Joint Tech Internship Community Program

Assignment 1

SUBMITTED BY :

JAHAGANAPATHI SUGUMAR

CANDIDATE ID : 2024060193

THE GIVEN TABLE :

Make Year	Brand	Variant	Mileage	Fuel	Transmission	Resale Price (INR)
2015	BMW	520D	80000	Diesel	Automatic	2500000
2016	Audi	A6	92000	Petrol	Automatic	1900000
2018	Mercedes Benz	E200	61000	Petrol	Automatic	3400000
2014	Skoda	Superb	95000	Petrol	Automatic	600000
2020	Benz	E200	35000	Petrol	Automatic	12000000

Feature :

From the given table the inputs are Make Year, Brand, Variant, Mileage, Fuel, Transmission. these input determines the price of the car. And price depends on the inputs.

LABEL :

From the given table the Label is the Resale price which is predicted with the help of its feature.

Prediction :

Prediction:-

Prediction refers to the process of using a trained model to estimate the output. This involves applying the learned relationship from the training data to new, unseen data or generate outcomes.

Outlier :

From the given table since there exist of one value which is different from the majority values it is univariate outlier

Make Year	Brand	Variant	Mileage	Fuel	Transmission	Resale Price (INR)
2015	BMW	520D	80000	Diesel	Automatic	2500000
2016	Audi	A6	92000	Petrol	Automatic	1900000
2018	Mercedes Benz	E200	61000	Petrol	Automatic	3400000
2014	Skoda	Superb	95000	Petrol	Automatic	600000
2020	Benz	E200	35000	Petrol	Automatic	12000000

Dictation can be done by the Z-Score if it is $\pm$ 3.29 *or beyond*

$$\text{Z-Score} = \frac{x_i - \overline{x}}{s}$$

$$\overline{x} \rightarrow mean \rightarrow 4{,}080{,}000$$
$$x_i \rightarrow outlier\ value \rightarrow 12000000$$
$$s \rightarrow standard\ deviation \rightarrow 4{,}063{,}200.71$$

Z-Score = $\frac{12000000 - 4{,}080{,}000}{4{,}063{,}200.71}$ = 1.95

it is not considered an outlier based on the Z-score analysis.

1) Domain knowledge 3) Math / Statistics
2) Visualization
$\Downarrow$
Two standard deviation

Outlier:-

→ Outlier are cases that have data values that are very different from the data values for the majority of cases in the data sets.

→ An outlier is an observation that is substantially different from the other observation

→ outliers are important because they can change the results of our data analysis

Types of outliers

univariate | multivariate

univariate:-

univariate outlier are cases that have an unusual value for a single variable.

univariate can be predicted with it Z-Score is ±3.29 or beyond

Z-score = (i-mean) / standard deviation

$$Z_i = \frac{x_i - \bar{x}}{s}$$

__/__/__

multivariate

→ multivariate outliers are cases that have an unusual combination of values for a number of variables

→ It can be detected by the Mahalanobis Distance (D2) Probability is less than 0.001

Mahalanobis distance:-

$$D_m = \sqrt{(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

where μ is mean

Σ is covariance mean

Generalization of Z-scores to multi-dimensional space.

→ Replace univariate mean with multivarite mean

→ Replace Standard deviation with covariance

Z-score

Mahal fun

$$Z_i = \frac{x_i - \bar{x}}{s} \longrightarrow \sqrt{(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

Training Data :

From the given table assuming that 80 percent of the data is training data.

Make Year	Brand	Variant	Mileage	Fuel	Transmission	Resale Price (INR)
2015	BMW	520D	80000	Diesel	Automatic	2500000
2016	Audi	A6	92000	Petrol	Automatic	1900000
2018	Mercedes Benz	E200	61000	Petrol	Automatic	3400000
2014	Skoda	Superb	95000	Petrol	Automatic	600000

Test Data:

From the given table assuming that 80 percent of the data is training data and remaining 20 percent is Test data.

Make Year	Brand	Variant	Mileage	Fuel	Transmission	Resale Price (INR)
2020	Benz	E200	35000	Petrol	Automatic	12000000

Validation Data :

Validation data:-

→ Validation data is used to evalulate the Performance of a model during training It is a subset of data sets.

Models:

Model:-

→ models are algorithm that allows computers to learn from & to make prediction based on data.

__/__/__

Types of models:-

1) Supervised Learning
2) unsupervisd.

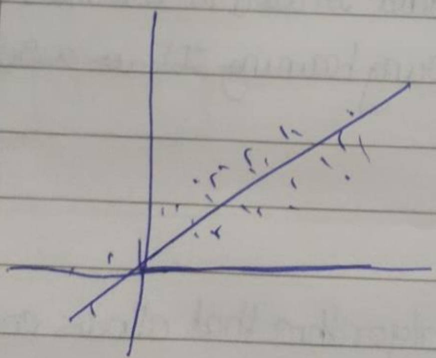
Supervised Learning:-
→ Regression
→ Classification

Regression:-

→ we try to make a relation between Dependent variable & Independent variable.

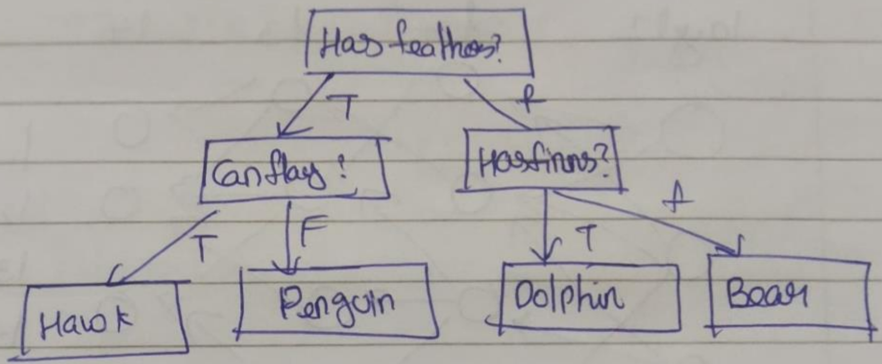
→ continuous output

I) Linear Regression



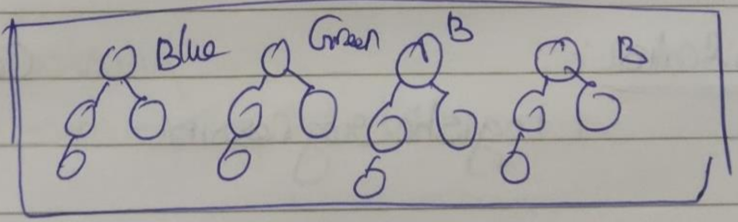
- Multiple Linear
- Polynomial ⇒ curve

II) Decision Tree



more nodes → more accurate

III) Random forest



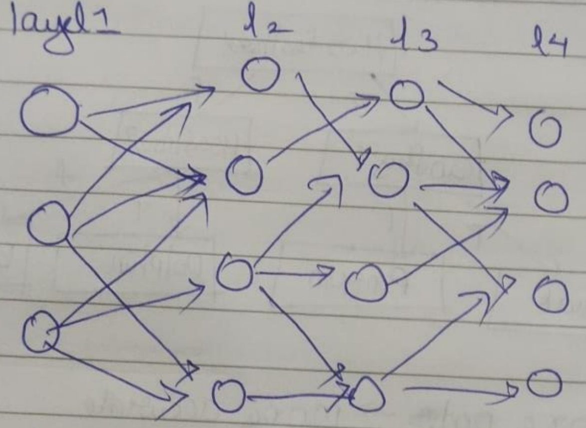
majority wins Model

↓

Blue

__/__/__

iv) Neural Network



l_1 = Input layer
l_2 = Hidden
l_3 = hidden
l_4 = output layer

hiddenlayer represent the function where the input goes through and leading the output.

<u>Classification</u> :- (discrete)

1. Logistic regression

→ To find the Probability of finite number of outcomes
→ varies between 0 & 1

II. Support Vector machine

→ N-dimensional space that distinctly classifies data Points.

III. Naives Bayes

→ Based on Bayes Theorem

$$\underbrace{P(B|A)}_{\text{Posterior Probability}} = \frac{\overbrace{P(A|B)}^{\text{Liklihood}}\ \overbrace{P(B)}^{\text{Prior Probability}}}{\underbrace{P(A)}_{\text{Predictor}}}$$

Un supervised Learning:-

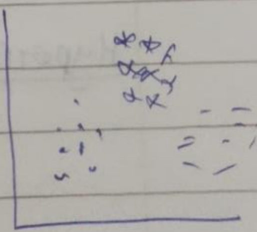
→ find Patterns from input data without reference to labeled outcomes

1) Clustering
2) Dimensionality Reduction

Clustering:-

Group data Points into clusters

> K-means
> Hierarchical
> Mean Shift
> Density-based



Hyperparameter :

<u>Hyperparameter:-</u>

→ Hyperparameter are Parameters whose Values are set before the learning Process begins

→ used to control the learning Process

common Hyperparameter:-

1) Learning Rate
2) Epochs
3) Batch size

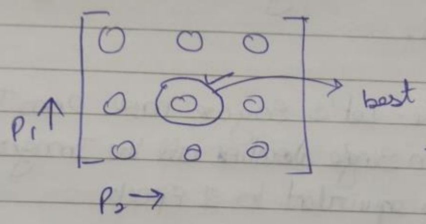
<u>Hyperparameter Tuning:-</u>

1) Grid Search
2) Random Search

HT refers to the Process of choosing the optimum set of hyperparameters for a machine learning model

Grid Search CV:-

$P_1 = \{1, 2, 3\}$
$P_2 = \{4, 5, 6\}$

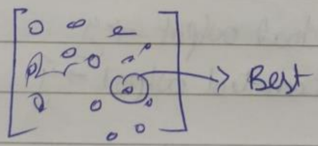


Time consuming

Random SearchCV:

→ randomly selects a specified number of combinations & evaluates

→ randomness makes the search process faster



faster

Epoch :

Epoch:-

→ Epoch is a complete iteration through the ~~entire~~ entire training dataset in one cycle for the training the model.

for Example:- Let's say we have 2000 Images as a dataset. In single iteration all the Images will be trained which is equivalent to 1 Epoch.

Loss Function:

loss function :-

→ loss function / cost function / error function

→ Loss function is a method of evaluating how well machine learning algorithm models given datasets.

→ actual output → y

→ Predicted output → $\hat{y}$

After training of the data the actual output & Predicted output should be same.

if it is not same it should be minimum

$\Downarrow$

model is trained properly

$$= \sum (y - \hat{y})^2$$

Learning Rate :

__/__/____

Learning Rate :-

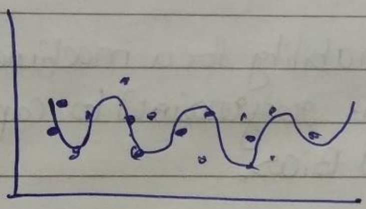
→ The learning rate controls how much to change the model in response to the estimated error each time the model weights are updated.

→ denoted by α or η

Overfitting :

→ can't find relationship between x & y

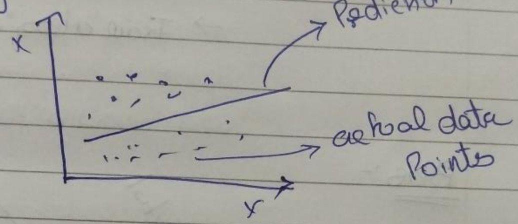
Overfitting:-



This is the case when the model is working well on training data but at time of testing the model is not performing well.

Underfitting :

underfitting:-



Since the line is far away from actual datasets means that the model is not able to learn from the data. → Poor on training data → model is to simple to understand data

→ can't find relationship between x & y

Overfitting:-

fix underfitting:-

→ By Increasing the model complexity

→ ~~Should be good relationship between~~ ~~X & Y~~

fix overfitting:-

→ Train with more data

→ remove features

→ Train a less complex model

Regularization :

_/_/_

Regularization:-

To ↓ Overfitting

Regularization avoids overfitting by adding a penalty to the model loss fonction.

Regularization = Lossfonction + Penalty

3 techniques

→ L_2 regularization

→ L_1

→ Elastic Net

L_2 regularization:-

Ridge Regression cost function = Loss fonction + $\frac{1}{2} \lambda \sum_{j=1}^{m} w_j^2$

where w is the slope of the line

λ control the strenght of regularization

if ~~λ controls the strength~~

$\lambda = 0$,

regularization will be eliminated

L_1 Regularization:-

Lasso Regression cost fonction = Loss fonction + $\lambda \sum_{j=1}^{m} |w_j|$

Elastic net Regularization

→ uses both L_1 & L_2

$$J(\beta_1, \beta_2, \ldots, \beta_n) = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} x_{ij} \beta_j\right)^2 + \lambda \Big(\underbrace{\alpha \sum_{j=1}^{m} |\beta_j|}_{L_1} + \underbrace{\frac{1-\alpha}{2} \sum_{j=1}^{m} \beta_j^2}_{L_2} \Big)$$

$\alpha \Rightarrow 0 \Rightarrow L_2$

$\alpha \Rightarrow 1 \Rightarrow L_1$

Cross-Validation :

Cross validation :-

→ It involves dividing the available data into multiple fold or subsets, using one of these folds as a validation set & training the model on the remaining folds.

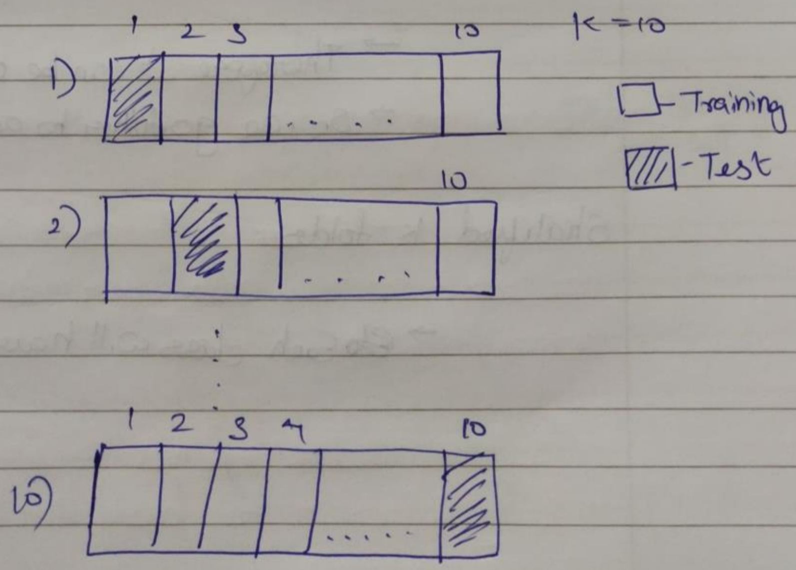
~~→ Ensure the model selection~~

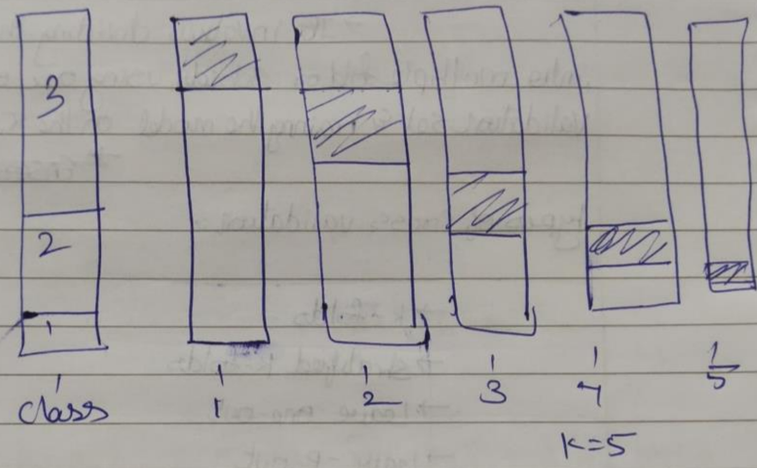
types of Cross validation :-

→ k-folds
→ stratified k-folds
→ Leave-one-out
→ Leave-P-out

k-folds :-

→ we divide the entire dataset into k-folds





disadvantage :-

→ Each iteration is not taking Equal representation of the class . !.

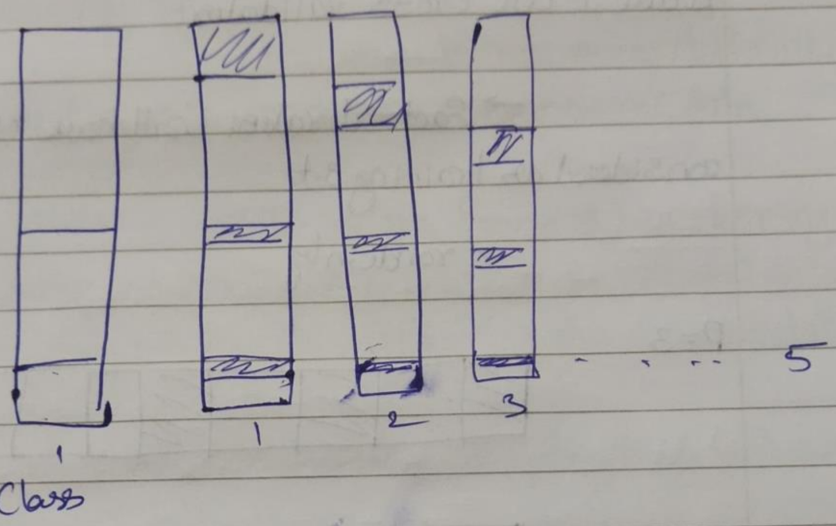
→ Therefore it can be over fitting again

→ So we go futher to anothe Type

Stratified k-folds:-

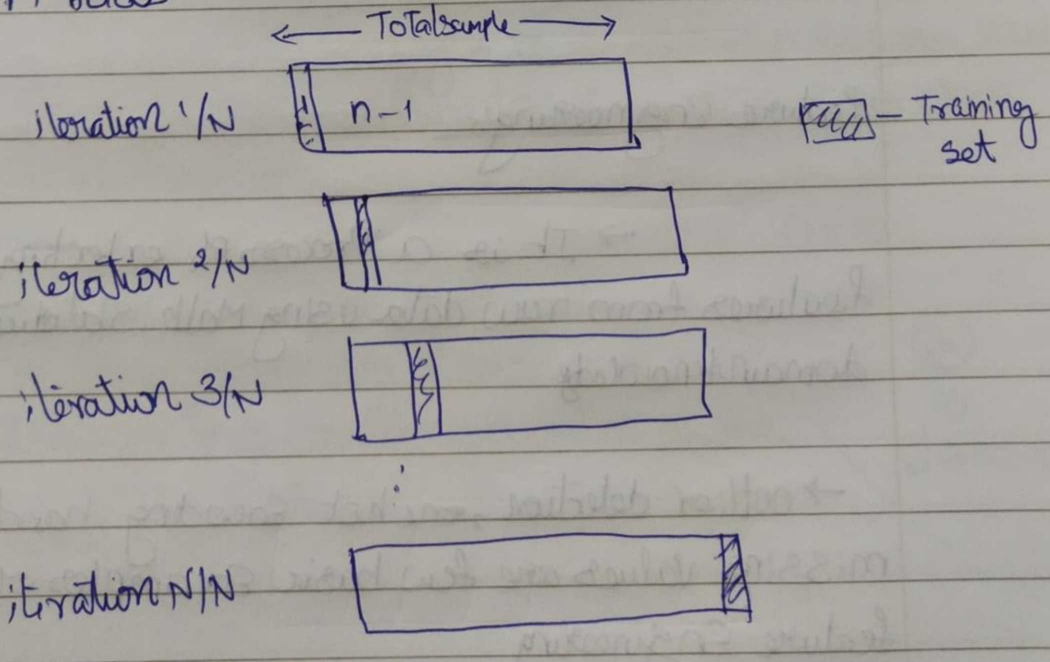
→ ~~Eto~~ Each class will have Equal representati

__/__/__



Leave one out cross validation :-

→ ToTal cross validation is divided into n folds

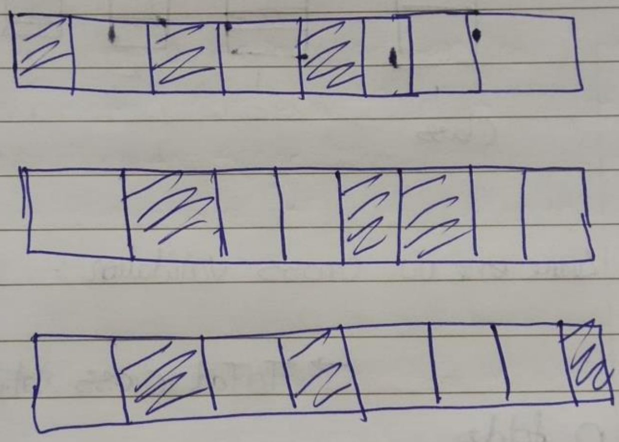


Leave P out cross validation

→ Each iteration will have No. P's will be considered as training set.

→ randomly

P=3



Feature Engineering :

Feature Engineering:-

→ It is a Process of extracting useful features from raw data using Math, Statistics and domain knowledge

→ outlier detection, one hot Encoding, handling missing values are few basic Examples of feature Engineering

Dimensionality Reduction :

Dimensionality Reduction:-

→ ML methods have some difficulty when dealing with such high-dimensional data

→ It is the Process of transforming high-dimensional data into a lower dimensional dataset that still Preserves the essence of the original data

→ used to reduce the no. of features

Types:-

1) Feature selection
2) feature Extraction

1) Feature selection

→ finding k of the total of n features that give us the most information & we discard the other (n-k) dimensions.

$n=10$

$10 \Rightarrow 7$

3

⇓

discarded

2) feature Extraction

→ finding a new set of k features that are the combination of the orginial n features

__/__/____

feature extraction

- → Principal component Analysis (PCA)
- → Linear Discriminant Analysis (LDA)

PCA:-

→ Introduced to karl Pearson

→ It works on the condition that while the data in a higher dimensional space is mapped to data in a lower dimension space, the variance of the data in the lower dimensional space should be minimum.

steps :-

→ constructs covariance matrix of the data

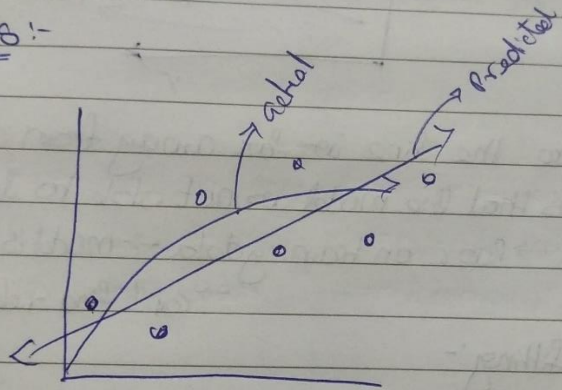
→ find the Eigen vectors

→ Eigenvectors corresponding to the largest eigenvalues are used to reconstruct a large ~~fraction~~ fraction of variance of the original data.

Bias :

→ Train a less complex model

Bias :-



The inability for a machine learning method (like linear regression) to capture the true relationship is called Bias.

→ It is the difference between the expected Prediction of our model & the true value.

→ These differences between actual or expected values and the Predicted values are known as Bias.

Low Bias :- The model will closely match the training dataset

High Bias :- The model will not match the training dataset

high Bias → underfitting.

Variance :

Variance:-

→ variance refers to the changes in the model when using diffrent Portions of the training or test data

→ Variance is the variability of the model that how much it is sensitive to another subset of the training dataset.

__/__/__

Low variance:-

→ Low variance means that the model is less sensitive to changes in the training data

→ case of underfitting

High variance:- High variance means that the model is very sensitive to changes in the traing data & can result in significant changes in the estimate of the target fuction when trained on different subsets

→ Case of overfitting.

high Bias, Low Variance ⇒ underfitting
high Var, Low Bias ⇒ overfitting
high var, high Bias ⇒ inconsistent & inaccurate Predictions

Low Bias, Low Variance ⇒ consistent & accurate Predictions.

Bias variance trade off:-

→ An model or Algorithm can't be more complex & less complex at the same time