Table of Contents

[Business Problem 1](#_Toc120529800)

[Data 2](#_Toc120529801)

[Files and description- 2](#_Toc120529802)

[data-pre-processing-v1.ipynb 2](#_Toc120529803)

[data-transformation-dpp0-v1.ipynb**:** 3](#_Toc120529804)

[data-modeling-v1.ipynb**:** --- 3](#_Toc120529805)

[dpp0-xgb-v1-final.ipynb:- 4](#_Toc120529806)

[New Hire Contribute to HSA – Using PHCE Table Results on Testing Set (30% of year 2021 data) | Year: **2021** 4](#_Toc120529807)

[New Hire Contribute to HSA – Using PHCE Table Results of Out-of-date Test | Year: **2020** 4](#_Toc120529808)

[New Hire Contribute to HSA – Using PHCE Table Results of Out-of-date Test | Year: **2022** 5](#_Toc120529809)

[New Hire Contribute to HSA – Using PHCE Table Results on Testing Set (30% of Data) 5](#_Toc120529810)

[Top 20 Features based on Feature Importance Score 5](#_Toc120529811)

[SHAP Values Summary Plot 6](#_Toc120529812)

[Top 15 Features based on SHAP Mean Values 7](#_Toc120529813)

Data Modeling: New Hire - Contribute to hsA

**Last Updated**: 26th November 2022

# Business Problem

A Health Savings Account (HSA) is a special bank account provided when somebody enroll in Bronze, Bronze Plus or Silver coverage level. It is a tax-free savings account. The aim of the project is to predict which new hires are likely to enroll in HDHP but are not likely to open and contribute to the HSA within 30 days. The prediction will help to identify participants who are not contributing to Health Savings Account and hence will help in taking appropriate action to ensure new hires who enroll in HDHP will also open and contribute to Health Savings Account.

Predict which new hires are likely to enroll in HDHP but are not likely to open and contribute to the HSA within 30 days.

For this exercise we tweaked the problem statement little bit, from "not likely to open and contribute to the HSA within **30 days from hire date**" to "not likely to open and contribute to the HSA within **30 days from first new hire enrollment date**". This is done due to some data constraint, the data source used does not contain hire/rehire date.

# Data

Data was fetched using SQL from Mike which was further pushed in AWS S3 bucket by him. Only **person\_hm\_choice\_enrolled\_integrated** table from **edh\_core\_health\_db** was used for this exercise.

Since all data cannot be pulled in single file, therefore two SQL queries were created to pull in parts which were further combined in Jupyter notebook using python code. SQL queries can be found in “HSA Data Pull SQL.txt” file.

Below are some AWS S3 location where raw, pre-processed data is stored –

* These two files are output of two SQL queries from HUE (~ rows)
  + s3://adl-core-dev-sagemaker-studio/external/amitmittal/New Hire - Contribute to HSA/Hype\_ML\_demographics\_2022\_06\_06\_HSA\_3x\_AM.csv
  + s3://adl-core-dev-sagemaker-studio/external/amitmittal/New Hire - Contribute to HSA/Hype\_ML\_demographics\_2022\_06\_06\_HSA\_4x\_AM.csv
* This file contains 1st level pre-processed data using above data (created target variable and brought data down to one row per ppt)
  + s3://adl-core-dev-sagemaker-studio/external/amitmittal/New Hire - Contribute to HSA/Hype\_ML\_2022\_06\_09\_PHCE\_HSA\_AM.csv
* These files contain 2nd level pre-processed data using above data. Each file is an output of unique data pre-processing steps used. Goal is to create some different sets of data pre-processing steps and use each data for modeling., and finally identify best dataset. Each file must be considered as separately, and should not be combined as they all represent same data only
  + In this file, no rollups done for categorical variable and missing values in categorical variables replaced with “Unknow”
  + s3://adl-core-dev-sagemaker-studio/external/amitmittal/New Hire - Contribute to HSA/Hype\_ML\_Partial\_Clean\_1\_HSA\_AM.csv
  + In this file, rollups are done for categorical variable and missing values in categorical variables replaced with “Unknow”
  + s3://adl-core-dev-sagemaker-studio/external/amitmittal/New Hire - Contribute to HSA/Hype\_ML\_Partial\_Clean\_2\_HSA\_AM.csv

## Files and description-

data-pre-processing-v1.ipynb:-----

* In this notebook, we will perform below steps
  1. Load raw data which is available in S3
  2. Define number of days to look forward to build target variable. For eg. it is 30 days for given problem statement
  3. Pre-process the above data, this will include below items
  4. Bring down data to one row per participant (platform\_person-internal\_id)
  5. Build target variable `label` using 30 days logic (X days logic)
  6. Handle few data exceptions
  7. Store pre-processed data in S3, which will further be transformed using another script

**S3 Path**- s3://adl-core-sagemaker-studio/external/artichauhan/HSA/train-data

Files used-

1. Hype\_ML\_demographics\_2022\_06\_06\_HSA\_3x\_AM.csv
2. Hype\_ML\_demographics\_2022\_06\_06\_HSA\_4x\_AM.csv

## data-transformation-dpp0-v1.ipynb**:**

* In this notebook, we will perform below steps
  1. Load pre-processed data
  2. Splitting the dataset into train (70%) and test (30%). Train set will be used to fit and transform the train set and transform test set
  3. Transform above data to make it ready for ML model. This step includes sub-steps
  4. Rolling up levels of categorical variables
  5. Imputing missing values of categorical variables by "Unknown"
  6. Imputing missing values of numeric variables by median based on employment status
  7. Handling data exceptions
  8. Visualize distribution of target variable w.r.t predictors
  9. Store tranformed training anf testing set in S3, which will be further used for ML modeling and evluation

**S3 path for preprocessed file used**- *s3://adl-core-sagemaker-studio/external/artichauhan/HSA/train-data/data.csv*

## data-modeling-v1.ipynb**:** ---

* In this notebook, we will perform below steps
  1. Load transformed training data and testing data
  2. Build multiple ML model using training data
  3. Hyper-parameter tuning of best model from above step to identify best parameters which give best results
  4. Evaluate model performance on various metrics and charts
  5. Saving the final (best) ML model in S3 which will be used to make inference on new data
  6. Interpreting Model using SHAP values

**S3 path for train set-** *s3://adl-core-sagemaker-studio/external/artichauhan/HSA-2/train/data.csv*

**S3 path for test set-** *s3://adl-core-sagemaker-studio/external/artichauhan/HSA-2/test/data.csv*

## dpp0-xgb-v1-final.ipynb:-

* In this notebook, below steps are performed
  1. Pre-process the new unseen raw data
  2. Transform above pre-processed data to make it ready for predictions
  3. Use pre-trained ML model to make predictions on above transformed data
* **S3 Path***- s3://adl-core-sagemaker-studio/external/artichauhan/HSA/test-data*
* **Files Used**-
  1. TBA3x\_CBA\_2020.csv
  2. TBA4x\_2020.csv

### New Hire Contribute to HSA – Using PHCE Table Results on Testing Set (30% of year 2021 data) | Year: **2021**

Best Algorithm –> XGBoost

Recall -> 73.1

Precision -> 73.1

Accuracy -> 73.1

AUC -> 81.6

AUPRC -> 84.1

F1 -> 73.1

Chart

Description automatically generatedChart, line chart, scatter chart

Description automatically generated Chart

Description automatically generated

### New Hire Contribute to HSA – Using PHCE Table Results of Out-of-date Test | Year: **2020**

Best Algorithm –> XGBoost

Recall -> 70.8

Precision -> 70.9

Accuracy -> 70.8

AUC -> 78.5

AUPRC -> 81.2

F1 -> 70.8

Chart

Description automatically generatedChart, line chart

Description automatically generated Chart

Description automatically generated

### New Hire Contribute to HSA – Using PHCE Table Results of Out-of-date Test | Year: **2022**

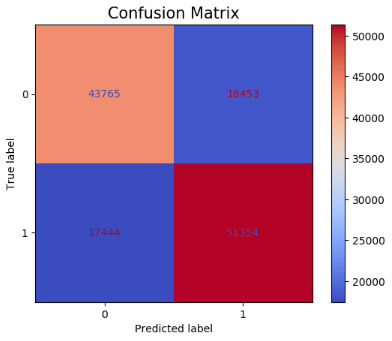
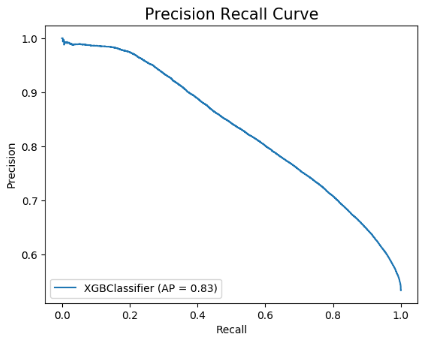
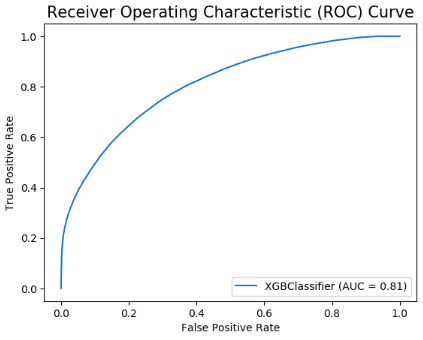
Chart

Description automatically generatedChart, line chart

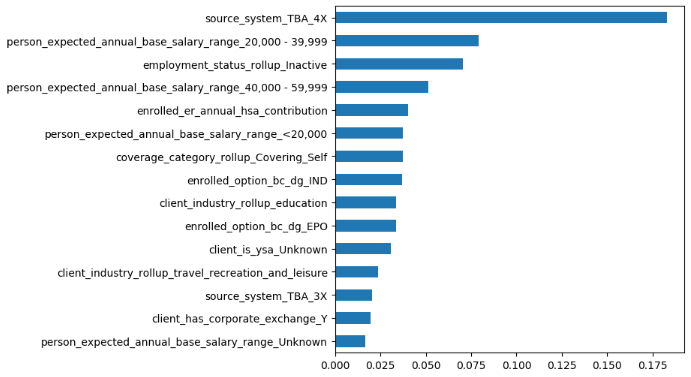
Description automatically generated Chart

Description automatically generated

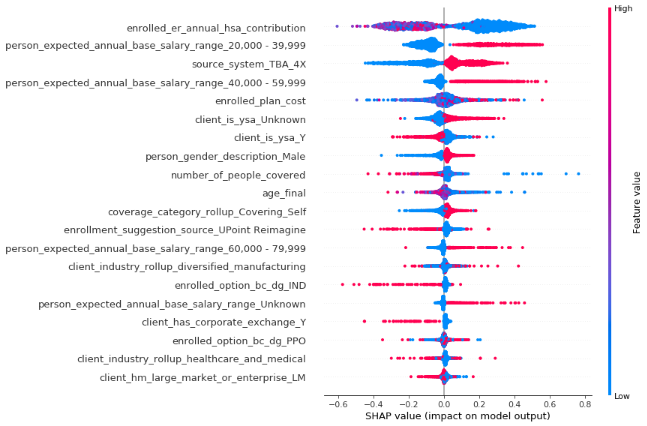
### New Hire Contribute to HSA – Using PHCE Table Results on Testing Set (30% of Data)



### Top 20 Features based on Feature Importance Score



### SHAP Values Summary Plot



### Top 15 Features based on SHAP Mean Values

