New Hire - Contribute to HSA

**Problem Statement**

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

**Goal**: Ensure new hires who enroll in HDHP will also open and contribute to HSA within 30 days.

**Data Source**

edh\_analytic\_solutions\_db.reporting\_ciba\_person\_hm\_choice\_enrolled\_integrated

**Questions**

* Do we need a very good predict model with less interpretability?
* Which target metrics we want to optimize?
  + Precision or Recall
* How is target variable defined?
  + 1 - Enrolled in HDHP and did not contribute to HSA within 30 days
  + 0 - Enrolled in HDHP and contributed to HSA within 30 days
  + 0 – Did not enroll in HDHP or/and did not contribute to HSA within 30 days
* 30 days from HDHP enrollment or hire date?

Questions from Data

* How to identify a new hire from the table?
  + mapped\_activity\_rollup\_level\_2=New Hire Enrollment
* How to identify participants enrolled in medical plan?
  + dwh\_plan\_brand\_code in ('MDCL', 'MDCL-MDCR', 'RTEE-MDCL') and
  + is\_no\_coverage\_option=0
* How to identify a participant enrolled in HDHP plan?
  + medical\_enrolled\_option\_type=HDHP and
  + is\_option\_hsa\_fixed=HSA and
  + enrolled in medical plan
    - dwh\_plan\_brand\_code in ('MDCL', 'MDCL-MDCR', 'RTEE-MDCL')
    - is\_no\_coverage\_option=0
* How to identify participant contributing in HSA?
  + enrolled\_ee\_annual\_hsa\_contribution > 0 or
  + ee\_hsa\_contribution\_flag = 1

Observations from Saurabh’s code on “Identifying new hires who are likely to enroll in HDHP medical plan but will not contribute to HSA within 30 days”

* HSAdata\_complete – This tables originally had a greater number of rows and columns
  + After importing this and doing pre-processing, it seems like Saurabh replaced original data in S3
* Dropping columns with the greatest number of NULL values
* Considering description columns instead of code columns
* Rolling-up categorical columns
* Creating Dummies
* Missing NULL values using K-Nearest Neighbor algorithm
  + After creating dummies of categorical variables, rows with missing values got mapped as 0’s i.e., by default they got imputed as some category
* Creating Age bins, creating dummies and imputing missing values
  + Issue while creating bins
* Target variables define, 1 -> participants who contributed to HSA, 0 -> not

From data it is not clear that participant contributed to HSA within 30 days of hire date

* Which data source is used in Juypter notebook?
  + From JIRA it seems Saurabh used edh\_analytic\_solutions\_db.reporting\_ciba\_person\_hm\_choice\_enrolled\_integrated
  + s3://adl-core-sagemaker-studio/external/Saurabh/HSAdata\_complete.csv
  + s3://adl-core-sagemaker-studio/external/Saurabh/HSD\_Main.csv
  + s3://adl-core-sagemaker-studio/external/Saurabh/HSdata\_Complete.csv

Data Prep

* Filter participants who are **new hire** and enrolled in **HDHP** type of **medical plan**
* Create target variable 1/0 type,
  + if participant not contributing to HSA within 30 days of hire date, then 1
  + else 0
* Create one row per client participant per enrolled year
* How to identify unique participant
  + Person internal id
  + Client Id
  + Platform Id
* Are we okay with person internal id coming 2 times in 2 different plan years?
* Which features to be used as independent variables?
  + Demographic
    - 'gender', 'marital\_status', 'is\_union','is\_rehire', 'is\_db\_elig', 'is\_dc\_elig', 'is\_hm\_elig', 'is\_hcm\_elig','mapped\_employment\_status\_description', 'mapped\_fullpart\_description','mapped\_permanent\_temporary\_description','mapped\_hourly\_salary\_description', 'mapped\_flex\_status\_description','client\_industry','coverage\_category\_group', ‘age\_bin’
  + S

TBA 4x, CBA - logic we have valid

TBA 3x (older version of TBA) - their data stored differently

Observations on data pre-processing after data received from Mike

* Number of rows coming after running the SQL at our end is different from number of rows we received from Mike
  + Data we received has lesser rows than expected
  + RCA – This possibly because of limitation of excel to have
* Number of rows changing after running same SQL on different days at our end
* As per the SQL and data received from Mike, it was expected that all people in the dataset had new hire enrollment in HDHP medical plan at some point of time in given plan year (2021). However, this does not seem to be the case. It was observed that there are 279,530 people in the dataset and 236,943 people had enrollment
  + This can happen because of some missing for those people or something else
* We started with 279,530 people and found only 236,943 people had new hire enrollment in HDHP medical plan in given plan year

Data Sources

Below two files contains raw data from HUE (given by Mike)

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

Below file contains pre-processed data (created target variables and 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

Below are two files which are further pre-processed from above step, these are partially cleaned

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

Above two files are considered as separate data pipelines and further pre-processing like imputing missing values for numerical variables and handling outliers done before passing data to the model.

# Hyper Parameter Tunning XGBoost

<https://www.analyticsvidhya.com/blog/2016/03/complete-guide-parameter-tuning-xgboost-with-codes-python/>

Baysian Optimization

* <https://towardsdatascience.com/bayesian-optimization-concept-explained-in-layman-terms-1d2bcdeaf12f>
* <https://distill.pub/2020/bayesian-optimization/>
* https://github.com/fmfn/BayesianOptimization
* <https://proceedings.neurips.cc/paper/2018/file/498f2c21688f6451d9f5fd09d53edda7-Paper.pdf#:~:text=Bayesian%20optimization%20is%20a%20sample-ef%EF%AC%81cient%20approach%20to%20global,heuristics%20%28acquisition%20functions%29%20to%20guide%20its%20search%20process> (from Arti)
* <https://stats.stackexchange.com/questions/492727/bayesian-optimization-with-xgb-cv-and-xgb-xgbclassifier-mismatch-between-auc-s>
* <https://analyticsindiamag.com/implementing-bayesian-optimization-on-xgboost-a-beginners-guide/>
* <https://ayguno.github.io/curious/bayesopt.html>

PipeLine in Python

* <https://towardsdatascience.com/pipelines-custom-transformers-in-scikit-learn-the-step-by-step-guide-with-python-code-4a7d9b068156>
* <https://towardsdatascience.com/step-by-step-tutorial-of-sci-kit-learn-pipeline-62402d5629b6>