Data Science Life Cycle

Leangseng Lim

Introduction

Name: Leangseng Lim

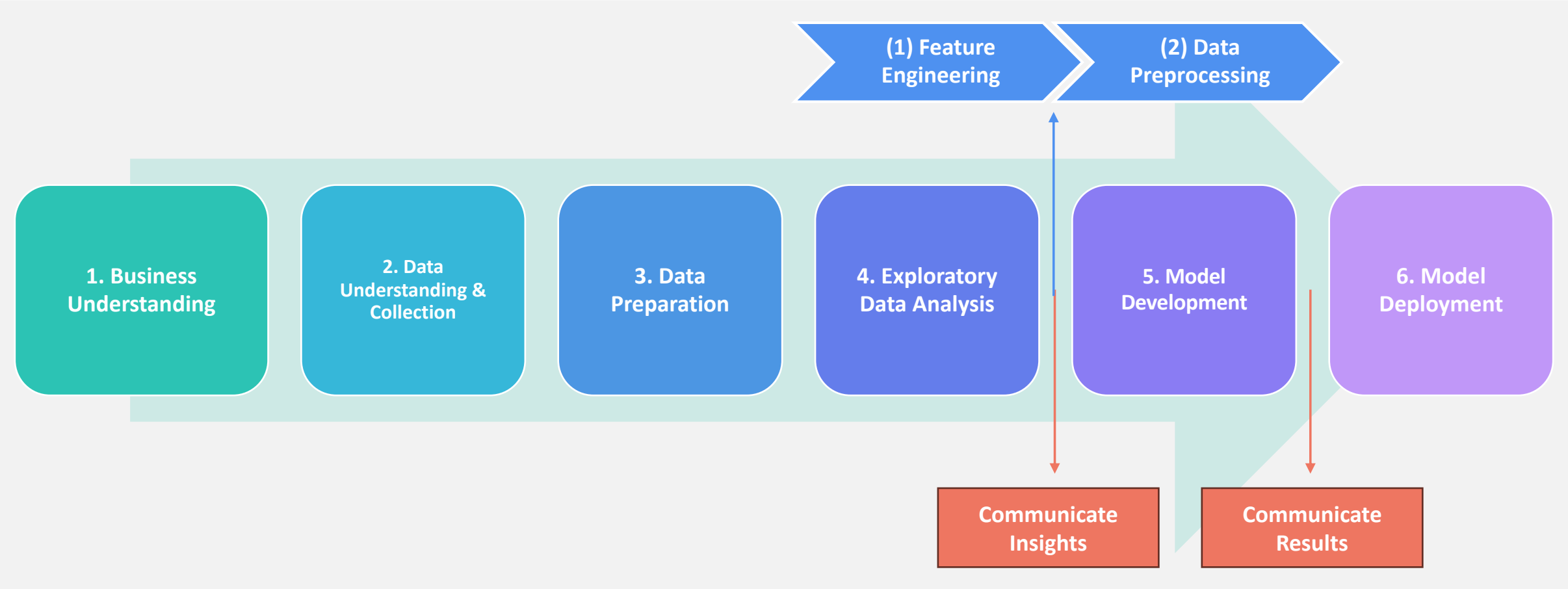
Role: Data Analytics Manager at Amret MFI

Agenda

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Exploratory Data Analysis
5. Model Development
6. Model Deployment

Data Science Life Cycle

The Data Science Lifecycle is an extensive step-by-step guide that illustrates how machine learning or data science project can be used to generate insights and predictions from data to accomplish a business objective.



1. Business Understanding

The key is to provide data science solution to business problem.

Key Questions:

- What are the issues that the business stakeholders are trying to solve?
- What is the use case?
- What is the end goal?

Use Case:

Problem Statement: They want to identify high value to low value clients for a preapproved self-service loan.

Model Objectives: The model is built to understand client patterns, and classify them from Grade A-E with the aims to:

- Measure level of customer engagement
- Measure customer loyalty
- Measure customers economic ability

2. Data Understanding & Data Collection

This is to understand the data using for the model and extract information to solve the problem.

Key Points:

- Understand the available data
- Where is the data sources?
- The granularity of the data you want to extract

- Is there any data dictionary or data catalog which you can use to learn about the definition of the data?
- Relational Database Systems:

- Is it One month data or Three months data? Or Weekly data?

3. Data Preparation

This is to prepare the data before doing data analysis and for modeling.

- Combine the extracted dataset
- Clean the data by finding missing values
- Fill the **missing values** by either mean or 0 based on business knowledge
- Detect **outliers** using boxplot or other statistical methods

Cust_ID	Region	Avg_InstAmount_last3months	ADB_1month
18557699	KM1	NULL	
18557789	KM1	NULL	1.11
18557899	KA1	NULL	0.14
18558079	PM1	NULL	1.01
18558119	1	707.85	756.3898

Improper Datatype

Null value

Missing value

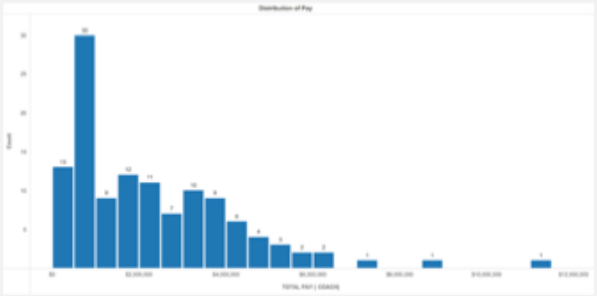
4. Exploratory Data Analysis

This step is crucial and is a-must do deeper analysis on the data for the problem you're trying to solve, and to provide insights on customer patterns.

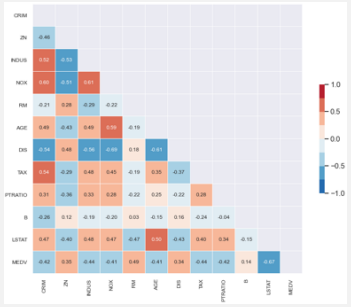
Goals:

- Understand how the data is distributed
- Identify patterns & Insights → think of the business and model objectives
    - What are the variables that we have to measure customer engagement? Customer loyalty? Customer Ability?

Histogram



Heatmap



Questions:

- How many deposit clients, saving, loan clients that we have?
- How many accounts and products do they have? Total, Avg?
- What are the different txn types that they usually perform in the past 3 months? Their avg. balance?

Feature Engineering, Data Preprocessing

(1) Feature Engineering

Model learns based on data and by squeezing the most out of data, you'll enable the model to learn better.

a) **Creating new features:**
- Combine multiple feature together into a new feature
- [AxB] : multiple the values of two features
- [A x B x C x D x E]: multiplying the values of 5 features

Ex: [Day of week, Hour] ==> [Hour of week]

b) **Feature Selection:**
- Correlation Coefficient
- Logistic Regression Coefficient
- Random Forest Feature Importance
- Recursive Feature Elimination

(2) Data Preprocessing

It involves transforming and normalizing raw data into numerical values for modeling.

a) **Data Transformation:**
- Normalization
  - Min Max Scalar
- Standardization
  - Standard Scalar
- Discretization
  - Binning, Clustering

b) **Data Reduction:** transforming data into lower dimension while preserving important information.
- Dimensionality Reduction Technique
  - PCA, T-SNE, etc.

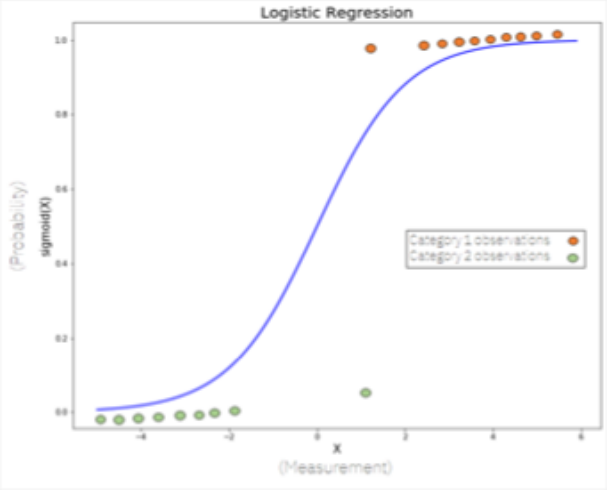
5. Model Development

(1) Clustering Model

Objective: to identify from high to low value deposit and saving clients.

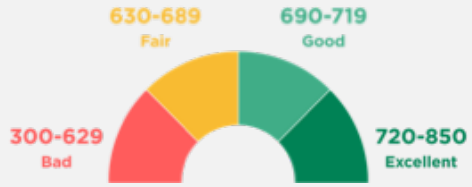
(2) Logistics Regression Model

Objective: to predict the propensity score of the eligible clients.



(3) Scorecard Model

Objective: to score clients based on their behavior from Grade A to Grade E.



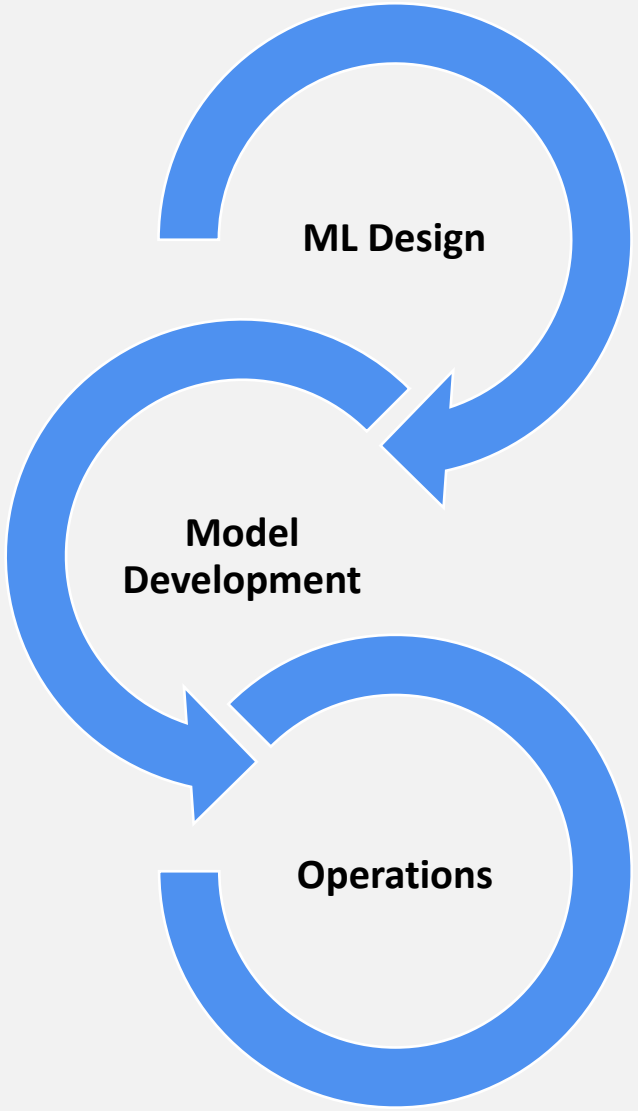
Attribute	Category	Scorecard Points
Age	Up to 25	80
	25 - 34	100
	35 - 40	150
	40 and above	200
Gender	Male	85
	Female	170
Salary	Up to 10000	100
	10001 - 30000	120
	30001 - 50000	140
	50001 - 70000	170
	70000 and above	200

6. Model Deployment

Communicate Results

A data scientist/analyst should be able to communicate and present your findings to Business Stakeholders.

- Simplify the concepts and explanation
- Focus on Key Business Metrics rather than model metrics



- Gather Requirements
- Business Understanding
- Data Understanding
- Data Collection

- Data Preparation & Preprocessing
- Feature Engineering
- Model design
- Model evaluation

- Batch/Realtime deployment
- Model Monitoring
- Triggering & Retraining the model
- Documentations

Thank you

Leangseng Lim

Email address: lleangseng.lim@gmail.com