Table of Contents

[LOA (Leave of Absence) 1](#_Toc120366613)

[Business Understanding- 1](#_Toc120366614)

[Data Understanding 2](#_Toc120366615)

[Data Preparation 2](#_Toc120366616)

[From Mike’s email- 2](#_Toc120366617)

[Data Source Used: 3](#_Toc120366618)

[Files and description- 3](#_Toc120366619)

[LOA-EDA.ipynb 3](#_Toc120366620)

[LOA-Pre-Processing.ipynb: 4](#_Toc120366621)

[Two demographics files are used in this notebook. 4](#_Toc120366622)

[***S3 path-*** *s3://adl-core-sagemaker-studio/external/artichauhan/LOA/loa\_script/Data/raw\_data/* 4](#_Toc120366623)

[***Two files are-*** 4](#_Toc120366624)

[*Hype\_ML\_demographics\_2022\_06\_01\_LOA\_AC.csv* 4](#_Toc120366625)

[*Hype\_ML\_demographics\_2022\_06\_01\_LOA\_ACTV\_AC.csv* 4](#_Toc120366626)

[LOA-Data-Transformation.ipynb 4](#_Toc120366627)

[LOA-Modelling.ipynb 4](#_Toc120366628)

[Baseline Model comparison on Balanced Dataset after SMOTE 4](#_Toc120366629)

[Classification Report 5](#_Toc120366630)

[ROC and Precision-Recall Curve 6](#_Toc120366631)

[Features Importance from XGBoost Model with default parameters 6](#_Toc120366632)

[Hyper-Parameter Tuning 7](#_Toc120366633)

[SHAP Values 7](#_Toc120366634)

# 

# LOA (Leave of Absence)

**Last Updated**- 26th November 2022

## Business Understanding-

Leave of Absence (LOA) is a period when the employee is away from the job., while maintaining the employment with the employer and can be paid or unpaid leave. There are multiple categories of LOA leaves that relates to Health and Welfare Plan. Government has also set up some rules and regulations for each type of LOA. Hence predicting who are likely to take LOA in next few months will help the organization proactively communicate the Health and Welfare plans to them. Alight provide comprehensive benefit package to support health, wealth, and wellbeing.

Knowing who is likely to take leave of absence will enable employers to proactively reach out with guidance on navigating the LOA process and explain how benefits coverage works while out on leave.

## Data Understanding

Identified the Leave of absence information from the *Mapped employment status code*.

Extracted the data from the Hue. Mike shared demographics file in AWS -S3

Analyzed the *Mapped employment status code* categories to create the label column

To create label column from *Mapped employment status code*, Active values are converted to 0 and all LOA categories to 1. Hence, label column includes 1 and 0.

## Data Preparation

Analyzed all individual columns with respect to the label column (EDA).

Duplicated values were removed.

Renamed the categories of few columns to reduce the duplicates and the frequency.

filled the missing values with median and unknow (column name) with respect to individual datatype of the columns.

Applied category encoder technique to convert the categorical data to numerical.

Applied the scaling technique convert different ranges to one single range.

Applied under/over sampling to make the label column balance. Over sampling worked better.

Train and test data was treated separately for missing values and outlier treatment.

dropped few columns which are not important for modeling and created final dataset.

**Note-** This use case was already performed by Harish once. Mike gave it to me to improve the modelling score. I checked Harish files in AWS to gain information. *After checking AWS S3 bucket and Harish’s folder, I could find only one csv file by the name HYPE\_ML\_demographics\_2022\_04\_14\_LOA\_HJ.csv*. Another document I had from Harish was a *Jupyer Notebook (Wellbeing-LOA-Harish.ipynb).* Later Mike provided me with the SQL file and the data (two demographic files) that I needed to use for modelling.

## From Mike’s email-

I ran the below SQL code to generate 2 files for this use case. One file contains all people on leave: **Location of the file in s3-** s3://adlcoresagemakerstudio/external/Hype/Hype\_ML\_demographics\_2022\_06\_01\_LOA\_AC.csv. (**This file contains all the employees with LOA.)**

The 2nd file contains 1M active people for the negative label:

**Location of the file in s3-**s3://adlcoresagemakerstudio/external/Hype/Hype\_ML\_demographics\_2022\_06\_01\_LOA\_ACTV\_AC.csv. **(This file contains all the Active employees.)**

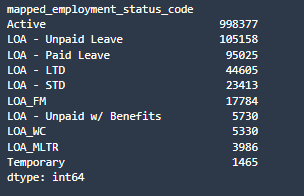
## Data Source Used:

Schema (Database)- edh analytics solutions db (QC)

Table- udp person idmapping (This table is used to get Person Internal Id)

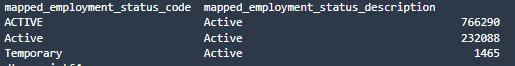
Table- participant integrated (This table is used to get other demographics columns)

**Number of counts for each LOA category-**



**Number of counts for Active category-**

Temporary status means that employee is on temporary leave. Hence, it is considered as Active.



## Files and description-

LOA-EDA.ipynb**:** In this notebook, performing visualize distribution of target variable w.r.t predictors. Performing EDA for understanding data. EDA is performed on rolled up data after combining both the file. Csv file used in this notebook is by the name “Hype\_LOA\_Active.csv”.

**S3 path**- *s3://adl-core-sagemaker-studio/external/artichauhan/LOA/loa\_script/Data/raw\_data/Hype\_LOA\_Active.csv*

LOA-Pre-Processing.ipynb:This file is created to join both the raw data files and remove the duplicates. Final data was stored as “preprocessed\_raw\_data.csv”.

Two demographics files are used in this notebook.

#### **S3 path-** s3://adl-core-sagemaker-studio/external/artichauhan/LOA/loa\_script/Data/raw\_data/

#### **Two files are-**

#### Hype\_ML\_demographics\_2022\_06\_01\_LOA\_AC.csv

#### Hype\_ML\_demographics\_2022\_06\_01\_LOA\_ACTV\_AC.csv

LOA-Data-Transformation.ipynb**:** In this notebook, pre-processed data is loaded from s3. Data is then rolled up to single category wherever there are multiple variants of one category. Further, data is split into train and test set. Other transformation is performed such as missing value imputation, outlier treatment separately on the train set and later on test set to avoid data leakage. Final train and test sets after cleaning are stored in s3.

**S3 path for preprocessed file used-** *s3://adl-core-sagemaker-studio/external/artichauhan/LOA/loa\_script/Data/raw\_data/preprocessed\_raw\_data.csv*

LOA-Modelling.ipynb**:** In this notebook, we will perform below steps

Load transformed training data and testing data

Build multiple ML model using training data

Hyper-parameter tuning of best model to identify best parameters which give good results

Evaluate model performance on various metrics and graphs such as ROC and PR curve.

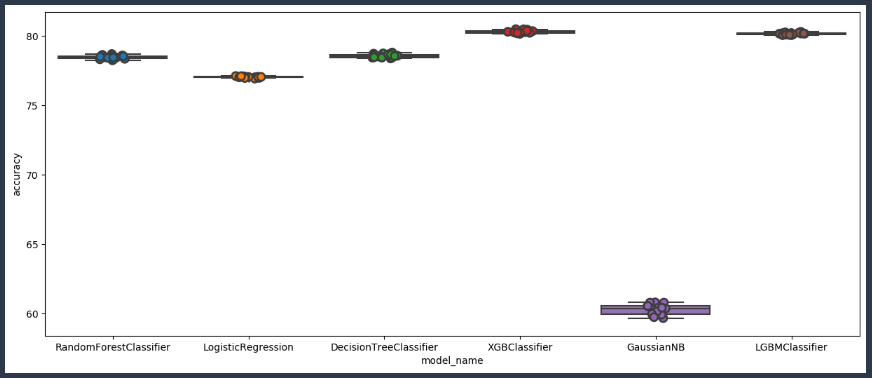
Saving the final (best) ML model in S3 which will be used to make inference on new data

Interpreting Model using SHAP values

**S3 path for train set-** *s3://adl-core-sagemaker-studio/external/artichauhan/LOA/loa\_script/Data/train\_data/transformed\_train\_data.csv*

**S3 path for test set-** *s3://adl-core-sagemaker-studio/external/artichauhan/LOA/loa\_script/Data/test\_data/transformed\_test\_data.csv (y\_test.csv is label column which is saved separately in the same location)*

## Baseline Model comparison on Balanced Dataset after SMOTE



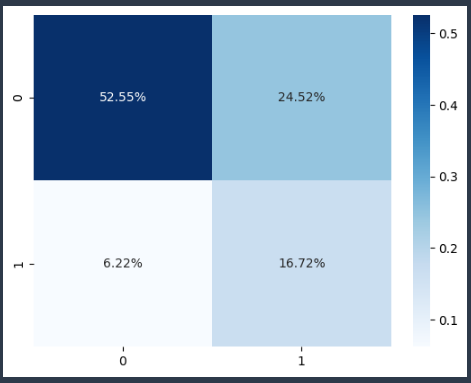
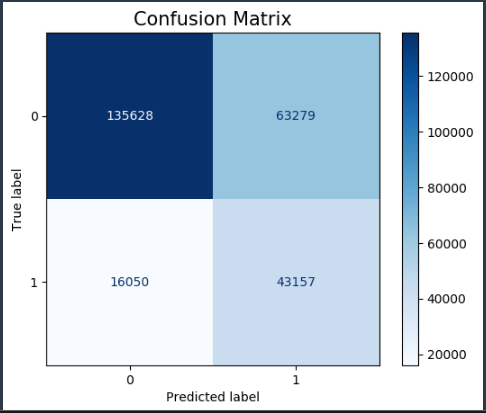
|  |  |  |
| --- | --- | --- |
| **Model Name** | **Mean Accuracy%** | **Std. Dev. Accuracy** |
| Decision Tree Classifier | 78.58 | 0.13 |
| Gaussian NB | 60.28 | 0.38 |
| LGBM Classifier | 80.17 | 0.07 |
| Logistic Regression | 77.05 | 0.05 |
| Random Forest Classifier | 78.45 | 0.13 |
| XGB Classifier | 80.32 | 0.11 |

* *SMOTE – Synthetic Minority Oversampling Technique.*
* Implemented an oversampling technique (SMOTE) to balance the dataset between the two target classes - active and loa.
* Ran different models on the balanced dataset and checked model accuracy,
* Mean accuracy is based on running K =15 folds cross validation.
* XGBoost gives us the highest accuracy (80.32%), LGBM a close second of (80.17%)

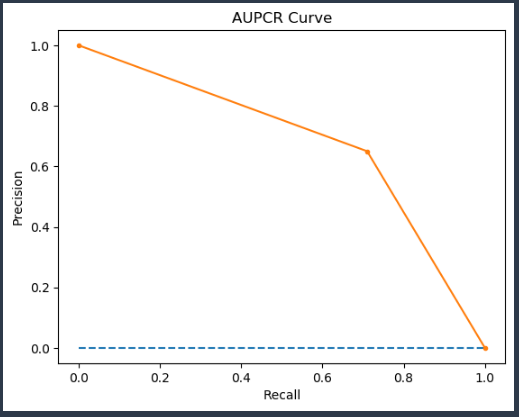
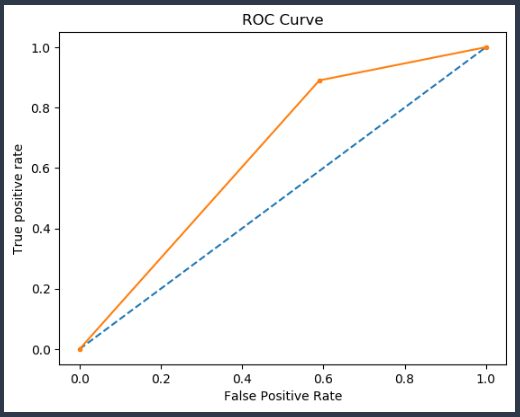
### Classification Report

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1-Score** |
| **0** | 0.89 | 0.68 | 0.77 |
| **1** | 0.41 | 0.73 | 0.52 |
| **Macro Average** | 0.65 | 0.71 | 0.65 |
| **Weighted Average** | 0.78 | 0.69 | 0.72 |
| **Accuracy** | | | 0.69 |

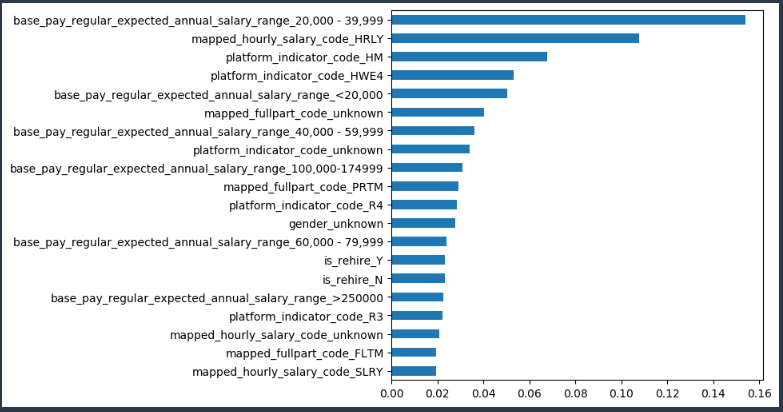
## Confusion Matrix



## ROC and Precision-Recall Curve



### Features Importance from XGBoost Model with default parameters



## Hyper-Parameter Tuning

On the balanced data, XGBoost was the model with the highest accuracy.

Bayesian Optimization (BO) – An approach that identifies hyperparameters best suited to generate a model with the highest accuracy possible.

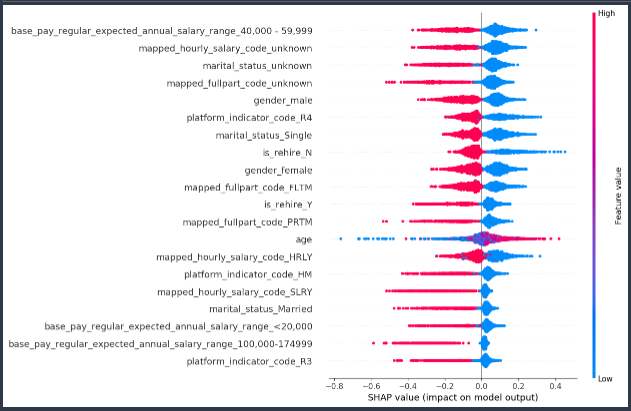
BO implemented on XGBoost.

Accuracy for XGBoost increased from 69% to 70% but recall for 1 call decreased from 73% to 71%.

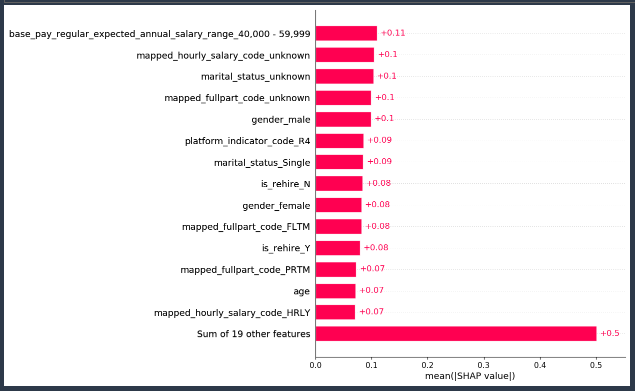
As we want to predict more accurately the LOA employees. We choose the default parameter model to be the best model.

## SHAP Values

SHAP stand for “Shapely Additive exPlanations”. Shap values are used to show feature importance. It also shows whether the feature has positive or negative impact on predictions.



Here, we can see top features who are contributing to the model output. For the topmost feature Red side is on negative, which means that lower the values of the variable higher the impact on the model output and vice-versa.



Feature importance based on mean shap values. Higher the value, more contribution towards model output.