IIITD | MACHINE LEARNING PROJECT

5 Takeaways

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

2 ML module projects are designed to enable you as a learner to workon real time industry scenarios, problems and data sets.

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

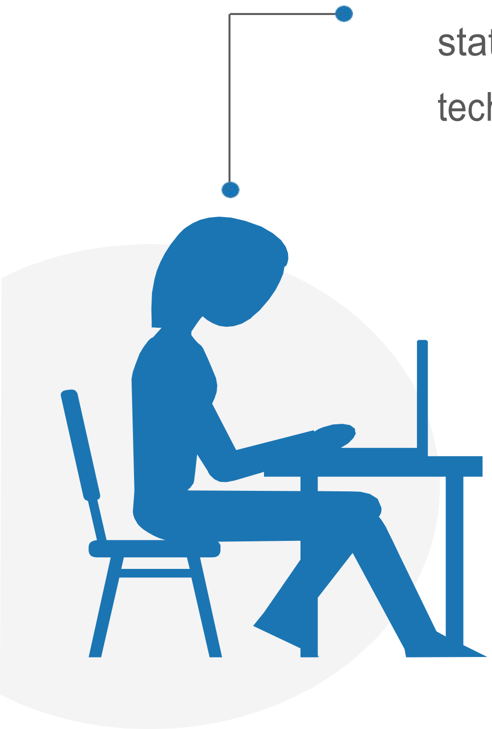
4 ML module projects are designed to be scored using a predefined rubric based system.

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

IIITD

MACHINE LEARNING PROJECT

SUPERVISED LEARNING



This project consists of industry based dataset and problem statement which can be solved using supervised learning techniques.

TOTAL SCORE 20

PROJECT BASED

TOTAL SCORE 20

DOMAIN: Telecom

DATA DESCRIPTION: Each row represents a customer, each column contains customer's attributes described on the column Metadata. The data set includes information about:

- Customers who left within the last month – the column is called Churn (target)
- Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
- Customer account information – how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges
- Demographic info about customers – gender, age range, and if they have partners and dependents

PROJECT OBJECTIVE: Build a model that will help to identify the potential customers who have a higher probability to churn. This help the company to understand the pinpoints and patterns of customer churn and will increase the focus on strategizing customer retention.

Steps to the project: [Total score: 20 points]

1. Import and warehouse data: [Score: 3 point]
   - Import the required libraries.
   - Create a database in MYSQL server using MYSQL workbench and Import the datasets 'TelecomCustomerChurn1' and 'TelecomCustomerChurn2' in it.
   - Connect to DB using mysql-connector-python package.
   - Import all the given datasets from MYSQL server and Explore shape and size.
   - Merge all datasets onto one and explore final shape and size.

2. Data cleansing and Exploratory data analysis: [Score: 8 point]

   Data Cleansing:
   - Check the percentage of missing values in each column of the data frame. Drop the missing values if there are any.
   - Check if there are any duplicate records in the dataset? If any drop them.
   - Drop the columns which you think redundant for the analysis.
   - Encode the categorical variables.
   - Write all the above steps in functions for modularity.

   Exploratory Data Analysis:
   - Perform detailed statistical analysis on the data.
   - Perform a detailed univariate, bivariate, and multivariate analysis with appropriate detailed comments after each analysis.

3. Data Preparation for model building: [Score: 2 point]
   - Store the target column (i.e. Churn) in the y variable and the rest of the columns in the X variable.
   - Split the dataset into two parts (i.e. 70% train and 30% test).
   - Standardize the columns using z-score scaling approach.

4. Model training, and testing: [Score: 6 point]
   - Train and test Logistic regression, KNN, and Naive Bayes models taught in the learning module.
   - Display the classification accuracies for train and test data.
   - Display and compare all the models designed with their train and test accuracies.
   - Select the final best trained model along with your detailed comments for selecting this model.

5. Conclusion and improvisation: [Score: 1 point]
   - Write your conclusion on the results.

“ Put yourself in the shoes of an actual ”

DATA SCIENTIST

THAT's YOU

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

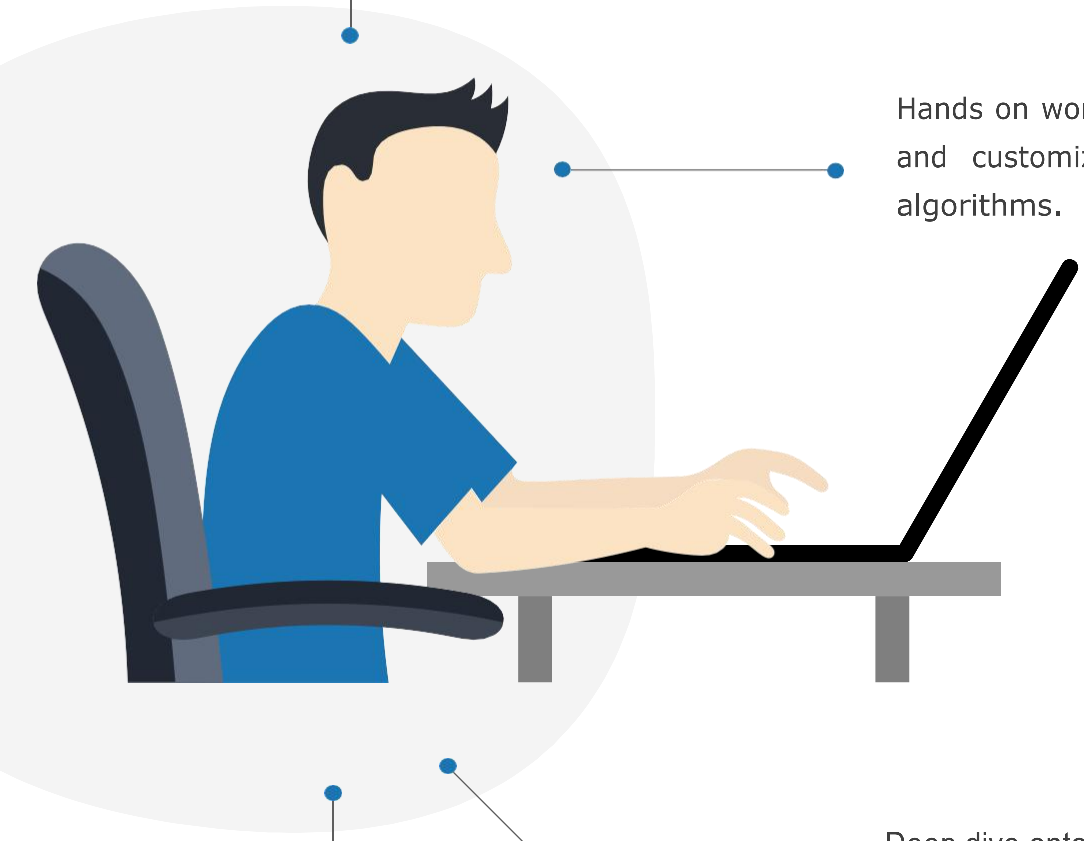
PLEASE NOTE

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 Machine learning model.
2. ML training:
Use the data to train and test a relevant ML model. Different ML models react differently and perform depending on quality of the data. Baseline your best performing model and store the learnings for future usage.

LEARNING OUTCOME

Hands on experience on data design, data cleansing, data warehousing and data manipulation required to drive an ML model.



Hands on working knowledge on all traditional and customized supervised techniques and algorithms.

Deep dive onto designing, and training a ML model.

Hands-on experience on testing the trained models and evaluating the models based on the appropriate performance metrics.

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