Lead data summary

1. Most opportunities generated by external accounts, while only 44 valid opportunities are generated by internal accounts. Not sure if we should split them. If we don't split, there are 4 columns within internal account table are 90% filled with null. If we split, the records is very few for the internal account.

2. Is there any data about owner table? This can contain very useful information. Or we just remove ownerID.

3. There are many accounts with the same parentID. The same parentID can occur 12 times in the Account table (This client grant Melbourne uni 12 times?).
Options:

- Fill up the parentID with yes or no.
- Calculate how many opportunities were created by its parentID before this opportunities.

4. Dealing with null value

Null value	
1% - 10%	Fill up with the average
11% - 40%	Clustering similarity
41% and above	Remove

5. Tasks table

We might only generate 1 useful attribute from this table. – The completed tasks by each opportunity. We assume more tasks are completed by the opportunity, which will have higher converted rate.

.

Account

If internal account and external account is analysed individually.
Split the internal account out and check them if they are in the account (internal). All the account labelled "internal" are in the Account (internal) (27 records).

Column name		Preprocessing
id		
RecordType.Name	Business Organization (99.9%) Business Organization+B2333 Administrative	REMOVE
Industry	18 industries	1. Fill the blank with average 2. factorise
Industry_Sub_Type__c	19 sub types Null value constitutes 38.2%	1. Fill the blank by using similarity clustering. 2. factorise
Business_Type__c	11 types	1. Fill the blank with average 2. factorise
Country__c	Australia constitutes of 72.3%	REMOVE
Is_External__c	External 100%	REMOVE if we analyse external acct and internal individually
ParentId	Parent accountID or ex accountID, but nothing related to parent company. If an accountID has a parentID, this company created opportunity before. We assume the record with parentID have higher conversion rate than those who don't.	1. YES/NO or 2. If parentID != null, parentID = count(the records with the same parentsID converted before this created time), else parentID = 0.

Account (internal)

Column name		Preprocessing
id		
RecordType.Name	University Department, Business Organization	

Industry	null > 90%	remove
Industry_Sub_Type__c	null > 90%	remove
Business_Type__c	null > 90%	remove
Country__c	null > 90%	remove
Is_External__c	Internal 100%	remove
ParentId	Parent accountID or ex accountID, but nothing related to parent company. If an accountID has a parentID, this company created opportunity before. We assume the record with parentID have higher conversion rate than those who don’t.	1. YES/NO or 2. If parentID != null, parentID = count(the records with the same parentsID converted before this created time), else parentID = 0.
UoM_Organisation_Level__c	null > 90%	remove

Opportunity

Column name		Preprocessing
Id		DELETE 9.9% opportunity without account ID.
StageName		Factorise// response?
Status_Reason__c		Fill the blank with the avg factorise
RecordType.Name		Remove(90%similar with the next one)
Final_Record_Type__c		Fill the blank with the avg factorise

RICE_Supported__c	DEF IS NEEDED	
CreatedDate		1. remove 2. Calculate the period Created – actual close (close date to fill up the null value)
CloseDate		
Actual_Close_Date__c		
Amount		1. use actual value 2. fill the blank with ?
Estimated_Project_Total_Value__c		
Booked_Revenue__c		
Actual_Project_Total_Value__c		
BD_Cluster__c	DEF IS NEEDED	
BD_Division__c	DEF IS NEEDED	
CE_Course_Audience_Type__c	Null > 90%	remove
CE_Course_Type__c	Null > 90%	remove
AccountId		Link to acct table
Customer_Contact__c		remove
Lead_Academic_contact__c		Yes/no or remove
Lead_Faculty__c		1. fill the null by cluster similarity 2. factorise
Lead_School__c		1. fill the null by cluster similarity 2. factorise
Lead_Department__c	Null > 70%	remove
Supporting_Faculty_1__c	Null > 90%	remove
Supporting_Faculty_2__c	Null > 90%	remove
OwnerId	Relate to owner table	Owner table ???
Parent_Opportunity__c	Null>80%	Yes/ no

Event

Column name		preprocessing
Id		
WhatId		primary id of Opportunity table
ActivityDate		Calculate the period between opportunity created - activity data
OwnerId	Owner table??	
Type		Factorise=(0: no?, 1:

		email, 2:phone)
EventSubtype	Event(100%)	remove