**Health Insurance Lead Prediction:**

Analysis and Approach:

* Analysed the data to know what each field meant
* Checked for certain patterns like If ‘Is\_Spouse’=’Yes’ , then Lower\_Age>18. But saw a few outliers where lower\_age=16,17. Tried removing these rows but dd not have any effect on model performance. Also tried calculating age difference between upper age and lower age for spouses and saw the it is greater than 15 years in many places. Again this feature did not add much value.
* Performed data cleaning. Removed ‘14+’ in holding policy duration and replaced it with 20 to indicate a value greater than 14.
* Imputed missing values using mode technique for few columns.
* Tried different encoding techniques for categorical variables: Frequency, Target, Label, OneHotEncoder
* Initially used pd.getDummies for all categorical variable, but these created sparse matrix and added a lot of features. Later, switched to other encoding techniques.
* Used transformation technique to transform the reco premium feature since it was right skewed
* Tried to create new features from existing features. Used binning technique for Upper\_age and Lower\_age and few other features, but upper and lower age binning added value. Hence used these in model training.
* Used visualization to know which features are adding values. Also tried SelectKBestFeature
* Tried different models, linear and tree based. Ensemble tree based gave a good score. Hyper parameter tuning did not boost performance.
* Used cross validation for model validation
* Used another approach using **NLP** technique to see if that gave performance boost. In this approach, combined all categorical variables in a single column and vectorized it using Tf-Idf. This did not add much value as well.
* Tried Neural network Sequential model as well but did not help much.
* The final model is giving roc-auc of 0.674 on train data and 0.64 on test data.