WNS Analytics Wizard 2018

Strategy - :

- The most important part before entering into any predictive modelling competitions is having a solid cross validation framework. I used a 5 fold CV and a 80:20 train test split.

- I started the competition with the following 3 modeling strategy -

  Approach 1: XGBoost with extensive hyper tuning

    - Simple modeling with XGBOOST performed well on both testing and training splits with a good ROC and f1 score, without any need of much feature engineering.

  Approach 2: Created 9 different XGBoost models for 9 different department

    - Each model performed decently well on its own split, but failed miserably on the test split.
    - Had to drop this option, wasted an entire day on it.

  Approach 3: Ensemble of XGBoost, ANN and Random Forest

    - The combination of the ensemble XGBoost, ANN and random forest had a decent performance, but it was kind of overfitting, as it didn't perform that well on the test split, as I was expecting.
    - Also due to time constraints, had to eventually leave the ensemble technique too, and I got back to plain XGBoost.

Feature Engineering - :

- Missing columns education and previous_year_rating were imputed with the most frequently occurring data in them. Filled Education with 'Bachelor's' and previous_year_rating with 3.0
- Performed one hot encoding on all the categorical columns.
- Generated all possible 2 degree pair of polynomial features.
- Since a lot of interaction of columns were generated, removed all those columns which had all zeros.

Hyperparameter tuning:

- This is the most important aspect of modeling. One should follow a good cross validation technique. I used a 5 fold CV.
- One should grid search intensively on all important parameters of the model. Specially in case of XGBoost , there are lot many parameters and sometimes it could become quite CPU intensive.
- I first started tuning 'min_child_weight' and 'max_depth'.When these 2 values are tuned, I moved into tuning 'colsample_bytree' , n_estimators 'and 'subsample'.
- One of the most important parameters which people often miss 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, you should tune in 'gamma' to avoid any overfitting.

Tips –:

- Take some time in the starting to understand and analyze the data thoroughly. This could help you in feature engineering your dataset efficiently.
- The most important thing in these types of competitions is having trust in your cross validation techniques.
- You should have thorough understanding of all the tuning parameters of your model. In such close competitions where winners are separated by very small margin,if you miss tuning any single parameter , your rank can be impacted significantly.
- To increase your score further, you can train you model on the entire training set, with the same tuned parameters, instead of just training on the training split. In most of the cases this technique gives a better result.