Table of Contents

[STD (Short Term Disability) 1](#_Toc120394539)

[Business Understanding- 1](#_Toc120394540)

[Data Understanding 1](#_Toc120394541)

[Data Preparation 1](#_Toc120394542)

[Data Source Used: 2](#_Toc120394543)

[Files and description- 2](#_Toc120394544)

[STD-Pre-Processing.ipynb: 2](#_Toc120394545)

[Three raw demographics files are used in this notebook whose path are described above. Final data is saved as *preprocessed\_raw\_data.csv.* 2](#_Toc120394546)

[STD-Data-Transformation.ipynb 2](#_Toc120394547)

[STD-Modelling.ipynb 3](#_Toc120394548)

[Baseline Model comparison on Balanced Dataset after SMOTE 3](#_Toc120394549)

[Classification Report XGBOOST 4](#_Toc120394550)

[Classification Report LGBM 4](#_Toc120394551)

[ROC and Precision-Recall Curve 5](#_Toc120394552)

[Features Importance from XGBoost Model with default parameters 5](#_Toc120394553)

[Bagging Classifier: Identify Best Threshold 6](#_Toc120394554)

[Confusion Matrix for optimal threshold 7](#_Toc120394555)

[SHAP Values 7](#_Toc120394556)

# 

# STD (Short Term Disability)

Last Updated- 26th November 2022

## Business Understanding-

Short Term Disability (STD) is a type of Leave of Absence (LOA). It is a period when the employee is away from the job due to disability., while maintaining the employment with the employer and can be paid or unpaid leave. STD plans enables the insurance provider to pay a percentage of the salary if a participants become temporarily disabled. This could be due to illness, injury, or pregnancy. It can be taken from 5 to 180 days. Once STD ends and employee is still not back then LTD (Long Term Disability) begins. Hence predicting who are likely to take STD 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 STD employee 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 and LOA categories other than STD are converted to 0 and all LOA-STD 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.

## Data Source Used:

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

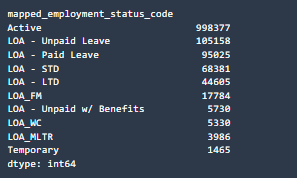
Table- participant integrated

Three different raw demographics files are used-

**Location of the file in s3-**

1. **LOA Employee Data-** s3://adl-core-sagemaker-studio/external/artichauhan/STD/Data/raw\_data/Hype\_ML\_demographics\_2022\_06\_01\_LOA\_AC.csv
2. **Active Employee Data-** s3://adl-core-sagemaker-studio/external/artichauhan/STD/Data/raw\_data/Hype\_ML\_demographics\_2022\_06\_01\_LOA\_ACTV\_AC.csv.
3. **STD Employee Data-** s3://adl-core-sagemaker-studio/external/artichauhan/STD/Data/raw\_data/Hype\_ML\_demographics\_2022\_08\_04\_STD\_AC.csv

**Number of counts for each category-**



## Files and description-

STD-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”.

Three raw demographics files are used in this notebook whose path are described above. Final data is saved as *preprocessed\_raw\_data.csv.*

STD-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/STD/Data/raw\_data/preprocessed\_raw\_data.csv*

STD-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.

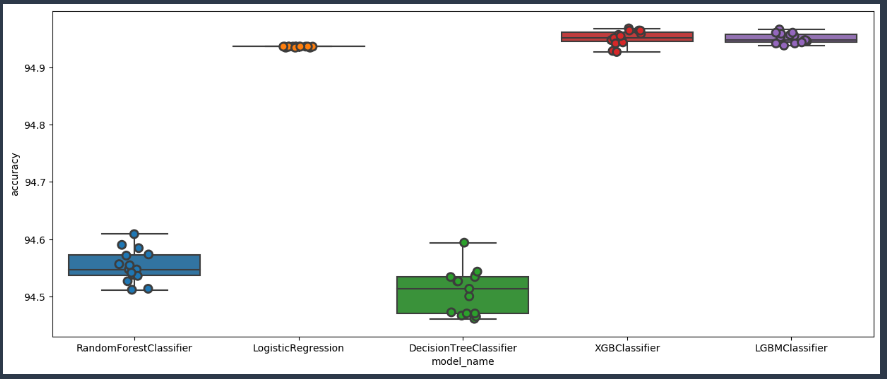
Saving the final (best) ML model in S3 which will be used to make inference on new data

Interpreting Model using SHAP values

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

**S3 path for test set-** *s3://adl-core-sagemaker-studio/external/artichauhan/STD/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 | 94.51 | 0.03 |
| LGBM Classifier | 94.95 | 0.08 |
| Logistic Regression | 94.94 | 0.0007 |
| Random Forest Classifier | 94.55 | 0.028 |
| XGB Classifier | 94.95 | 0.01 |

* *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 and LGBM gives us the highest accuracy (94.95%).

### Classification Report XGBOOST

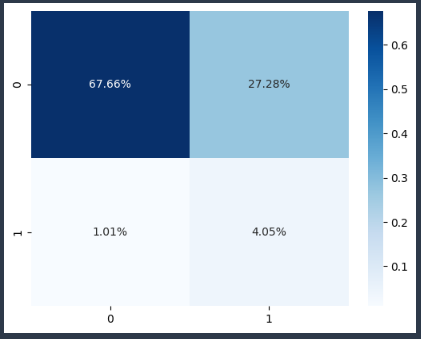
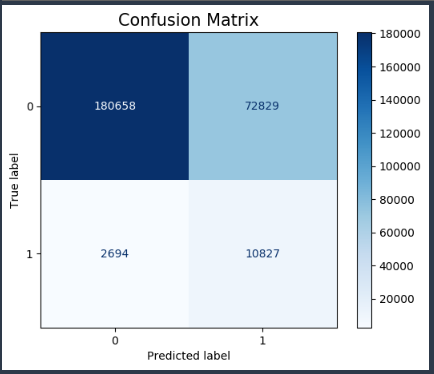
|  |  |  |  |
| --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1-Score** |
| **0** | 0.99 | 0.71 | 0.83 |
| **1** | 0.13 | 0.80 | 0.22 |
| **Macro Average** | 0.56 | 0.76 | 0.52 |
| **Weighted Average** | 0.94 | 0.72 | 0.80 |
| **Accuracy** | | | 0.72 |

### Classification Report LGBM

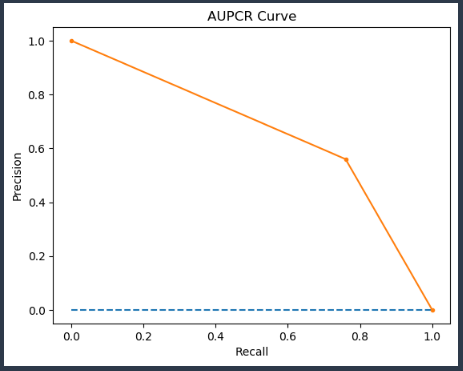
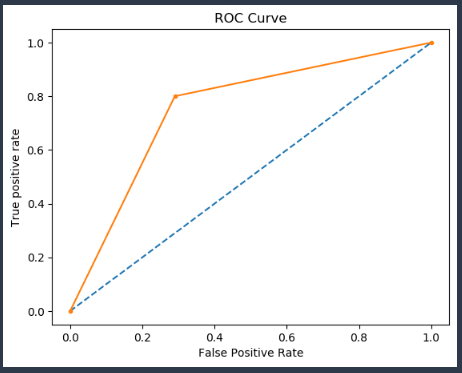
|  |  |  |  |
| --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1-Score** |
| **0** | 0.99 | 0.69 | 0.81 |
| **1** | 0.12 | 0.82 | 0.21 |
| **Macro Average** | 0.55 | 0.75 | 0.51 |
| **Weighted Average** | 0.94 | 0.70 | 0.78 |
| **Accuracy** | | | 0.70 |

After comparing all the above metrices for XGBoost and LGBM, we found that XGBoost is better with F1 score of 52%.

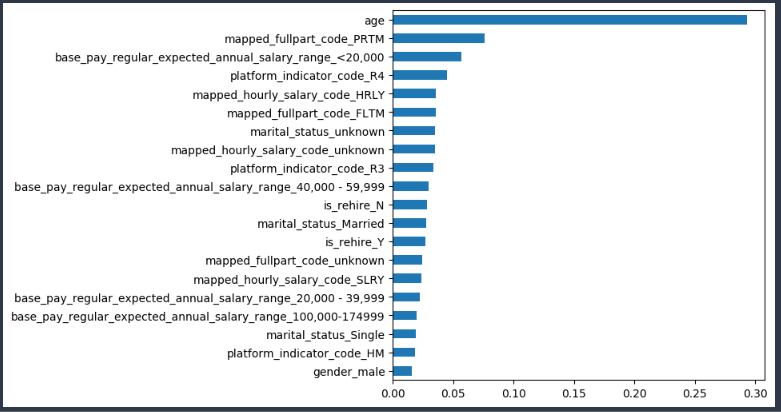
Confusion Matrix XGBoost



## ROC and Precision-Recall Curve



### Features Importance from XGBoost Model with default parameters



## Bagging Classifier: Identify Best Threshold

**Bagging Classifier**- A Bagging classifier is an ensemble meta-estimator that fits base classifiers each on random subsets of the original dataset and then aggregate their individual predictions (either by voting or by averaging) to form a final prediction.

Instead of performing Bayesian Optimization (BO) which is computationally expensive, a simple and straightforward approach to improving the performance of a classifier that predicts probabilities on an imbalanced classification problem is to tune the threshold used to map probabilities to class labels.

Though, accuracy for XGBoost remain 70% but, it increased True Positive predictions at threshold 0.476. Optimal threshold was identified for best Recall.

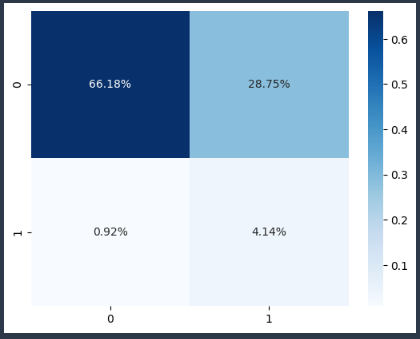
As we want to predict more accurately the STD employees. It is better to go with threshold 0.476 for predicting probabilities for new dataset and later convert these probabilities to class label.

Blog link to understand threshold-

[*https://machinelearningmastery.com/threshold-moving-for-imbalanced-classification/*](https://machinelearningmastery.com/threshold-moving-for-imbalanced-classification/)

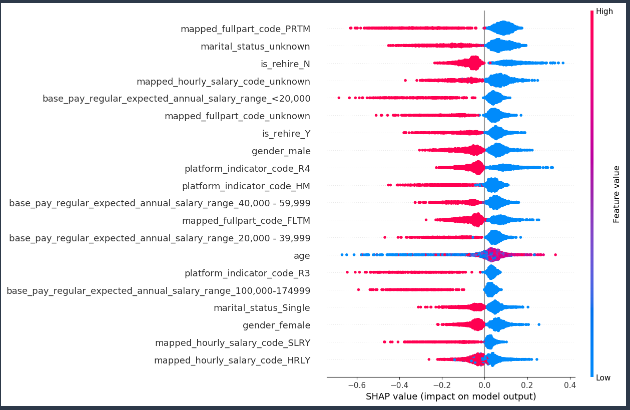
### Confusion Matrix for optimal threshold

|  |  |
| --- | --- |
| **Status** | **Predicted** |
| True Positive (TP) | 11057 |
| False Negative (FN) | 2464 |
| False Positive (FP) | 76774 |
| True Negative (TN) | 176713 |

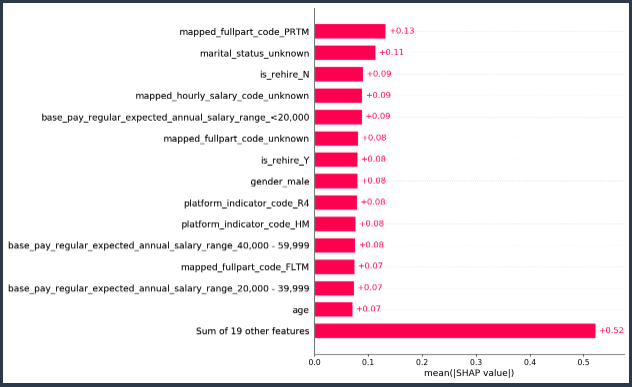


## SHAP Values

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.