OBJECTIVE

To predict customers who fill-up the form, will be converted into lead or not.

Approached this problem in very traditional way.

Exploratory Data Analysis

Looking at the data it was evident that there was lot of missing values in few of the features like

Health Indicator	11691
Holding_Policy_Duration	20251
Holding_Policy_Type	20251

Secondly data type was not correct for many features and Few of the features were not carrying any information.

Most important that I have to take care of the categorical features which were string type and the Class imbalance issue.

So, from basic EDA it was very clear that before even making the model I have to do Missing value imputation as I cannot afford to lose that much of information.

Also, I have to get rid of the redundant variables which are not carrying information.
And last, I have to remove outliers to make sure data is in proper form before entry in the model.

Handling Categorical Variable::

Label Encoding

Used labeled encoding over one hot encoding as it was giving better result.

Secondly as the labels were limited so it was affecting the ML model in anyway in creating a wrong information.

Missing Value Analysis:

For missing value analysis initially used the predictive modelling but as the accuracy came out to be very low so decided to go with imputing using sklearn Simple imputer library.

Oversampling:

Used Smote+ENN for oversampling because of the way it creates synthetic samples in the backend.

Feature Importance:

Used ‘SelectfromModel’ library to get the best features from using random forest model.

After testing the features decided to go with the subset of features rather than all.

Also use chi-square test for categorical variables and correlation matrix for numeric.

Model:

Created various classification model using Random Forest, Gradient boosting, XGBoost, CatBoost, Light GBM. As the auc_roc was dropping on the validation set so decided to go with the Ensemble model.

To check the mean auc_roc score used K fold cross validation on the validation set.

Hyper Parameter Tuning

Tuned Light GBM and Random forest model for the best auc_roc score.

Right am using the hyper tuned Ensemble model of Light GBM, Random Forest and Cat Boost