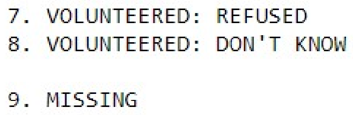
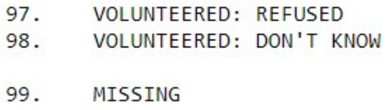
Before starting to train the model, we need to analyze the data do preprocessing. First of all, I focused on the missing values of the data. In some columns, the ratio of missing values may be quite high. So, I prefer the eliminate the columns which have missing values of more than fifty percent ratio. In each column, the missing values are labeled with different values as shown in the pictures. To eliminate these columns easily, I firstly grouped the columns whose missing values are labeled in the same way. Then, I gathered the columns that do not have missing values of more than fifty percent ratio.



```
7. VOLUNTEERED: REFUSED
8. VOLUNTEERED: DON'T KNOW

9. MISSING
```



```
97.    VOLUNTEERED: REFUSED
98.    VOLUNTEERED: DON'T KNOW

99.    MISSING
```

After that, I detected categorical and ordered columns. I eliminated the columns “D2027” and “D2028” because there are too many categories of these columns and there is not enough information the cluster these categories.

Then, I grouped the unknown data of the columns. For example, in the first picture, there are 3 different values (7, 8 and 9) to refer to a missing value. In column “D2003” there is an abnormal value (96) that is not indicated in the questionnaire script. I also added this value to the group.

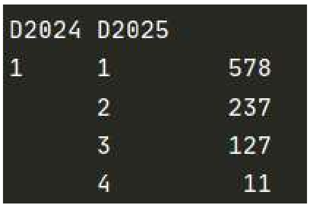
Even though I extracted some columns from the data because of missing values, there are still some missing values to handle in some columns. Especially ordered and numerical columns, we should decrease the ratio of these values. To do that, I looked over the data names and detected some potential columns which can be related to each other.

1. D2021(number in household in total) & D2022(number of children in household under age 18) & D2023(number in household under age 6):

These three columns are numerical columns. Logically, D2021 >= D2022 >= D2023 and also data shows these equations are satisfied. So, I filled in the missing values in these columns. When randomly assigning the missing values, I protected these equations. The details can be seen in the coding part.

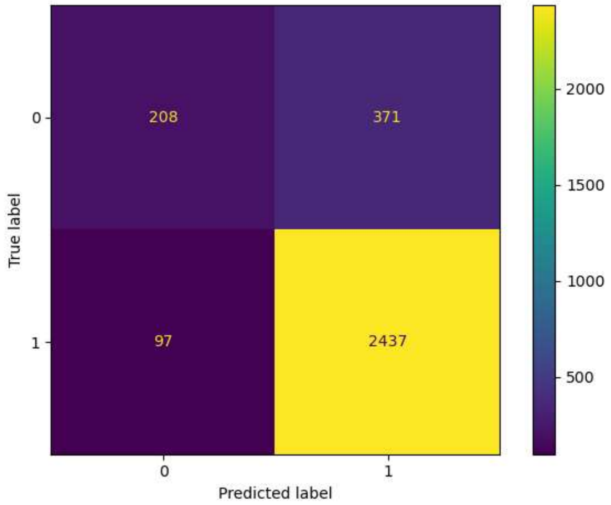
2. D2024(religious services attendance) & D2025(religiosity)

Both columns are ordered and represent the people’s belief in religion. When eliminating missing values, the correlation between these columns is greater than 0.6. When filling the missing values in the columns, I firstly looked at the columns by grouping them. For example, the picture shows how the values in “D2025” are distributed in individuals with 1 in “D2024”. It is seen that individuals with 1 in “D2024” chose 1 and 2 in “D2025” mostly. So, for the missing value in “D2025” if it is chosen 1 in “D2024”, I randomly assigned 1 or 2 to “D2025”.



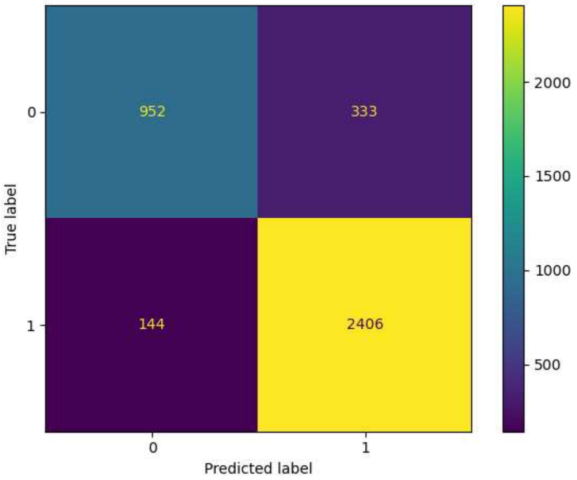
```
D2024 D2025
1     1        578
      2        237
      3        127
      4         11
```

After handling missing data, I trained some machine learning models on data such as decision tree, random forest and XGBoost. I observed that although the accuracy scores (0.77-0.85) are high for all models, AUC scores are relatively low (0.61-0.65). Thus, I plotted the confusion matrix of the XGBoost model (because it is better than others).



The confusion matrix demonstrates that the model failed to detect label 0 because the model estimated Label 0 wrong more than fifty percent. Maybe the model cannot learn Label 0 effectively. Label 1 is approximately 5 times as large as label 0. We may oversample Label 0 and train the model with oversampled data.

I increased Label 0 to be one-half of Label 1. Then, I trained the data in XGBoost. The AUC score increased from 0.66 to 0.84 which is quite successful. The confusion matrix is as follows. In Label 0, true labels are 3 times higher than false labels.


True label
Predicted label