Approach :

- After loading the data, I started to explore the target as well as the independent variable.
- I observed that that in region_code there are many categories which have less that ten value counts, so we decide to replace them all with 'OTHERS', for the train as well as test data
- Now we fill the NaN values
- And create a dummies columns for all categorical columns, I am not using here the get_dummies finction because it is not guaranteed to have same categories in training and testing in realistic dataset, that's why I create a set containing the common categories between test and train dataset; and then create the encoding for only them
- So I finally have 527 columns,
- And I started with feature engineering
- IO created new features from holding policy duration, is_spuce, lower and upper age, Reco_policy_premium
- I used multiple ML models as well as simple ANN model and XGBoost gave the maximum score so I used that final model
- I did model tuning but there was not much improvement so I used vanilla model.