TOPSTECHNOLOGIES
Training | Outsourcing | Placement | Study Abroad

Data Science Assignments

MODULE-1: Python

1. Write a python program to sum of the first n positive integers.
2. Write a Python program to count occurrences of a substring in a string.
3. Write a Python program to count the occurrences of each word in a given sentence.
4. Write a Python program to get a single string from two given strings, separated by a space and swap the first two characters of each string.
5. Write a Python program to add 'ing' at the end of a given string (length should be at least 3).
   If the given string already ends with 'ing' then add 'ly' instead If the string length of the given string is less than 3, leave it unchanged
6. Write a Python program to find the first appearance of the substring 'not' and 'poor' from a given string, if 'not' follows the 'poor', replace the whole 'not'...'poor' substring with 'good'.
   Return the resulting string
7. Program to find Greatest Common Divisor of two numbers.
   For example, the GCD of 20 and 28 is 4 and GCD of 98 and 56 is 14.
8. Write a Python program to check whether a list contains a sublist.
9. Write a Python program to find the second smallest number in a list.
10. Write a Python program to get unique values from a list.
11. Write a Python program to unzip a list of tuples into individual lists.
12. Write a Python program to convert a list of tuples into a dictionary
13. Write a Python program to sort a dictionary (ascending /descending) by value.
14. Write a Python program to find the highest 3 values in a dictionary.
15. Given a number n, write a python program to make and print the list of Fibonacci series up to n.
    Input : n=7
    Hint : first 7 numbers in the series
    Expected output :
    First few Fibonacci numbers are 0, 1, 1, 2, 3, 5, 8, 13
16. Counting the frequencies in a list using a dictionary in Python.
    Input : [1, 1, 1, 5, 5, 3, 1, 3, 3, 1,4, 4, 4, 2, 2, 2, 2]
    Expected output : 1 : 5 , 2 : 4 , 3 : 3 , 4 : 3 , 5 : 2

17. Write a python program using function to find the sum of odd series and even series

    Odd series: 12/ 1! + 32/ 3! + 52/ 5!+……n

    Even series: 22/ 2! + 42/ 4! + 62/ 6!+……n

18. Python Program to Find Factorial of Number Using Recursion
19. Write a Python function that takes a list and returns a new list with unique elements of the first list.
20. Mini project :

    Problem Statement : Password Generator

    Make a program to generate a strong password using the input given by the user. To generate a password, randomly take some words from the user input and then include numbers, special characters and capital letters to generate the password. Also, keep a check that password length is more than 8 characters.

    Note: Include Exception handling wherever required. Also, make a 'User' class and store the details like user id, name and password of each user as a tuple.

MODULE-2: Excel

On the dataset as discussed by the class mentor Do as directed:

1.
   - Find Mean , Median , Mode of the desired Columns of the Data Set
   - Use Sum , SumIF, and SumIFS get the desired output from the Dataset
   - Use Count, CountA and CountIFS , Print a Table of the outputs
   - By Using VLOOKUP, HLOOKUP and XLOOKUP make as compressed dataset.
   - Perform a logical test using IF, IFS ,IFNA,INDEX Functions
   - Perform a logical test on DATE and TIME Functions
2. Create a Pivot Tables of each from the above results
3. Do the visualization of the Dataset using the Graphs of Excel.
4. Make a Productivity Dashboard of a Pvt. Ltd Company.

MODULE-3: Statistics

1. For the given data find mean, standard deviation and variance in excel.
2. From the given data, take a sample and find mean, standard deviation and variance for population in excel. Also, validate Central Limit Theorem in excel and python as discussed in the class.
3. Theory Task: Estimate mean for the problem statement discussed by your mentor.
4. Theory Task: Validate Hypothesis for the problem statement discussed by your mentor.
5. Perform One-way and two-way ANOVA on the dataset as discussed in class using formulae as well as ANOVA function in excel.
6. Perform t-test on the dataset as discussed in class using t-test function in excel.

MODULE-4: Data Analysis with Python

Project 1:

Problem Statement: Data Analysis

The dataset contains more of 10, 000 rows and more than 10 columns which contains features of the car and its (MSRP) manufacturer's suggested retail price. Clean the data and analyse it making it ready for modelling.

MODULE-5: Machine Learning with Python

Project 2:

Problem Statement: House Price Prediction

The data contains 1460 training data points and 80 features that might help to predict the selling price of a house.

Note: Save each resultant model by different name so that you can compare all of them in your final conclusion.

Do as directed

1. Perform EDA.
2. Build a Simple Linear Regression model to predict the Sale price of the house. Use Area as the independent variable
3. Build Multiple Linear Regression model to predict Sale price of the house.
4. Use dimensionality reduction technique PCA/LDA and build Multiple Linear Regression model to predict Sale price of the house.
5. Build a model using Lasso and Ridge regression to reduce model complexity.
6. Build SVR model to predict Sale price of the house.
7. Build Decision Tree Regressor to predict Sale price of the house.
8. Build Random Forest Regression model to predict Sale price of the house.
9. Use GridsearchCV and RandomizedsearchCV for tuning hyperparameters and fit your model on the optimal parameters.
10. Model Selection: Evaluate and compare performance of all the models to find the best model.

Project 3 :

Problem Statement: Predicting Customer Churn

The data is centred on customer churn, the rate at which a commercial customer will leave the commercial platform that they are currently a (paying) customer of a telecommunications company.

Do as directed

1. Perform EDA.

2. Check if there is an imbalance in data. If there is an imbalance in data, resolve it.
3. Build a Logistic Regression classification model which will predict whether a customer is at risk to churn from the platform.
4. Build Naive Bayes model which will predict whether a customer is at risk to churn from the platform.
5. Build a K-nearest classifier which will predict whether a customer is at risk to churn from the platform.
6. Find optimal parameters for the algorithm through GridSearchCV and build SVC model which will predict whether a customer is at risk to churn from the platform.
7. Find optimal parameters for the algorithm through GridSearchCV and build a Decision tree which will predict whether a customer is at risk to churn from the platform.
8. Find optimal parameters for the algorithm through RandomSearchCV and build a Random which will predict whether a customer is at risk to churn from the platform.
9. Model Selection: Evaluate and compare performance of all the models to find the best model.

Project 4 :

Problem Statement: Customer Segmentation

Customer Personality Analysis is a detailed analysis of a company's ideal customers. It helps a business to better understand its customers and makes it easier for them to modify products according to the specific needs, behaviours and concerns of different types of customer segments.

Do as directed

1. Find Customer segments using K-means clustering algorithm.
2. Find Customer segments using Hierarchical clustering algorithm.
3. Compare the Clusters and conclude your analysis.

Project 5 :

Problem Statement: Forecast number of Passengers

This dataset provides monthly totals of US airline passengers from 1949 to 1960. Analyse the data and forecast the number of passengers in the airline.

Case Study 1:

Build model based on the Recommendation System as discussed by your class mentor.

Case Study 2:

Build a model based on Reinforcement learning as discussed by your class mentor.

MODULE-6: Deep Learning

Project 6:

Problem Statement (Revisited) : Predicting Customer Churn

The data is centred on customer churn, the rate at which a commercial customer will leave the commercial platform that they are currently a (paying) customer of a telecommunications company.

Build ANN to predict Sale price of the house.

Project 7:

Problem Statement : Digit Recogniser

The MNIST dataset is an inbuilt that consists of images of digits from a variety of scanned documents. Each image is a 28X28 pixel square. In this dataset 60,000 images are used to train the model and 10,000 images are used to test the model. There are 10 digits (0 to 9) or 10 classes to predict.

Build CNN to predict the digit.

MODULE-7: Advance Deep Learning

Project 8:

Problem Statement : Sentiment Analysis

The imdb dataset is an inbuilt that consists of reviews given by the viewer.

Do as directed

1. Build RNN to Analyse the review.
2. Build LSTM to Analyse the review.
3. Build GRU to Analyse the review.
4. Model Selection: Evaluate and compare performance of all the models to find the best model suited for this application.

MODULE-8: Natural Language processing

Project 9:

Problem Statement: Build a text classifier model for emotion detection in text.

The Dataset consists of real life emotions of the people of the US, where the data is in the csv format and you have to build a model.

Project 10:

Problem Statement : Text Classification (Average Word2vec)

This is a dataset for binary sentiment classification containing substantially more data than previous benchmark datasets. We provide a set of 25,000 highly polar movie reviews for training and 25,000 for testing.

MODULE-9: Computer Vision

Project 11: Avengers Face Detection

The Dataset contains around 50 cropped face images of each avenger.

Chris Evans (Captain America)

Chris Hemsworth (Thor)

Mark Ruffalo (Hulk)

Robert DowneyJr (The Iron man)

Scarlett Johansson (Black Widow)

MODULE-10: Big Data

Project 12: FBI Data analysis of Crime using Hadoop

The Data Contains the details of the FBI Crime Branch, which is in the csv format. Analyse the data using hive, hbase, MySQL and Python