ABC Bank Ltd.

**Model Development Document**

**0**

V1.0

2021-02-18

Table of Contents

[1. EXECUTIVE SUMMARY 3](#_Toc42208125)

[2. MODEL PERFORMANCE SUMMARY 3](#_Toc42208126)

[3. DATASET 3](#_Toc42208127)

[4. EDA 3](#_Toc42208128)

[5. Methodology Overview 3](#_Toc42208129)

[5.1. Data Preparation 4](#_Toc42208130)

[5.2. Feature Transformation 4](#_Toc42208131)

[5.3. Model Tuning 4](#_Toc42208132)

[5.4. Model Performance Evaluation 4](#_Toc42208133)

[5.5. Model Stability 5](#_Toc42208134)

[6. Model Details 5](#_Toc42208135)

[6.1. Model Hyperparameters 5](#_Toc42208136)

[6.2. Important Features 5](#_Toc42208137)

[6.3. Model Performance 5](#_Toc42208138)

[6.4. Model Stability 5](#_Toc42208139)

[6.5. Model Performance by Number of Features 6](#_Toc42208140)

[7. Annexure-1 Detailed EDA 7](#_Toc42208141)

## 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% |
| OOS Test | 1990 | 74.98% |

## DATASET

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

|  |  |  |  |
| --- | --- | --- | --- |
| **Dataset** | **Size** | **Features** | **Purpose** |
| Train | 16090 | 70 | Model training |
| Validation | 1920 | 70 | Hyperparameter tuning |
| OOS Test | 1990 | 70 | OOS performance estimation |

## EDA

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

## Methodology Overview

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.

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

**Figure-1**: Model Development Pipeline

## 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).

## Model Performance Evaluation

Performance of trained models is compared on validation dataset using different statistics. Final HP combination and the resultant final model is selected on basis of performance on the validation dataset.

## Model Stability

Model stability is checked by detecting drift/shift in features between train, validation and test dataset. This is done by computing Stability Index at model and individual feature level to identify if model is stable or not.

## Model Details

## Model Hyperparameters

Following is a summary of key model hyperparameters:

|  |  |
| --- | --- |
| **Statistics** | **Value** |
|  |  |

**Table**: Model Hyperparameters

## Important Features

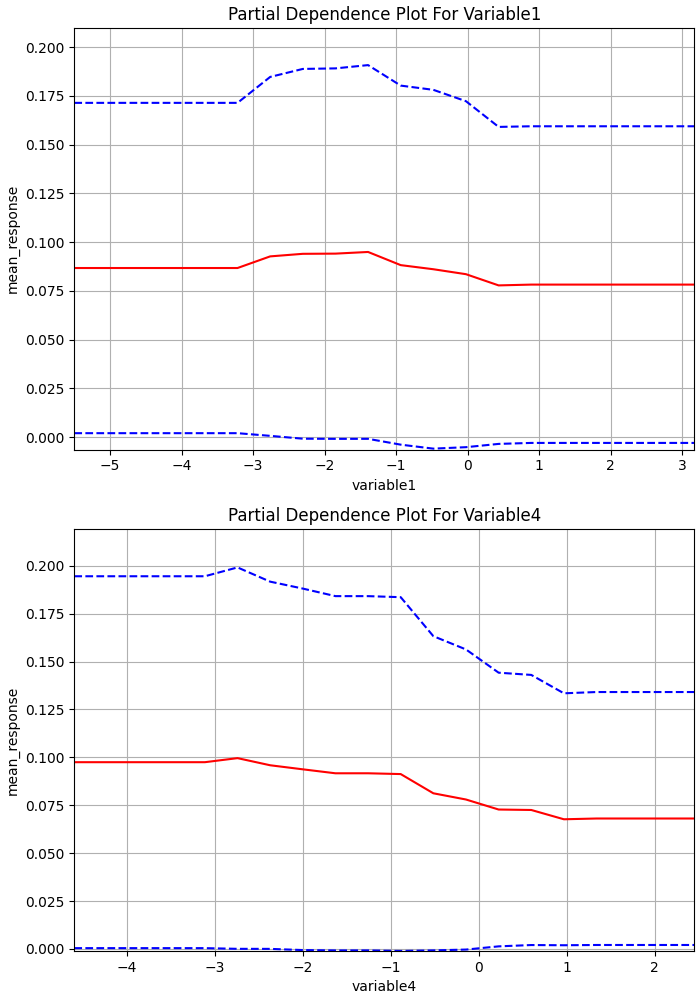
Following is a list of important features for the model:

variablerelative\_importancescaled\_importancepercentageVariable11195.51.00.04Variable41177.820.990.04Variable201086.690.910.04Variable3961.240.80.03Variable52793.180.660.03Variable18767.410.640.03Variable8764.320.640.03Variable6732.730.610.02Variable10731.080.610.02Variable17683.080.570.02Variable15667.820.560.02Variable16663.760.560.02Variable12644.810.540.02Variable5566.870.470.02Variable9554.650.460.02Variable38553.110.460.02Variable19532.290.450.02Variable23532.00.450.02Variable25526.030.440.02Variable27524.520.440.02Variable28496.890.420.02Variable26495.540.410.02Variable21489.80.410.02Variable39479.290.40.02Variable24470.640.390.02Variable45468.110.390.02Variable2465.80.390.02Variable66458.210.380.02Variable64430.40.360.01Variable13421.90.350.01Variable11405.510.340.01Variable59404.020.340.01Variable48394.00.330.01Variable14375.110.310.01Variable53372.940.310.01Variable61371.530.310.01Variable36365.270.310.01Variable31356.740.30.01Variable34354.470.30.01Variable30346.610.290.01Variable33346.340.290.01Variable56342.350.290.01Variable41330.710.280.01Variable55318.130.270.01Variable42314.630.260.01Variable46309.190.260.01Variable40306.230.260.01Variable57306.050.260.01Variable44298.250.250.01Variable22285.490.240.01Variable37284.980.240.01Variable47278.620.230.01Variable58272.660.230.01Variable29272.650.230.01Variable35260.070.220.01Variable32246.290.210.01Variable54222.40.190.01Variable62198.60.170.01Variable43198.320.170.01Variable49144.410.120.0Variable63140.770.120.0Variable7139.750.120.0Variable51132.610.110.0Variable50129.730.110.0Variable67115.060.10.0Variable6066.020.060.0Variable6938.550.030.0Variable6525.590.020.0Variable6817.120.010.0

**Table**: Global Variable Importance Table

**Figure**: Global Variable Importance Graph

Partial Dependence Plots for Top-5 variables are shown below:

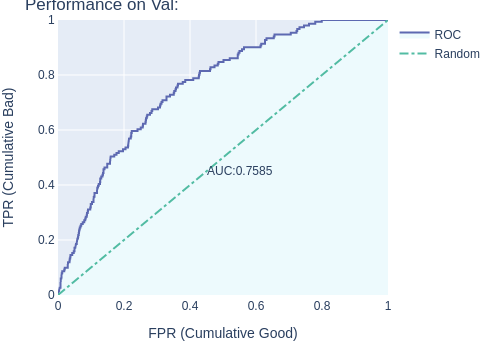


**Figure**: Partial Dependence Plot for Top Variables

## Model Performance

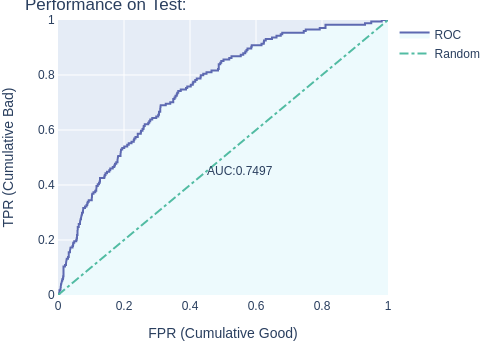
Following are the model performance statistics on validation and OOS test dataset:

**Validation dataset**



**Figure**: Model Performance on Validation dataset

**Test dataset**



**Figure**: Model Performance on Test dataset

## Model Stability

Following are model stability statistics:

**Train vs. Validation dataset**

**Figure**: Model Stability (Val vs. Train dataset)

**Validation vs. Test dataset**

**Figure**: Model Stability (Val vs. Test dataset)

## Model Performance by Number of Features

Following is a summary of model performance, if it is replaced with a model with subset of important features. This may be used to identify if final model’s performance maybe matched with a simpler model with less number of features.

## Annexure-1 Detailed EDA