EDB Assignment

SQL

Question 1: Walk us through your logic for a query to get the names of all companies with no projects.

Ans 1:

```
select company_name
FROM company C where C.company_id not in (select company_id from project);
```

(Get all the company_id from project which are not present in company table using sub-query)

Ans 2:

```
SELECT company_name
FROM
        company c
FULL OUTER JOIN project p
    ON c.company_id = p.company_id
WHERE
        project_id IS NULL;
```

(Apply full outer join so all the rows having null from both table will appear, select rows where project_id is null)

Question 2: Walk us through your logic to get projects above $1M project value that have more than 3 contributors.

```
WITH project_contributors AS (
        SELECT project_id, COUNT(contributor_employee_id)
        FROM project_member
        GROUP BY project_id HAVING COUNT(contributor_employee_id)>3)
SELECT p.project_id FROM project p inner join project_contributors pc on
p.project_id = pc.project_id
WHERE project_value > 1
;
```

(Use common table expressions to get project_id having more than 3 contributors and use that table joining with project to get projects with >1 M project value)

Question 3: Walk us through your logic to write a query to find the top 10th Employee in absolute Project Value contribution.

```
SELECT TOP 1 Project_Value_Contribution
FROM ( SELECT DISTINCT TOP 10 Project_Value_Contribution
          FROM project_member1 ORDER BY Project_Value_Contribution DESC )
AS temp ORDER BY Project_Value_Contribution;
```

(Use inner query to get top 10 records in descending order and using main query get the first record which represent top 10th record.)

Machine Learning

1. A business user would like to get insights into the financial health of companies who are EDB's customers. They are worried about the risk of these companies being unable to complete their projects due to financial reasons and wish to implement more stringent financial checks and controls on riskier companies.
a. What are some questions you would ask the business user to scope this into a data science problem?

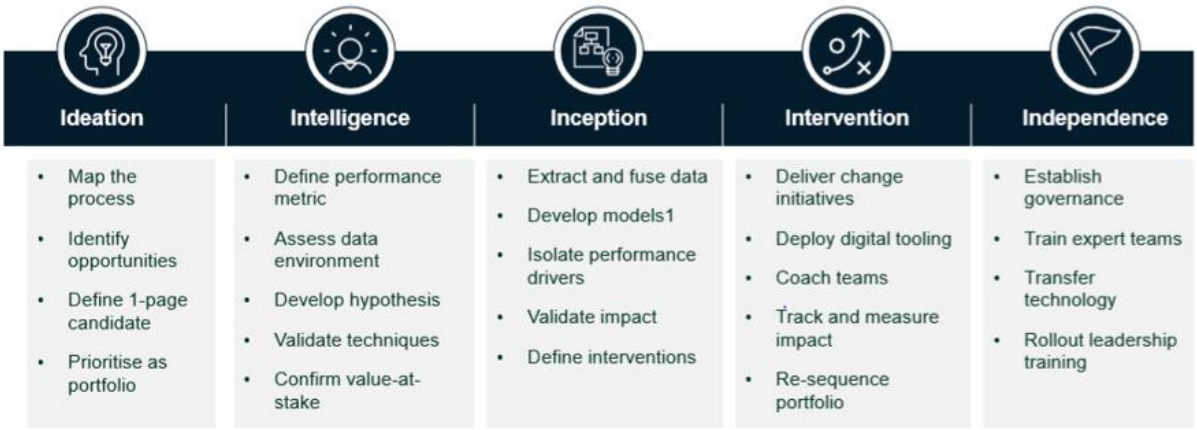
Objective: Determine the risk of the companies who are EDB's customers being unable to complete their projects by monitoring their financial health.

To breakthrough the problem, below questions needs to be answered in structured way.
Before scoping the data, it is necessary to know the purpose of the project and current pain points, where Business Understanding comes into picture. For example:

1) How is the financial health of the companies monitored currently?
2) Know your stakeholders
3) Explain with example how risk is identified and mitigated using current process and how it impacts EDB.
4) Many other methods are available like 5 whys, root cause analysis to know business user's requirement and priorities.

From above question, problem can be formulated, and deliverables can be identified.

It is necessary to scope the project in terms of process, timeline, and time-to-time deliverables.
For example, 5i method is one of the methods used as best practice approach for delivery.



Business Financials and information:

1) Understanding of the financial data.

(i) Is there any rule of thumb process they follow to determine low-risk or high-risk company?

(ii) Do all the companies belong to same sector or they belong to different sector? Since different sector companies have different type of workings and financial health so to get the **sense of segmentation** – we need to understand the sector company working for. Ex. Manufacturing companies would require lot of working capital and initial investment compared to purely service-based companies where capital investment is smaller. (This column can be added in current data)

Data related questions:

(i) How much data is available? Define granularity of data (yearly, half-yearly risk data). How much history data is available?

(ii) What each record / raw represents into data set? Does it represent different companies? Or do we expect to see same company records for multiple years (if any)?

(iii) Any third-party source which should be considered from business understanding.

(iv) Understanding of each variable. Getting sense of variable which are important from business perspective.

(v) What is the unit of the quantitative numbers available in Database? This is important to treat each data points into comparable parameter.

(vi) Since multiple functional currencies involved, it is also important to know company's operations in which currencies so that we can also understand about FX rates movement volatility impacts on earnings.

(vii) Missing value imputations:
E.g. Goodwill column has 95% value missing. Can we safely replace empty value with 0?

(viii) Redundancies:
E.g., Out of 160 rows, 40 rows are duplicate. Since company name is not given, can we assume that they are redundant and can be dropped?

(ix) Abnormalities:
E.g. Interest coverage ratio is negative (maximum -237) , still Risk is low. Why?

2. You have been given a dataset on (fictitious) companies' financial data. Please build a model to predict the financial risk of these companies in 2021. Show us all the steps you've taken, and assessments or assumptions you made in the process.

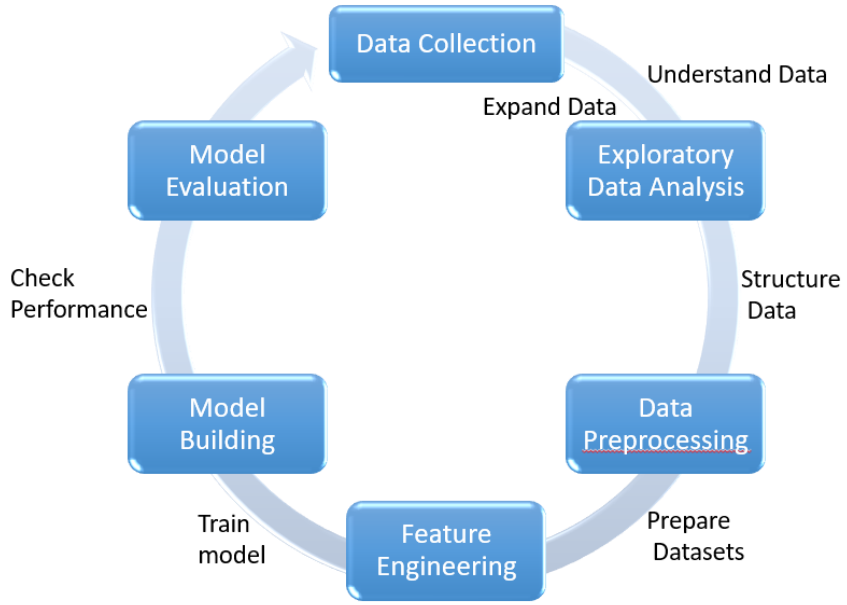
a. How would you construct the training data set?

b. What exploratory data analysis steps would you take?

c. How would you measure accuracy?

d. What modelling strategies would you consider?

e. What are the pros and cons of them?



At different points in Python code, training has been done using 2 strategies.

1) Train_test_split with 80-20 ratio. But this method validates the test data only once. So train the model more precisely, the same process should be done multiple time. That is where Kfold and Stratified KFold comes into picture.
2) In this dataset since all the categories are evenly divided KFold and StratifiedKFold can be used interchangeably. I have used Kfold.

➢ Complex model takes more training time if number of KFold is high.

Univariate EDA is performed using pandas_profiling.
Also, functions are written to find out missing value, general information of data and to remove duplicates.

Since data is mostly numeric, bivariate analysis has been done with focusing on target variable.

Missing values are replaced with 0 assuming that values are not present.

3 modelling techniques for classification are used: Logistic Regression, CatBoost Classification and XGBoost Classification.

3. Assuming that for this case study, you have built a classification model and achieved an accuracy of 96 percent.
a. Is this a good result?
b. What can you do about it?

Since the dataset is imbalanced with high risk to low-risk ratio of 1:5, so even if we achieve 96% accuracy, it is possible that model is biased towards low-risk predictions and never identifies high risk.
There are better metrics available to monitor the results like:

- AUC- ROC score : Separability between classes
- Precision-Recall :

- Confusion matrix : To see correctly and incorrectly identified results
- F1 Score: Suppose that classifier A has a higher recall, and classifier B has higher precision. In this case, the F1-scores for both the classifiers can be used to determine which one produces better results.

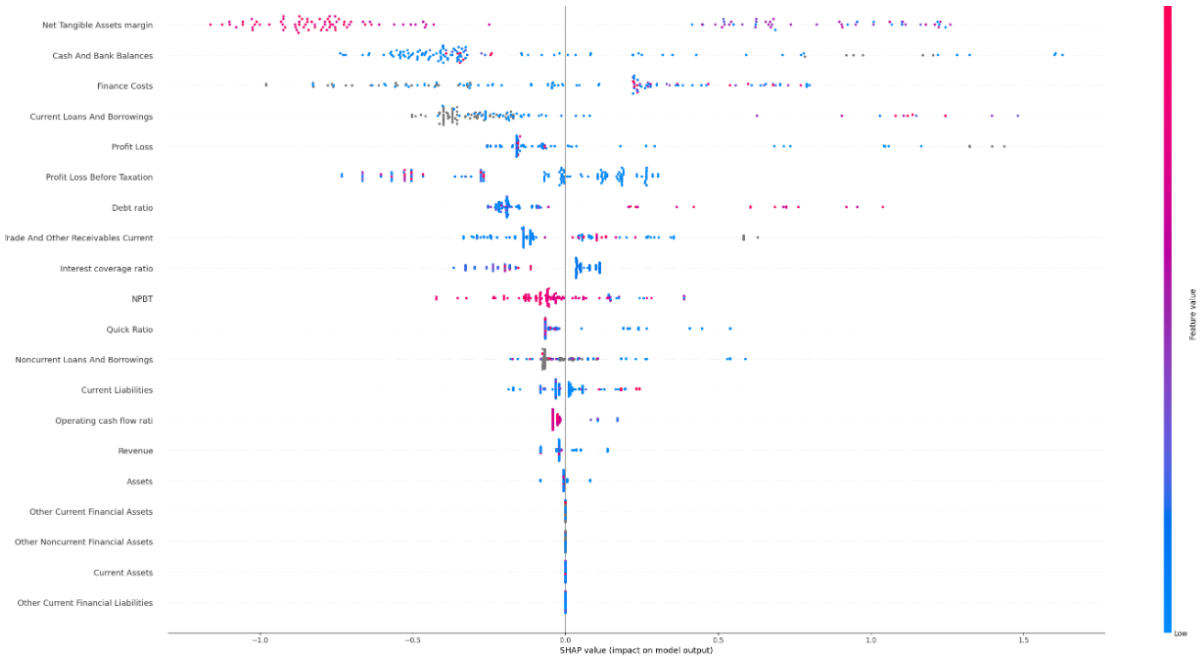
	Model	Accuracy	F1	AUC-ROC
0	Logistic Regression	0.808333	0.000000	0.485000
1	CatBoost	0.925000	0.918919	0.925000
2	XGBoost	0.975000	0.926829	0.965000

XGBoost gives better results compared to other algorithms.

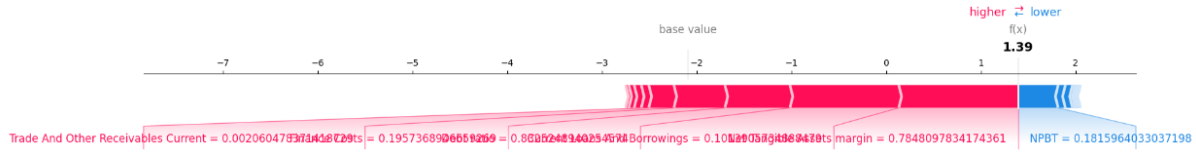
4. The business user is keen to use your model's predictions. However, he/she has to justify to his/her own stakeholders why a certain company has been given a high risk rating (and therefore warranting more stringent checks).
a. What can you do to further help the business user?
b. How would you advise the business user?

In order to explain the results of the model at global and local level, feature importance is used:

Specifically, Shap package can help to understand most impacting factors overall and at row-level.



Features importance for Overall model
Top 3 important factors: 1) Net tangible asset margin 2) Cash and Bank balances 3) Current loans and borrowings



Feature Importance shown for specific company.
For example, raw number 105 is high risk due to High current Loans and Borrowings and low Net Tangible Assets margin and high debt ratio.