Credit Card Lead Prediction

Jainita Fulwadwa

Approach

- The Machine learning method used for the given problem involves Random Forest Classification.
- The above is selected as is highly robust to outliers, imbalanced data and continuous data in different scales.
- Is it observed that the data is skewed and hence a robust classifier needs to be selected

Feature Engineering

- All the categorical variables except Region_Code are encoded using one hot encoding.
- The cardinality of the above features is low hence the one hot encoding is an exceptable encoding method
- Region_Code has around 35 label and cannot be encoded with one hot encoding
- Region_Code is encoded using word embeddings where 35 labels are mapped an embedding vector of size 7 as 2^7 can hold more than sufficient information for 35 labels

Over Sampling

- The data is imbalanced with only around 11% of target lable showing lead in credit cards.

- To balance the data SMOTTeck is used to generate synthetic data for less frequent lables and attain a ration of 1:2

Feature Selection

- With pearson correlation it shows that there is no significant correlations amongst features other than the word embeddings.

- Hence all the features are chosen to target variable

Hyperparameter Tuning

- Random forest classifiers have various hyperparameters, which should be tuned to attain better accuracy
- RandomisedSearchCV is used to pick out best set of hyperparamaters chosen randomly
- With cross validation it is ensured that the model is not fitted to one test set for the tuning and generalisation

Metrics

- As the data is imbalanced the roc_auc_Score is used to caclulate the model performance.

- In the spitted test it is observed to have 90% roc_auc_score with