Short Essay Questions

Question 1 Solution:

Data Driven:

Data Driven can be used in organizations to make decisions based on the datasets. Data Driven strategy effectively helps to identify what is going on in the organization and how to improve the day to day operations within organization. Data Driven strategy helps organizations to lower their cost and increase the revenue as intelligent decisions are made by analyzing the data.

Data Driven Making Process Stages:

The following are stages for the overall implementation on data driven strategies in the organization:

- **Identify business objectives:**

Every organization has their own set of goals which may be short term or long term based on their strategies like increase of revenue or increase of sale of items in large volume. The goals help to make strategies to identify trends in the data and also help in manipulation or mining of data as well to extract useful information.

- **Data preparation:**

Data Entry operators are usually hired in organizations to keep track of different records. Software are used in organizations to store the sales data as well. These datasets are combined so that they can help to identify relevant trends in the data. Data is kept cleaned and missing values are imputed before analysis.

- **Data Analysis and Visualizations:**

Data is analyzed by using python or R programming languages etc. They can help to manipulate the data to identify insights in the data. For visualization purposes most of the organizations use Tableau or Microsoft Power BI software. Various types of Maps, Graphs and Charts can be made to get information from the raw data.

- **Advanced Insights:**

Various statistical tests can are applied to the datasets to identify the advanced insights in the datasets. Null hypothesis and Alternate hypothesis are created to find answer for the research questions developed to identify the trends. Various Machine Learning approaches are also applied to the datasets like supervised machine learning or unsupervised machine learning.

- **Data Driven Decisions:**

When all the data mining steps are finished then it is a time to implement the decisions based on the insights. Recommendations are given to the Top

Management by the Data Scientists and it is their responsibility to implement new strategies in the organization. They may be recommendation about the marketing of specific products or to offer sales to the targeted customers.

Question 2 Solution:

Theory of Change vs Logic Model

Organizations use approaches like Theories of Change, Logic Models, Log Frames and many others based on the project or problem they want to solve. We can understand advantages and disadvantages of each problem to help organizations to decide which approach is best for organization.

Theory of Change:

Theory of Change are flow charts, diagrams or description of the changes in a process that we like to see. It starts from the initial state, goes in several processes and finally goes to its end state with the desired result which we would like to see.

Logic Model:

Logic Model is designed as a table which follow steps like a flowchart. It describes the project needs, how the project goes through the stages and the final result we want to achieve.

Differences between Theory of Change or Logic Model:

- Theory of Change takes more time than Logic Model as it is more complex and requires more resources.

- Theory of change capture more complexities of relationship than Logic Model. If we are working in a project where internal and external factors are considered for final stage then Theory of Change is better than Logic Model as Logic Model follows a straightforward path.

- A Theory of Change works well when the project is in early stages of development. It can do backward mapping, identify the more better results and then move back. Logic Model can't suggest the changes going in the process. Logic Models are better when the requirements are pretty clear.

- A Theory of Change display overall understanding of flow which is not present in Logic Model.

- For long term projects, Theory of Change is preferable while for short term projects, Logic Model is preferred.

- A Theory of Change explain why a change will occur while Logic Model explains what will occur.

- Theory of Change shows the big picture with all possible pathways to reach the end goal while Logic Model shows just the pathway that our program deals with.

- Theory of Change is massive and complex while Logic Model is neat and tidy.

- Theory of Change can capture those pathways which are even not related to our program like complex social, economic, political and institutional processes. Logic Model is linear and all activities leads to output and there are no cyclic processes or feedback loops.

Question 3 Solution:

Identify the two-step-based approach for parametric models:

Parametric Models are a learning models that summarizes data with a set of fixed size like independent of the number of training examples. It doesn't matter how much data is given to parametric model and it will not change its mind about how many parameters need. Parametric models make large assumptions about the mapping of the input variables to the output variable and in turn are faster to train, require less data but may not be as powerful.

The algorithms involve two steps:

1. Select a form for the function.
2. Learn the coefficients for the function from the training data.

$$b0 + b1*x1 + b2*x2 = 0$$

Where b0, b1 and b2 are the coefficients of the line that control the intercept and slope, and x1 and x2 are two input variables.

It is assumed that the functional form of a line greatly simplifies the learning process. Now, all we need to do is estimate the coefficients of the line equation and we have a predictive model for the problem. Often the assumed functional form is a linear combination of the input variables and as such parametric machine learning algorithms are often also called "linear machine learning algorithms". The problem in parametric models is, the actual unknown underlying function may not be a linear function like a line. It could be almost a line and require some little transformation of the input data to work right. Or it could be nothing like a line in which case the assumption is wrong and the approach will produce poor results. Some more examples of parametric machine learning algorithms include Regression, Linear Discriminant Analysis, Perceptron, Naive Bayes, Simple Neural Networks, Linear Support Vector Machines and Logistic Regression.

Benefits of two steps Parametric Machine Learning Algorithms are that these methods are easier to understand and interpret results. Parametric models are very fast to learn from data. They do not require as much training data and can work well even if the fit to the data is not perfect.

Limitations of two step Parametric Machine Learning Algorithms are that by choosing a functional form these methods are highly constrained to the specified form. Parametric Models are more suited to simpler problems. In practice these models are unlikely to match the underlying mapping function

Question 4 Solution:

Bias:

Bias is the difference between the expected value of an estimator and the true value being estimated. Bias error occur when we simplify our model too much so that target functions are easily approximate. If we resample the data it can affect bias. If the average prediction values are significantly different from the true value based on the sample data, the model has a high level of bias. Every model starts with some level of bias, because bias results from assumptions in the model that make the target function easier to capture. Bias occur due to under-fitting . Linear algorithms often has high bias while non-linear models have low bias.

Variance:

Variance measures how much a random variable differs from its expected value. Variance is based on a single training set. Variance measures the inconsistency of different predictions using different training sets and it is not a measure of overall accuracy related to the model. Variance can cause over-fitting in the model, it will show random noise in training data instead of the target function. Variance shows the inconsistency in the predicted values with respect to each other. Variance can lead to over-fitting

Goal for bias-variance trade-off:

Bias and variance are components of reducible error. The trade-off is well known: increasing bias decreases variance, and increasing variance decreases bia Let's suppose some cases like high-bias & high-variance, high-bias & low-variance, high-variance & low-bias and low-bias & low-variance.

- **High Bias & High Variance:**

If a model predicts values which are significantly different from the actual values then we can say that model is highly biased and if model predicts values which shows too much inconsistency in predicted values with each other then the model has high variance. These models have very bad performance.

- **High Bias & Low Variance:**

If a model predicts values which are significantly different from the actual values then we can say that model is highly biased and if model predicts values which shows consistency in predicted values with each other then the model has low variance. These models also have very bad performance.

- **High Variance and Low Bias:**

  If model predicts values which shows too much inconsistency in predicted values with each other then the model has high variance If a model predicts values which are similar to the actual values then we can say that model is low biased. These models have little better performance.

- **Low Variance and Low Bias:**

  if model predicts values which shows consistency in predicted values with each other then the model has low variance If a model predicts values which are similar to the actual values then we can say that model is low biased. These models also have better performance among all bias-variance trade-off. This is the ideal approach for any model.

Question 4:

A) What is the meaning of Test MSE?

Test MSE is the average of all the squared differences between the actual values and the estimated values of the testing set by the model.

B) What is the meaning of Training MSE?

Training MSE is the average of all the squared differences between the actual values and the estimated values of the training set by the model.

C) What means small MSE in our simulations?

Small Mean Squared Error means that the average of all the squared differences between the actual values and the estimated values of the training set and testing set is small by the model in our simulation. It means that our model predicts values nearly equal to the actual values.

D) What represent the horizontal dotted line?

Horizontal dotted line in the simulation represents the moderate value for bias-variance trade-off. It means model performs well when both test mse and training mse are close to the dotted horizontal line.

E) Provide an interpretation of the two simulation figures.

- **Interpretation of Simulation of figure 1:**

In figure 1, at initial stage, Training mse starts at 1.5 with flexibility = 2 and test mse starts at 2.1 with flexibility = 2. Training mse and Test mse have best bias-variance trade-off at flexibility = 5 and then Training mse decreases to 0.1 with flexibility over 20 and Testing mse increases to 2.5 with flexibility over 20.

- **Interpretation of Simulation of figure 2:**

In figure 2, at initial stage, Training mse starts at 1 with flexibility = 2 and test mse starts at 1.1 with flexibility = 2. Training mse and Test mse have best bias-variance trade-off at flexibility = 3.5 and then Training mse decreases to 0.0 with flexibility over 20 and Testing mse increases to 2.4 with flexibility over 20.

Question 5:

a. How is the green curve performance in terms of variance level?

Green curve has high variance level as compared to blue curve and orange curve.

b. b. How is the orange curve performance in terms of variance level?

Orange curve has low variance as compared to green curve and blue curve.