Football Hackathon Approach

by Team 'Heritage is Real'

Team Members:
Alyal Samal (alyalsonu@gmail.com)
Ashis Kumar Parida (ashiskumarparida73@gmail.com)

Data Wrangling:

1. There are more than 50 columns which have more than 80% Null Values and some of them even had 100% Null Values.
2. There are columns which seem to have continuous values, but in reality, they have multimodal distribution. (Especially Raw vars)
3. There are plenty of columns with skewed Distribution. (Especially Derived and Ratio Vars)

Approach :

1. For Columns with 100% Null Values, we decide to drop them. And for the Rest of the Columns, we adopted the mean Imputation Technique.
2. We decided to apply Label Encoding to specific Raw Vars which had Multimodal Distribution.
3. We applied Quantile Transformation to specific 'Derived' and 'Ratio' Columns which had skewed distribution.
4. Then we modelled the data using mainly 3 kinds of Regressors, i.e, XGBoost, CatBoost and LGBMRegressor. We also tried Blending Various Model Outputs. But We finalised a stacked model of CatBoost and LGBMRegressor.
5. A more Significant Model was CatBoost, where we specifically provided input for the specific columns to be considered categorical columns.