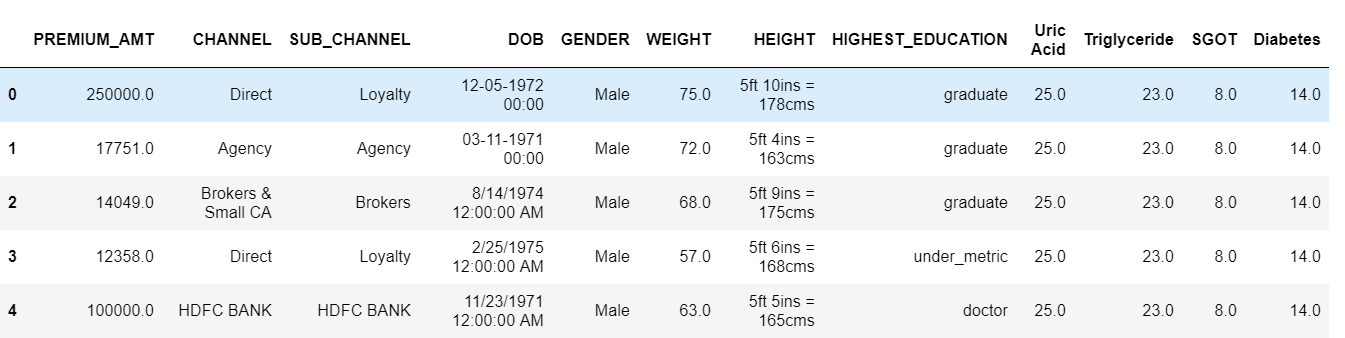
**OVERVIEW OF MODELING APPROACH**

We started with creating target variable using RATE\_UP\_REASONS as surrogate. Applicants who were rated up due to medical reasons are encoded as **1** and the rest are encoded as **0** . So all the applicants with TARGET ==1 are those who was diagnosed with some medical condition after being sent for medical screening and TARGET==0 are those cases who came out clean after medical screening .

We wanted to use DIGILYTICS dataset also for modeling purpose and because of this geographic scope of this project was limited to only 'MUMBAI', 'DELHI', 'PUNE', 'BANGALORE','KOLKATA' and 'CHANDIGARH'. For these 6 cities we had 24789 records/data points out of which 2344 are diagnosed with medical conditions.

|  |  |  |
| --- | --- | --- |
| Total records | TARGET==1 | TARGET==0 |
| 24789 | 2344 | 22445 |

Next step was to merge HDFC dataset and DIGILYTICS dataset .We defined cohorts based on [CITY ,GENDER, AGE GROUPS ] in HDFC dataset and merged DIGILYTICS dataset .



Once we had merged data , the next step was to do necessary data sanitation check and feature engineering , which includes the generation of new features using existing information . For example, BMI was computed using WEIGHT and HEIGHT etc.Similarly we have created many more features .

Next step was model fitting and since out dataset has primarily categorical features we decided go with tree based approach .We started with simple Decision tree and then tried ensambles ( gradient boosting)

But model was not generalizing well enough possibly due to class imbalance.

Next step was to treat class imbalance and for that we decided to go with SMOTE which is synthetic minority oversampling technique.We increased the minority class proportion to 0.30 in the dataset by using SMOTE .We referred to [this paper](https://jair.org/index.php/jair/article/view/10302/24590)  for reference.

After dealing with class imbalance , again we trained the model and noticed significant improvement. We used 80 percent for training and 20 percent for validation . We tried to minimize log-loss and maximize the RECALL for our model. We achieved the following performance metrics:

{'learn': {'Accuracy': 0.9414475467606397,

'Logloss': 0.17106071399785838,

'Recall': 0.7332073887489504},

'validation': {'AUC': 0.9042634159221197,

'Accuracy': 0.926273942898446,

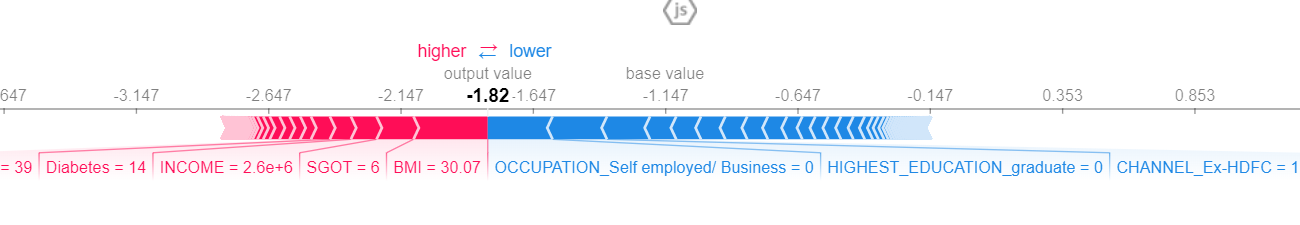
'Logloss': 0.23193646700437145,

'Recall': 0.6839080459770115}}

Further we tried to understand how our model is making predictions by examining SHAP values which is

A newly proposed tool, called **SHAP** (SHapley Additive exPlanation) **values**, allowed us to build a complex tree based model capable of making highly accurate predictions for which customers were at risk, while still allowing for an individual-level interpretation of the factors that made each of these customers .

For example ,



Red colored blocks represents the features which are increasing the overall probability while Blue color represents the opposite(i.e decrease the probability).it can be seen that this particular applicant has high BMI , SGOT and due which it has a high chance of coming out positive if sent for screening . Model predicted the same.

More on SHAP values:

<https://medium.com/civis-analytics/demystifying-black-box-models-with-shap-value-analysis-3e20b536fc80>