ABC Bank Ltd.

Cognext model

amol

22-08-2022

Table of contents

EXECUTIVE SUMMARY

This document covers the model development process for **XGBoost__2__AutoML__20210218__195405** model. The model is a classification model that uses xgboost with input data consisting of 20000 observations and 70 features. The model achieves Auto of **75.84%** on validation dataset and **74.98%** on Out-of-Sample (OOS) test dataset.

MODEL PERFORMANCE SUMMARY

Dataset	Size	Auto
Validation	1920	75.84%
OSS Test	1990	74.98%

DATASET

Following dataset were used for model training, tuning and OOS performance estimation:

Dataset	Size	Features	Purpose
Train	1690	70	Model training
Validation	1920	70	Hyperparameter tuning
OSS Test	1990	70	OOS performance estimation

EDA

Following is a summary of input data. Refer Annexure-1 for detailed EDA.

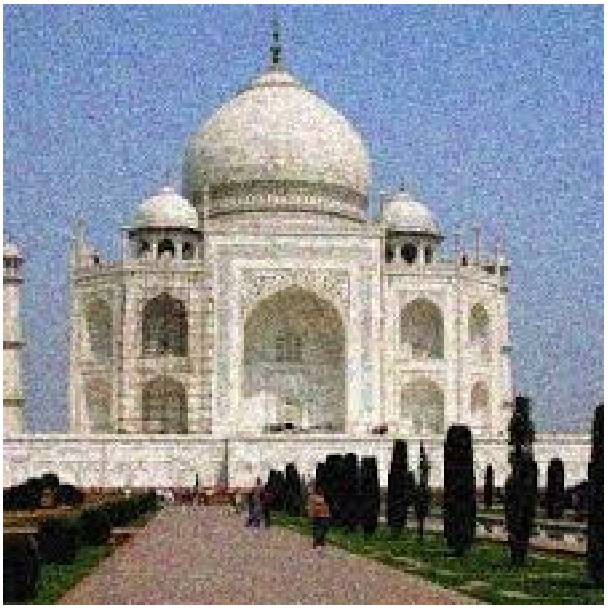
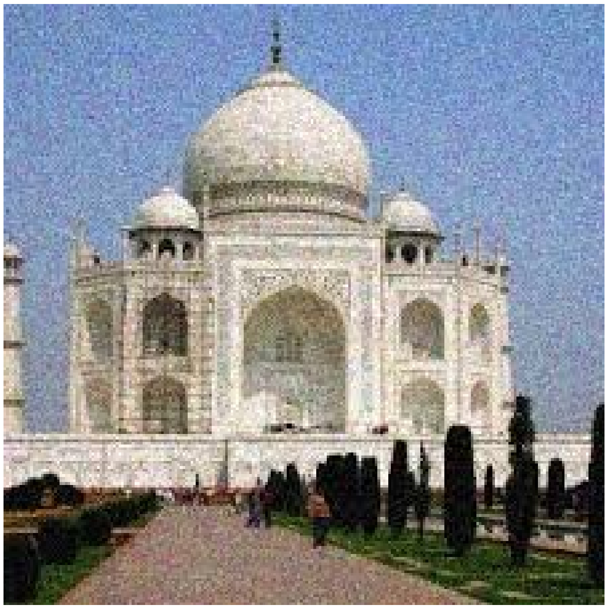


Figure 1: Image

Methodology Overview

XGBoost XGBoost is a fast and efficient implementation of gradient boosting algorithm. Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees.

Dataset	Size	Auto
Validation	1920	75.84%
OSS Test	1990	74.98%



Following is a summary of steps performed to train the model:

Data Preparation

The dataset is randomly split into train, validation and holdout test datasets. Train data is used for model fitting. Validation dataset is used for model tuning i.e. finding the optimal combination of hyperparameters that provide the best fit on a given dataset. Holdout test dataset is used to arrive at an unbiased estimate of OOS performance of the model.

Feature Transformation

Typically all features are converted into numeric features. This is a mandatory transformation for many algorithms such as XGBoost.

Model Tuning

Various models are fitted to the train dataset with multiple combination of hyperparameters (HP). These HP typically control model capacity (large capacity models will provide better fit on train data but may fail to generalize to OOS dataset), model complexity (typically models with larger capacity are also more complex) and model generalization (to prevent overfitting to train data).