Univ.AI

Customer Relationship Management

By Team Noobs- Aparna Wani, Divyanshu Torlikonda, Ravindra Manne, Vaishnavi M R

Motivation and About the Project

A French Telecom company Orange wants to predict the propensity of customers to switch providers (churn), buy new products or services (appetency), or buy upgrades or add-ons proposed to them to make the sale more profitable (up-selling). They want to do this from some data which they have collected and want to use this to build a model that can predict the customers' churn, appetency, and up-selling probability.

Data

Train and Test data have 230 variables with 50000 data points in 190 numerical features and 40 categorical features.

Labels

We are predicting the churn. So, it is our label

References

Data(original): https://kdd.org/kdd-cup/view/kdd-cup-2009/Data
Data(drive): https://drive.google.com/drive/folders/1qB2iqa1

Model

1.Filter out the columns with more than 70% missing values and categorical columns with more than 10 unique values.
2.Replace the missing values with the median if the column is numerical and with the most frequent value if the column is categorical.
3.Encode the categorical values and scale the whole data.
4.Remove the numerical columns with high correlation and see that the outlier does not affect our model.
5.Now split the data to train and test.
6.Now fit the train data with the following models:
   a. Logistic Regression
   b. Decision Trees
   c. Bagging
   d. Random Forest
   e. Random Forest using Smote
   f. Balanced Random Forest
   g. AdaBoost
7.Calculate all the classification metrics such as accuracy score, precision score, recall score, f1-score, and AUC score.
8.Compare the Metrics to select the most suitable model for our Data set.

Conclusion

1)The data is highly imbalanced with lots of missing values and outliers.
2)Due to the lack of documentation on the data, could not relate to the data set from the domain perspective.
3)Observed Multi-Collinearity in the data.
4)Balanced Random Forest performs best
5)As expected, Tree-based Models performed well compared to Logistic Regression.
6)Due to constraints on computational power, could not explore the power of GridSearchCV.
7)There is a scope for improvement in the performance of the model.

Future Scope

1)Work on improving the model performance by
   a. Treating the outliers.
   b. Identifying the multi-collinearity between the categorical variables.
   c. Exploring GridSearchCV with full potential.
   d. Explore feature selection methods.
2) Extend the model to predict Appetency and Upselling.
3) Learning to deal with large datasets.

Results

Model	AUC Score	F1-Score	Precision	Recall	Accuracy
Logistic Regression	0.62	0.17	0.10	0.59	0.57
Decision Trees using RandomOverSampler	0.59	0.18	0.11	0.51	0.65
Bagging using RandomOverSampler	0.52	0.13	0.08	0.50	0.53
Random Forest using RandomOverSampler	0.65	0.19	0.13	0.36	0.78
Random Forest using Smote	0.60	0.16	0.10	0.34	0.74
Balanced Random Forest	0.67	0.19	0.11	0.62	0.62
AdaBoost	0.56	0.15	0.11	0.28	0.77