Design Document for Jobathon

**Approach on High level**

* Data pre-processing followed by model training and tuning is used.
* A standalone Catboost ensemble classifier was trained with blending on seeds using StratifiedKFold cross validation.
* Using the above model it was determined that what features and hyperparameters are important which were used in further step.
* Finally a Stacked ensemble classifier with combination of Xgboost , Catboost and Lightgbm is used along with tuned hyperparameter tuning and feature engineering information provided by standalone catboost model.

**Feature engineering**

* Out of all features, Credit\_product was the most important feature followed by occupation and vintage.
* Hence mean, max, min, mode, count, nunique type of aggregation for Credit\_product, Occupation on other remaining categorical variables from level of 1 till 4 was applied for adding more features.
* Adding features using frequency encoding on aggregation of Credit\_product with others resulted in around 10% boost in accuracy.
* Missing value imputation using *mode* before frequency encoding resulted in low auc score so mode was avoided.
* Feature encoding using Label Encoder was found to be more beneficial as compared to One Hot Encoding hence label encoding was preferred .
* Discretization on numerical features using KBinsDiscretizer, boxcox transformation, normalization seems to have no or little effect on test auc\_score.

**Final model**

* Final model used is Stacked ensemble classifier with combination of Xgboost , Catboost and Lightgbm along with a hyperparameter **scale\_pos\_weight** tuned to 1.78, scale\_pos\_weight is used for imbalanced dataset.
* A learning rate of 0.03 was found to be beneficial for Catboost classifier.
* Stacked model consist of of Xgboost , Catboost and Lightgbm as base models and Catboost as meta model.
* Here Catboost among base models is used with auc\_roc\_score for prediction during the modelling phase which improved test accuracy.

**Standalone Catboost ensemble**

* This model is not final model but used for 70% of the time for training and the information provided by it was used for final modelling , it uses blending on seed with 3 ensemble model.
* Weights are assigned to each model based on the auc score in such a way that least weight is assigned to a model with highest auc score to penalize the imbalance of class which worked well.
* With the below model highest auc score on leaderboard achieved was 0.872717051326166 with the similar above configuration except encoding used was target encoding.

Author: Abhishek Soni

Email: [abhisheksoni@hotmail.com](mailto:abhisheksoni@hotmail.com)

Contact No. : +919772169889

Current Company: Verizon