IIITD

ADVANCED MACHINE LEARNING PROJECT

5 Takeaways

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

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

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

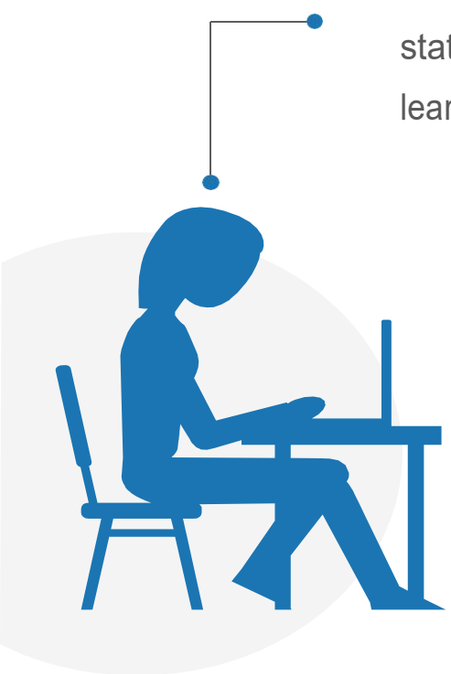
4. Advanced ML module projects are designed to be scored using a predefined rubric based system.

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

IIITD

ADVANCED MACHINE LEARNING PROJECT

SUPERVISED LEARNING



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

TOTAL SCORE 20

Google Store App Rating Prediction

CONTEXT:
The Play Store apps data has enormous potential to drive app-making businesses to success. However, many apps are being developed every single day and only a few of them become profitable. It is important for developers to be able to predict the success of their app and incorporate features which makes an app successful. We can collect app data and user ratings from the app stores and use it to extract insightful information. A machine learning model can be used to predict rating for a given app, which can be used to estimate success and scope of improvement.

PROJECT OBJECTIVE:
The Goal is to predict the rating for an app based on the given input features like size, number of downloads etc.

DATA DESCRIPTION: Web scraped data of 10k Play Store apps for analyzing the Android market. Each app (row) has values for category, rating, size, and more.

ATTRIBUTE INFORMATION:

1. **App:** Application name
2. **Category:** Category the app belongs to
3. **Rating:** Overall user rating of the app
4. **Reviews:** Number of user reviews for the app
5. **Size:** Size of the app
6. **Installs:** Number of user downloads/installs for the app
7. **Type:** Paid or Free
8. **Price:** Price of the app
9. **Content Rating:** Age group the app is targeted at - Children / Mature 21+ / Adult
10. **Genres:** An app can belong to multiple genres (apart from its main category). For eg, a musical family game will belong to Music, Game, Family genres.
11. **Last Updated:** Date when the app was last updated on Play Store
12. **Current Ver:** Current version of the app available on Play Store
13. **Android Ver:** Min required Android version

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

1. Import required libraries and read the data: [Score: 1 point]
   - Import the required libraries and read the dataset.
   - Check the first few samples, shape, info of the data and try to familiarize yourself with different features.

2. Data cleansing and Exploratory data analysis: [Score: 10 points]
   - Check summary statistics of the dataset. List out the columns that need to be worked upon for model building.
   - Check if there are any duplicate records in the dataset? if any drop them.
   - Check the unique categories of the column 'Category', Is there are any invalid category? If yes drop them.
   - Check if there are missing values present in the column Rating, If any? drop them and Convert ratings to high and low categories(>3.5 is high rest low) and store it in a new column ‘Rating_category’.
   - Check the distribution of the newly created column 'Rating_category' and comment on the distribution.
   - Convert the column "Reviews" to numeric datatype and check the presence of outliers in the column and handle the outliers using transformation approach.
   - The column 'Size' contains alphanumeric values, handle the non numeric data and convert the column into suitable datatype. (hint: Replace M with 1 million and K with 1 thousand, and drop/impute the entries where size='Varies with device').
   - Check the column 'Installs', handle the unwanted characters and convert the column into suitable dataype.
   - Check the column 'Price', remove the unwanted characters and convert the column into suitable datatype.

3. Data Preparation for model building: [Score: 2 points]
   - Drop the columns which you think redundant for the analysis.(suggestion: drop column 'rating', since we created a new feature from it (i.e. rating_category) will use that as target.)
   - For the target column 'Rating_category' Replace 'high' as 1 and 'low' as 0.
   - Encode the categorical columns.
   - Segregate the target and independent features.
   - Split the dataset into train and test.
   - Standardize the data, so that the values are within a particular range.

4. Model training, and testing: [Score: 5 points]
   - Write a function to fit and print the model predictions, input parameters would be model, train, and test data.
   - Use the above function and train a Decision tree, Random Forest, Bagging, Boosting, and Stacked Classifier models and make predictions on test data and evaluate the models.

5. Conclusion and improvisation: [Score: 2 point]
   - Compare and write your conclusions and steps to be taken in future in order to improve the accuracy of the model.

" 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.

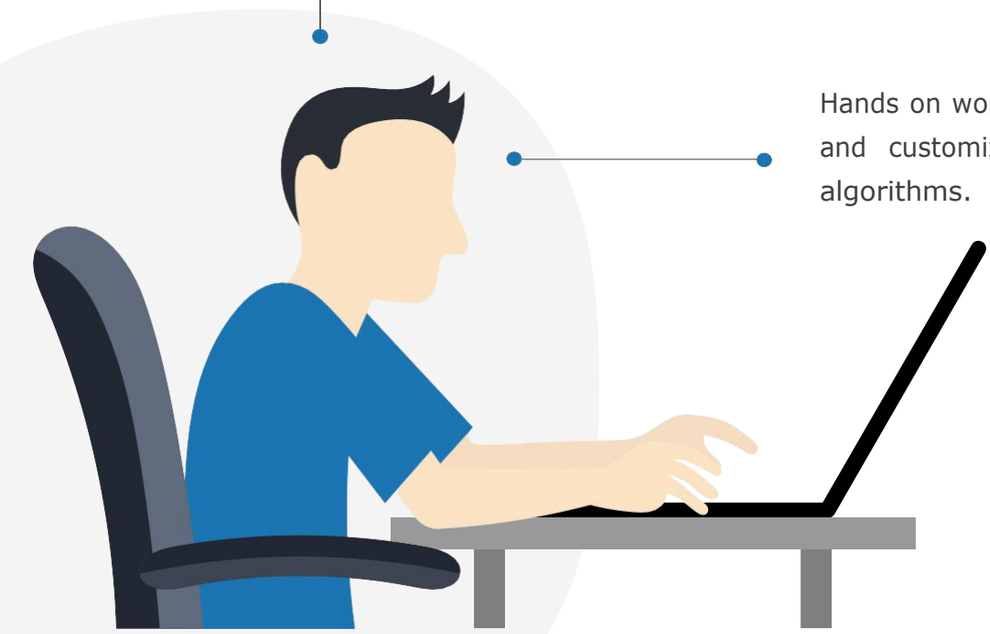
PLEASE NOTE

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 advanced ML model.



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

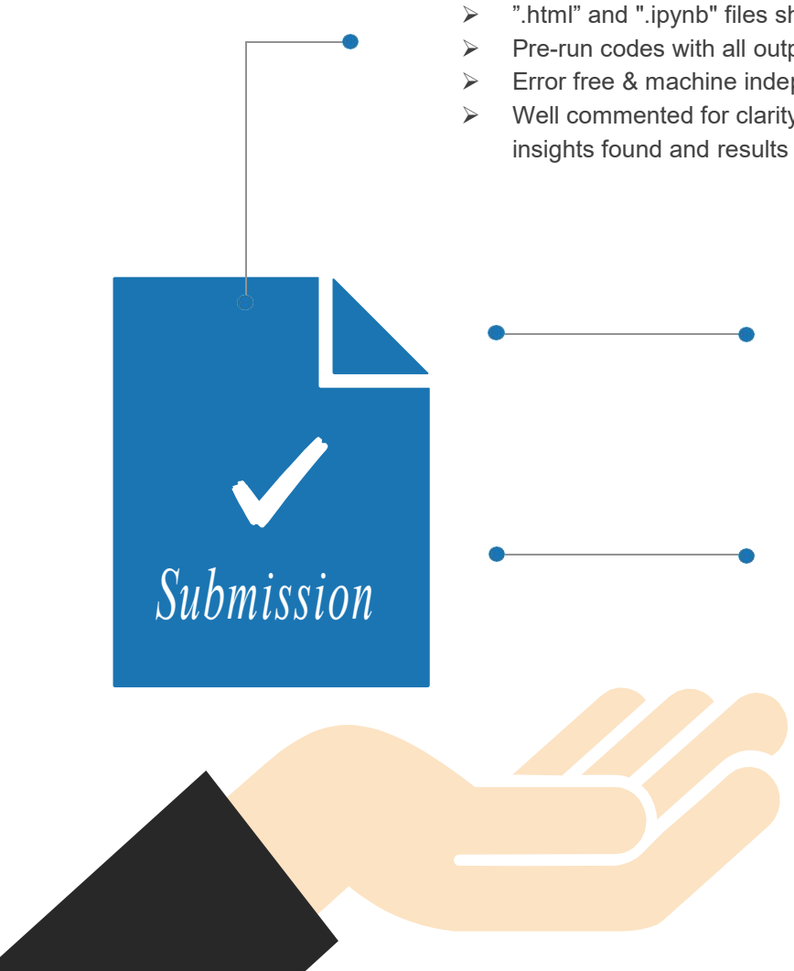
Deep dive onto designing, and training a advanced ML model.

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.

HAPPY LEARNING