Exploratory Data Analysis

1. We have 14 features Id, 8+1 categorical variable(Response is binary), 4 continuous variable.
    - Categorical --> City_Code, Region_Code(Nominal), Accommodation_Type(Binary), Reco_Insurance_Type(Binary), Is_Spouse(Binary), Health Indicator, Holding_Policy_Type, Reco_Policy_Cat, Response
    - Continuous --> Upper_Age, Lower_Age, Holding_Policy_Duration, Reco_Policy_Premium
2. We have null values Health Indicator(22%), Holding_Policy_Duration(40%) and Holding_Policy_Type(40%)
3. Company have reach upto 36 different cities.
4. People from cities C1, C2, C3 and C4 have filled most forms i.e either these locations are close to FinMan office or they have ran a advertisement there.
5. Some locations have very few leads i.e either of them is possible
    - They are impressed with the previous services provided by the company and want to continue with them.
    - or friends or family member has recommended the company policy
6. C1, C2, C3 and C4 are big cities i.e its populous and people are prosperous as they can pay premium of almost 15k .
7. Most of the people from C1 who filled the form have their own house.
8. Having own house is one of the big indicator of being prosperous.
9. People from C1, C2, C3 and C4 are more educated and have capacity to take health insurance thats why they filled the form.
10. Dataset is almost balanced in terms of Accommodation_type.
11. In case of joint insurance type -
    - upper_age = 75
    - lower_Age = 22
    - is_spouse = No
    - Upper_Age can be fathers age, lower_age can be child's age.
12. Most of them are recommended with Individual insurance type.
13. Youngsters are recommended more with individual edition
14. Old age people are recommended with Joint edition
15. Lower_Age of joint insurance type is multimodal because it is possible that father is taking joint edition for whole family. So, father's age is 50(upper_age) and child age is 20 or 30 (lower_age)
16. Most of the married lives in own house.
17. Most of the people are in X1 and X2 health indicator.
18. On binning the age (code)
    A. 0 to 30 --> Kids (1)
    B. 30 to 48 --> Young (2)
    C. 48 to 65 --> Old (3)
    D. Otherwise Very old (4)
19. Most of the people living in their own house have taken policy for more than 14 years (almost 3 times of rented).
20. People living in rented house prefer to take policy for 1-2 years.
21. Holding Policy 3 is preferred if individuals if taken for more than 14 years.
22. From the analysis we got to know that - **Health Indicator is Ordinal, Holding_Policy_Type is Nominal, Reco_Policy_Cat is Ordinal**
23. This is a unbalanced class problem.
24. Premium is almost normally distributed
25. Company want to market mainly policy 18, 21 and 22
26. In health Indicator X1 represents good health and X9 is bad health.

Approach

Preprocessing

1. We have null values in Health indicator, Holding policy duration and Holding Policy type.
2. Is Null data representing any information? --> `YES`

Null in Health Indicator

- It means data is unavailable. Now in what cases data can be unavailable?
- Data collection team deleted that data due to security reasons.
- Customer didn't bring the medical report.
- Providing medical report is not mandatory.
- Company was not collecting (not made mandatory) at the time of subscription.

Holding policy duration / type

- They are the new customers for the company as they have never took any policy before.
- We will represent them with a new value.

1. Since the dataset is unbalanced so we will use either oversampling or undersampling.

Baseline Model (VERSION 1)

- Fill NA with -999
- One Hot Encode all categorical data
- Performed oversampling with SMOTE and got AUC = 0.85 for random forest.

Feature engineering (VERSION 2)

- `14+` in Holding_Policy_Duration will be represented by `15`
- Remove Region_code for now
- Encode accomodation type as-
    - 0 --> Rented
    - 1 --> Owned
- Encode insurance type as-
    - 0 --> Individual
    - 1 --> Joint
- Encode is_spouse type as-
    - 0 --> No
    - 1 --> Yes
- Fill NA in Health indicator with X0 and one hot encode it.
- One hot encode city_code and drop city_code
- Split the data.(**BREAKPOINT 1**)
- Use standard scaler for reco_policy_premium as its almost normally distributed.
- Use MinMax scaler for upper_Age and lower_age as they don't have outlier but we need to scale it in between 0 and 1.
- Oversample the train set.
- With XGB classifier we get AUC of 0.83

  ## Trying to improve model (VERSION 3)
- Clone the dataframe upto **BREAKPOINT 1** to a new variable.
- Use log tranformation for all upper_age, lower_age and reco_policy_premium.
- Build the model with oversampling.
- we will get AUC = 0.83 but on the public score board the score improved to 61%

  # In VERSION 12 doing some more feature engineering
    - Encode city based on bussiness value -

```
city_value = {
    '1':['C29', 'C35', 'C34', 'C32', 'C20', 'C4'],
    '2':['C23', 'C25', 'C33', 'C16','C6', 'C13'],
    '3':['C9', 'C14', 'C28', 'C27', 'C18', 'C10'],
    '4':['C5', 'C11', 'C3','C17', 'C22', 'C31'],
    '5':['C8', 'C12', 'C21', 'C7', 'C19', 'C2'],
    '6':['C15','C24', 'C26', 'C36', 'C1', 'C30'],
}
```

- Remaining will be same as **VERSION 2** upto **BREAKPOINT 1**
- create a variable to represent city and accomodation together say living_lux
- create a variable that represent both type and premium of recomended policy.
- and a meta variable, sum of all the data in the row except Response.
- Build the model with oversampling.
- **With XGB classifier we will get AUC = 0.88, 3% increase from baseline model.**

In [ ]: