**About me:**

I am Anupam Misra. Currently I am pursuing PGPDS at Praxis Business school, Kolkata. I have 31 months’ experience at Reliance Industries limited working at their largest manufacturing division at Jamnagar.

**Initiation:**

The necessary libraries are loaded.

**Data:**

The data is uploaded to Google Colab and stored in dataframes

**Preprocessing:**

There is a lot of missing data in three columns.

1. For two columns, ‘Holding\_Policy\_Duration’ and ‘Holding\_Policy\_Type’ unavailable data is interpreted as the person had not subscribed to any Health insurance. So their value is set to 0.
2. For some joint account holders, their joint account holder is not their spouse. The joint account holder is assumed to be their child.
3. If a person owns and house and has individual health insurances for each member of the family, I can assume that he is wealthy. On the other hand, if a person lives in a rented apartment and has a joint family health insurance we can assert that he would be careful with his expenses.
4. All the categorical data is converted to numerical data through label encoding or simple coding.
5. For the missing data in Health indicator column, I could not find any dependent feature so I used KNN Imputer to fill those values with relation to closest values in other features.

**Machine learning model:**

I tried different classification models for classifying the data. The standalone base estimator which had highest ROC\_AUC score was KNN with 7 nearest neighbors and KD tree. KD tree outperformed ball tree because there aren’t so many features to create the requirement of hyper spheres.

After that I tried several bagging and boosting algorithms and Random Forest performed really well because of the way the data is structured. It puts weights on classes\_subsamples inversely proportional to the class frequencies and bootstrap samples taken for each tree. This helps in balancing out the unbalanced class problem with the data. I had also tried using SMOTE and Tomek links to create new data points but it didn’t help in predicting unseen data. Also, Random forest performed better on the holdout set. Hence I settled on Random Forest for using as the final model.

**Further discussion:**

I would want to really meet the business analysts and the managers to know more about the data and capture more features. Machine learning is a closed loop process. By exploring the importance of age features and health indicators we could segment customers more accurately. Also by knowing more details about the policy holders, their premium calculation and their policy category we could better analyse the data and understand the different types of policies and on what basis they are offered to certain customers.