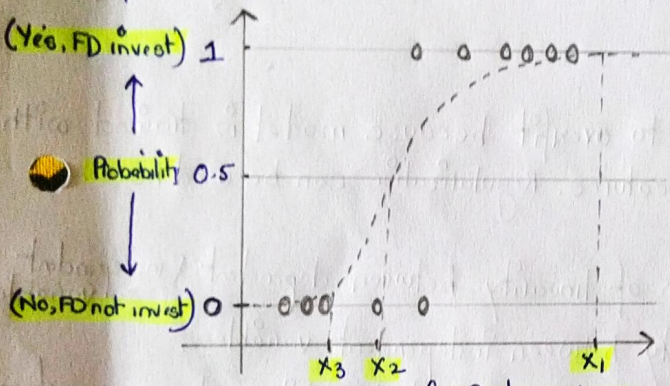
INTERVIEW DAY - LOGISTIC REGRESSION (LR), Sigmoid function, Advantage, Disadvantage, Assumption.

Q) What is logistic regression? Give an example.

Ans - Logistic regression predict the categories (categorical values) instead of continuous values like in simple / Multiple linear regression.

- LR is a special type of Generalized Linear Model (GLM).
- So consider a dataset where response variable falls into one of the two categories. LR models the probability that Y belong to particular category.

Example → If a customer will invest in Fixed Deposit or Not?



→ LR fits "S" shaped logistic function.

→ The curve goes from 0 to 1, tells us what probability that the customer will invest in fixed deposit.

→ If the customer age is high (x_1), he/she will invest in fixed deposit (high probability)

→ If the customer age is low (x_3), he/she will not invest in FD (Low probability).

→ If the customer age is intermediate (x_2), there is 50% chance will invest in FD.

So, in short if the probability is greater than 0.5/50%, he/she will invest in FD.

Normally, linear regression ⇒ $y = b_0 + b_1 * x$; add sigmoid fn on Left side.

$$\ln\left(\frac{P}{1-P}\right) = b_0 + b_1 x$$, known as logistic function/logit fn.

$\frac{P}{1-P}$ = odds, that can take any value between 0 and ∞, then apply log.

$\log\left(\frac{P}{1-P}\right)$ also known as logit function. Also, odd = $\frac{P(\text{event})}{1-P(\text{event})} = \frac{\text{Probability event occurs}}{\text{probability event not occur}}$

Q) How to create S shape? i.e, Sigmoid function/Logistic function.

Ans - Sigmoid function is a mathematical function having a characteristic that can take any real values and map it between 0 to 1, shaped like letter "S".

- Sigmoid function also known as logistic function.

Formula, $\boxed{y = \frac{1}{1+e^{-z}}}$

- So if value of z goes to positive infinity then predicted value of y will become 1 and if goes to negative infinity then predicted value of y become 0.

- If the outcome of sigmoid function is more than 0.5 then we classify label as class 1 and if it is less than 0.5 then we classify it as class 0.

Q) Give the advantages and disadvantages of Logistic Regression?

Ans - ADVANTAGES OF LR —

i) It makes no assumption about distribution of target class.
ii) It measure how appropiate a predictor size (coefficient size) & degree of association (positive/negative)
iii) It can interpret model coefficient as indicator of feature importance.
iv) LR is less inclined to overfit but it can overfit in high dimensional dataset. We may consider (L1 and L2) technique to avoid overfitting.
v) This algorithim can easily be extended to multi-class classification (more than 2 class target) using softmax classifier, known as Multiclass logistic regression.

DISADVANTAGES OF LR —

i) On high dimensional data, LR tends to overfit because model is trained with little training data and lots of features. Regularization can be used here.
ii) LR is sensitive to outliers.
iii) Major limitation of LR is assumption of linearity between dependent & independent variable.
iv) LR require average/no multicollinearity between independent variables.
v) Non linear problem cannot be solved with LR since it has linear decision surface. So transformation of non linear feature is required, so data become linearly seperate

---

Q) What are the assumption in Logistic Regression?

Ans - Assumptions of LR are —

i) Outcomes should be categorical in binary or multinomial.
ii) There should be linear relationship between Independent and dependent variable.
iii) There should be NO influential value (extreme/outliers) in predictors.
iv) There should be NO high ~~intercorrelated~~ intercorrelation/multicollinearity among the predictor.

INTERVIEW DAY – , ODDS, log(odds), Odds Ratio, Logs(Odds Ratio)

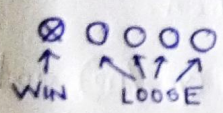
Q) What is Odd Ratio? How it is different from probability?

Ans –

$$\text{Odds} = \frac{\text{Something happening}}{\text{Something not happening}}$$

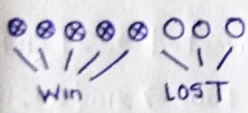
$$\text{Probability} = \frac{\text{Something happening}}{\text{everything that could happen}}$$

Example 1 – We might say that the odds in favour of my team winning the game are 1 to 4 times.



$\text{odds} = \frac{1}{4} = 0.25$, $\text{Probability} = \frac{1}{5} = 0.20$

Example 2 – we may say that odds in favour of my team winning the game 5 to 3.



$\text{odds} = \frac{5}{3} = 1.7$, $\text{Probability} = \frac{5}{8} = 0.625$

Odds are not probability. Odds are ratio of something happening divide by Something not happening. Probability is ratio of something happening divide by everything that could happen. In many case, $\text{odds} = \frac{P}{1-P}$

It is derived from, $\text{odds(win)} = \frac{\text{Probability (win)}}{\text{Probability (loose)}} = \frac{\text{Probability(win)}}{1 - \text{Probability (loose)}} = \frac{P}{1-P}$

Q) what is log(odds)? Given an example. And why we added logs.

Ans – Log of odds solve the problem of symmetry.

Suppose my team is bad,

odd = 1/4 = 0.25

more worse, 1/8 = 0.125.

more more worse, 1/16 = 0.06

Full worse, $\frac{0}{\text{anything}} = 0$

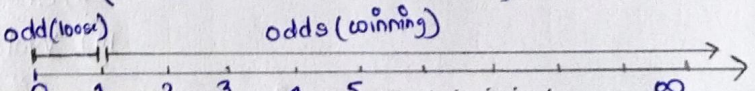
Suppose my team is good,

odds = 5/3 = 1.7, improve 9/3 = 3.

More improve 27/3 = 9, will go to ∞.

So, range of winning if team is strong is 1 to ∞.

(Numerator > Denominator)



Range will be 0 to 1 (Denominator > Numerator)

So asymetry make it difficult to compare odd for or against my team winning.

Example, if odds are against 1 to 6, then odds = 1/6 = 0.17 but if in favour, 6/1 = 6.

Taking logs of this odd can solve and make everything symmetry.

If odds are against 1 to 6 = log(odds) = log(1/6) = log(0.17) = -1.79.

If odds are against 6 to 1 = log(odds) = log(6/1) = log(6) = 1.79.

$$\log(\text{odds}) = \log\left(\frac{P}{1-P}\right)$$

this is known as the log of ratio of the probabilities, is called logit function, and form basics of logistic regression. Logs make the range $-\infty$ to ∞, and midpoint is 0.

Q) What is Odd Ratio and log (Odd Ratio)? Give an example.

Ans - When people say about odd Ratio, they are talking about ratio of odds.

$$\text{Odd Ratio} = \frac{\text{Odd}_1}{\text{Odd}_2} = \frac{\frac{xx}{0000}}{\frac{000}{x}} = \frac{2/1}{3/1} = 0.17$$

So when we calculate odd Ratio of something.

- If denominator is larger than numerator, odd ratio goes from 0 to 1.
- If numerator is larger than denominator, odd ratio goes from 1 to ∞.

Therefore taking, logs (odd Ratio) will make things symmetrical.

Example, if there is any relationship between Investment in FD (yes/no) and age of customer (high, age ≥ 60 & low, age < 60)

Age	INVESTED FD: Yes	INVESTED FD: No
High	23	117
Low	6	210

Total people → 356.

Total FD invested → 29 (23+6), Not Invested → 327 (117+210)

Age, High total → 140 (23+117), Total Low → 216 (6+210)

Given a person of high age, odds of FD invested = $\frac{23}{117}$.

Given a person of low age, odds of FD invested = $\frac{6}{210}$.

odds ratio = $\frac{23/117}{6/210} = \frac{0.2}{0.03} = 6.88$, so odds ratio tells us that the odds are 6.88 times greater that someone with high age will invested in Fixed deposit.

So odd's ratio / log (odd's ratio) tells us if there is strong / weak relationship between two variables, like whether or not Age (Low/high) increase odds of FD investment.

Larger value means that ~~rotated~~ age is a good predictor of FD.

Smaller value means that age is not a good predictor of FD investment.

3 ways to determine odds / log (odd ratio) → i) Fisher Exact Test

ii) Chi-Square test (to calculate p value)

iii) Wald test (calculate Confidence Interval, pvalue)

Q) Even if Odd formula is ratio, how it is different from Odds Ratio?

Ans - If we take random numbers, that all those odds up to 100 (for example) and use them to calculate odds it will not be normally distributed.

So, if we log (odds) it will become normally distributed.

And how it is different from Odds Ratio is explained above.

INTERVIEW DAY – Probability, Likelihood, Maximum likelihood

Q) What is the difference between probability and likelihood?

Ans – Probability is the percentage that a success occur. For example, tossing a coin. 0.5 is the probability of success.

– Likelihood is the conditional probability. Same example, we toss coin 10 times and suppose that we get 7 success and 3 failed.

Likelihood (0.5 | 7) = 0.1171.

Meaning, 0.117 is the probability that above event will happen (7 success out of 10 trials) by knowing probability of one success is 0.5 (toss one time).

Therefore,

Likelihood, is the probability (conditional probability) of a event (a set of succes) occur by knowing the probability of a success occur.

Probability, is the percentage that a success occur.

Another example, in a cricket match, a coin is tossed and one captain calls head and win the toss.

Now, what is the probability that winning captain will elect to bat?

1/2, either they will elect to bat or bowl, so probability is straight 50%.

But the commentator are discussing what is the likelihood that "winning Captain" will elect to bat.

Now that likelihood will not straight to 50% because likelihood depends on type of pitch, strengths and weakness of team, weather etc.

So, probability is straight up mathematics whereas likelihood is a function of many parameters and conditions.

Q) What is Maximum likelihood?

Ans – The goal of Maximum likelihood is to find the optimal way to fit a distribution to the data.

– There are lots of different distributions like normal or exponential.

– The reason we want to find / fit a distribution to data is it can be easier to work and more general.

Q) So how Maximum Likelihood work?

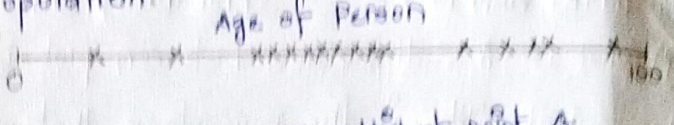
Ans – Normally in case of Normal distribution, there are lots of things.

i) We expect most of the measurement to be close to the average / mean.

ii) We expect measurement to be relatively symmetrical around the mean.

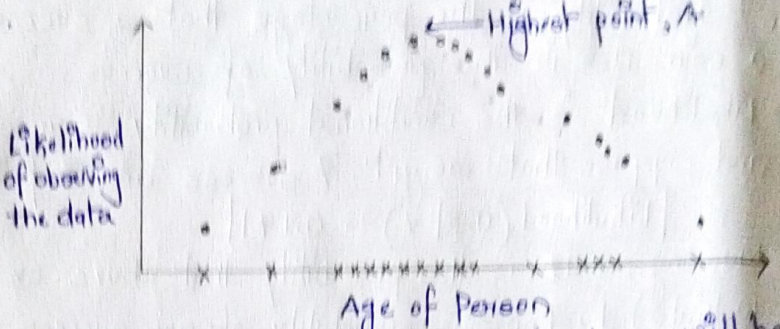
For example, average age of the population.

First, lets plot the data on axis



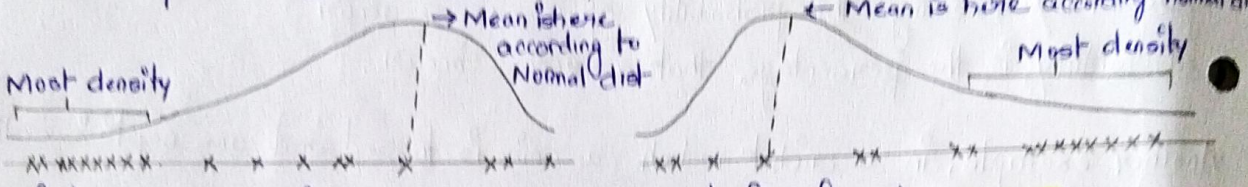
Likelihood plot on the data.

We want to locate "maximize likelihood" by observing the Age of a person i.e, which is point A



There can be other scenario as well, but we have to settle out where will be the center things, i.e, most of the values should be near the average (ideal)

Two possible scenario where likelihood will be low.



Unfortunately, most of the values are measured far from the mean.

So, maximum likelihood find the most optimal way to find distribution of given data.

Q) What is the role of Maximum likelihood in logistic regression?

Ans. Logistic regression uses Maximum likelihood for parameter estimation.

- In this a probability distribution for target variables (class label) must be assumed and then a likelihood function defines that calculated the probability of observing the outcome given the input data and the model.
- This function can then be optimized to find the set of parameters that results in the largest sum likelihood over training set.
- In MLE, we wish to maximize the conditional probability of observing the data (X) given a specific distribution and its parameters.

INTERVIEW DAY – REGULARIZATION, L1, L2, Elastic Net

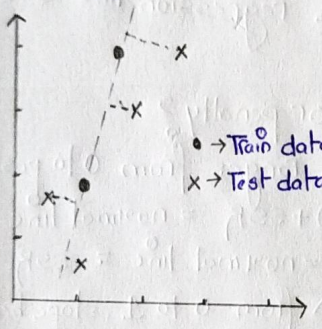
Q) How to address Overfitting in regression?

Ans – Overfitting, is the model which overfit in training data & performs poorly in test.
- One way could be reduce number of features in the model but there will be loss of information due to discarding. Model will not have benefit of all info.
- So regularization can be used, regularization is the process which regularize or shrink the coefficient towards zero.

Type of Regularization – i) Lasso Regularization (L1) ii) Ridge regularization (L2)
iii) Elastic Net Regularization.

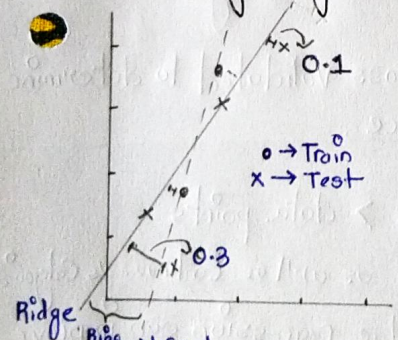
Q) What is Ridge Regularization (L2 Regularization)?

Ans – Suppose we have two datapoints and want to build linear regression.



- Since the line overlaps two data points, minimum sum of square = 0.
- So in train set, SSR = 0 but test set = very high.
- This means line is having high variance, in other word line is overfit in training dataset.
- So ridge regression comes here, main idea is to find a new line that doesn't fit training data well.
- In other word, we introduce a small amount of bias into line fit to the data.

Q) So how Ridge regression work?



- In normal, it try to reduce/minimize sum of square error.
- In ridge, it minimize = sum of square residual + $\lambda (\text{slope})^2$.
  slope², means adds a penalty
  λ, determine how severe the penalty is.

$y = mx + c$, eg size = y-axis intercept + slope (weight)
↑ slope ↑ slope/coefficient

Sum of Square (Ridge) = $(0.3)^2 + 0 + 0 + (0.1)^2 = 0.09 + 0.01 = 0.10$

Q) Give an example how normal & Ridge regression?

Ans – Let take example of FD take up and Age.

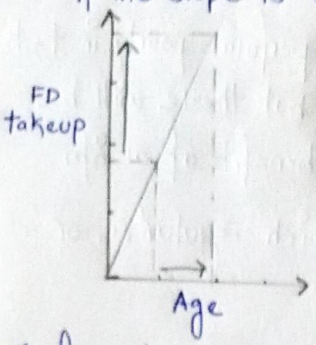
Normal equation, FD = 0.4 + 1.3 (age)
here sum of square = 0, slope = 1.3
for now $\lambda = 1$, penalty = $0 + 1(1.3)^2 = 0 + 1.69 = 1.69$

Ridge eqn is, FD = 0.9 + 0.8 (age)
here, SSR = 0.10, slope = 0.8
for now $\lambda = 1$, penalty = $0.10 + (0.8)^2$
$= 0.10 + 0.64 = 0.74$

So if we want to minimize sum of square residual plus ridge regression penalty, choose ridge regression over normal line because Ridge has small bias due to penalty has less variance.

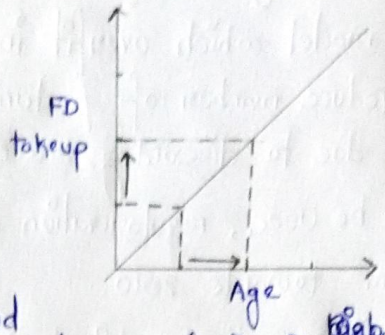
Q) Why Ridge regression is working better than least Square method?

Ans - If the slope is steeper



For every one unit increase in age then prediction for FD take up over two units.

If the slope is normal.



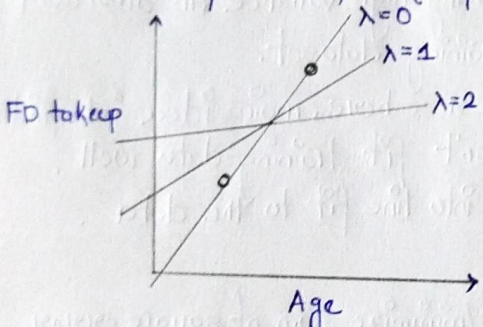
Line suggest that for every one unit increase in Age then there is one unit increase in FD takeup.

So from above, when the slope is ~~steeper~~ good / slope of line is ~~small~~ high, then predictions of FD takeup is less sensitive to change in FD take up. normal

Therefore in case of normal vs Ridge, least Square line is much more steeper than ridge line. Ridge regression penalty resulted in a line that has a smaller slope which means that predictions made with Ridge regression line are less sensitive to Age than the least Square line.

Q) How Ridge regression tries to minimize the error ? or penalty?

Ans → Penalty = SSR + λ(slope)^2, λ can be any value ranges from 0 to positive infinity



when $\lambda = 0$, penalty = 0 + SSR = normal line

there Ridge regression = normal line = SSR

As we increase the λ from 0 to 1, slope become less steeper

$\lambda = 2$, & slope become less steeper than previous

So larger we become λ, our predichon for FD takeup become less and less sensitive to Age.

Optimal value of λ → Use bunch of values of λ and use cross validated to determine which one result in the lowest variance.

Q) So when to use Ridge Regression?

Ans - Ridge regression is used when parameter / features > data points

- Ridge regression works with both Continuous Vs Continuous as well as Continuous vs Categorical
- In short when sample size are relatively small then Ridge regression can improve predictions (reduce variance) by making prediction less sensitive to training data.
- In general, ridge regression penalty shrinks but contains all the parameters.

Ridge regression = sum of squared residuals + λ(slope)^2

Ridge regression penalty itself is λ times sum of square parameters except Y intercept & λ is determined using Cross Validation.

Q) What is Lasso Regularization / L1 Regularization?

Ans → Ridge Regression = sum of square residuals + $\lambda (\text{slope})^2$

Lasso Regression = sum of square residuals + $\lambda * |\text{slope}|$

Here also, λ can be any value from 0 to ∞ and determined by cross validation.

Q) So what is the difference between L1 and L2 Regularization?

Ans

L1/Lasso Regularization	L2/Ridge Regularization
i) Eqn, sum of square residuals + $\lambda(\text{slope})^2$	i) Eqn, sum of square residuals + $\lambda * \|\text{slope}\|$
ii) Lasso can shrink slope all way to 0.	ii) Ridge can shrink the slope close to 0.
iii) Lasso, variable value can be equal to 0	iii) Ridge, variable value can be shrink to very small but cannot be 0.

Example, FD takeup = slope + β_1(Age) + β_2(occupation) + β_3(name) + β_4(weight)

So for FD takeup, 2 good variables → Age of a Customer, Occupation (Government/Non)

2 bad variables → Name of a Customer, Weight of a Customer.

So in L1/lasso, value will shrink and useless variable become 0. Only useful variable remain

In L2/ridge, value will shrink but never ≠0. So all 4 will remain regardless of useless.

Since, lasso regression can exclude useless variable from equation, it is little better than Ridge at reducing variance in a model that contain useless variables.

- In contrast, Ridge regression tends to do little better when most variable are useful.

Q) what is Elastic Net Regularization?

Ans - When a dataset include millions of parameters and when we have millions of parameters, we almost need to use some sort of regularize to esimate them. However, variable in those model might be useful/useless, we don't know. So in those case, we use Elastic Net Regularization which is combination of both Lasso (Some Variable are useful) or Ridge (all variables are useful).

Elastic Net = sum of square residuals + $\lambda_1 |var_1| + \cdots |var_n| + \lambda_2 (var_1)^2 + \cdots + (var_n)^2$

λ_1 → lasso penalty, λ_2 → ridge penalty, cross validation for different combination λ_1 & λ_2 to find best value.

$\lambda_1 = 0$ and $\lambda_2 = 0$, original least square. $\lambda_1 \neq 0$ and $\lambda_2 = 0$, lasso

$\lambda_1 \neq 0$ and $\lambda_2 \neq 0$, elastic net $\lambda_1 = 0$ and $\lambda_2 \neq 0$, ridge

- Elastic Net is good when we have highly correlated variables.

Elastic Net = sum of square residuals + $\lambda_1 |var_1| \cdots |var_n|$ + $\lambda_2 (var_1)^2 \cdots (var_n)^2$

$\lambda_1 |var_1| \cdots |var_n|$: Lasso penalty

$\lambda_2 (var_1)^2 \cdots (var_n)^2$: Ridge penalty

INTERVIEW DAY - WOE, IV

Q) What is Weight of Evidence(WOE)?

Ans - WOE tells the predictive power of an independent variables in relation to the dependent variable. It is generally described as a measure of the separation of good customers and bad customers.

- Good customers refers to the customer who paid loan back and bad customers who defaulted to loan.

$$WOE = \ln\left[\frac{\text{Distribution of Good}}{\text{Distribution of Bad}}\right]$$

ln → Natural log.

Distribution of Good → % Good customers in particular group.

Distribution of Bad → % Bad customers in particular group.

- Positive WOE → Distribution of Good > Distribution of Bad.
- Negative WOE → Distribution of Good < Distribution of Bad.

In general terms, $WOE = \ln\left(\frac{\%\text{ of non-events}}{\%\text{ of events}}\right)$

Steps to calculate WOE - 1) For continuous variable, split data into bins.
2) Calculate the number of event and non events in each bins.
3) Calculate the % of events and % non events in each bins.
4) Calculate WOE by taking log of division (% non event & % event)

NOTE → For categorical variable, we don't need to split the data.

Q) What are the rules related to WOE?

Ans - 1) Each bins should have atleast 5% of observations.
2) Each bins should be non-zero for both non-events and events.
3) WOE should be distinct for each category. Similar group should be aggregated.
4) WOE should be monotonic, either growing or decreasing with grouping.
5) Missing values should be binned seperately.

Q) Why combine categories with similar WOE?

Ans - Combine categories with similar WOE and then create new categories of an independent variable with continuous WOE values. It is because the categories with similar WOE have almost same proportion of events and non-events. In other word, the behaviour of both the categories is same.

To check correct binning with WOE -
1) WOE should be monotonic i.e, either growing or decreasing with bins. We can plot WOE and check linearity on graph.
11) Perform WOE transformation & check with logistic regression output

Q) What is Information Value (IV)?

Ans - IV is one of technique to select important variable in a predictive model. It help us to rank variable on the basis of importance.

$$\text{Information Value (IV)} = \Sigma \left(\frac{\% \text{ of non-events}}{\% \text{ of events}} \right) * \text{WOE}$$

INFORMATION VALUE (IV)	VARIABLE PREDICTIVENESS
Less than 0.02	Not useful for prediction (Not useful for modelling)
0.02 to 0.1	Weak predictive power (weak relation, to Good)
0.1 to 0.3	Medium predictive power (medium strength relation)
0.3 to 0.5	Strong predictive power (strong strength relation)
>0.5	Suspicious predictive power (check once again)

- IV is not a good selection method when we have multi class classification (not binary)
- Random forest can detect non linear relationship very well so selecting variable via Information value & using in RF might not produce robust model.

Q) What are the advantages of WOE & IV?

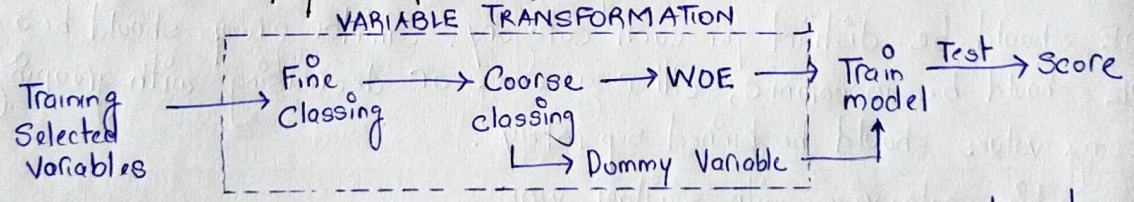
Ans - ① Main practical use of WOE is for encoding, where we can replace the classes with their associated value. For example, suppose in a dataset we found we can replace "Male" with 0.98 and "Female" with -1.58.

② Another positive outcome of WOE is to reduce the number of columns of the input used for training a model. Imagine we have a categorical ~~values~~ variable with 10 different classes, one hot encoding will give 10 different columns. Using WOE, classes can replaced by their associated WOE values.

③ As for IV, it provide relationship between independent and dependent variables.

Q) What are the work flow of WOE?

Ans -



- Fine classing → Applied to all continuous variable and those discrete variable with high cardinality. This is the process of initial binning into typically between 20 and 50 fine granular bins.
  To summarize create 10/20 bins for continuous independent variable and calculate WOE & IV of a variable.
- Coarse classing → Combine adjacent categories with similar WOE scores.