IIITD

ADVANCED MACHINE LEARNING PROJECT

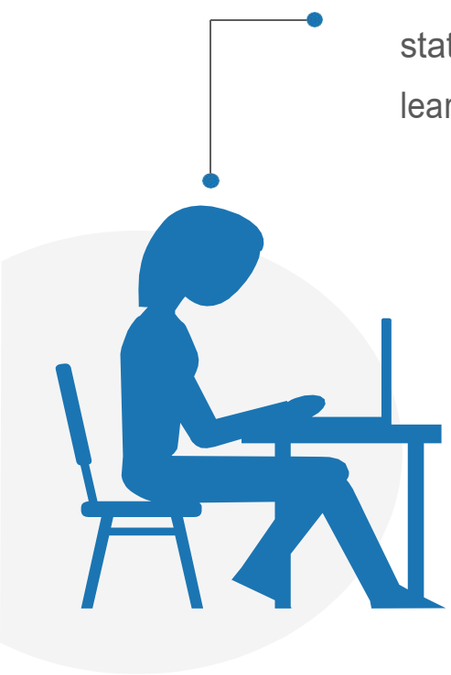
5 Takeaways

1. AML module projects are designed to have a detailed hands on to integrate theoretical knowledge with actual practical implementations.
2. AML module projects are designed to enable you as a learner to work on real time industry scenarios, problems and data sets.
3. AML module projects are designed to enable you simulating the designed solution using ML techniques onto python technology platform.
4. AML module projects are designed to be scored using a predefined rubric based system.
5. AML module projects are designed to enhance your learning above and beyond. Hence, it might require you to 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

PROJECT BASED

TOTAL SCORE 20

Bank Loan Defaulter Prediction

DOMIAN: Finance and Banking

CONTEXT:
There seems to be no end to bad loans in the country. According to the Reserve Bank of India, the overall bad loans as of March 2021 stood at INR 8.35 lakh crore, compared to INR 8.96 lakh crore in March 2020.
Banks run into losses when a customer doesn't pay their loans on time. Because of this, every year, banks have losses in crores, and this also impacts the country's economic growth to a large extend.
In This is project, we will look at various attributes such as funded amount, term, interest rates, loan amount, balance, etc. to predict if a customer will be a loan defaulter or not.

PROJECT OBJECTIVE: The Goal is to predict if a customer will be a loan defaulter or not based on the given input features such as funded amount, term, interest rate etc.

DATASET DESCRIPTION:
Data set has around 67463 instances and 35 features and includes target column as Loan Status.(1=Defaulter and 0=Non Defaulters)
Dataset Source: https://www.kaggle.com/datasets/sachinsarkar/deloitte-hackathon

ATTRIBUTES:
1. **ID:** unique ID of representative.
2. **Loan Amount:** loan amount applied.
3. **Funded Amount:** loan amount funded.
4. **Funded Amount Investor:** loan amount approved by the investors.
5. **Term**: term of loan (in months).
6. **Batch Enrolled:** batch numbers to representatives.
7. **Interest Rate:** interest rate (%) on loan.
8. **Grade:** grade by the bank.
9. **Sub Grade:** sub-grade by the bank.
10. **Employment Duration:** duration.
11. **Home Ownership:** Owner ship of home.
12. **Verification Status:** Income verification by the bank.
13. **Payment Plan:** if any payment plan has started against loan.
14. **Loan Title:** loan title provided.
15. **Debit to Income:** ratio of representative's total monthly debt repayment divided by self reported monthly income excluding mortgage.
16. **Delinquency - two years:** number of 30+ days delinquency in past 2 years.
17. **Inquires- six months:** total number of inquiries in last 6 months.
18. **Open Account:** number of open credit line in representative's credit line.
19. **Public Record:** number of derogatory public records.
20. **Revolving Balance:** total credit revolving balance.
21. **Revolving Utilities:** amount of credit a representative is using relative to revolving_balance.
22. **Total Accounts:** total number of credit lines available in representatives credit line.
23. **Initial List Status:** unique listing status of the loan - W(Waiting), F(Forwarded).
24. **Total Received Interest:** total interest received till date.
25. **Total Received Late Fee:** total late fee received till date.
26. **Recoveries:** post charge off gross recovery.
27. **Collection Recovery Fee:** post charge off collection fee.
28. **Collection 12 months Medical:** total collections in last 12 months excluding medical collections.
29. **Application Type:** indicates when the representative is an individual or joint.
30. **Last week Pay:** indicates how long (in weeks) a representative has paid EMI after batch enrolled.
31. **Accounts Delinquent:** number of accounts on which the representative is delinquent.
32. **Total Collection Amount:** total collection amount ever owed.
33. **Total Current Balance:** total current balance from all accounts.
34. **Total Revolving Credit Limit:** total revolving credit limit.
35. **Loan Status:** 1 = Defaulter, 0 = Non Defaulters.

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

1. Read the dataset and display some information about the dataset. [Score: 1 point]
   - Import 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: 3 points]
   - Check if there are any duplicate records in the dataset? if any drop them, and check the percentage of missing values, if any? treat them with appropriate methods.
   - Check summary statistics of the dataset, and write your key observations.
   - Drop the columns which you think redundant for the analysis.
   - Perform necessary univariate and multivariate analysis.
   - Check the distribution of the target column 'Loan Status', and comment on the class distribution.

3. Data preparation for model building: [Score: 5 points]
   - Segregate the target and independent features.
   - Encode the categorical data.
   - Handle the imbalanced data using oversampling or under sampling approach, and check the distribution of the re-sampled target class.
   - Split the data into train and test.
   - Select the K best features using wrapper or embedded methods.(Hint: refer to the sklearn documentation and try different methods for feature selection https://scikit-learn.org/stable/modules/feature_selection.html)

4. Model Building and evaluation: [Score: 5 points]
   - Build a base model using the Original Imbalanced data.
   - Try multiple models and tune their hyperparameters with appropriate methods and report the best performing model(use balanced data).

5. Pipeline: [Score: 5 points]
   - Build a pipeline and put all the possible steps in the pipeline and fit the pipeline on train data and get predictions on the test data.

6. Conclusions: [Score: 1 point]
   - Compare the evaluation metrics of the base model and the tuned model and write your conclusion. Mention the steps taken to improve the performance 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.

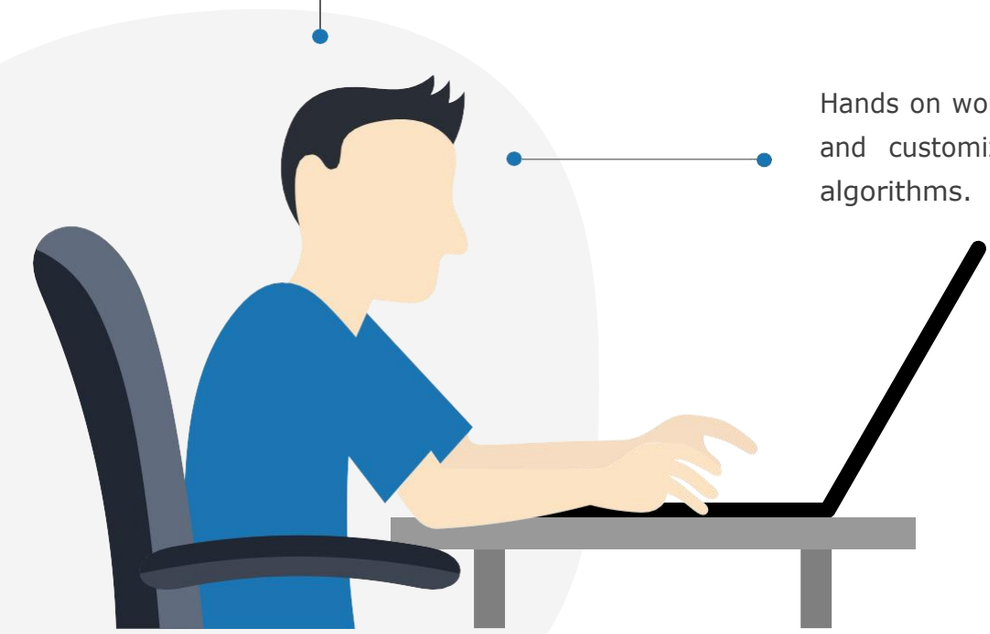
PLEASE NOTE

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 AML model.



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

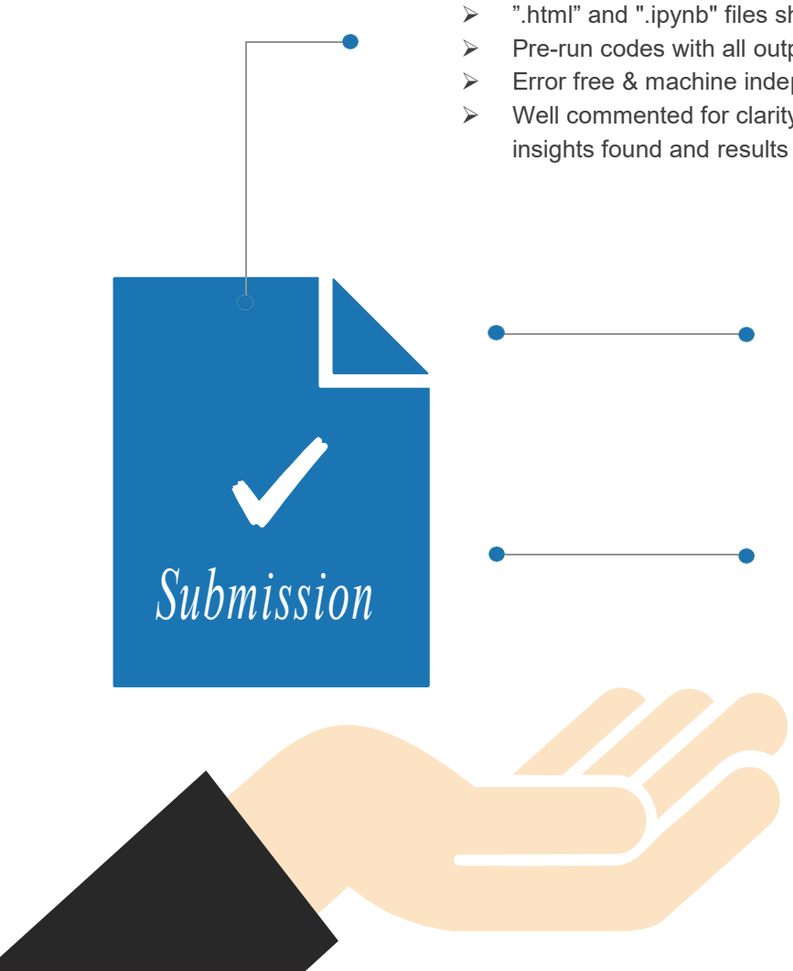
Deep dive onto designing, and training a AML model, handling imbalanced data, feature selection and hyperparameter tuning.

Hands-on experience on handling imbalanced data, testing the trained models, evaluating the models based on the appropriate performance metrics and selecting the best features and best combination of hyperparameters.

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