Santander Customer Transaction Prediction

Abstract

Santander Bank has created a kaggle competition identifying how significant are other predictor variables in its database to identify whether a customer would perform a transaction or not. This binary classification problem has many anonymized predictor variables which are continuous in nature and some have good correlation with the target. I have evaluated our model based on evaluation metric Area under ROC curve between the predicted probability and the observed target as expected by the competition. The Kaggle summary of many important work done in areas of Exploratory Data Analysis, identification of presence of fake data present in train and test data, LGBM model with cumulative analysis of aggregating each variable's impact on increasing validation accuracy has been documented in kaggle summary notebook. Due to the observed unbalance in dataset, I have incorporated addition of some acceptable synthetic data and have performed hyperparameter tuning on LGBM model, CNN model(0.9333), found LGBM model using h20 automl for new dataset and have done Hyperparameter tuning on that as well. Additionally I have incorporated AutoML from H2O for checking which models are best performing.

Introduction

Santander Bank has a few confidential predictors that can predict if a customer will make a specific transaction the day in question based on the data available for each customer.Santander has provided an anonymized dataset with 200 predictor variables and confidential continuous numerical values associated with unique IDs. The goal of this competition is to associate each ID and respective given information to the binary target variable and predict if a customer will make a transaction or not. The values in each attribute are distributed more or less normally and hence is taken as authentic with little or no adulteration.Various machine learning algorithms render different results and the performance of the algorithm is measured in terms of area under the ROC curve of these models.However, one thing that caught my eye was the fact that the distribution of the number of unique values (across features) is significantly different between training set and test set.

Methods

Exploratory data Analysis: EDA of the dataset showed a relatively similar to normal distribution of all variables however the presence of a normal distribution in terms of mean, min, max, median and standard deviation values of rows suggested presence of features that are not authentic and synthetic. I have documented this solution by kagglers of removing the fake features from model to predict auc and get high accuracy using LGBM

Cumulative LGBM: Every variable has its own impact on auc scores, the new dataset bereft of fake values has been incorporated and stepwise cumulation has been performed to increase auc.

Addition of Synthetic Data : I have added 25000 of random data to get better idea of population from randomness added to mean of every variable.This model has given auc score of 1.

Results

Model	Validation AUC
LGBM Base	0.900462
Cumulative LGBM	0.9236
CNN (implemented on own)	0.933
LGBM(without synthetic features)	0.8896
LGBM(with synthetic features)	1
H2O AutoMl(LGBM)	0.9781

Discussion

While presence of synthetic data is highly debatable one way to look at it is removing imbalance from database to get idea of population, other methods like bootstrapping can be used to generate better results as well. The removal of fake data from data is

also one really good method to maintain data authenticity. A method to balance data with removal of fake features and padding with features of randomness is a good practice and further research has to be done to achieve a balance. Neural networks give good auc scores and Hyperparameter tuning of neural networks can generate even better results. Also applying transfer learning from humongous labelled data is feasible for institutions like Santander to map current trends while incorporating previous historical data

References

Basic Synthetic value generator formula:
https://www.sciencedirect.com/topics/computer-science/synthetic-data
https://towardsdatascience.com/https-medium-com-faizanahemad-generating-synthetic-classification-data-using-scikit-1590c1632922
https://arxiv.org/pdf/1801.06397.pdf
Grid Search:
https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html
AUC Function:
https://stackoverflow.com/questions/41032551/how-to-compute-receiving-operating-characteristic-roc-and-auc-in-keras
Kaggle:
https://www.kaggle.com/mlisovyi/lightgbm-hyperparameter-optimisation-lb-0-761
https://www.kaggle.com/c/santander-customer-transaction-prediction
https://www.kaggle.com/allunia/santander-customer-transaction-eda
https://www.kaggle.com/fayzur/lgb-bayesian-parameters-finding-rank-average
https://www.kaggle.com/fayzur/lightgbm-customer-transaction-prediction
https://www.kaggle.com/frtgnn/elo-eda-lgbm
CNN:
https://en.wikipedia.org/wiki/Convolutional_neural_network

Scope

The project was undertaken by only 1 person and best models that give highest results have been performed with help of CNN, addition of synthetic data from randomness and fake data elimination analysis, LGBM, Hyperparameter tuning of LGBM and H2O models were submitted as part of the project.

Context

The kaggle summary has been taken with code from kaggle website from various kernels cited previously. However data loading, a few nuances in data have been performed from Google colab documentation and stack overflow with modification. All codes have some similarity to the work done previously on many python files available on internet with tailor to problem at hand.