The assumption behind using KNN for missing values is that a point value can be approximated by the values of the points that are closest to it, based on other variables.

A new sample is imputed by finding the samples in the training set "closest" to it and averages these nearby points to fill in the value.

The distances between the datapoint to be imputed and all the rest of the datapoint is computed. K nearest neighbours are selected and the missing value is imputed based on the values from K nearest neighbours

14 Moizzah Asif - Machine Learning and Decision Making © University of South Wales

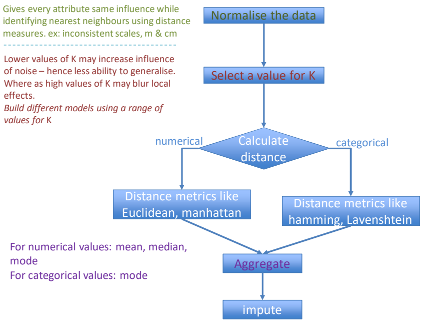
14

University of South Wales Prifysgol De Cymru

1. Basics of Machine Learning (ML)
2. The Machine Learning Process
i. *Data Preparation*

Missing Values

Data Imputation – K-NN



15 Moizzah Asif - Machine Learning and Decision Making © University of South Wales

15

University of South Wales Prifysgol De Cymru

1. Basics of Machine Learning (ML)
2. The Machine Learning Process
i. *Data Preparation*

Missing Values

Critical thoughts on K-NN

- particularly useful for dealing with all kinds of missing data
  - It can be used for data that are continuous, discrete, ordinal and categorical
- the imputed data are confined to be within the range of the training set values

- an impractical choice in environments where predictions need to be made rapid
  - becomes significantly slower as the volume of data increases
- entire training set is required every time a missing value needs to be imputed

16 Moizzah Asif - Machine Learning and Decision Making © University of South Wales

16

University of South Wales Prifysgol De Cymru

1. Basics of Machine Learning (ML)
2. The Machine Learning Process
i. *Data Preparation*

Missing Values

Back to the case of *missing values*

The *total_bedrooms* attribute in today's python exercise – what can we possibly do …

1. Remove the whole column.

```
housing.dropna('total_bedrooms', axis = 1)
```

2. Remove the rows with missing values

```
housing.dropna('total_bedrooms', axis = 0)
```

3. Set the values to some value (zero, mean, median, etc.)

```
housing['total_bedrooms'].fillna(0,inplace=True)
```

- Find mean, median and mode first, and then replace the value with '0' in the above loc (line of code)

17 Moizzah Asif - Machine Learning and Decision Making © University of South Wales

17

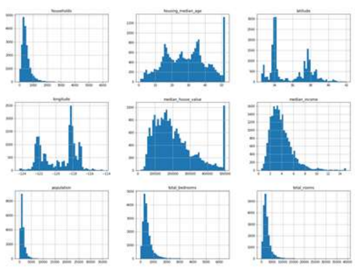
University of South Wales Prifysgol De Cymru

1. Basics of Machine Learning (ML)
2. The Machine Learning Process
i. *Data Preparation*

Transformations

Skewed Data

A roughly symmetric distribution is un-skewed.



18 Moizzah Asif - Machine Learning and Decision Making © University of South Wales

18

University of South Wales
Prifysgol De Cymru

1. Basics of Machine Learning (ML)
2. The Machine Learning Process
   i. *Data Preparation*

For simplicity we will use a readily available data source to get hands on experience on with some data cleaning, pre-processing and descriptive tasks.

Please refer to additional resources pdf file for setting up python and Jupyter notebook on your personal computers.

The lab is already equipped with everything you need.

Let's move to python script of the week.

19

Transformations

University of South Wales
Prifysgol De Cymru

1. Basics of Machine Learning (ML)
2. The Machine Learning Process
   i. *Data Preparation*

Skewed Data

A general rule of thumb to consider is that skewed data whose *ratio of the **highest value to the lowest value is greater than 20*** have significant skewness.

Replacing the data with the log, square root, or inverse may help to remove the skew. This is one form and reason of **data transformation**

let's take the example of population from housing data being used in python script.

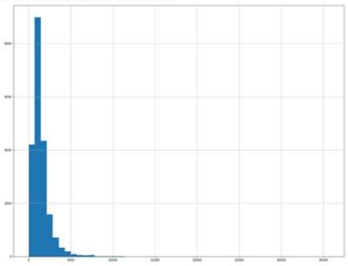
20

Transformations

University of South Wales
Prifysgol De Cymru

1. Basics of Machine Learning (ML)
2. The Machine Learning Process
   i. *Data Preparation*

```
In [21]: housing['population'].hist(bins=50, figsize=(20,15))
         # mpl.pyplot.show()
Out[21]: <matplotlib.axes._subplots.AxesSubplot at 0x7f8c39ee4d30>
```



```
In [17]: housing['population'].describe()
Out[17]: count    20640.000000
         mean      1425.476744
         std       1132.462122
         min          3.000000
         25%        787.000000
         50%       1166.000000
         75%       1725.000000
         max      35682.000000
         Name: population, dtype: float64
```

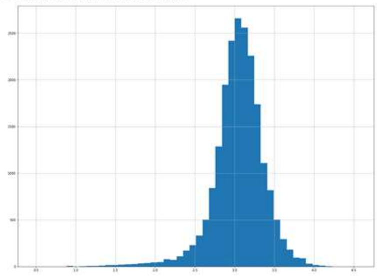

21

Transformations

University of South Wales
Prifysgol De Cymru

1. Basics of Machine Learning (ML)
2. The Machine Learning Process
   i. *Data Preparation*

```
In [22]: import numpy as np
In [23]: (np.log10(housing['population'])).hist(bins=50, figsize=(20,15))
Out[23]: <matplotlib.axes._subplots.AxesSubplot at 0x7f8c19142320>
```



log() for natural log
log10() base 10 log
log2() base 2 log

22

Transformations

University of South Wales
Prifysgol De Cymru

1. Basics of Machine Learning (ML)
2. The Machine Learning Process
   i. *Data Preparation*

Skewed Data

Box-Cox transformation (1964)

Statistical methods can also be used to empirically identify an appropriate transformation.

George Box and David Cox (1964) developed a procedure to identify an appropriate exponent (Lambda) to use to transform data into a "normal shape."

Box and Cox propose a family of transformations that are indexed by a parameter, denoted as λ:

$$x = \begin{cases} \dfrac{x^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log(x) & \text{if } \lambda = 0 \end{cases}$$

23

Transformations

University of South Wales
Prifysgol De Cymru

1. Basics of Machine Learning (ML)
2. The Machine Learning Process
   i. *Data Preparation*

Box-Cox transformation (1964)

the Box-Cox power transformation searches from $-5 \geq \lambda \leq 5$ until the best value is found

the method checks for the smallest standard deviation

absolutely necessary to always check the transformed data for normality using a probability plot.

the Box-Cox Power transformation only works if all the data is positive and greater than 0

24

University of South Wales Prifysgol De Cymru

Transformations

Outliers

1. Basics of Machine Learning (ML)
2. The Machine Learning Process
   i. *Data Preparation*

University of South Wales Prifysgol De Cymru

Transformations

Scaling

1. Basics of Machine Learning (ML)
2. The Machine Learning Process
   i. *Data Preparation*

Generally ML algorithms don't perform well when the input numerical attributes have different scales.

Ex: housing data used today

Feature	Range	
	Min	Max
total_rooms	6	39320
median_income	0	15

University of South Wales Prifysgol De Cymru

Transformations

Scaling

1. Basics of Machine Learning (ML)
2. The Machine Learning Process
   i. *Data Preparation*

Two most common way of scaling:

- min-max scaling/normalisation
  Values are shifted and rescaled so they end up ranging from 0 to 1
- standardisation
  subtracts the mean value from each index and divide by standard deviation, resulting into new distribution have unit variance.

University of South Wales Prifysgol De Cymru

Feature Engineering

1. Basics of Machine Learning (ML)
2. The Machine Learning Process
   i. *Data Preparation*
   ii. *Feature Engineering*

The question of How to represent your data for a particular application is known as feature engineering.

Correct feature engineering depends on several factors:

- some encoding may work well for one type of model, whereas be poor for others
- Relationship between predictor and the target variable
  - Ex: which representation to use for date predictor when the data has a seasonal trend: just year, just month, both, day of the year?

University of South Wales Prifysgol De Cymru

Feature Engineering

Feature Extraction

1. Basics of Machine Learning (ML)
2. The Machine Learning Process
   i. *Data Preparation*
   ii. *Feature Engineering*

These methods reduce the data by generating a smaller set of predictors

they capture a majority of the information in the original variables.

fewer variables can be used that provide reasonable fidelity to the original data.

For most data reduction techniques, the new predictors are functions of the original predictors

University of South Wales Prifysgol De Cymru

Feature Engineering

Feature extraction

1. Basics of Machine Learning (ML)
2. The Machine Learning Process
   i. *Data Preparation*
   ii. *Feature Engineering*

Ex:

- combinations of predictors can sometimes be more effective than their individual value

$$f(x_1, x_2, x_3) = y$$

$$\downarrow$$

$$f\left(\frac{x_1}{x_2}, x_3\right) = y$$

Or

$$f(x_1, (x_2 x_3)) = y$$

Feature Engineering – categorical variables

There are other ways to handle categorical variables other than as shown in python example.

One-Hot-Encoding

- Also known as creating dummy variables.
- Categorical features are replaced by one or more new features which can have numerical values.
- Any number of categories can be represented by introducing one new feature per category.

31

Before one hot encoding:

```
housing.head()
```

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_proximity
0	-122.23	37.88	41.0	880.0	129.0	322.0	126.0	8.3252	452600.0	NEAR BAY
1	-122.22	37.86	21.0	7099.0	1106.0	2401.0	1138.0	8.3014	358500.0	NEAR BAY
2	-122.24	37.85	52.0	1467.0	190.0	496.0	177.0	7.2574	352100.0	NEAR BAY
3	-122.25	37.85	52.0	1274.0	235.0	558.0	219.0	5.6431	341300.0	NEAR BAY
4	-122.25	37.85	52.0	1627.0	280.0	565.0	259.0	3.8462	342200.0	NEAR BAY

After one hot encoding:

```
housing_dummy.tail()
```

ds	median_income	median_house_value	ocean_proximity_<1H OCEAN	ocean_proximity_INLAND	ocean_proximity_ISLAND	ocean_proximity_NEAR BAY	ocean_proximity_NEAR OCEAN
).0	1.5603	78100.0	0	1	0	0	0
1.0	2.5568	77100.0	0	1	0	0	0
3.0	1.7000	92300.0	0	1	0	0	0
3.0	1.8672	84700.0	0	1	0	0	0
).0	2.3886	89400.0	0	1	0	0	0

32

Feature Engineering - Reduction

Data Reduction

PCA – Principal Component Analysis

Finds linear combinations of predictors, called PCs.

First PC captures variability of all possible linear combinations of predictors.

Subsequent PCs are derived to capture remaining variability while being uncorrelated to all previous PCs.

33

Feature Engineering - Reduction

PCA

$$PC_j = (a_{j1} \times Predictor_1) + (a_{j2} \times Predictor_2) + \ldots + (a_{jP} \times Predictor_P)$$

where:

P is the number of predictors
a_{ji} coefficients – called component weights

Critical thoughts

- Creates uncorrelated components

- Seeks predictor set variations, and disregards any further understanding
  - Blind to the response variable and predictor relation to response

34

Feature Engineering - selection

Removing predictors/features
Fewer predictors mean less computational time and complexity for ML models.

To remove or not to remove ⁉

Remove if:

1. two predictors are highly correlated, discard one
2. Some models could do better with removing predictors which have degenerate distributions
3. Zero variance predictor - single unique value; some models may be unaffected whereas other may find them problematic – discard discard!
4. Near zero variance predictor - handful unique values, and may have a single value for the vast majority of samples

35

Feature Engineering - selection

Understanding near zero variance predictors - I

Suppose you are doing a text mining task.

1. using keywords as predictors
2. some keywords occur in a small group of documents and remain unused other wise

Number of Documents	Keyword_1 unique count values
523	0
6	2
1	3
1	6

3. Of the 531 documents , only four documents with keywords

36

University of South Wales Prifysgol De Cymru

1. Basics of Machine Learning (ML)
2. The Machine Learning Process
   i. Data Preparation
   ii. Feature Engineering

Feature Engineering - selection

Understanding near zero variance predictors - II

4. Of the 531 documents , only four unique count values
5. Majority has zero count – (523) do not have any key words.
6. six documents have a count of *two, one documents each have a count of three and six* respectively
7. The percentage of unique values to total number of documents is 0.8%
8. A resampling of the data might not even have any keywords at all, this resampling may put undue influence eon the model

University of South Wales Prifysgol De Cymru

1. Basics of Machine Learning (ML)
2. The Machine Learning Process
   i. Data Preparation
   ii. Feature Engineering

Feature Engineering - selection

Diagnosing near zero variance predictors - I

1. Check whether the number of unique point relative to data/sample is small
2. Itself, this may be not a concern unless the frequency of these values in severely disproportionate
3. To check the disproportion, check the ratio of the most common frequency to the second most common frequency
   Ex:
   523 documents had a frequency of 0, and then 6 documents had another unique frequency; so, 523/6 = 87. The ratio is rather high and indicates a strong imbalance!

University of South Wales Prifysgol De Cymru

1. Basics of Machine Learning (ML)
2. The Machine Learning Process
   i. Data Preparation
   ii. Feature Engineering

Feature Engineering - selection

Diagnosing near zero variance predictors - II

Based on the discussion, we can have two rules of thumb to diagnose as follows:

1. The fraction of unique values over the sample is is low (say 10%) – in our example was 0.8%
2. The ratio of frequency of the most prevalent value to the frequency of the second most prevalent value is large (say ~20)

University of South Wales Prifysgol De Cymru

1. Basics of Machine Learning (ML)
2. The Machine Learning Process
   i. Data Preparation
   ii. Feature Engineering

Feature Engineering - selection

Correlation between predictors

Collinearity - the situation where a pair of predictor variables have a substantial correlation with each other

Multicollinearity – correlation between multiple predictor at once

University of South Wales Prifysgol De Cymru

1. Basics of Machine Learning (ML)
2. The Machine Learning Process
   i. Data Preparation
   ii. Feature Engineering

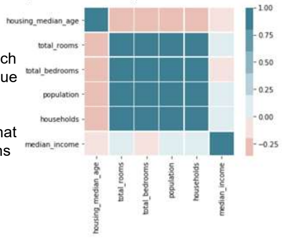
Feature Engineering - selection

Correlation between predictors

The target variable in housing dataset is 'median_house values'

Let's draw a correlation plot between predictor variables.

more heuristic approach to dealing with this issue is to remove the minimum number of predictors to ensure that all pairwise correlations are below a certain threshold



University of South Wales Prifysgol De Cymru

1. Basics of Machine Learning (ML)
2. The Machine Learning Process
   i. Data Preparation
   ii. Feature Engineering

Feature Engineering - selection

Correlation between predictors

The method just discussed tackles collinearity in two dimensions, but can still impact some models' performance. Following set of rules can be helpful:

1. Calculate the correlation matrix of the predictors.
2. Determine the two predictors associated with the largest absolute pairwise correlation (call them predictors A and B).
3. Determine the average correlation between A and the other variables.
4. Do the same for predictor B.
5. If A has a larger average correlation, remove it; otherwise, remove predictor B.
6. Repeat Steps 2–4 until no absolute correlations are above the threshold.

Feature Engineering - selection

University of South Wales Prifysgol De Cymru

1. Basics of Machine Learning (ML)
2. The Machine Learning Process
   i. Data Preparation
   ii. Feature Engineering

Correlation between predictors

The method just discussed tackles collinearity in two dimensions, but can still impact some models' performance. Following set of rules can be helpful:

1. Calculate the correlation matrix of the predictors.
2. Determine the two predictors associated with the largest absolute pairwise
3. correlation (call them predictors A and B).
4. Determine the average correlation between A and the other variables.
5. Do the same for predictor B.
6. If A has a larger average correlation, remove it; otherwise, remove predictor B.
7. Repeat Steps 2–4 until no absolute correlations are above the threshold.

43

Feature Engineering

University of South Wales Prifysgol De Cymru

1. Basics of Machine Learning (ML)
2. The Machine Learning Process
   i. Data Preparation
   ii. Feature Engineering

Adding Features

The high correlation among features: total_rooms, total_bedrooms, population and households is expected.

But should we go ahead with the algorithm discussed previously and start dropping features?

NO
Remember: correlation just shows linear dependence

We shouldn't just blindly follow the correlation rule. In our example we can use the high correlation features to create new features!

44

Feature Engineering

University of South Wales Prifysgol De Cymru

1. Basics of Machine Learning (ML)
2. The Machine Learning Process
   i. Data Preparation
   ii. Feature Engineering

Adding Features

A categorical feature can be decomposed into a set of more specific variables. For ex:

From housing data ' ocean_proximity' can be broken down into various features using different approaches and some threshold values. Some of them are as follows:

1. *near_water, inland, island*
2. *near_water, inland, coast_type (bay,ocean,island)*
3. *inland, near_ocean_or_bay, island, <1hocean*

45

Feature Engineering

University of South Wales Prifysgol De Cymru

1. Basics of Machine Learning (ML)
2. The Machine Learning Process
   i. Data Preparation
   ii. Feature Engineering

Adding Features and binning

Binning - Take a numeric predictor and pre-categorize or "bin" it into two or more groups

Binning can be used to create new features or can be used to simply categorise features as they are.

There could be certain drawbacks to binning continuous data when being used as features for ML models:
1. there can be a significant loss of performance in the model
2. there is a loss of precision in the predictions when the predictors are categorised
3. may lead to a high rate of false positives according to some research

46

Feature Engineering

University of South Wales Prifysgol De Cymru

1. Basics of Machine Learning (ML)
2. The Machine Learning Process
   i. Data Preparation
   ii. Feature Engineering

Binning, discretization, and linear models

However, if used on continuous data, binning can improve the performance of linear ML models.
Linear models only model linear relationship between predictors and target variables.

on continuous data, binning of the feature splits it into multiple features/ categorical feature

We will now follow the example provided in lecture resource.

The example draws comparison between linear regression (a linear model) and a decision tree regression (a non linear model)

47