Capstone Session - 1
Loan Default Prediction

Agenda

1) Loan Default Prediction Problem Statement
2) Solution Steps Walkthrough
3) General Best Practices
4) Q&A

Loan Default Prediction Dataset

- The Home Equity dataset (HMEQ) contains baseline and loan performance information for **5,960 recent home equity loans**.
- **The target (BAD) is a binary variable** that indicates whether an applicant has ultimately defaulted or has been severely delinquent.
- This adverse outcome occurred in **1,189 cases (20 percent)**.
- **12 input variables** were registered for each applicant.

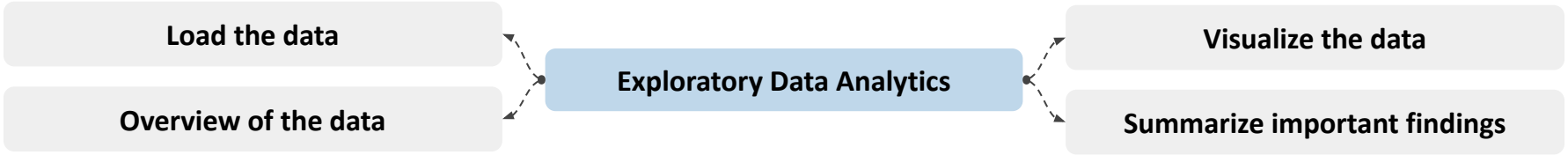
Problem Definition and Objective

- A bank's consumer credit department aims to simplify the decision-making process for home equity lines of credit to be accepted.
- To do this, they will adopt the Equal Credit Opportunity Act guidelines to establish an empirically derived and statistically sound model for credit scoring.
- The model will be based on the data obtained via the existing loan underwriting process from recent applicants who have been given credit.
- The model will be built from predictive modeling techniques, but the model created must be interpretable enough to provide a justification for any adverse behavior (rejections).

Objective:

- Build a classification model to predict clients who are likely to default on their loan and give recommendations to the bank on the important features to consider while approving a loan.

Solution Step 1



Example of questions that can be answered by EDA:

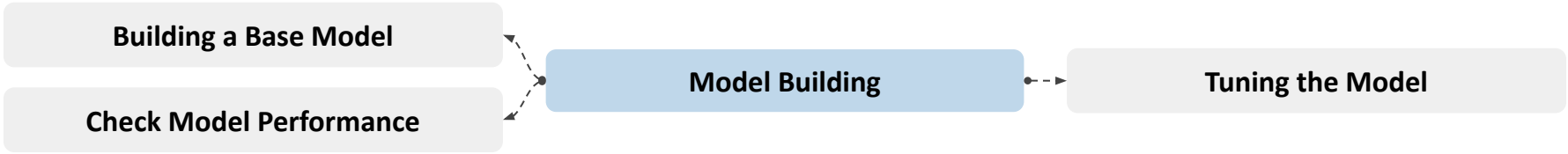
1. What is the shape of the dataset?
2. Are there any missing values in the dataset?
3. How does the distribution of years at present job "YOJ" vary across the dataset?
4. Is there a relationship between the REASON variable and the proportion of applicants who defaulted on their loan?
5. Do applicants who default have a significantly different mortgage amount compared to those who repay their loan?

Solution Step 2

Example of questions that can be answered by Data Preprocessing:

1. Are there any patterns in the missing values for particular variables?
2. How to identify the same?
3. How to impute missing values for numerical and categorical variables?
4. Do we need to normalise the data before splitting into train and test set?

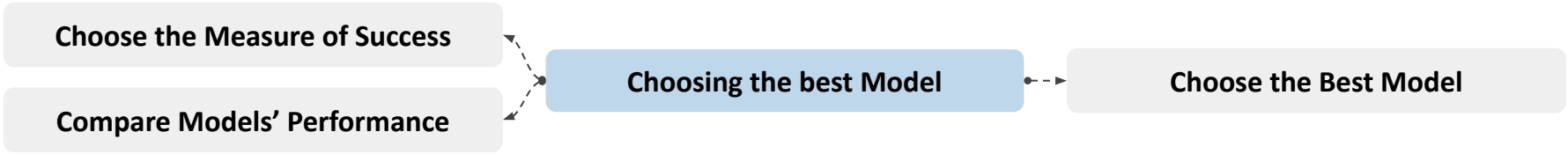
Solution Step 3



Example of questions that can be answered by Model Building:

1. Which algorithm can used to predict the target variables?
2. Is there any significant difference between the results linear and non linear algorithms
3. Does linear algorithm is easy to interpret as compared to non linear algorithms
4. How to choose the parameters to optimise the algorithm

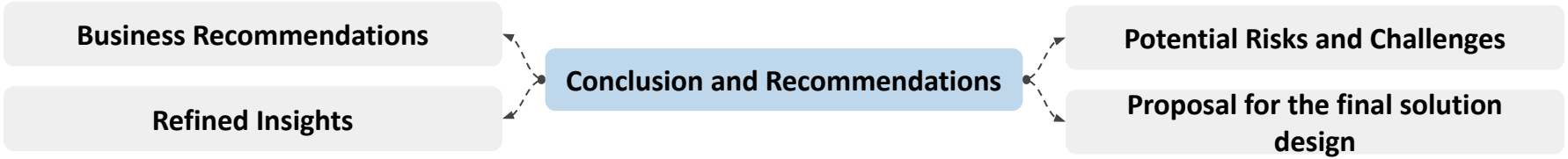
Solution Step 4



Example of questions that can be answered by Model Selection:

1. What is the metric (Measure of Success) for this business problem?
2. Should we consider the trade-off between the interpretability and model performance to choose the best model

Solution Step 5



Example of questions that can be answered by Conclusion and Recommendation:

1. What are the refined insights from EDA and model building?
2. What observations and insights can be drawn from the confusion matrix and classification report?
3. Is the model performance good enough for deployment in production?
4. What is proposal for final solution design? What are expected benefits and costs (assume numbers) of this solution design?

General Best Practices

Some of the best practices for submission:

- **Address all key questions in the rubric**: Make sure to read the rubric carefully and understand all the requirements. Address all the key questions asked in the rubric in your submission.
- **Provide observations and insights**: Provide observation and insight for every important output, such as plots, summary statistics, missing values detection and treatment. This will help to make your work more understandable and actionable.
- **Explain your design steps**: Explain the steps you took to design your solution approach. This will help the reader to understand the overview of your solution approach and how you arrived at your final model.
- **Implement both linear and non-linear algorithms**: Try to implement both linear and non-linear algorithms and provide observations from both the techniques. This will help to compare and contrast the performance of each technique and make better decisions.

General Best Practices

Some of the best practices for submission:

- **Improve the model by removing variables or pruning the model**: Always try to improve the model by removing variables or pruning the model. This will help to avoid overfitting and make your model more generalizable.
- **Select the performance metric that best fits the business objectives**: Choose the performance metric that best fits the business objectives. This will help to ensure that your model is relevant and useful to the business.
- **Interpret potential benefits from the model**: Provide an interpretation of potential benefits from the model. This will help the business to determine the next steps and make informed decisions based on your work.

Q&A