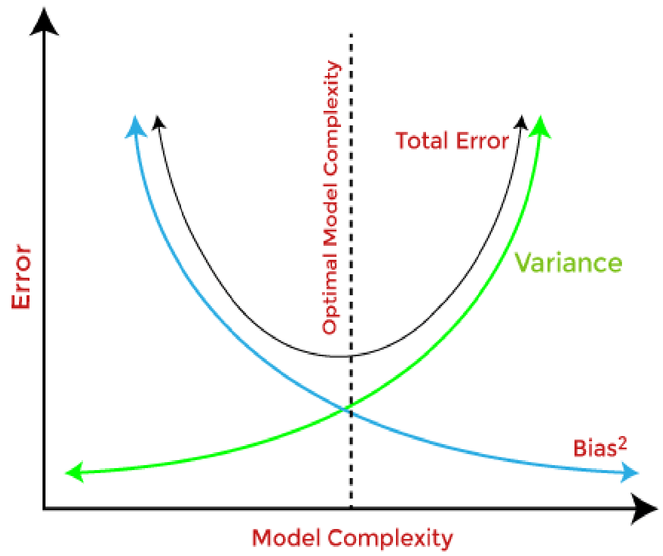
CMR INSTITUTE OF TECHNOLOGY	USN											CMRIT

Internal Assesment Test - II

Sub:	**Exploratory Data Analysis for Business**							Code:	**22MBABA304**
Date:	**07-03-2024**	Duration:	**90 mins**	Max Marks:	**50**	Sem:	**III**	Branch:	**MBA**

SET- III – Answer Key

		Marks	OBE CO	OBE RBT
	Part A - Answer Any Two Full Questions (2* 20 = 40 marks)			
1 (a)	**Define the term 'Overfitting'.** A statistical model is said to be overfitted when the model does not make accurate predictions on testing data. When a model gets trained with so much data, it starts learning from the noise and inaccurate data entries in our data set. And when testing with test data results in High variance. Then the model does not categorize the data correctly, because of too many details and noise. The causes of overfitting are the non-parametric and non-linear methods because these types of machine learning algorithms have more freedom in building the model based on the dataset and therefore they can really build unrealistic models. A solution to avoid overfitting is using a linear algorithm if we have linear data or using the parameters like the maximal depth if we are using decision trees. In a nutshell, Overfitting is a problem where the evaluation of machine learning algorithms on training data is different from unseen data.	[03]	CO3	L1
(b)	**Discuss the importance of linear regression.** Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between a dependent variable and one or more independent features. When the number of the independent feature, is 1 then it is known as Univariate Linear regression, and in the case of more than one feature, it is known as multivariate linear regression. The interpretability of linear regression is a notable strength. The model's equation provides clear coefficients that elucidate the impact of each independent variable on the dependent variable, facilitating a deeper understanding of the underlying dynamics. Its simplicity is a virtue, as linear regression is transparent, easy to implement, and serves as a foundational concept for more complex algorithms. Linear regression is not merely a predictive tool; it forms the basis for various advanced models. Techniques like regularization and support vector machines draw inspiration from linear regression, expanding its utility. Additionally, linear regression is a cornerstone in assumption testing, enabling researchers to validate key assumptions about the data.	[07]	CO3	L2
(c)	**Analyze the concept of bias-variance trade-off in the context of linear regression.** **Bias-Variance Trade-Off** While building the machine learning model, it is really important to take care of bias and variance in order to avoid overfitting and underfitting in the model. If the model is very simple with fewer parameters, it may have low variance and high bias. Whereas, if the model has a large number of parameters, it will have high variance and low bias. So, it is required to make a balance between bias and variance errors, and this balance between the bias error and variance error is known as **the Bias-Variance trade-off.**	[10]	CO3	L4



For an accurate prediction of the model, algorithms need a low variance and low bias. But this is not possible because bias and variance are related to each other:

- If we decrease the variance, it will increase the bias.
- If we decrease the bias, it will increase the variance.

Bias-Variance trade-off is a central issue in supervised learning. Ideally, we need a model that accurately captures the regularities in training data and simultaneously generalizes well with the unseen dataset. Unfortunately, doing this is not possible simultaneously. Because a high variance algorithm may perform well with training data, but it may lead to overfitting to noisy data. Whereas, high bias algorithm generates a much simple model that may not even capture important regularities in the data. So, we need to find a sweet spot between bias and variance to make an optimal model.

Hence, the ***Bias-Variance trade-off is about finding the sweet spot to make a balance between bias and variance errors.***

2 (a) **Define L1 Regularization.** [03] CO3 L1

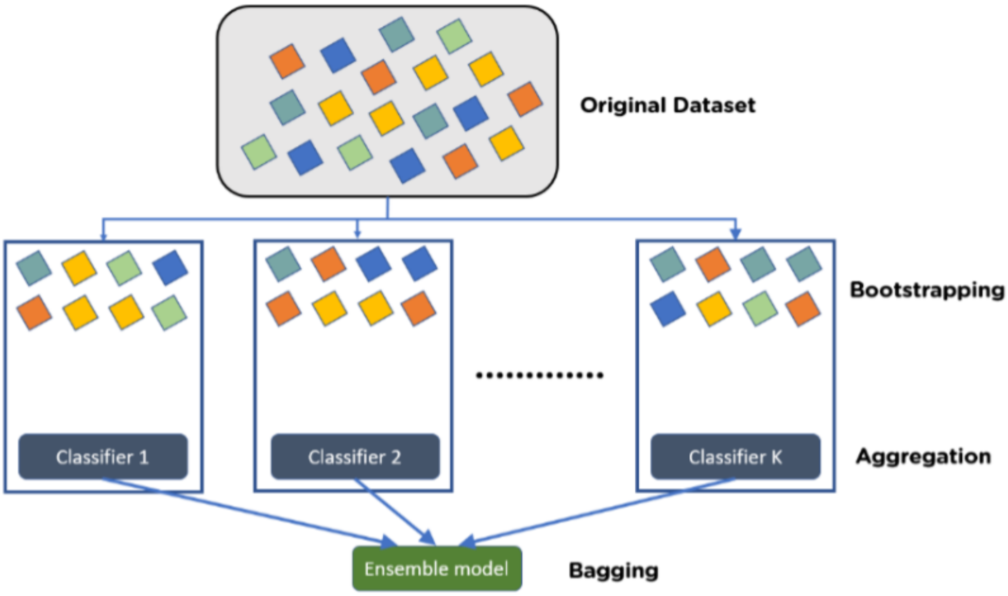
A regression model which uses the **L1 Regularization** technique is called **LASSO(Least Absolute Shrinkage and Selection Operator)** regression. **Lasso Regression** adds the *"absolute value of magnitude"* of the coefficient as a penalty term to the loss function(L). Lasso regression also helps us achieve feature selection by penalizing the weights to approximately equal to zero if that feature does not serve any purpose in the model.

(b) **Write the concept of bagging and examine its role in improving the performance of tree-based models.** [07] CO4 L3

Bagging, also known as Bootstrap aggregating, is an ensemble learning technique that helps to improve the performance and accuracy of machine learning algorithms. It is used to deal with bias-variance trade-offs and reduces the variance of a prediction model. Bagging avoids overfitting of data and is used for both regression and classification models, specifically for decision tree algorithms.

Steps to Perform Bagging

- Consider there are n observations and m features in the training set. You need to select a random sample from the training dataset without replacement
- A subset of m features is chosen randomly to create a model using sample observations
- The feature offering the best split out of the lot is used to split the nodes
- The tree is grown, so you have the best root nodes
- The above steps are repeated n times. It aggregates the output of individual decision trees to give the best prediction



(c)	**Outline the importance regression shrinkage methods.**	[10]	CO3	L4

Ever have a question that, “*Why is Linear Regression giving me such good accuracy on the training set but a low accuracy on the test set in spite of adding all the available dependent features to the model*?”

The question above seems inexplicable to many people but is answered by a concept called **overfitting** in which your model, in addition to learning the data, also learns the noise present in it. Hence learning the training points, a bit too perfectly.

How do you solve it?

This is where **shrinkage methods** (also known as regularization) come in play. These methods apply a penalty term to the Loss function used in the model. Minimizing the loss function is equal to maximizing the accuracy. To understand this better, we need to go into the depths of Loss function in **Linear Regression**.

Linear Regression uses Least Squares to calculate the minimum error between the actual values and the predicted values. The aim is to minimize the squared difference between the actual and predicted values to draw the best possible regression curve for the best prediction accuracy.

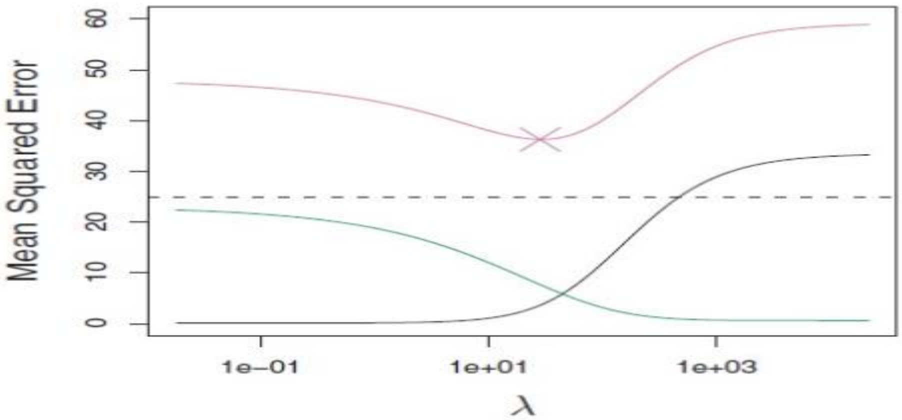
Now, what does shrinking do?

Shrinking the coefficient estimates significantly reduces their variance. When we perform shrinking, we essentially bring the coefficient estimates closer to 0.

The need for shrinkage method arises due to the issues of underfitting or overfitting the data. When we want to minimize the mean error (Mean Squared Error(MSE) in case of Linear Regression), we need to optimize the bias-variance trade-off.

What is this bias-variance trade-off?

The bias-variance trade-off indicates the level of underfitting or overfitting of the data with respect to the Linear Regression model applied to it. A high bias-low variance means the model is underfitted and a low bias-high variance means that the model is overfitted. We need to trade-off between bias and variance to achieve the perfect combination for the minimum Mean Squared Error as shown by the graph below.



In this figure, the green curve is variance, the black curve is squared bias and the purple curve is the MSE. Lambda is the regularization parameter which will be covered later.

How do we use shrinking methods?

The best known shrinking methods are **Ridge Regression** and **Lasso Regression** which are often used in place of Linear Regression.

Ridge Regression, like Linear Regression, aims to minimize the Residual Sum of Squares(RSS) but with a slight change. As we know, Linear Regression estimates the coefficients using the values that minimize the following equation:

$$\mathrm{RSS} = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2.$$

Ridge Regression adds a penalty term to this, lambda, to shrink the coefficients to 0 :

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = \mathrm{RSS} + \lambda\sum_{j=1}^{p}\beta_j^2,$$

Ridge Regression's advantage over Linear Regression is that it capitalizes on the bias-variance trade-off. As λ increases, the coefficients shrink more towards 0.

Ridge Regression has a major disadvantage that it includes all p predictors in the output model regardless of the value of their coefficients which can be challenging for a model with huge number of features. This disadvantage is overcome by Lasso Regression which performs variable selection. Lasso Regression uses L-1 penalty as compared to Ridge Regression's L-2 penalty which instead of squaring the coefficient, takes its absolute value as shown below :

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = \mathrm{RSS} + \lambda\sum_{j=1}^{p}|\beta_j|.$$

Ridge Regression brings the value of coefficients close to 0 whereas Lasso Regression forces some of the coefficient values to be exactly equal to 0. It is important to optimize the value of λ in Lasso Regression as well to reduce the MSE error.

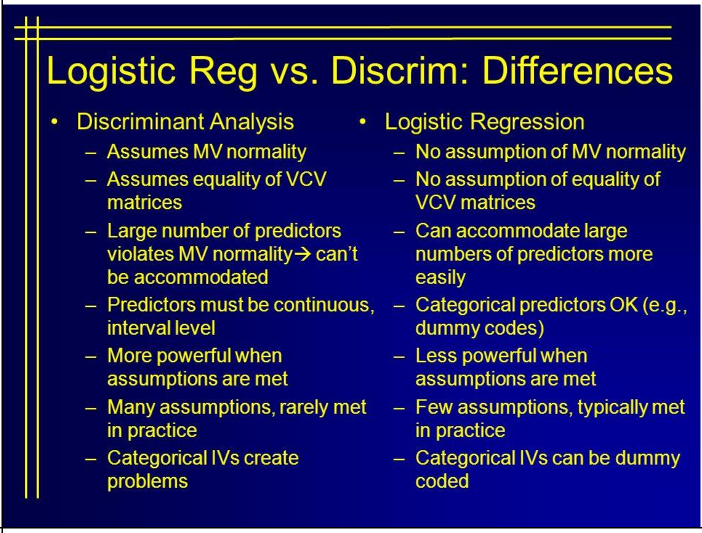
3 (a)	**Define logistics regression.** Logistic regression is a supervised machine learning algorithm that accomplishes binary classification tasks by predicting the probability of an outcome, event, or observation. The model delivers a binary or dichotomous outcome limited to two possible outcomes: yes/no, 0/1, or true/false. Logical regression analyzes the relationship between one or more independent variables and classifies data into discrete classes. It is extensively used in predictive modeling, where the model estimates the mathematical probability of whether an instance belongs to a specific category or not. For example, 0 – represents a negative class; 1 – represents a positive class. Logistic regression is commonly used in binary classification problems where the outcome variable reveals either of the two categories (0 and 1).	[03]	CO4	L1
(b)	**Analyze the process of constructing a multiclass SVM.**	[07]	CO4	L4

In its most basic type, SVM doesn't support multiclass classification. For multiclass classification, the same principle is utilized after breaking down the multi-classification problem into smaller subproblems, all of which are binary classification problems.

The popular methods that are used to perform multi-classification on the problem statements using SVM (multiclass support vector machines) are as follows:

- One vs One (OVO) approach
- Directed Acyclic Graph (DAG) approach
- One vs All (OVA) approach

(c) **Compare and contrast discriminant analysis with logistic regression for classification.** [10] CO4 L2



Part B - Compulsory (01*10=10 marks) – CASE STUDY

4 **Critically evaluate the advantages and limitations of Support Vector Machines in real-world scenarios.** [10] CO4 L5

Advantages

- Performs well at classifying non-linear data
- Optimizing margins can help reduce the overfitting of data and allow for capacity control
- Learning without a local minima
- There are many kernels (transformations) that could be used to fit the data unlike any other algorithm
- Often provides sparse solutions
- Performs well on data sets that have many attributes, even if there are relatively very few cases on which to train the model

Limitations

- Choice of the right Kernel
  - The number of possible kernels is infinite and can make it hard to choose the right one
  - Most software uses a few kernels that generalize to many situations, but no kernel generalizes to every situation
- Can be computationally intensive
  - Algorithms can be complex
- Can overfit the model

Course Outcomes (Cos)		PO1	PO2	PO3	PO4	PO5	PSO1	PSO2	PSO3	PSO4
CO1:	Understand data mining and its importance.									
CO2:	Apply knowledge of research design for business problems.									
CO3:	Analyse the cause and effect relationship between the variables from the analysis.				1a, 1b,		1c, 2a		2c	
CO4:	Evaluate regression and decision tree based method to solve business problems.		2b ,3a		3b			3c		4

*****