Data Science for Managerial Decisions (MB 511)
Building Machine Learning Solutions

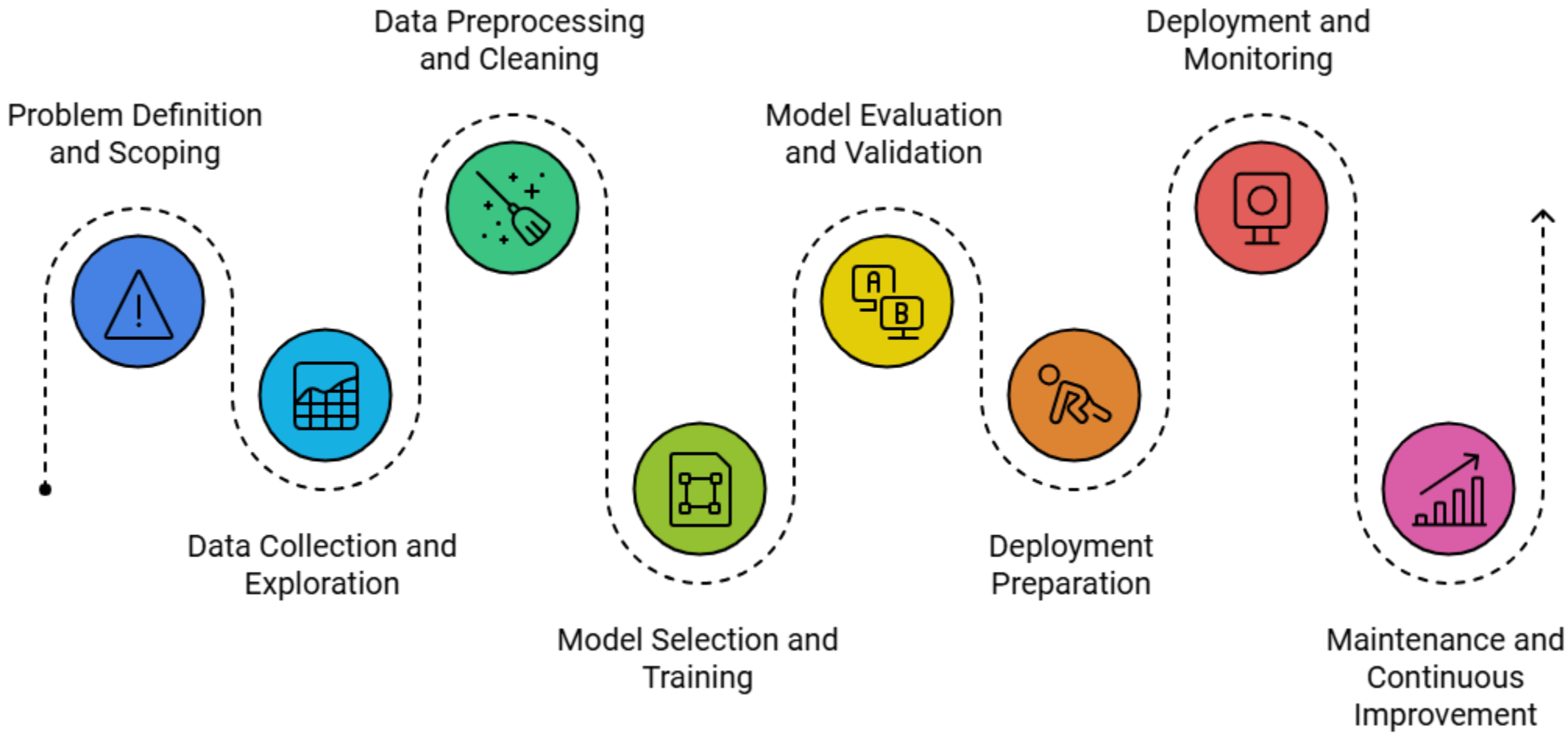
Instructor
Anant Prakash Awasthi

EMBA Program
MB-511

Agenda for today:

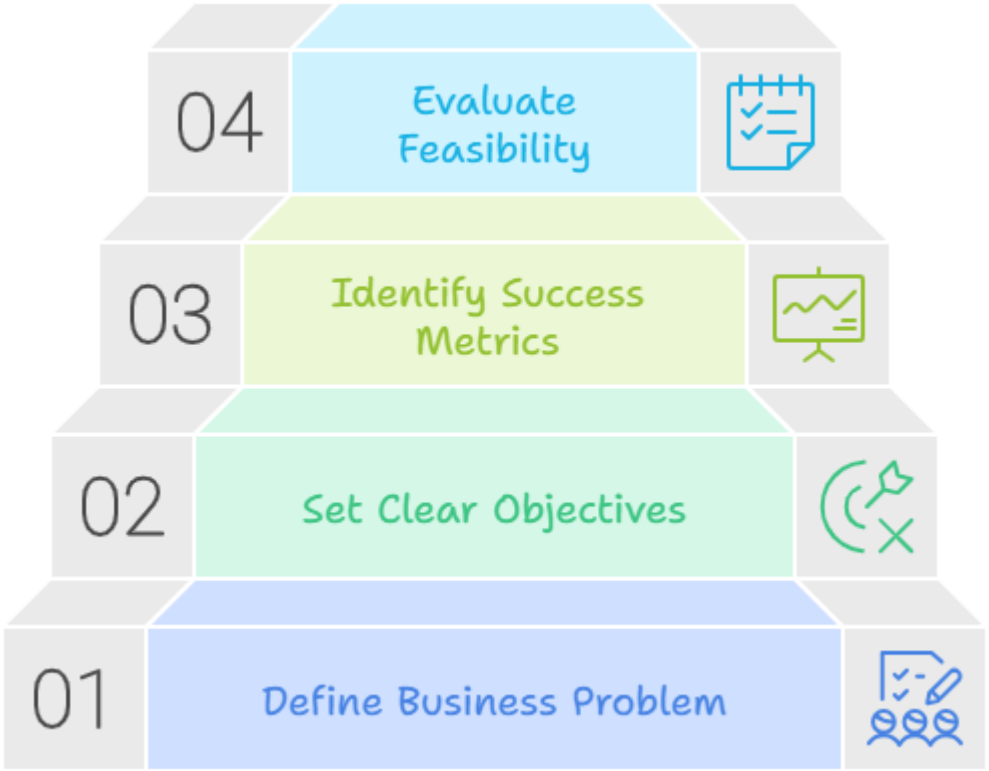
- Framework for building Data Science (Machine Learning) Solution
- Introduction to Kaggle for building the solutions (https://www.kaggle.com/)
- Bank Marketing Case Study

Machine Learning Solution Framework



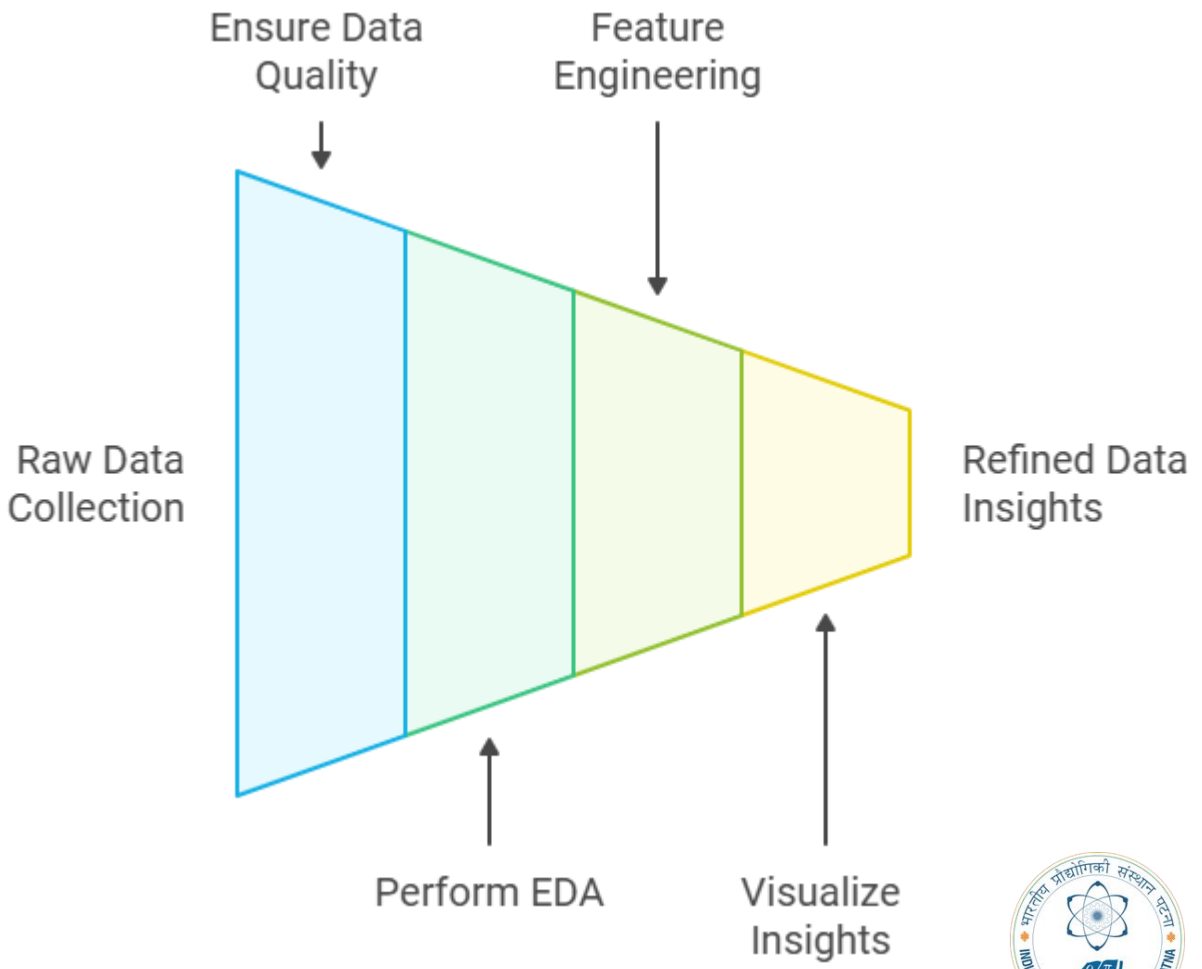
Problem Definition and Scoping

- **Define the Business Problem:**
  - Collaborate with stakeholders to fully understand the challenge
  - Identify the specific issue machine learning will address (e.g., predicting customer churn, detecting fraud)
- **Set Clear Objectives and Goals:**
  - Outline desired outcomes (e.g., reduce churn by 10%, improve accuracy of fraud detection)
  - Ensure alignment with business priorities and resources
- **Identify Success Metrics:**
  - Define quantitative metrics to measure model performance and success
  - Examples: accuracy, recall, precision, cost savings, revenue increase, or industry-specific KPIs
- **Evaluate Feasibility:**
  - Assess whether machine learning is appropriate given the data, budget, and time constraints
  - Determine potential constraints, such as data availability, computing power, or regulatory requirements



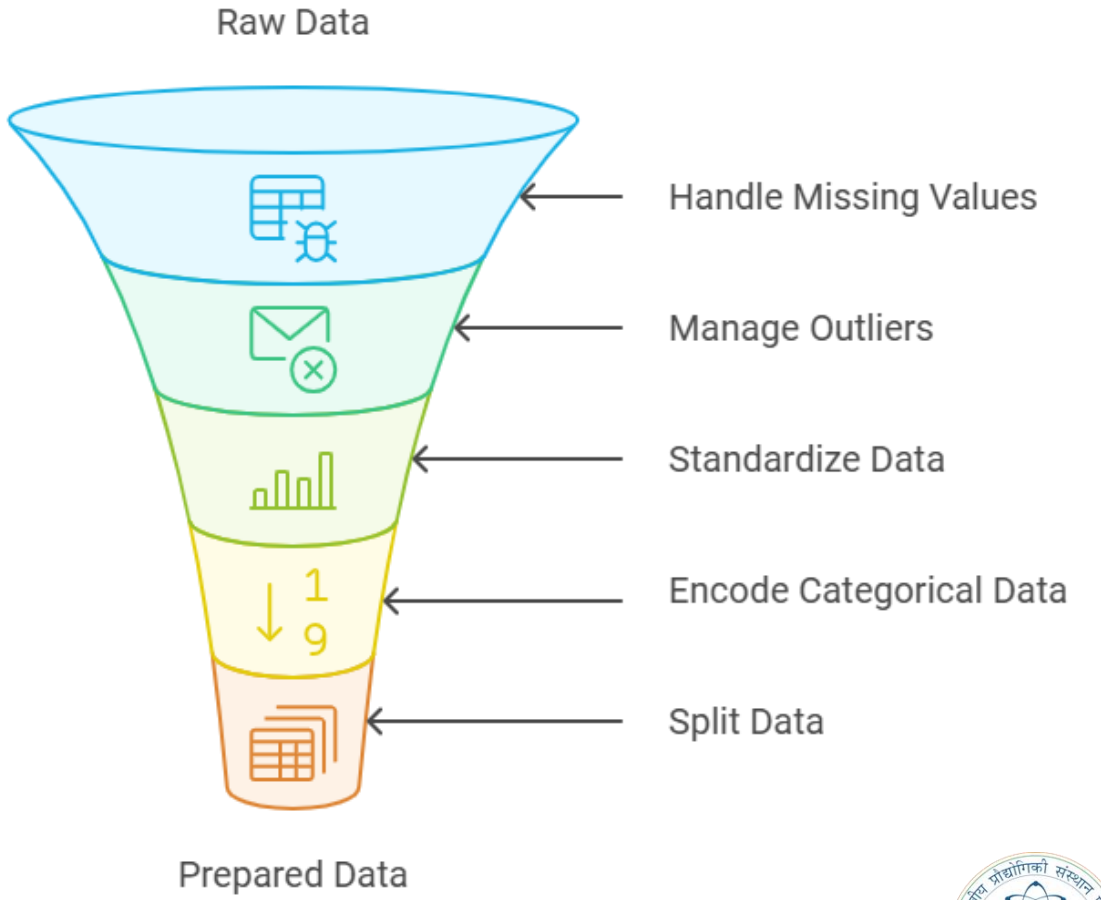
Data Collection and Exploration

- **Collect Relevant Data:**
  - Gather data from all necessary sources (e.g., databases, APIs, sensors, third-party datasets)
  - Ensure data quality, completeness, and relevance to the problem
- **Perform Exploratory Data Analysis (EDA):**
  - Use statistical summaries and visualizations to understand data distributions and relationships
  - Identify initial patterns, trends, and potential outliers to inform feature selection
- **Feature Engineering:**
  - Create or transform features that can improve model performance
  - Examples: creating new variables from dates, grouping categories, or calculating ratios
- **Visualize Insights:**
  - Use charts, graphs, and heatmaps to reveal data relationships and guide decision-making
  - Help stakeholders understand the data and provide input on key features to focus on



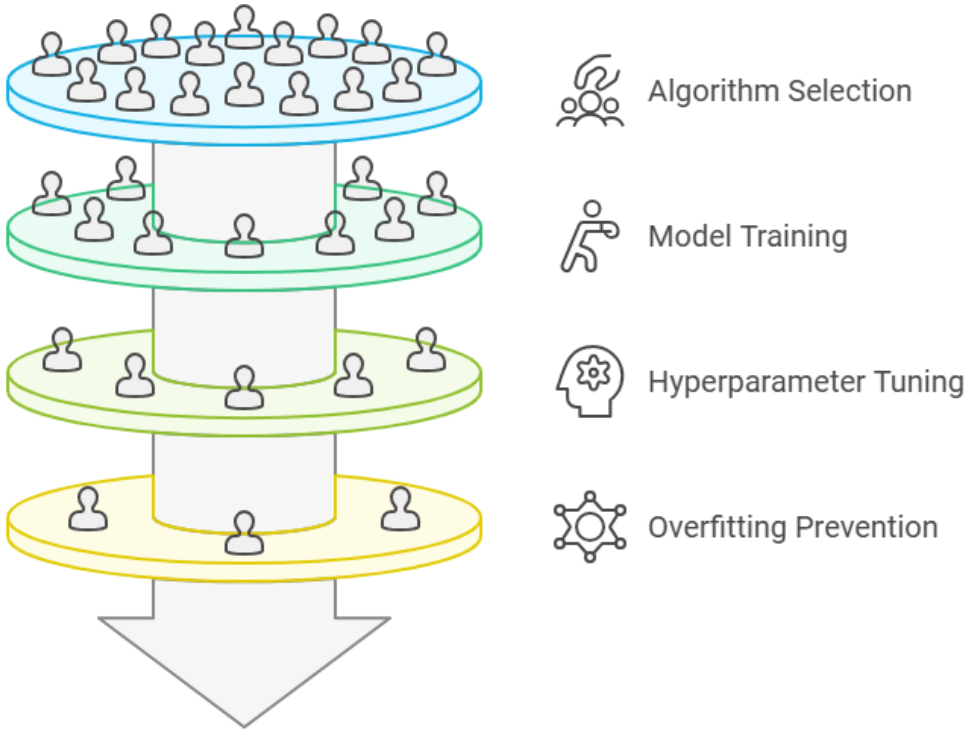
Data Preprocessing and Cleaning

- **Handle Missing Values and Outliers:**
  - Address missing data through methods like imputation, deletion, or interpolation
  - Identify and manage outliers to prevent skewed model results
- **Standardize and Transform Data:**
  - Apply scaling (e.g., normalization, standardization) to numerical features for consistency
  - Encode categorical variables into numerical formats (e.g., one-hot encoding, label encoding)
- **Data Splitting:**
  - Divide data into Training, Validation, and Test sets for model development and evaluation
  - Ensure data is representative across splits to prevent bias and enable reliable model assessment
- **Data Consistency and Integrity:**
  - Check for and correct inconsistencies or duplicates that could distort model predictions
  - Maintain data integrity to ensure high-quality inputs for model training



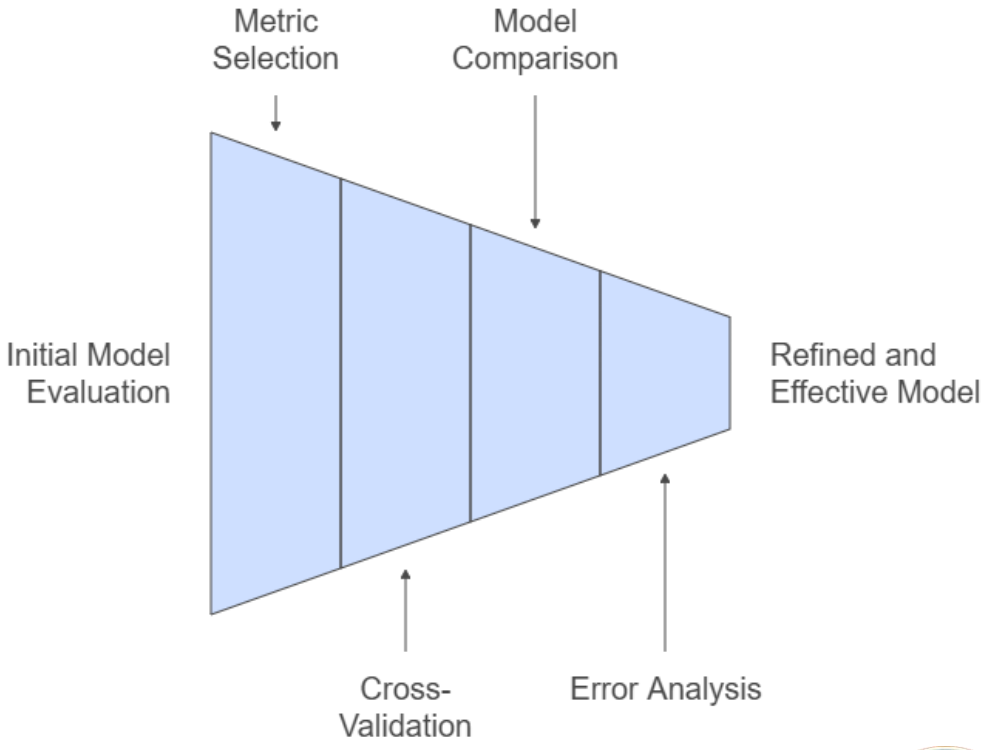
Model Selection and Training

- **Choose the Right Algorithm:**
  - Select algorithms based on the problem type (e.g., regression, classification, clustering)
  - Consider algorithm complexity, interpretability, and suitability for the data size
- **Train the Model:**
  - Use the training data to teach the model patterns and relationships within the data
  - Ensure a strong foundation by iterating and refining the training process as needed
- **Hyperparameter Tuning:**
  - Optimize model performance by adjusting hyperparameters (e.g., learning rate, depth of trees)
  - Use tuning methods like grid search, random search, or Bayesian optimization to find the best settings
- **Avoid Overfitting:**
  - Use techniques like regularization and early stopping to prevent the model from memorizing training data instead of generalizing to new data



Model Evaluation and Validation

- **Evaluate Model Performance:**
  - Use relevant metrics to assess model effectiveness, depending on the task
    - ❖ Classification: Accuracy, Precision, Recall, F1 Score
    - ❖ Regression: RMSE, MAE, R-squared
  - Match metrics to business goals to ensure meaningful evaluation
- **Cross-Validation:**
  - Apply techniques like k-fold cross-validation to test model consistency across different data subsets
  - Helps detect overfitting or underfitting and ensures that the model performs well on unseen data
- **Compare and Select the Best Model:**
  - Test multiple models and select the one with the best performance on key metrics
  - Consider interpretability, efficiency, and business relevance when choosing the final model
- **Error Analysis and Iteration:**
  - Analyze model errors to identify improvement areas (e.g., misclassified examples, high residuals)
  - Refine model or features based on insights from errors and repeat evaluation as needed

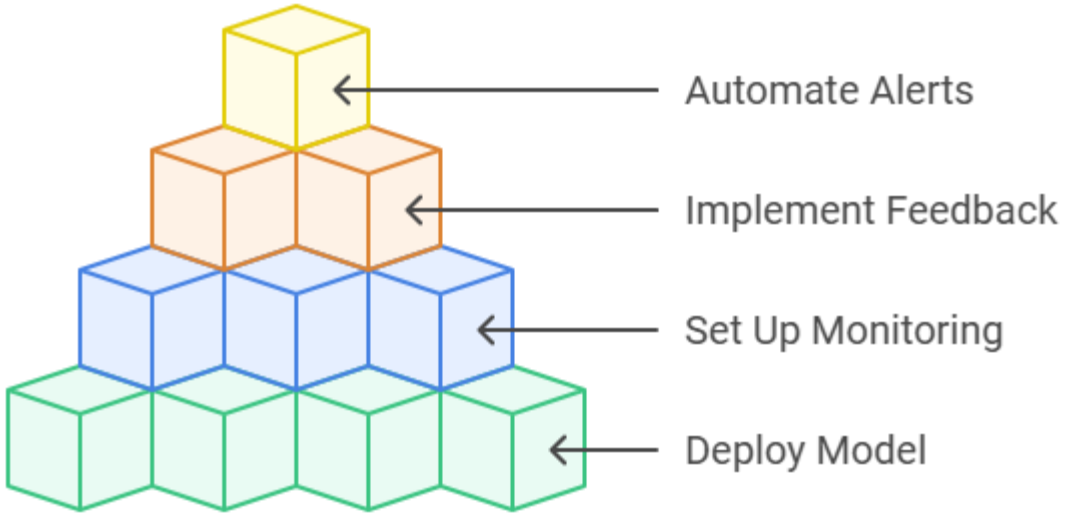


Deployment Preparation

- **Optimize Model for Production:**
  - Streamline model to improve efficiency (e.g., reduce complexity, optimize code)
  - Ensure model meets performance and latency requirements for deployment
- **Select Deployment Strategy:**
  - Choose the deployment approach based on business needs:
    - ❖ Real-time: Instant predictions, ideal for customer-facing applications
    - ❖ Batch Processing: Periodic predictions for bulk data, suited for offline analysis
    - ❖ Edge Deployment: Model runs on devices locally, useful for IoT or mobile applications
- **Establish Deployment Environment:**
  - Set up infrastructure, such as cloud platforms (AWS, Azure, GCP) or on-premises servers
  - Ensure compatibility and scalability of infrastructure for production workloads
- **Prepare for Model Versioning:**
  - Implement version control for models to manage updates and track changes over time
  - Allows easy rollback to previous versions if needed, ensuring stability

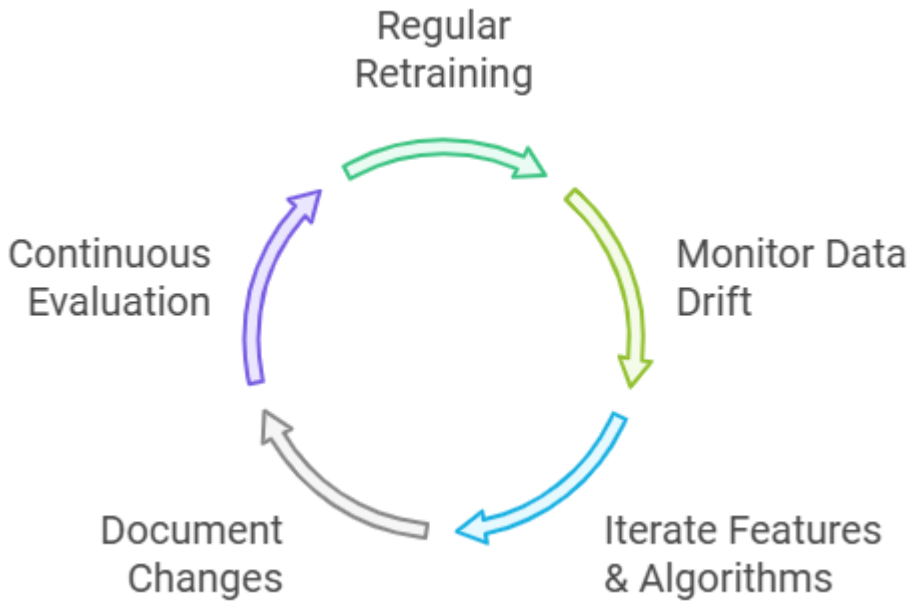
Deployment and Monitoring

- **Deploy the Model to Production:**
  - Launch the model in the production environment (cloud, on-premises, or edge)
  - Conduct final tests to ensure the model performs as expected under real-world conditions
- **Set Up Monitoring:**
  - Track key performance metrics such as:
    - ❖ Accuracy: Measure predictive quality over time
    - ❖ Latency: Ensure model response times meet user requirements
    - ❖ Data Drift: Detect shifts in input data distributions that may affect model performance
- **Implement a Feedback Loop:**
  - Collect user feedback and new data to assess model impact and accuracy
  - Use this information for continuous model updates or adjustments
- **Automate Alerts and Retraining:**
  - Set up alerts to notify of any performance degradation or data anomalies
  - Automate retraining triggers if needed to keep the model updated and effective



Maintenance and Continuous Improvement

- **Regular Model Retraining:**
  - Periodically retrain the model with new data to maintain accuracy and relevance
  - Ensure model adapts to evolving patterns and trends over time
- **Monitor for Data Drift and Model Degradation:**
  - Track data changes over time to identify shifts (data drift) that could affect model performance
  - Set up processes to detect and address model degradation promptly
- **Iterate on Model Features and Algorithms:**
  - Regularly evaluate feature engineering and consider new, relevant features based on domain insights
  - Experiment with updated algorithms or methods to improve model performance
- **Document Changes and Improvements:**
  - Maintain thorough documentation of model updates, feature changes, and retraining cycles
  - Helps ensure transparency, accountability, and easy handoff within the team
- **Engage in Continuous Evaluation:**
  - Regularly assess model performance in line with business goals and metrics
  - Collect stakeholder feedback to align the model with changing business needs and objectives



Introduction to Kaggle for building the solutions

Bank Marketing Case Study

Dataset - https://www.kaggle.com/datasets/janiobachmann/bank-marketing-dataset
Code - https://www.kaggle.com/code/janiobachmann/bank-marketing-campaign-opening-a-term-deposit

Introduction:

- Context
    - Find the best strategies to improve for the next marketing campaign. How can the financial institution have a greater effectiveness for future marketing campaigns? In order to answer this, we have to analyze the last marketing campaign the bank performed and identify the patterns that will help us find conclusions in order to develop future strategies.
- Source
    - [Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014

References:

1. A data-driven approach to predict the success of bank telemarketing
2. https://arxiv.org/abs/1503.04344
3. https://www.kaggle.com/code/aleksandradeis/bank-marketing-analysis
4. https://www.kaggle.com/code/mammadabbasli/bank-marketing-campaign
5. https://www.kaggle.com/code/varunsaikanuri/bank-term-deposit-marketing-analysis
6. https://www.kaggle.com/code/enesztrk/bank-credit-analysis-classification
7. https://www.kaggle.com/code/goldens/classification-review-with-python
8. https://www.kaggle.com/code/ludovicocuoghi/how-to-make-a-successful-marketing-campaign
9. https://www.kaggle.com/code/kareemellithy/bank-market-deposit-prediction-xgboost
10. https://www.kaggle.com/code/gcmadhan/bank-campaign-eda-classification-83-accu

EMBA Program
MB-511

Have a question?

Feel Free to Reach out at

- +91-88846-92929 (WhatsApp)
- anant.awasthi@outlook.com (E-Mail)