**Lead data summary**

q1. only focus on Closed won closed lost and do prediction on other rows(Probability).

2. how to split data into train/test? 70%, 80%

Cross-valid over fitting

First iteration:

Factorize all the columns (One hot encoding)?

business insights (clusters which have higher converted rate)

change response to 0, 1(probability)

2nd : important activity features

Q3. 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.

Internal account- education

Q for alec. 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

Q 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. | Q  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 education?? |
| **Industry\_Sub\_Type\_\_c** | null > 90% | **☑️**Remove null |
| **Business\_Type\_\_c** | null > 90% | Remove NOT FOR PROFIT |
| **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.  REMOVE FIRST |
| **UoM\_Organisation\_Level\_\_c** | null > 90% | remove |

**Opportunity**

|  |  |  |
| --- | --- | --- |
| **Column name** |  | **Preprocessing** |
| **Id** |  | Q. DELETE 9.9% opportunity without account ID.  REMOVE FIRST  (See if account attributes are important, if yes, add back) |
| **StageName(response)** |  | Closed won/ closed lose |
| **Status\_Reason\_\_c** |  | Fill the blank with the avg  factorise |
| **RecordType.Name** |  | Fill the blank with the avg  Factorise with the next one |
| **☑️ Final\_Record\_Type\_\_c** |  | Remove (90% similar with the last one) |
| **RICE\_Supported\_\_c** | DEF IS NEEDED  Research innovation - - |  |
| **CreatedDate** |  | 1. remove  2. change to `201911(without date)  `separate Year, month -one hot encoding  2. Calculate the period  Created – actual close (close date to fill up the null value)  Closed win/lost:  Close – actual close  Predicting data:  Today-create date |
| **CloseDate** |  |
| **Actual\_Close\_Date\_\_c** |  |
| **Amount** | Initial estimate | **☑️**Remove first  (clustering in 2nd modelling)  Q. priorities  1. booked  2. estimated  Closed lost -  Closed won - |
| **Estimated\_Project\_Total\_Value\_\_c** |  |
| **Booked\_Revenue\_\_c** | Final figure on the contract (null>40) |
| **Actual\_Project\_Total\_Value\_\_c** | (null>40) |
| **BD\_Cluster\_\_c** | Business developer | **☑️**Remove first  ASK ALEX if it is important |
| **BD\_Division\_\_c** | Business developer | **☑️**Remove first  (successful rate for each team) |
| **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** |  | **☑️**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 (bd manager)** | Relate to owner table | **☑️**remove |
| **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 date  Remove first  Add it back in the 2nd modelling |
| **OwnerId** | Owner table?? | remove |
| **Type** |  | Factorise= (0: no?, 1: email, 2:phone) |
| **EventSubtype** | Event(100%) | remove |

Add interaction term, cross features

One hot encoding record types

<https://developers.google.com/machine-learning/crash-course/feature-crosses/video-lecture>

2nd cleaning

1. converted rate increased

For those opportunities which implied they are won, change them to won opportunities.

If StageName is not CloseWon, but statusReason is won, post award and then we change StageName to CloseWon.

If StageName is PostAward, change it to CloseWon

2. change create date to create year.

3. delete close date. 50% is randomly imputed number, such as 48434

4. one-hot encoding

5. 0 instead of knn in account id/ faculty

3rd

1. add create month

2.falcuty, create month, accounteid, RIC, business type,

4th

1. Change createDate (string) to createDate (dateTime)

CreatedDate, CloseDate change string to time, seperate year, month

2. retrieve Actual\_Close\_Date\_\_c， change to 0,1

3. delete

Those opportunities with “Is\_external” empty, do not have relative account. This means these opportunities have 8 features empty. And 81% of these opportunities is closed loss

4. retrieve Industry\_Sub\_Type\_\_c

Use Industry to fill up the empty

5. the empty values of status reason are filled up with stageName

6. retrieve leadSchool, fill up with “NotGiven”