Hello Sprocket Central Pty Ltd,

My name is Eric and I am a data analyst in KPMG. The following is the result after our team inspected the dataset you provided. And there are several questions we need to discuss with you.

I will show the problems group by the tables (Customer Demographic, Customer Addresses, Transaction data in the past three months).

1. Customer Demographic

* Customer ID
Problems: There are 4 ids are not in the customer demographic file (4001,4002,4003,5034)
Our solution: Please ensure that all tables are from the same period. Only customers in the Customer Master list will be used as a training set for our model.

* DOB
Problems: DOB contains 87 missing values and also have one outlier data
Our solution: Filling the missing cells by the most frequent date. Also, drop the outlier from the dataset

* Job_titile:
Problems: Job_title contains 506 missing values.
Our solution: Filling the missing cells by the “unknown” label.

* Job_industry_catagory
Problems: Job_industry_category contains 656 missing values.
Our solution: Filling the missing cells by the “unknown” label.

* Default
Problems: The column contains unreadable data.
Our solution: Discuss with the the client, need more details or drop it if this column is not necessary for further analysis.

* Tenure
Problems: The column contains 87 missing values and there are 55 outliers (over 2 standard deviation).
Our solutions: Assign a mean tenure value to missing cells. Also, about the outliers, after inspecting, even though the values are over 2 standard deviations, the values still see reasonable. Therefore, if the data aren’t occurred by typing mistake, we can keep it.

* Gender
Problems: There are 3 unregulated input M, F, Femal) and Unknown label □U□.
Our solutions: Convert the unregulated input to the regular class (Male, Female). Also, need further information about label “U”.

2. Customer Addresses

* Postcode
Problems: The data type is ‘int’ but I think ‘object’ is better data type for this column.
Our solutions: Convert the data type to object

* Property_valuation
Problems: 154 outliers in this column.
Our solutions: After dive into the column, the value in outlier table is all 1. It seems the data is reasonable and it’s not typo. We need to discuss this with the client.

3. Transaction data
* Online_order
Problems: 360 missing values.
Our solutions: Evenly distributed to each exist class (1, 0) or assign a new label to missing values.

* Brand, product_line, product_class, product_size
Problems: all of the columns above have 197 missing values.
Solutions: assigned a new label “Unknown” to each column.

* Standard_cost
Problems: there are 1122 outliers. And all the data are in the right side of whole dataset. It cause the data skewed to the right
Our solutions: We can have two dataset, one remain the outliers, the other remove all outliers. And compare two results.

Moving forward, the team will continue with the data cleaning, standardization and transformation process for the purpose of model analysis.

Regards,
Eric Wang