A linear regression model can be used to make predictions for **continuous variables**.

A Logistic Regression model can be used to make predictions in cases where the output is a **categorical variable**.

Basic concepts related to logistic regression. Broadly speaking, the topics that will be covered in this session are:

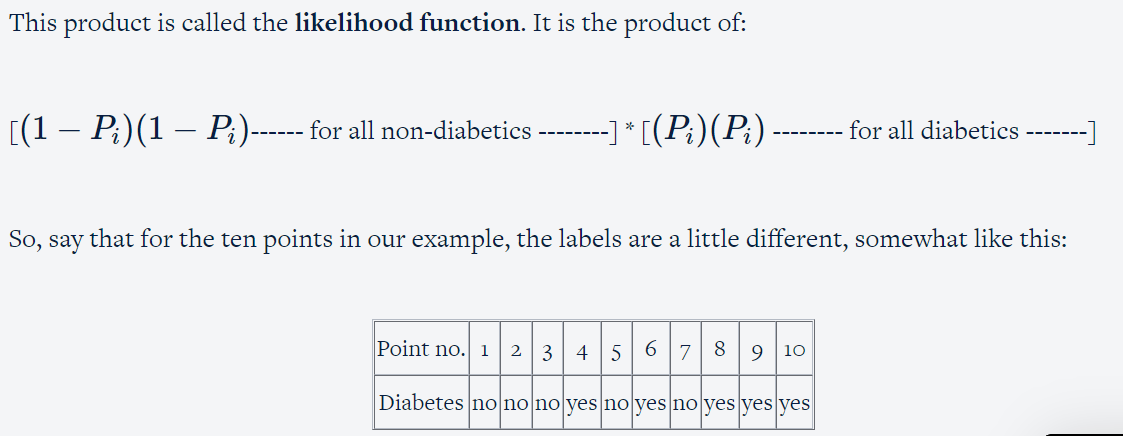
* Binary classification
* Sigmoid function
* Likelihood function
* Building a logistic regression model in Python
* Odds and log odds

In the sigmoid curve, as you can see, you have low values for a lot of points, then the values rise all of a sudden, after which you have a lot of high values. In a straight line though, the values rise from low to high very uniformly, and hence, the “boundary” region, the one where the probabilities transition from high to low is not present.

We saw what a sigmoid function is and why it is a good choice for modelling the probability of a class.

Best Fit sigmoid curve: finding the B0 and B1 in the equation : 1/(1+e^-(B0+B1x)

So, the best fitting combination of β0 and β1 will be the one which maximises the product:



In this case, the likelihood would be equal to

(1−P1)(1−P2)(1−P3)(1−P4)(1−P6)(P5)(P7)(P8)(P9)(P10)

By trying different values of β0 and β1, you can manipulate the shape of the sigmoid curve. At some combination of β0 and β1, the 'likelihood' (length of yellow bars) will be maximised.

How do you find the optimal values of β0 and β1 such that the likelihood function is maximized?

The optimisation methods used to do that maximum likelihood estimation, or MLE.

(Introduction to maximum likelihood estimation (MLE), MLE for continuous (normal) and discrete probability distributions  (Bernoulli distribution and logistic regression) ,Optimising MLE cost functions using gradient descent , Alternate optimisation methods: Newton-Raphson method ) ----**Things to know in MLE**

**Logistic Regression in Python**

In python, logistic regression can be implemented using libraries such as SKLearn and statsmodels, though looking at the coefficients and the model summary is easier using statsmodels.

You can find the optimum values of β0 and β1 using the python code:

**Odds and Log Odds**

So far, you’ve seen this equation for logistic regression:

P=1/(1+e−(β0+β1x))

Odds (p/1-p) = e^(B0+B1x)

Log odds ln(p/1-p) = B0+B1x

**Log odds is easy to interpret because as there is a linear relationship between log odds with the variables multiplied by coefficients. As the positive and more significant coefficients, the higher the log odds and vice versa.**

you learnt that in order to find the **best-fit sigmoid curve**, you need to vary β0 and β1 until you get the combination of beta values that maximises the **likelihood.**

This process, where you vary the betas until you find the best fit curve for the probability of diabetes, is called **logistic regression.**

VIF  calculates how well one independent variable is explained by all the other independent variables combined. And its formula is given as:

VIFi=11−Ri2

where 'i' refers to the ith variable which is being represented as a combination of rest of the independent variables.

1. In the logistic regression we are calculating the probabilities of the segmentation but not actual segmentation. we need to determine the cuttoff to create a segmentation.

2. After calculating the probabilities while determining the key decision cutoff we minimize the prediction error to find the best cutoff decision, which give rise to accurate logistic model.

3. In the logistic regression the we have determined the probabilities for a binomial classification but we have multinomial classification as well.

**Multivariate Logistic Regression (Model Evaluation)**'.

Metrics for logistic regression model evaluation:

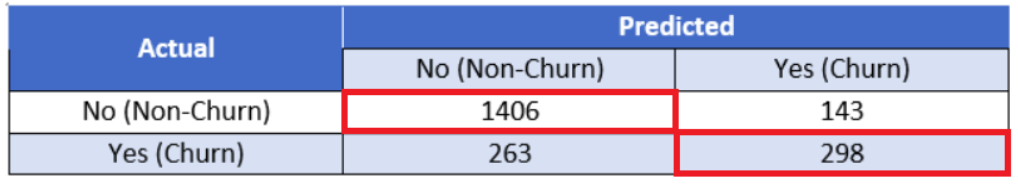
1. Accuracy: True predictions/total predictions = TP+TN/(TP+TN+FP+FN)….this we can determine from the confusion matrix.

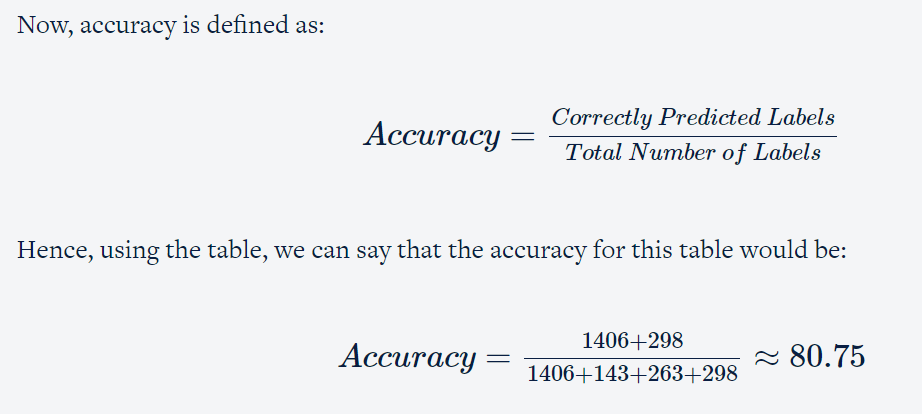
**# Create confusion matrix**

confusion = metrics.confusion\_matrix(y\_train\_pred\_final.Churn, y\_train\_pred\_final.predicted)

**# Calculate accuracy**

print(metrics.accuracy\_score(y\_train\_pred\_final.Churn, y\_train\_pred\_final.predicted))

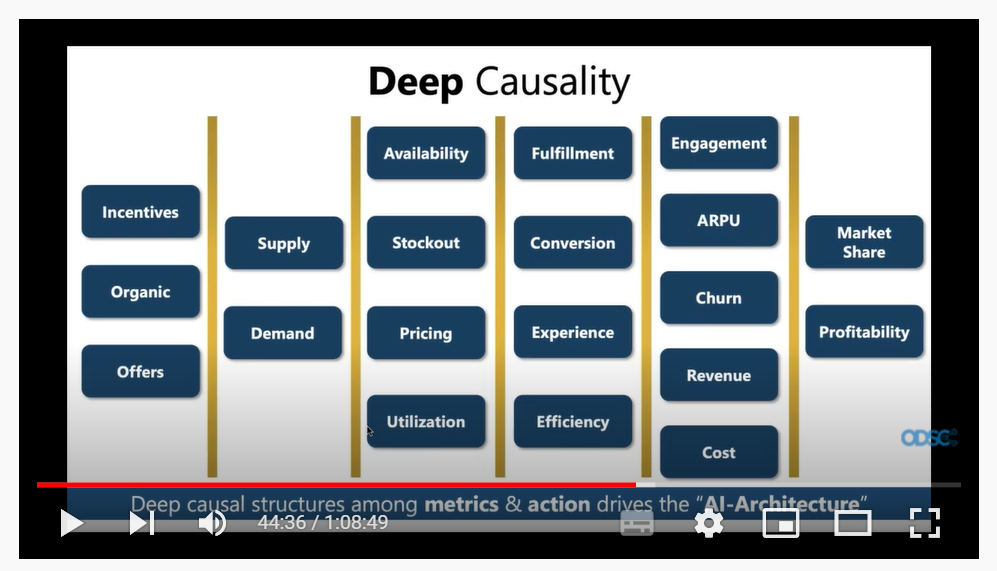


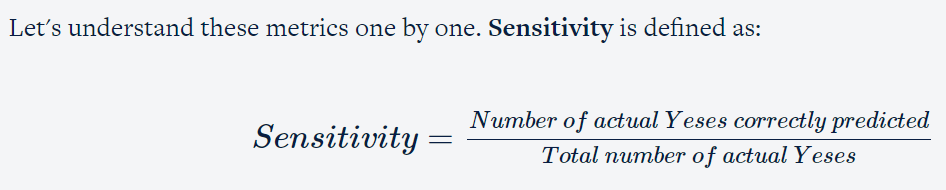


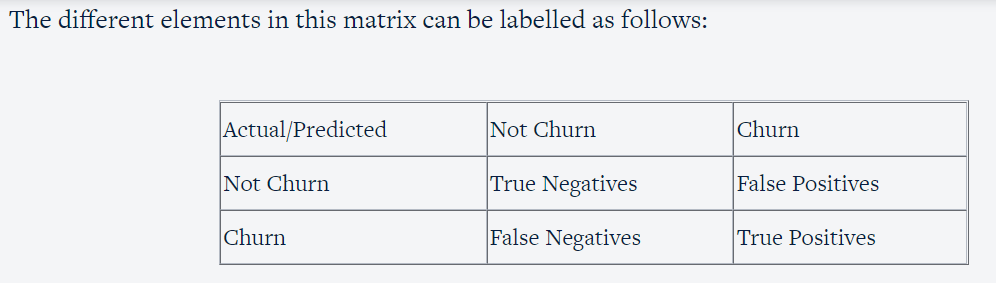
Few more metrics beyond accuracy that are essential to evaluate the performance of a logistic regression model are:

Sensitivity, specificity and Roc Curve, precision and recall.

We need to find out the optimal scenario where the model will perform the best based on the use case. Finally, once you've chosen the optimal scenario based on the evaluation metrics, you'll finally go on and make predictions on the test dataset and see how your model performs there as well.







Sensitivity/recall: TP/(FN+TP) specificity : TN/(TN+FP), Accuracy : TN+TP/(TN+TP+FN+FP)

False Positivity : FP/(FN+TP)

Positive Predictive Rate/Precision: TP/(TP+FP)

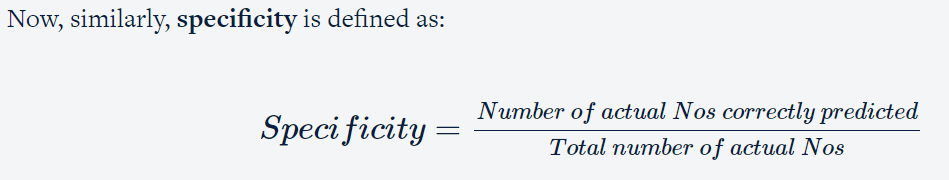
Negative Predictive Rate: TN(TN+FN)

True Positive Rate = Sensitivity

False Positive Rate = 1-specificity

**ROC shows the plot between TPR and FPR and this is used for evaluation of the models in classification problems. (Maximize TPR and Minimize FPR).**

So, now that you have understood what these terms are, you'll now learn about **Receiver Operating Characteristic (ROC) Curves** which show the **tradeoff between the True Positive Rate (TPR) and the False Positive Rate (FPR)**. And as was established from the formulas above, TPR and FPR are nothing but sensitivity and (1 - specificity), so it can also be looked at as a tradeoff between sensitivity and specificity.



**ROC Curve:**

Churn rate: churned/(churned+nonchurn)

True positive rate: actual true and predicted as True/(actual true predicted as true+actual true predicted as false)

False Positive Rate: Actual true but predicted as false(actual true predicted as true+actual true predicted as false)

1. Data cleaning and preparation
   * Combining three dataframes
   * Handling categorical variables
     + Mapping categorical variables to integers
     + Dummy variable creation
   * Handling missing values
2. Test-train split and scaling
3. Model Building
   * Feature elimination based on correlations
   * Feature selection using RFE (Coarse Tuning)
   * Manual feature elimination (using p-values and VIFs)
4. Model Evaluation
   * Accuracy
   * Sensitivity and Specificity ----1st View
   * Optimal cut-off using ROC curve
   * Precision and Recall----2nd view
   * Optimal cut-off using 2nd view
5. Predictions on the test set
6. To evaluate the logistic regression model we have multiple parameters like accuracy,sensitivity(Recall),specificity,Positive predictive Rate(precision), Negative prediction Rate, False Positivity, ROC curve, precision vs recall curve.
7. Depending on the scenario and to maximize the business output we utilize the evaluation metric for decision making and there is no single parameter to evaluate but a combination of multiple parameters and a tradeoff should be made for our decision making purposes.
8. Since the predicted values are probabilistic , so there exists a of FP’s and FN’s.

What are the two main differences between logistic regression and linear regression?

Logistic regression is a classification problem, whereas linear regression is a predicting problem.

Logistic regression target variable is categorical variable whereas linear regression target variable is a continuous variable.

**Suggested Answer:**

The two main important differences between logistic and linear regression are: 1. Dependent/response variable in linear regression is continuous whereas, in logistic regression, it is the discrete type. 2. Cost function in linear regression minimise the error term Sum(Actual(Y)-Predicted(Y))^2 but logistic regression uses maximum likelihood method for maximising probabilities.



**Nuances of Logistic Regression - Sample Selection**

However, even before you start building a model, you have to decide what kind of data would be appropriate for building it.

selecting the right sample is essential for solving any business problem. As discussed in the lecture, there are major errors you should be on the lookout for while selecting a sample. These include:

1. **Cyclical** or**seasonal fluctuations** in the business that need to be taken care of while building the samples. E.g. Diwali sales, economic ups and downs, etc.
2. The sample should be **representative of the population** on which the model will be applied in the future.
3. For **rare events samples**, the sample should be balanced before it is used for modelling

**Nuances of Logistic Regression - Segmentation**

**Nuances of Logistic Regression - Variable Transformation-I**

Dummy variable creation, standardising scales of continuous variables, etc. These processes are generally referred to as variable transformation

There are some pros and cons of transforming variables to dummies. Creating dummies for **categorical variables** is very straightforward. You can directly create n-1 new variables from an existing categorical variable if it has n levels. But for **continuous variables**, you would be required to do some kind of EDA analysis for binning the variables.

The**major advantage**offered by dummies especially for continuous variables is that they make the model stable. In other words, small variations in the variables would not have a very big impact on a model that was made using dummies, but they would still have a sizeable impact on a model built using continuous variables as is.

On the other side, there are some **major disadvantages** that exist. E.g. if you change the continuous variable to dummies, all the data will be compressed into very few categories and that might result in **data clumping**.

**Nuances of Logistic Regression - Variable Transformation-II**

So, creating dummy variables is one way of transforming variables. Let’s now move on to another technique commonly used for transforming variables — **Weight of evidence (WOE) analysis**.

To summarise, you learnt three important things in this lecture:

1. Calculating woe values for fine binning and coarse binning
2. The importance of woe for fine binning and coarse binning
3. The usage of woe transformation

**WOE** can be calculated using the following equation:

 WOE = ln(good in the bucketTotal  Good )−ln(bad in the bucketTotal  bad )

Or, it can be expressed as:

WOE=ln(PercentageofGoodPercentageofBad)

There are two main advantages of WOE:

1. WOE reflects group identity: This means it captures the general trend of distribution of good and bad customers. E.g. the difference between customers with 30% credit card utilisation and 45% credit card utilisation is not the same as the difference between customers with 45% credit card utilisation and customers with 60% credit card utilisation. This is captured by transforming the variable credit card utilisation using WOE.
2. WOE helps you in treating missing values logically for both types of variables — categorical and continuous. E.g. in the credit card case, if you replace the continuous variable credit card utilisation with WOE values, you would replace all categories mentioned above (0%-45%, 45% - 60%, etc.) with certain specific values, and that would include the category "missing" as well, which would also be replaced with a WOE value.

The pros and cons of a WOE transformation are similar to dummy variables.

Pros: The model becomes more stable because small changes in the continuous variables will not impact the input so much.

Cons: You may end up doing some score clumping.

So, **information value** can be calculated using the following expression:

IV=WOE∗ (Good in the bucketTotal Good−Bad in the BucketTotal Bad)

Or it can be expressed as:

IV=WOE ∗(Percentageofgoodinthebucket−Percentageofbadinthebucket)

It is an important indicator of **predictive power**.

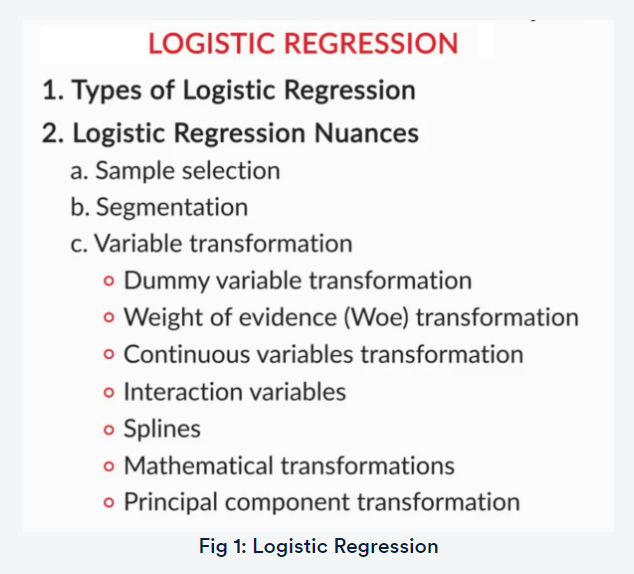
Mainly, it helps you understand how the binning of variables should be done. The binning should be done such that the WOE trend across bins is monotonic — either increasing all the time or decreasing all the time. But one more thing that needs to be taken care of is that IV (information value) should be high.

Will learn about some advanced transformation techniques such as spline transformation, interaction variables, mathematical transformation and principal component transformation. These transformations are hardly used in the model exercise.

**Comprehension  2: Missing Value -WOE**

You saw that NA values can be treated with WOE values. However, you can replace the NA bucket with a bucket which shows similar woe values.

Industry relevance of logistic regression and its applicability. Apart from its applicability, you learnt new techniques such as sample selection, segmentation and variable transformation in order to improve the model performance.



How is Logistic Regression Used in Industries?

What are your top three takeaways from this session?

1. In the logistic regression, few techniques like sample selection, segmentation and variable transformation helps in improving the model performance.
2. The selected sample to build a model should represent the population.
3. Segmentation of variables and building the child models can help in improving the model performance as compared to parent model we built without segmentation. Variable transformation using weight of evidence is a meaning full way of classification depending on the nature of data helps in better prediction which improves the model performance as compared to just dummy variable creation.

**Challenges in Logistic Regression:**

1. Low event rate: 99%vs1% or 95%vs5% of event occurring rate, example in the case of fraud in a credit card business, rare incident events, defaults etc.

We could increase the event rate by identifying the pre indicators of a fraud,heating up of an engine and multiple missed payments along with default can be used to increase the default rate.

1. Missing Value:

Imputation using woe

Imputation using median

Imputation using mean

Imputation using predictive pattern

Markov chen monte carlo imputation method

Expectation maximization imputation method

1. Truncated data:

**model evaluation**

Model evaluation measures, such as accuracy, sensitivity, specificity, KS statistic, etc. Now, let's look at some more measures commonly used for model evaluation.

Model evaluation methods are classified broadly into 3 different types:

1. Discriminatory power(default vs no default)
   * + 1. KS statistics
       2. Gini ((Area under ROC is gini) and ROC are used for evaluation)
       3. Rank ordering
       4. sensitivity
       5. specificity
2. Accuracy
3. Stability:
4. Performance stability
5. Variable stability

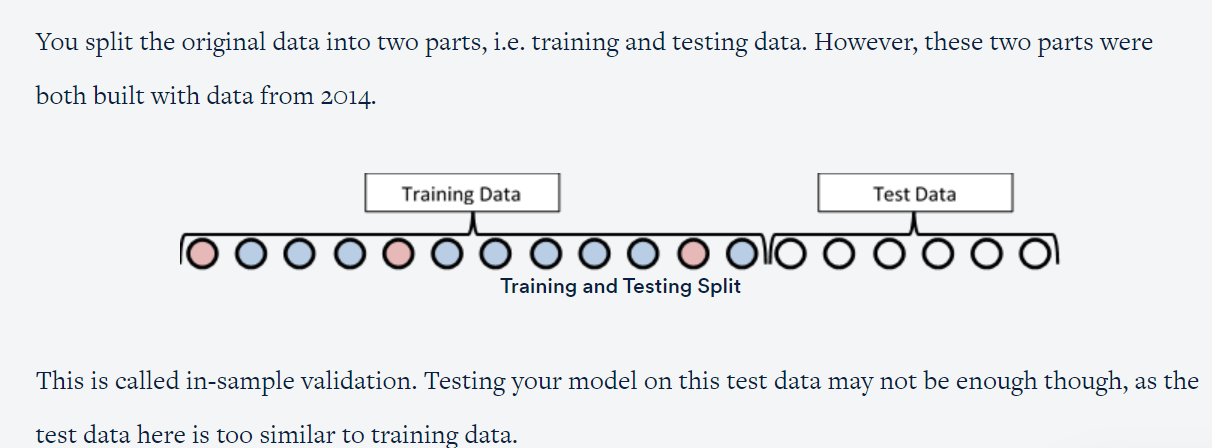
b.1 Variable distribution stability

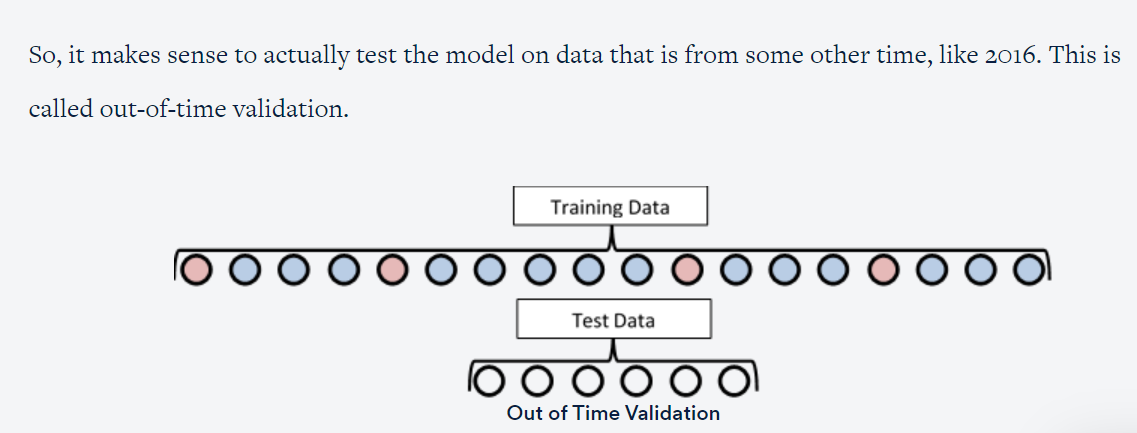
b.2 Population stability Index

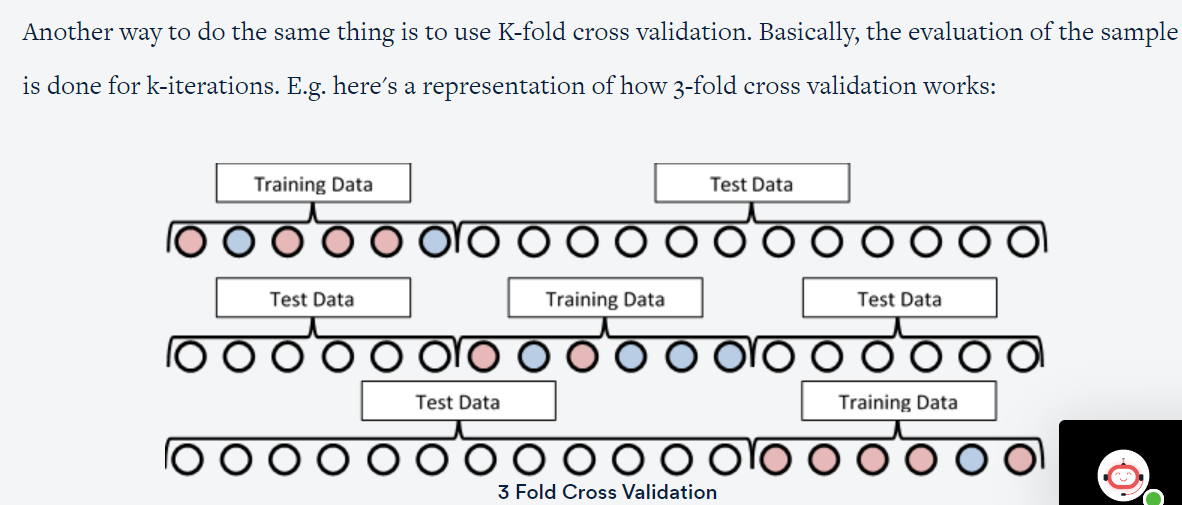
Obviously, a good model will be stable. A model is considered stable if it has:

1. **Performance Stability:** Results of in-sample validation approximately match those of out-of-time validation
2. **Variable Stability:** The sample used for model building hasn't changed too much and has the same general characteristics

**Model Validation** and Importance of Stability:







**Model Tracking or Model Governance:**

Once the model is build evaluated with metrics like sensitivity,specificity,accuracy,ginicoefficient, roc curve etc and validated with in sample, out sample or k-fold and if the model comes out to be stable then its good to be deployed in production for decision making.

However, there is a need for tracking the model to know if the model is still good, thus the process of tracking over time is called model tracking.

* + Check for variable distribution
  + Check for variable stability
  + Performance stability

**Lost tracking?**

1. Recalibrate by rebuilding the model using the same variables as in the original.
2. Recalibration of output with overlay keeping weights of the coefficients same and same variables
3. Rebuilt the model
4. While building a model we need to consider simple nuances like sample selection, segmentation, variable transformation etc which helps us with better predictive power, accuracy, performance and stability over time.
5. Once evaluated , we need to validate the model for in sample , out sample or k-cross validation steps to understand the model performance, accuracy and stability. If this seems to be alright we can go for deploy in the production to make the decisions.
6. However there is still need for tracking the model over time as the model may degrade for that we need to recalibrate the coefficients or rebuild the model with same original coefficients with new sample data.

While building a model we need to consider simple nuances like sample selection, segmentation, variable transformation, helps in better predictive power, accuracy, performance and stability over time. we need to validate the model for in sample, out sample or k-cross validation steps to know model performance, accuracy and stability. If this seems to be alright we can go for deploy in the production to make the decisions. However there is still need for tracking the model over time as it may degrade, need to recalibrate the coefficients or rebuild the model with original variables with new sample data.

# Subjective Questions - I

The following questions related to logistic regression are asked quite frequently in interviews.

arrow

**Q1. What is a logistic function? What is the range of values of a logistic function?**

The logistic function is as defined below:

                                                                           f(z)=1(1+e−z)

The values of a logistic function will range from 0 to 1. The values of Z will vary from −∞ to +∞.

arrow

**Q2. Why is logistic regression very popular/widely used?**

Logistic regression is famous because it can convert the values of logits (log-odds), which can range from −∞ to +∞ to a range between 0 and 1. As logistic functions output the probability of occurrence of an event, it can be applied to many real-life scenarios. It is for this reason that the logistic regression model is very popular. Another reason why logistic fairs in comparison to linear regression is that it is able to handle the categorical variables.

arrow

**Q3. What is the formula for the logistic regression function?**

In general, the formula for logistic regression is given by the following expression:

                                                      f(z)=1(1+e−(β0+β1X1+β2X2+….+βkXk))

arrow

**Q4. How can the probability of a logistic regression model be expressed as a conditional probability?**

The conditional probability can be given as:

                           P(Discrete value of target variable|X1,X2,X3….Xk)

It is the probability of the target variable to take up a discrete value (either 0 or 1 in case of binary classification problems) when the values of independent variables are given. For example, the probability an employee will attrite (target variable) given his attributes such as his age, salary, KRA’s, etc.

arrow

**Q5. What are odds?**

It is the ratio of the probability of an event occurring to the probability of the event not occurring. For example, let’s assume that the probability of winning a lottery is 0.01. Then, the probability of not winning is 1 - 0.01 = 0.99.

Now, as per the definition,

The odds of winning the lottery = (Probability of winning)/(Probability of not winning)

The odds of winning the lottery = 0.01/0.99

Hence, the odds of winning the lottery is 1 to 99, and the odds of not winning the lottery is 99 to 1

arrow

**Q6. Why can’t linear regression be used in place of logistic regression for binary classification?**

The reasons why linear regressions cannot be used in case of binary classification are as follows:

Distribution of error terms: The distribution of data in the case of linear and logistic regression is different. Linear regression assumes that error terms are normally distributed. In the case of binary classification, this assumption does not hold true.

Model output: In linear regression, the output is continuous. In the case of binary classification, an output of a continuous value does not make sense. For binary classification problems, linear regression may predict values that can go beyond 0 and 1. If we want the output in the form of probabilities, which can be mapped to two different classes, then its range should be restricted to 0 and 1. As the logistic regression model can output probabilities with logistic/sigmoid function, it is preferred over linear regression.

Variance of Residual errors: Linear regression assumes that the variance of random errors is constant. This assumption is also violated in the case of logistic regression

arrow

**Q7. What is the likelihood function?**

The likelihood function is the joint probability of observing the data. For example, let’s assume that a coin is tossed 100 times and you want to know the probability of getting 60 heads from the tosses. This example follows the binomial distribution formula.

**p = Probability of heads from a single coin toss**

**n = 100 (the number of coin tosses)**

**x = 60 (the number of heads – success)**

**n - x = 40 (the number of tails)**

**Pr (X=60 | n = 100, p)**

The likelihood function is the probability that the number of heads received is 60 in a trail of 100 coin tosses, where the probability of heads received in each coin toss is p. Here the coin toss result follows a binomial distribution.

This can be reframed as follows:

**Pr(X=60|n=100, p) =**c×p60×(1−p)100−60

**c = constant**

**p = unknown parameter**

The likelihood function gives the probability of observing the results using unknown parameters.

arrow

**Q8. What are the outputs of the logistic model and the logistic function?**

The logistic model outputs the logits, i.e. log odds; and the logistic function outputs the probabilities.

Logistic model=β0+β1X1+β2X2+β3X3+...+βnXn

The output of the same will be logits.

Logistic function=f(z)=1(1+e−(β0+β1X1+β2X2+β3X3+...+βnXn))

The output, in this case, will be the probabilities

arrow

**Q9. How to interpret the results of a logistic regression model? Or, what are the meanings of the different betas in a logistic regression model?**

β0 is the baseline in a logistic regression model. It is the log odds for an instance when all the attributes (X1,X2,X3,...,Xn) are zero. In practical scenarios, the probability of all the attributes being zero is very low. In another interpretation, β0 is the log odds for an instance when none of the attributes is taken into consideration.

All the other Betas are the values by which the log odds change by a unit change in a particular attribute by keeping all other attributes fixed or unchanged (control variables).

arrow

**Q10.  What is odds ratio?**

Odds ratio is the ratio of odds between two groups. For example, let’s assume that you are trying to ascertain the effectiveness of a medicine. You administered this medicine to the ‘intervention’ group and a placebo to the ‘control’ group.

                                           Odds Ratio (OR)=Odds of the Intervention GroupOdds of the Control Group

**Interpretation**

* If odds ratio = 1, then there is no difference between the intervention group and the control group.
* If the odds ratio is greater than 1, then the odds of the intervention group is greater than the odds of the control group.
* If the odds ratio is less than 1, then the odds of the intervention group is smaller than the odds of the control group.

arrow

**Q11. What is the formula for calculating odds ratio?**

The formula can be given as:

ORX1,X0=e∑i=1 to K . βi(X1i−X0i)

In the formula above, X1 and X0 stand for two different groups for which the odds ratio needs to be calculated. X1i stands for the instance ‘i’ in group X1. Xoi stands for the instance ‘i’ in group X0.β0 stands for the coefficient of the logistic regression model. Note that the baseline is not included in this formula.

# Subjective Questions - II

The following questions are related to the maximum likelihood estimator that are asked frequently in interviews.

arrow

**Q1. What is the Maximum Likelihood Estimator (MLE)?**

The MLE chooses those sets of unknown parameters (estimator) that maximise the likelihood function. The method to find the MLE is to use calculus and setting the derivative of the logistic function with respect to an unknown parameter to zero, and solving it will give the MLE. For a binomial model, this will be easy, but for a logistic model, the calculations are complex. Computer programs are used for deriving MLE for logistic models.

(Here’s another approach to answering the question.)

MLE is a statistical approach to estimate the parameters of a mathematical model. MLE and ordinary square estimation give the same results for linear regression if the dependent variable is assumed to be normally distributed. MLE does not assume anything about independent variables.

arrow

**Q2. What are the different methods of MLE and when is each method preferred?**

In the case of logistic regression, there are two approaches to MLE. They are conditional and unconditional methods. Conditional and unconditional methods are algorithms that use different likelihood functions. The unconditional formula employs the joint probability of positives (for example, churn) and negatives (for example, non-churn). The conditional formula is the ratio of the probability of observed data to the probability of all possible configurations.

The unconditional method is preferred if the number of parameters is lower compared to the number of instances. If the number of parameters is high compared to the number of instances, then conditional MLE is to be preferred. Statisticians suggest that conditional MLE is to be used when in doubt. Conditional MLE will always provide unbiased results.

arrow

**Q3. What are the advantages and disadvantages of conditional and unconditional methods of MLE?**

Conditional methods do not estimate unwanted parameters. Unconditional methods estimate the values of unwanted parameters also. Unconditional formulas can directly be developed with joint probabilities. This cannot be done with conditional probability. If the number of parameters is high relative to the number of instances, then the unconditional method will give biased results. Conditional results will be unbiased in such cases.

arrow

**Q4. What is the output of a standard MLE program?**

The output of a standard MLE program is as follows:

**Maximised likelihood value:** This is the numerical value obtained by replacing the unknown parameter values in the likelihood function with the MLE parameter estimator.

**Estimated variance-covariance matrix:** The diagonal of this matrix consists of the estimated variances of the ML estimates. The off-diagonal consists of the covariances of the pairs of the ML estimates

arrow

**Q5. Why can’t we use Mean Square Error (MSE) as a cost function for logistic regression?**

In logistic regression, we use the sigmoid function and perform a non-linear transformation to obtain the probabilities. Squaring this non-linear transformation will lead to non-convexity with local minimums. Finding the global minimum in such cases using gradient descent is not possible. Due to this reason, MSE is not suitable for logistic regression. Cross-entropy or log loss is used as a cost function for logistic regression. In the cost function for logistic regression, the confident wrong predictions are penalised heavily. The confident right predictions are rewarded less. By optimising this cost function, convergence is achieved.

# Subjective Questions - III

The following interview questions are based on the evaluation metrics used in logistic regression.

arrow

**Q1. What is accuracy?**

Accuracy is the number of correct predictions out of all predictions made.

Accuracy=True Positives+True NegativesTotal Number of Predictions

arrow

**Q2. Why is accuracy not a good measure for classification problems?**

Accuracy is not a good measure for classification problems because it gives equal importance to both true positives and true negatives. However, this may not be the case in most business problems. For example, in the case of cancer prediction, declaring cancer as benign is more serious than wrongly informing the patient that he is suffering from cancer. Accuracy gives equal importance to both cases and cannot differentiate between them.

arrow

**Q3. What is the importance of a baseline in a classification problem?**

Most classification problems deal with imbalanced datasets. Examples include telecom churn, employee attrition, cancer prediction, fraud detection, online advertisement targeting, and so on. In all these problems, the number of the positive classes will be very low when compared to the negative classes. In some cases, it is common to have positive classes that are less than 1% of the total sample. In such cases, an accuracy of 99% may sound very good but, in reality, it may not be.

Here, the negatives are 99%, and hence, the baseline will remain the same. If the algorithms predict all the instances as negative, then also the accuracy will be 99%. In this case, all the positives will be predicted wrongly, which is very important for any business. Even though all the positives are predicted wrongly, an accuracy of 99% is achieved. So, the baseline is very important, and the algorithm needs to be evaluated relative to the baseline.

arrow

**Q4. What are false positives and false negatives?**

False positives are those cases in which the negatives are wrongly predicted as positives. For example, predicting that a customer will churn when, in fact, he is not churning.

False negatives are those cases in which the positives are wrongly predicted as negatives. For example, predicting that a customer will not churn when, in fact, he churns.

arrow

**Q5. What are the true positive rate (TPR), true negative rate (TNR), false positive rate (FPR), and false negative rate (FNR)?**

TPR refers to the ratio of correctly classified positives to the total number of positive instances in the data.

TPR=TPTP+FN

TNR refers to the ratio of correctly classified negatives to the total number of negative instances in the data.

TNR=TNTN+FP

FPR refers to the ratio of negatives in the data which are incorrectly classified as positives to the total number of negative instances in the data.

FPR=FPTN+FP

FNR refers to the ratio of positive instances in the data which are incorrectly classified as negatives to the total number of positive instances in the data.

FNR=FNTP+FN

arrow

**Q6. What are sensitivity and specificity?**

Specificity is the same as true negative rate, or it is equal to 1 – false positive rate. It tells you out of all the actual '0' labels, how many were correctly predicted.

Specificity=TNTN+FP

Sensitivity is the true positive rate. It tells you out of all the actual '1' labels, how many were correctly predicted.

Sensitivity=TPTP+FN

arrow

**Q7. What are precision and recall?**

Precision is the proportion of true positives out of predicted positives. To put it in another way, it is the accuracy of the prediction. It is also known as the ‘positive predictive value’.

Precision=TPTP+FP

Recall is the same as the true positive rate (TPR) or the sensitivity.

Recall=TPTP+FN

arrow

**Q8. What is F-measure?**

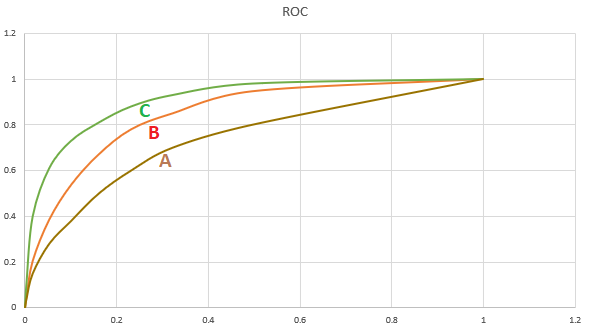
It is the harmonic mean of precision and recall. In some cases, there will be a trade-off between the precision and the recall. In such cases, the F-measure will drop. It will be high when both the precision and the recall are high. Depending on the business case at hand and the goal of data analytics, an appropriate metric should be selected.

F−measure=2×Precision×RecallPrecision+Recall

arrow

**Q9. Explain the use of ROC curves and the AUC of an ROC Curve.**

An ROC (Receiver Operating Characteristic) curve illustrates the performance of a binary classification model. It is basically a TPR versus FPR (true positive rate versus false positive rate) curve for all the threshold values ranging from 0 to 1. In an ROC curve, each point in the ROC space will be associated with a different confusion matrix. A diagonal line from the bottom-left to the top-right on the ROC graph represents random guessing. The Area Under the Curve (AUC) signifies how good the classifier model is. If the value for AUC is high (near 1), then the model is working satisfactorily, whereas if the value is low (around 0.5), then the model is not working properly and just guessing randomly. From the image below, curve C (green) is the best ROC curve among the three and curve A (brown) is the worst ROC curve among the three.



arrow

**Q10. How to choose a cutoff point in case of a logistic regression model?**

The cutoff point depends on the business objective. Depending on the goals of your business, the cutoff point needs to be selected. For example, let’s consider loan defaults. If the business objective is to reduce the loss, then the specificity needs to be high. If the aim is to increase the profits, then it is an entirely different matter. It may not be the case that profits will increase by avoiding giving loans to all predicted default cases. But it may be the case that the business has to disburse loans to default cases that are slightly less risky to increase the profits. In such a case, a different cutoff point, which maximises profit, will be required. In most of the instances, businesses will operate around many constraints. The cutoff point that satisfies the business objective will not be the same with and without limitations. The cutoff point needs to be selected considering all these points. If the business context doesn't matter much and you want to create a balanced model, then you use an ROC curve to see the tradeoff between sensitivity and specificity and accordingly choose an optimal cutoff point where both these values along with accuracy are decent.