**Novartis Server hack Challenge**

**APPROACH:**

**DATA:**

Given use case is binary classification problem with imbalanced data containing 1068 records of class 0 and 22788 records of class 1. Train data consisted of 23856 rows and Test had 15903 rows. Raw dataset contained 16 independent features

For creating best model, I tried with three variants of the dataset:

* Raw data with 16 features of input
* Modified data with Feature Engineering on Date Column
* Modified data with dropping Date Column – Gave best results

**Model Selection:**

* Selected Random Forest, XGBoost models as starting point as ensembling tend to work well with Imbalanced dataset when compared to other models
* Also tried with Artificial Neural Networks (ANN) but didn’t perform well enough when compared to ensemble models
* XGBoost algorithm with parameters tuning gave best Recall score when compared to other models

**Cross Validation:**

I used 5-fold Cross Validation and a 75:25 train test split

**Feature Engineering:**

* Transformed Date column to weekend column which checks whether the given date is weekend/weekday, and dropped it as it didn’t improve results significantly

Handling Missing Values:

* Only column ‘X\_12’ has more than 100 missing values, didn’t perform any imputation strategy as XGBoost known to handle missing values

**Hyperparameter Tuning:**

* This is the most important aspect of modeling. One should follow a good cross validation technique. I have used 5-fold CV.
* I have used GridSearchCV intensively to hypertune the model parameters as there are lot many parameters and it is quite CPU intensive
* I started tuning with ‘min\_child\_weight’ and ‘max\_depth’. When these 2 paremeters are tuned, I moved into tuning ‘colsample\_bytree’, ‘n\_estimators‘ and ‘subsample’
* One of the most important parameters in case of imbalanced dataset is ‘scale\_pos\_weight’. This parameter should be tuned quite carefully, as this often leads to overfitting the data
* Finally, tuned ‘gamma’ to avoid any overfitting
* Selected ‘recall’ as scoring method to find best parameters using GridSearchCV