greatlearning
Power Ahead

AIML | MODULE PROJECT

greatlearning
Power Ahead

5 Takeaways

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.

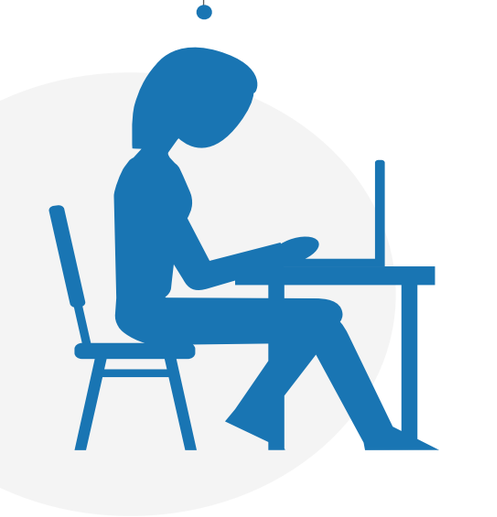
AIML | MODULE PROJECT

greatlearning
Power Ahead

APPLIED STATISTICS

PART I

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



PART III

AIML module project part III consists of industry based dataset which can be solved using statistical learnings

PART II

AIML module project part II consists of industry based dataset which can be solved using statistical learnings

TOTAL SCORE 60

PART ONE | QUESTION BASED | TOTAL SCORE | 15

Please answer the following questions with all relevant assumptions, explanation and details. **[Total Score: 15 Points]**

1. **Question:** Please refer the table below to answer below questions:

Planned to purchase Product A	Actually placed and order for Product A - Yes	Actually placed and order for Product A - No	Total
Yes	400	100	500
No	200	1300	1500
Total	600	1400	2000

   1. Refer to the above table and find the joint probability of the people who planned to purchase and actually placed an order.
   2. Refer to the above table and find the joint probability of the people who planned to purchase and actually placed an order, given that people planned to purchase.

2. **Question:** An electrical manufacturing company conducts quality checks at specified periods on the products it manufactures. Historically, the failure rate for the manufactured item is 5%. Suppose a random sample of 10 manufactured items is selected. Answer the following questions.
   A. Probability that none of the items are defective?
   B. Probability that exactly one of the items is defective?
   C. Probability that two or fewer of the items are defective?
   D. Probability that three or more of the items are defective ?

3. **Question:** A car salesman sells on an average 3 cars per week.
   A. Probability that in a given week he will sell some cars.
   B. Probability that in a given week he will sell 2 or more but less than 5 cars.
   C. Plot the poisson distribution function for cumulative probability of cars sold per-week vs number of cars sold per-week.

4. **Question:** Accuracy in understanding orders for a speech based bot at a restaurant is important for the Company X which has designed, marketed and launched the product for a contactless delivery due to the COVID-19 pandemic. Recognition accuracy that measures the percentage of orders that are taken correctly is 86.8%. Suppose that you place order with the bot and two friends of yours independently place orders with the same bot. Answer the following questions.
   A. What is the probability that all three orders will be recognised correctly?
   B. What is the probability that none of the three orders will be recognised correctly?
   C. What is the probability that at least two of the three orders will be recognised correctly?

5. **Question:** A group of 300 professionals sat for a competitive exam. The results show the information of marks obtained by them have a mean of 60 and a standard deviation of 12. The pattern of marks follows a normal distribution. Answer the following questions.
   A. What is the percentage of students who score more than 80.
   B. What is the percentage of students who score less than 50.
   C. What should be the distinction mark if the highest 10% of students are to be awarded distinction?

6. **Question:** Explain 1 real life industry scenario [other than the ones mentioned above] where you can use the concepts learnt in this module of Applied statistics to get a data driven business solution.

PART TWO | PROJECT BASED | TOTAL SCORE | 15

- **DOMAIN:** Sports
- **CONTEXT:** Company X manages the men's top professional basketball division of the American league system. The dataset contains information on all the teams that have participated in all the past tournaments. It has data about how many baskets each team scored, conceded, how many times they came within the first 2 positions, how many tournaments they have qualified, their best position in the past, etc.
- **DATA DESCRIPTION:** Basketball.csv - The data set contains information on all the teams so far participated in all the past tournaments.

    - ATTRIBUTE INFORMATION:
        1. Team: Team's name
        2. Tournament: Number of played tournaments.
        3. Score: Team's score so far.
        4. PlayedGames: Games played by the team so far.
        5. WonGames: Games won by the team so far.
        6. DrawnGames: Games drawn by the team so far.
        7. LostGames: Games lost by the team so far.
        8. BasketScored: Basket scored by the team so far.
        9. BasketGiven: Basket scored against the team so far.
        10. TournamentChampion: How many times the team was a champion of the tournaments so far.
        11. Runner-up: How many times the team was a runners-up of the tournaments so far.
        12. TeamLaunch: Year the team was launched on professional basketball.
        13. HighestPositionHeld: Highest position held by the team amongst all the tournaments played.

- **PROJECT OBJECTIVE:** Company's management wants to invest on proposal on managing some of the best teams in the league. The analytics department has been assigned with a task of creating a report on the performance shown by the teams. Some of the older teams are already in contract with competitors. Hence Company X wants to understand which teams they can approach which will be a deal win for them.

Steps and tasks: **[Total Score: 15 points]**

1. Read the data set, clean the data and prepare a final dataset to be used for analysis.
2. Perform detailed statistical analysis and EDA using univariate, bi-variate and multivariate EDA techniques to get a data driven insights on recommending which teams they can approach which will be a deal win for them.. Also as a data and statistics expert you have to develop a detailed performance report using this data.

   Hint: Use statistical techniques and visualisation techniques to come up with useful metrics and reporting. Find out the best performing team, oldest team, team with highest goals, team with lowest performance etc. and many more. These are just random examples please use your best analytical approach to build this report. You can mix match columns to create new ones which can be used for better analysis. Create your own features if required. Be highly experimental and analytical here to find hidden patterns. Use graphical interactive libraries to enable you to publish interactive plots in python.
3. Please include any improvements or suggestions to the association management on quality, quantity, variety, velocity, veracity etc. on the data points collected by the association to perform a better data analysis in future.

PART THREE | 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.

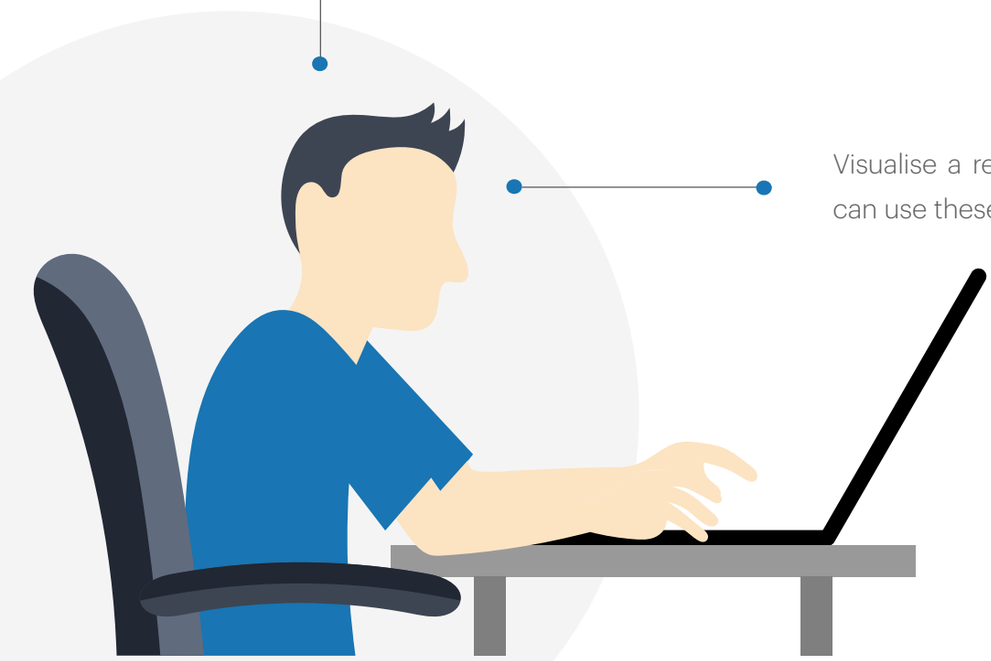
- **PROJECT OBJECTIVE:** Analyse the data of the various companies from the given dataset and perform the tasks that are specified in the below steps. Draw insights from the various attributes that are present in the dataset, plot distributions, state hypotheses and draw conclusions from the dataset.

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 OperatingState 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 Result 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 concepts and knowledge 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