greatlearning
Power Ahead

AIML | MODULE PROJECT

1 AIML module projects are designed to have a detailed hands on to integrate theoretical knowledge with actual practical implementations.

2 AIML module projects are designed to enable you as a learner to work on realtime industry scenarios, problems and datasets.

3 AIML module projects are designed to enable you simulating the designed solution using AIML techniques onto python technology platform.

4 AIML module projects are designed to be scored using a predefined rubric based system.

5 AIML module projects are designed to enhance your learning above and beyond. Hence, it might require you to experiment, research, self learn and implement.

AIML | MODULE PROJECT

STATISTICAL LEARNINGS

PROJECT

AIML module project consists of industry based problems framed as detailed questions which can be solved using statistical learnings

TOTAL SCORE 30

PROJECT BASED

TOTAL SCORE 30

- **DOMAIN:** Startup ecosystem
- **CONTEXT:** Company X is a EU online publisher focusing on the startups industry. The company specifically reports on the business related to technology news, analysis of emerging trends and profiling of new tech businesses and products. Their event i.e. Startup Battlefield is the world's pre-eminent startup competition. Startup Battlefield features 15-30 top early stage startups pitching top judges in front of a vast live audience, present in person and online.
- **DATA DESCRIPTION:** CompanyX_EU.csv - Each row in the dataset is a Start-up company and the columns describe the company. ATTRIBUTE INFORMATION:
    1. Startup: Name of the company
    2. Product: Actual product
    3. Funding: Funds raised by the company in USD
    4. Event: The event the company participated in
    5. Result: Described by Contestant, Finalist, Audience choice, Winner or Runner up
    6. OperatingState: Current status of the company, Operating ,Closed, Acquired or IPO

    *Dataset has been downloaded from the internet. All the credit for the dataset goes to the original creator of the data.
- **PROJECT OBJECTIVE:** Analyse the data of the various companies from the given dataset and perform the tasks that are specified in the below steps. Draw insights from the various attributes that are present in the dataset, plot distributions, state hypotheses and draw conclusions from the dataset.

Steps and tasks: **[Total Score: 30 points]**

1. Data warehouse:
    - Read the CSV file.
2. Data exploration:
    - Check the datatypes of each attribute.
    - Check for null values in the attributes.
3. Data preprocessing & visualisation:
    - Drop the null values.
    - Convert the 'Funding' features to a numerical value.
    - Plot box plot for funds in million.
    - Get the lower fence from the box plot.
    - Check number of outliers greater than upper fence.
    - Drop the values that are greater than upper fence.
    - Plot the box plot after dropping the values.
    - Check frequency of the status features classes.
    - Plot a distribution plot for Funds in million.
    - Plot distribution plots for companies still operating and companies that closed.
4. Statistical analysis:
    - Is there any significant difference between Funds raised by companies that are still operating vs companies that closed down?

      Write the null hypothesis and alternative hypothesis.

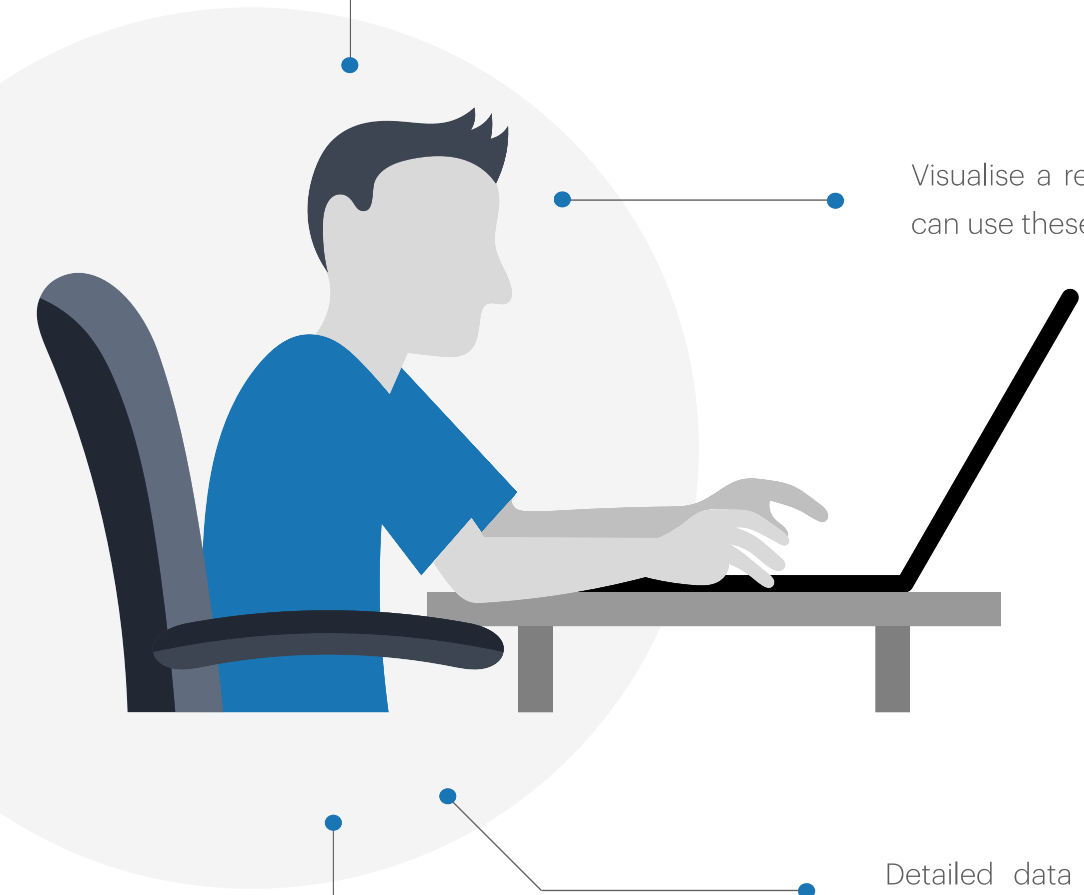
      Test for significance and conclusion
    - Make a copy of the original data frame.
    - Check frequency distribution of outcome variable.
    - Calculate percentage of winners that are still operating and percentage of contestants that are still operating
    - Write your hypothesis comparing the proportion of companies that are operating between winners and contestants:

      Write the null hypothesis and alternative hypothesis.

      Test for significance and conclusion
    - Check distribution of the Event variable.
    - Select only the Event that has disrupt keyword from 2013 onwards.
    - Write and perform your hypothesis along with significance test comparing the funds raised by companies across NY, SF and EU events from 2013 onwards.
    - Plot the distribution plot comparing the 3 city events.
5. Write your observations on improvements or suggestions on quality, quantity, variety, velocity, veracity etc. on the data points collected to perform a better data analysis.

LEARNING OUTCOME

Using the core statistical hypothesis to make business decisions in order to solve realtime problem statement.



Visualise a realtime industry scenario where one can use these statistical concepts.

Detailed data analysis and number crunching using statistics.

Exhaustive report building using EDA and visualisation techniques to help the business take decisions using insights from the data.

“ *Put yourself in the shoes of an actual* ”

DATA SCIENTIST

THAT's YOU

Assume that you are working at the company which has received the above problem statement from internal/external client. Finding the best solution for the problem statement will enhance the business/ operations for your organisation/project. You are responsible for the complete delivery. Put your best analytical thinking hat to squeeze the raw data into relevant insights and later into an AIML working model.

PLEASE NOTE

Designing a data driven decision product typically traces the following process:

1. Data and insights:
   Warehouse the relevant data. Clean and validate the data as per the the functional requirements of the problem statement. Capture and validate all possible insights from the data as per the the functional requirements of the problem statement. Please remember there will be numerous ways to achieve this. Sticking to relevance is of utmost importance. Pre-process the data which can be used for relevant AIML model.
2. AIML training:
   Use the data to train and test a relevant AIML model. Tune the model to achieve the best possible learnings out of the data. This is an iterative process where your knowledge on the above data can help to debug and improvise. Different AIML models react differently and perform depending on quality of the data. Baseline your best performing model and store the learnings for future usage.
3. AIML end product:
   Design a trigger or user interface for the business to use the designed AIML model for future usage. Maintain, support and keep the model/ product updated by continuous improvement/training. These are generally triggered by time, business or change in data.

IMPORTANT POINTERS

Project should be submitted as a single “.html” and “.ipynb” file. Follow the below best practices where your submission should be:

- ”.html” and ".ipynb" files should be an exact match.
- Pre-run codes with all outputs intact.
- Error free & machine independent i.e. run on any machine without adding any extra code.
- Well commented for clarity on code designed, assumptions made, approach taken, insights found and results obtained.

Project should be submitted on or before the deadline given by the program office.

Project submission should be an original work from you as a learner. If any percentage of plagiarism found in the submission, the project will not be evaluated and no score will be given.

greatlearning
Power Ahead

HAPPY LEARNING