Package ‘PatientLevelPrediction’

April 26, 2024

Type Package

Title Developing patient level prediction using data in the OMOP Common Data Model

Version 6.3.7.9999

Date 2024-04-04

Maintainer Jenna Reps <jreps@its.jnj.com>

Description A user friendly way to create patient level prediction models using the OMOP common data model. Given a
cohort of interest and an outcome of interest, the package can use data in the
OMOP Common Data Model to build a large set of features. These features can then
be assessed to fit a predictive model using a number of machine learning algorithms.
Several performance measures are implemented for model evaluation.

License Apache License 2.0

URL https://ohdsi.github.io/PatientLevelPrediction, https://github.com/OHDSI/PatientLevelPrediction

BugReports https://github.com/OHDSI/PatientLevelPrediction/issues

VignetteBuilder knitr

Depends R (>= 4.0.0)

Imports Andromeda,
Cyclops (>= 3.0.0),
DatabaseConnector (>= 6.0.0),
dplyr,
FeatureExtraction (>= 3.0.0),
ggplot2,
gridExtra,
Matrix,
memuse,
mgcv,
ParallelLogger (>= 2.0.0),
polspline,
pROC,
PRROC,
reticulate (>= 1.30),
rlang,
SqlRender (>= 1.1.3),
survival,

tidyr,
utils

Suggests AUC,
BigKnn (>= 1.0.0),
devtools,
Eunomia,
IterativeHardThresholding,
knitr,
markdown,
Metrics,
parallel,
plyr,
pool,
readr,
ResourceSelection,
ResultModelManager (>= 0.2.0),
rmarkdown,
RSQLite,
scoring,
ShinyAppBuilder (>= 1.1.1),
survminer,
testthat,
withr,
xgboost (> 1.3.2.1),
lightgbm

Remotes ohdsi/BigKnn,
ohdsi/Eunomia,
ohdsi/FeatureExtraction,
ohdsi/ShinyAppBuilder,
ohdsi/ResultModelManager,

RoxygenNote 7.3.1

Encoding UTF-8

Contents

`accuracy` *Calculate the accuracy*

Description

Calculate the accuracy

Usage

```
accuracy(TP, TN, FN, FP)
```

Arguments

`TP`	Number of true positives
`TN`	Number of true negatives
`FN`	Number of false negatives
`FP`	Number of false positives

Details

Calculate the accuracy

Value

accuracy value

`addDiagnosePlpToDatabase`
Insert a diagnostic result into a PLP result schema database

Description

This function inserts a diagnostic result into the result schema

Usage

```
addDiagnosePlpToDatabase(
  diagnosePlp,
  connectionDetails,
  databaseSchemaSettings,
  cohortDefinitions,
  databaseList = NULL,
  overWriteIfExists = T
)
```

Arguments

`diagnosePlp`	An object of class `diagnosePlp`
`connectionDetails`	A connection details created by using the function `createConnectionDetails` in the `DatabaseConnector` package.
`databaseSchemaSettings`	A object created by `createDatabaseSchemaSettings` with all the settings specifying the result tables
`cohortDefinitions`	A set of one or more cohorts extracted using ROhdsiWebApi::exportCohortDefinitionSet()
`databaseList`	(Optional) If you wish to overwrite the settings in the plp object use `createdatabaseList` to specify the databases
`overWriteIfExists`	(default: T) Whether to delete existing results and overwrite them

Details

This function can be used to upload a diagnostic result into a database

Value

Returns NULL but uploads the diagnostic into the database schema specified in databaseSchemaSettings

addMultipleDiagnosePlpToDatabase

Insert mutliple diagnosePlp results saved to a directory into a PLP result schema database

Description

This function inserts diagnosePlp results into the result schema

Usage

```
addMultipleDiagnosePlpToDatabase(
  connectionDetails,
  databaseSchemaSettings,
  cohortDefinitions,
  databaseList = NULL,
  resultLocation
)
```

Arguments

`connectionDetails`	A connection details created by using the function `createConnectionDetails` in the `DatabaseConnector` package.
`databaseSchemaSettings`	A object created by `createDatabaseSchemaSettings` with all the settings specifying the result tables
`cohortDefinitions`	(list) A list of cohortDefinitions (each list must contain: name, id)
`databaseList`	(Optional) ...
`resultLocation`	The location of the diagnostic results

Details

This function can be used to upload diagnosePlp results into a database

Value

Returns NULL but uploads multiple diagnosePlp results into the database schema specified in databaseSchemaSettings

addMultipleRunPlpToDatabase

Populate the PatientLevelPrediction results tables

Description

This function formats and uploads results that have been generated via an ATLAS prediction package into a database

Usage

```
addMultipleRunPlpToDatabase(
  connectionDetails,
  databaseSchemaSettings = createDatabaseSchemaSettings(resultSchema = "main"),
  cohortDefinitions,
  databaseList = NULL,
  resultLocation = NULL,
  resultLocationVector,
  modelSaveLocation
)
```

Arguments

`connectionDetails`	A connection details created by using the function `createConnectionDetails` in the `DatabaseConnector` package.
`databaseSchemaSettings`	A object created by `createDatabaseSchemaSettings` with all the settings specifying the result tables
`cohortDefinitions`	A set of one or more cohorts extracted using ROhdsiWebApi::exportCohortDefinitionSet()
`databaseList`	(Optional) A list created by `createDatabaseList` to specify the databases
`resultLocation`	(string) location of directory where the main package results were saved
`resultLocationVector`	(only used when resultLocation is missing) a vector of locations with development or validation results
`modelSaveLocation`	The location of the file system for saving the models in a subdirectory

Details

This function can be used upload PatientLevelPrediction results into a database

Value

Returns NULL but uploads all the results in resultLocation to the PatientLevelPrediction result tables in resultSchema

addRunPlpToDatabase *Function to add the run plp (development or validation) to database*

Description

This function adds a runPlp or external validation result into a database

Usage

```
addRunPlpToDatabase(
  runPlp,
  connectionDetails,
  databaseSchemaSettings,
  cohortDefinitions,
  modelSaveLocation,
  databaseList = NULL
)
```

Arguments

runPlp	An object of class `runPlp` or class `externalValidatePlp`
connectionDetails	A connection details created by using the function `createConnectionDetails` in the `DatabaseConnector` package.
databaseSchemaSettings	A object created by `createDatabaseSchemaSettings` with all the settings specifying the result tables
cohortDefinitions	A set of one or more cohorts extracted using ROhdsiWebApi::exportCohortDefinitionSet()
modelSaveLocation	The location of the directory that models will be saved to
databaseList	(Optional) If you want to change the database name then used `createDatabaseList` to specify the database settings but use the same cdmDatabaseId was model development/validation

Details

This function is used when inserting results into the PatientLevelPrediction database results schema

Value

Returns a data.frame with the database details

`averagePrecision` *Calculate the average precision*

Description

Calculate the average precision

Usage

```
averagePrecision(prediction)
```

Arguments

`prediction`	A prediction object

Details

Calculates the average precision from a predition object

Value

The average precision

`brierScore` *brierScore*

Description

brierScore

Usage

```
brierScore(prediction)
```

Arguments

`prediction`	A prediction object

Details

Calculates the brierScore from prediction object

Value

A list containing the brier score and the scaled brier score

`calibrationLine` *calibrationLine*

Description

calibrationLine

Usage

```
calibrationLine(prediction, numberOfStrata = 10)
```

Arguments

`prediction`	A prediction object
`numberOfStrata`	The number of groups to split the prediction into

Details

Calculates the calibration from prediction object

`computeAuc` *Compute the area under the ROC curve*

Description

Compute the area under the ROC curve

Usage

```
computeAuc(prediction, confidenceInterval = FALSE)
```

Arguments

`prediction`	A prediction object as generated using the `predict` functions.
`confidenceInterval`	Should 95 percebt confidence intervals be computed?

Details

Computes the area under the ROC curve for the predicted probabilities, given the true observed outcomes.

computeGridPerformance *Computes grid performance with a specified performance function*

Description

Computes grid performance with a specified performance function

Usage

```
computeGridPerformance(prediction, param, performanceFunct = "computeAuc")
```

Arguments

prediction	a dataframe with predictions and outcomeCount per rowId
param	a list of hyperparameters
performanceFunct	a string specifying which performance function to use . Default “'compute_AUC'“

Value

A list with overview of the performance

configurePython *Sets up a virtual environment to use for PLP (can be conda or python)*

Description

Sets up a virtual environment to use for PLP (can be conda or python)

Usage

```
configurePython(envname = "PLP", envtype = NULL, condaPythonVersion = "3.11")
```

Arguments

envname	A string for the name of the virtual environment (default is 'PLP')
envtype	An option for specifying the environment as'conda' or 'python'. If NULL then the default is 'conda' for windows users and 'python' for non-windows users
condaPythonVersion	String, Python version to use when creating a conda environment

Details

This function creates a virtual environment that can be used by PatientLevelPrediction and installs all the required package dependancies. If using python, pip must be set up.

covariateSummary *covariateSummary*

Description

Summarises the covariateData to calculate the mean and standard deviation per covaraite if the labels are input it also stratifies this by class label and if the trainRowIds and testRowIds specifying the patients in the train/test sets respectively are input, these values are also stratified by train and test set

Usage

```
covariateSummary(
  covariateData,
  cohort,
  labels = NULL,
  strata = NULL,
  variableImportance = NULL,
  featureEngineering = NULL
)
```

Arguments

`covariateData`	The covariateData part of the plpData that is extracted using `getPlpData`
`cohort`	The patient cohort to calculate the summary
`labels`	A data.frame with the columns rowId and outcomeCount
`strata`	A data.frame containing the columns rowId, strataName
`variableImportance`	A data.frame with the columns covariateId and value (the variable importance value)
`featureEngineering`	(currently not used) A function or list of functions specifying any feature engineering to create covariates before summarising

Details

The function calculates various metrics to measure the performance of the model

Value

A data.frame containing: CovariateCount CovariateMean and CovariateStDev plus these values for any specified stratification

createCohortCovariateSettings

Extracts covariates based on cohorts

Description

Extracts covariates based on cohorts

Usage

```
createCohortCovariateSettings(
  cohortName,
  settingId,
  cohortDatabaseSchema,
  cohortTable,
  cohortId,
  startDay = -30,
  endDay = 0,
  count = F,
  ageInteraction = F,
  lnAgeInteraction = F,
  analysisId = 456
)
```

Arguments

`cohortName`	Name for the cohort
`settingId`	A unique id for the covariate time and
`cohortDatabaseSchema`	The schema of the database with the cohort
`cohortTable`	the table name that contains the covariate cohort
`cohortId`	cohort id for the covariate cohort
`startDay`	The number of days prior to index to start observing the cohort
`endDay`	The number of days prior to index to stop observing the cohort
`count`	If FALSE the covariate value is binary (1 means cohort occurred between index+startDay and index+endDay, 0 means it did not) If TRUE then the covariate value is the number of unique cohort_start_dates between index+startDay and index+endDay
`ageInteraction`	If TRUE multiple covariate value by the patient's age in years
`lnAgeInteraction`	If TRUE multiple covariate value by the log of the patient's age in years
`analysisId`	The analysisId for the covariate

Details

The user specifies a cohort and time period and then a covariate is constructed whether they are in the cohort during the time periods relative to target population cohort index

Value

An object of class covariateSettings specifying how to create the cohort covariate with the covariateId cohortId x 100000 + settingId x 1000 + analysisId

---

`createDatabaseDetails` *Create a setting that holds the details about the cdmDatabase connection for data extraction*

Description

Create a setting that holds the details about the cdmDatabase connection for data extraction

Usage

```
createDatabaseDetails(
  connectionDetails,
  cdmDatabaseSchema,
  cdmDatabaseName,
  cdmDatabaseId,
  tempEmulationSchema = cdmDatabaseSchema,
  cohortDatabaseSchema = cdmDatabaseSchema,
  cohortTable = "cohort",
  outcomeDatabaseSchema = cdmDatabaseSchema,
  outcomeTable = "cohort",
  targetId = NULL,
  outcomeIds = NULL,
  cdmVersion = 5,
  cohortId = NULL
)
```

Arguments

`connectionDetails`	An R object of type `connectionDetails` created using the function `createConnectionDetails` in the `DatabaseConnector` package.
`cdmDatabaseSchema`	The name of the database schema that contains the OMOP CDM instance. Requires read permissions to this database. On SQL Server, this should specifiy both the database and the schema, so for example 'cdm_instance.dbo'.
`cdmDatabaseName`	A string with the name of the database - this is used in the shiny app and when externally validating models to name the result list and to specify the folder name when saving validation results (defaults to cdmDatabaseSchema if not specified)
`cdmDatabaseId`	A string with a unique identifier for the database and version - this is stored in the plp object for future reference and used by the shiny app (defaults to cdmDatabaseSchema if not specified)
`tempEmulationSchema`	For dmbs like Oracle only: the name of the database schema where you want all temporary tables to be managed. Requires create/insert permissions to this database.

`cohortDatabaseSchema`	The name of the database schema that is the location where the target cohorts are available. Requires read permissions to this database.
`cohortTable`	The tablename that contains the target cohorts. Expectation is cohortTable has format of COHORT table: COHORT_DEFINITION_ID, SUBJECT_ID, COHORT_START_DATE, COHORT_END_DATE.
`outcomeDatabaseSchema`	The name of the database schema that is the location where the data used to define the outcome cohorts is available. Requires read permissions to this database.
`outcomeTable`	The tablename that contains the outcome cohorts. Expectation is outcomeTable has format of COHORT table: COHORT_DEFINITION_ID, SUBJECT_ID, COHORT_START_DATE, COHORT_END_DATE.
`targetId`	An integer specifying the cohort id for the target cohort
`outcomeIds`	A single integer or vector of integers specifying the cohort ids for the outcome cohorts
`cdmVersion`	Define the OMOP CDM version used: currently support "4" and "5".
`cohortId`	(depreciated: use targetId) old input for the target cohort id

Details

This function simply stores the settings for communicating with the cdmDatabase when extracting the target cohort and outcomes

Value

A list with the the database specific settings (this is used by the runMultiplePlp function and the skeleton packages)

`createDatabaseList` *Create a list with the database details and database meta data entries*

Description

This function creates a list with the database details and database meta data entries used in the study

Usage

```
createDatabaseList(cdmDatabaseSchemas, cdmDatabaseNames, databaseRefIds = NULL)
```

Arguments

`cdmDatabaseSchemas`	(string vector) A vector of the cdmDatabaseSchemas used in the study - if the schemas are not unique per database please also specify databaseRefId
`cdmDatabaseNames`	Sharable names for the databases
`databaseRefIds`	(string vector) Unique database identifiers - what you specified as cdmDatabaseId in `PatientLevelPrediction::createDatabaseDetails()` when developing the models

Details

This function is used when inserting database details into the PatientLevelPrediction database results schema

Value

Returns a data.frame with the database details

`createDatabaseSchemaSettings` *Create the PatientLevelPrediction database result schema settings*

Description

This function specifies where the results schema is and lets you pick a different schema for the cohorts and databases

Usage

```
createDatabaseSchemaSettings(
  resultSchema = "main",
  tablePrefix = "",
  targetDialect = "sqlite",
  tempEmulationSchema = getOption("sqlRenderTempEmulationSchema"),
  cohortDefinitionSchema = resultSchema,
  tablePrefixCohortDefinitionTables = tablePrefix,
  databaseDefinitionSchema = resultSchema,
  tablePrefixDatabaseDefinitionTables = tablePrefix
)
```

Arguments

`resultSchema`	(string) The name of the database schema with the result tables.
`tablePrefix`	(string) A string that appends to the PatientLevelPrediction result tables
`targetDialect`	(string) The database management system being used
`tempEmulationSchema`	(string) The temp schema used when the database management system is oracle
`cohortDefinitionSchema`	(string) The name of the database schema with the cohort definition tables (defaults to resultSchema).
`tablePrefixCohortDefinitionTables`	(string) A string that appends to the cohort definition tables
`databaseDefinitionSchema`	(string) The name of the database schema with the database definition tables (defaults to resultSchema).
`tablePrefixDatabaseDefinitionTables`	(string) A string that appends to the database definition tables

Details

This function can be used to specify the database settings used to upload PatientLevelPrediction results into a database

Value

Returns a list of class 'plpDatabaseResultSchema' with all the database settings

---

`createDefaultExecuteSettings`
Creates default list of settings specifying what parts of runPlp to execute

---

Description

Creates default list of settings specifying what parts of runPlp to execute

Usage

```
createDefaultExecuteSettings()
```

Details

runs split, preprocess, model development and covariate summary

Value

list with TRUE for split, preprocess, model development and covariate summary

---

`createDefaultSplitSetting`
Create the settings for defining how the plpData are split into test/validation/train sets using default splitting functions (either random stratified by outcome, time or subject splitting)

Description

Create the settings for defining how the plpData are split into test/validation/train sets using default splitting functions (either random stratified by outcome, time or subject splitting)

Usage

```
createDefaultSplitSetting(
  testFraction = 0.25,
  trainFraction = 0.75,
  splitSeed = sample(1e+05, 1),
  nfold = 3,
  type = "stratified"
)
```

Arguments

`testFraction`	(numeric) A real number between 0 and 1 indicating the test set fraction of the data
`trainFraction`	(numeric) A real number between 0 and 1 indicating the train set fraction of the data. If not set train is equal to 1 - test
`splitSeed`	(numeric) A seed to use when splitting the data for reproducibility (if not set a random number will be generated)
`nfold`	(numeric) An integer > 1 specifying the number of folds used in cross validation
`type`	(character) Choice of: • 'stratified' Each data point is randomly assigned into the test or a train fold set but this is done stratified such that the outcome rate is consistent in each partition • 'time' Older data are assigned into the training set and newer data are assigned into the test set • 'subject' Data are partitioned by subject, if a subject is in the data more than once, all the data points for the subject are assigned either into the test data or into the train data (not both).

Details

Returns an object of class `splitSettings` that specifies the splitting function that will be called and the settings

Value

An object of class `splitSettings`

`createExecuteSettings` *Creates list of settings specifying what parts of runPlp to execute*

Description

Creates list of settings specifying what parts of runPlp to execute

Usage

```
createExecuteSettings(
  runSplitData = F,
  runSampleData = F,
  runfeatureEngineering = F,
  runPreprocessData = F,
  runModelDevelopment = F,
  runCovariateSummary = F
)
```

Arguments

`runSplitData`	TRUE or FALSE whether to split data into train/test
`runSampleData`	TRUE or FALSE whether to over or under sample
`runfeatureEngineering`	TRUE or FALSE whether to do feature engineering
`runPreprocessData`	TRUE or FALSE whether to do preprocessing
`runModelDevelopment`	TRUE or FALSE whether to develop the model
`runCovariateSummary`	TRUE or FALSE whether to create covariate summary

Details

define what parts of runPlp to execute

Value

list with TRUE/FALSE for each part of runPlp

`createFeatureEngineeringSettings` *Create the settings for defining any feature engineering that will be done*

Description

Create the settings for defining any feature engineering that will be done

Usage

```
createFeatureEngineeringSettings(type = "none")
```

Arguments

`type` (character) Choice of:

- 'none' No feature engineering - this is the default

Details

Returns an object of class `featureEngineeringSettings` that specifies the sampling function that will be called and the settings

Value

An object of class `featureEngineeringSettings`

createLearningCurve *createLearningCurve*

Description

Creates a learning curve object, which can be plotted using the `plotLearningCurve()` function.

Usage

```
createLearningCurve(
  plpData,
  outcomeId,
  parallel = T,
  cores = 4,
  modelSettings,
  saveDirectory = getwd(),
  analysisId = "learningCurve",
  populationSettings = createStudyPopulationSettings(),
  splitSettings = createDefaultSplitSetting(),
  trainFractions = c(0.25, 0.5, 0.75),
  trainEvents = NULL,
  sampleSettings = createSampleSettings(),
  featureEngineeringSettings = createFeatureEngineeringSettings(),
 preprocessSettings = createPreprocessSettings(minFraction = 0.001, normalize = T),
  logSettings = createLogSettings(),
  executeSettings = createExecuteSettings(runSplitData = T, runSampleData = F,
    runfeatureEngineering = F, runPreprocessData = T, runModelDevelopment = T,
    runCovariateSummary = F)
)
```

Arguments

`plpData`	An object of type `plpData` - the patient level prediction data extracted from the CDM.
`outcomeId`	(integer) The ID of the outcome.
`parallel`	Whether to run the code in parallel
`cores`	The number of computer cores to use if running in parallel
`modelSettings`	An object of class `modelSettings` created using one of the function: • `setLassoLogisticRegression()` A lasso logistic regression model • `setGradientBoostingMachine()` A gradient boosting machine • `setAdaBoost()` An ada boost model • `setRandomForest()` A random forest model • `setDecisionTree()` A decision tree model • `setKNN()` A KNN model
`saveDirectory`	The path to the directory where the results will be saved (if NULL uses working directory)
`analysisId`	(integer) Identifier for the analysis. It is used to create, e.g., the result folder. Default is a timestamp.

`populationSettings`	An object of type `populationSettings` created using `createStudyPopulationSettings` that specifies how the data class labels are defined and addition any exclusions to apply to the plpData cohort
`splitSettings`	An object of type `splitSettings` that specifies how to split the data into train/validation/test. The default settings can be created using `createDefaultSplitSetting`.
`trainFractions`	A list of training fractions to create models for. Note, providing `trainEvents` will override your input to `trainFractions`.
`trainEvents`	Events have shown to be determinant of model performance. Therefore, it is recommended to provide `trainEvents` rather than `trainFractions`. Note, providing `trainEvents` will override your input to `trainFractions`. The format should be as follows: • `c(500, 1000, 1500)` - a list of training events
`sampleSettings`	An object of type `sampleSettings` that specifies any under/over sampling to be done. The default is none.
`featureEngineeringSettings`	An object of `featureEngineeringSettings` specifying any feature engineering to be learned (using the train data)
`preprocessSettings`	An object of `preprocessSettings`. This setting specifies the minimum fraction of target population who must have a covariate for it to be included in the model training and whether to normalise the covariates before training
`logSettings`	An object of `logSettings` created using `createLogSettings` specifying how the logging is done
`executeSettings`	An object of `executeSettings` specifying which parts of the analysis to run

Value

A learning curve object containing the various performance measures obtained by the model for each training set fraction. It can be plotted using `plotLearningCurve`.

Examples

```
## Not run:
# define model
modelSettings = PatientLevelPrediction::setLassoLogisticRegression()

# create learning curve
learningCurve <- PatientLevelPrediction::createLearningCurve(population,
                                                             plpData,
                                                             modelSettings)
# plot learning curve
PatientLevelPrediction::plotLearningCurve(learningCurve)

## End(Not run)
```

createLogSettings *Create the settings for logging the progression of the analysis*

Description

Create the settings for logging the progression of the analysis

Usage

```
createLogSettings(verbosity = "DEBUG", timeStamp = T, logName = "runPlp Log")
```

Arguments

`verbosity`	Sets the level of the verbosity. If the log level is at or higher in priority than the logger threshold, a message will print. The levels are: • DEBUG Highest verbosity showing all debug statements • TRACE Showing information about start and end of steps • INFO Show informative information (Default) • WARN Show warning messages • ERROR Show error messages • FATAL Be silent except for fatal errors
`timeStamp`	If TRUE a timestamp will be added to each logging statement. Automatically switched on for TRACE level.
`logName`	A string reference for the logger

Details

Returns an object of class `logSettings` that specifies the logger settings

Value

An object of class `logSettings`

createModelDesign *Specify settings for deceloping a single model*

Description

Specify settings for deceloping a single model

Usage

```
createModelDesign(
  targetId,
  outcomeId,
  restrictPlpDataSettings = createRestrictPlpDataSettings(),
  populationSettings = createStudyPopulationSettings(),
  covariateSettings = FeatureExtraction::createDefaultCovariateSettings(),
  featureEngineeringSettings = NULL,
  sampleSettings = NULL,
  preprocessSettings = NULL,
  modelSettings = NULL,
 splitSettings = createDefaultSplitSetting(type = "stratified", testFraction = 0.25,
    trainFraction = 0.75, splitSeed = 123, nfold = 3),
  runCovariateSummary = T
)
```

Arguments

`targetId`	The id of the target cohort that will be used for data extraction (e.g., the ATLAS id)
`outcomeId`	The id of the outcome that will be used for data extraction (e.g., the ATLAS id)
`restrictPlpDataSettings`	The settings specifying the extra restriction settings when extracting the data created using `createRestrictPlpDataSettings()`.
`populationSettings`	The population settings specified by `createStudyPopulationSettings()`
`covariateSettings`	The covariate settings, this can be a list or a single `'covariateSetting'` object.
`featureEngineeringSettings`	Either NULL or an object of class `featureEngineeringSettings` specifying any feature engineering used during model development
`sampleSettings`	Either NULL or an object of class `sampleSettings` with the over/under sampling settings used for model development
`preprocessSettings`	Either NULL or an object of class `preprocessSettings` created using `createPreprocessingSetti`
`modelSettings`	The model settings such as `setLassoLogisticRegression()`
`splitSettings`	The train/validation/test splitting used by all analyses created using `createDefaultSplitSetting()`
`runCovariateSummary`	Whether to run the covariateSummary

Details

This specifies a single analysis for developing as single model

Value

A list with analysis settings used to develop a single prediction model

createPlpResultTables *Create the results tables to store PatientLevelPrediction models and results into a database*

Description

This function executes a large set of SQL statements to create tables that can store models and results

Usage

```
createPlpResultTables(
  connectionDetails,
  targetDialect = "postgresql",
  resultSchema,
  deleteTables = T,
  createTables = T,
  tablePrefix = "",
  tempEmulationSchema = getOption("sqlRenderTempEmulationSchema"),
  testFile = NULL
)
```

Arguments

`connectionDetails`	The database connection details
`targetDialect`	The database management system being used
`resultSchema`	The name of the database schema that the result tables will be created.
`deleteTables`	If true any existing tables matching the PatientLevelPrediction result tables names will be deleted
`createTables`	If true the PatientLevelPrediction result tables will be created
`tablePrefix`	A string that appends to the PatientLevelPrediction result tables
`tempEmulationSchema`	The temp schema used when the database management system is oracle
`testFile`	(used for testing) The location of an sql file with the table creation code

Details

This function can be used to create (or delete) PatientLevelPrediction result tables

Value

Returns NULL but creates the required tables into the specified database schema(s).

createPreprocessSettings

Create the settings for preprocessing the trainData.

Description

Create the settings for preprocessing the trainData.

Usage

```
createPreprocessSettings(
  minFraction = 0.001,
  normalize = TRUE,
  removeRedundancy = TRUE
)
```

Arguments

minFraction	The minimum fraction of target population who must have a covariate for it to be included in the model training
normalize	Whether to normalise the covariates before training (Default: TRUE)
removeRedundancy	Whether to remove redundant features (Default: TRUE)

Details

Returns an object of class `preprocessingSettings` that specifies how to preprocess the training data

Value

An object of class `preprocessingSettings`

createRandomForestFeatureSelection

Create the settings for random foreat based feature selection

Description

Create the settings for random foreat based feature selection

Usage

```
createRandomForestFeatureSelection(ntrees = 2000, maxDepth = 17)
```

Arguments

ntrees	number of tree in forest
maxDepth	MAx depth of each tree

Details

Returns an object of class `featureEngineeringSettings` that specifies the sampling function that will be called and the settings

Value

An object of class `featureEngineeringSettings`

`createRestrictPlpDataSettings`
createRestrictPlpDataSettings define extra restriction settings when calling getPlpData

---

Description

This function creates the settings used to restrict the target cohort when calling getPlpData

Usage

```
createRestrictPlpDataSettings(
  studyStartDate = "",
  studyEndDate = "",
  firstExposureOnly = F,
  washoutPeriod = 0,
  sampleSize = NULL
)
```

Arguments

`studyStartDate`	A calendar date specifying the minimum date that a cohort index date can appear. Date format is 'yyyymmdd'.
`studyEndDate`	A calendar date specifying the maximum date that a cohort index date can appear. Date format is 'yyyymmdd'. Important: the study end data is also used to truncate risk windows, meaning no outcomes beyond the study end date will be considered.
`firstExposureOnly`	Should only the first exposure per subject be included? Note that this is typically done in the `createStudyPopulation` function, but can already be done here for efficiency reasons.
`washoutPeriod`	The mininum required continuous observation time prior to index date for a person to be included in the at risk cohort. Note that this is typically done in the `createStudyPopulation` function, but can already be done here for efficiency reasons.
`sampleSize`	If not NULL, the number of people to sample from the target cohort

Details

Users need to specify the extra restrictions to apply when downloading the target cohort

Value

A setting object of class `restrictPlpDataSettings` containing a list getPlpData extra settings `createSampleSettings` *Create the settings for defining how the trainData from* `splitData` *are sampled using default sample functions.*

Description

Create the settings for defining how the trainData from `splitData` are sampled using default sample functions.

Usage

```
createSampleSettings(
  type = "none",
  numberOutcomestoNonOutcomes = 1,
  sampleSeed = sample(10000, 1)
)
```

Arguments

`type`	(character) Choice of: • 'none' No sampling is applied - this is the default • 'underSample' Undersample the non-outcome class to make the data more ballanced • 'overSample' Oversample the outcome class by adding in each outcome multiple times
`numberOutcomestoNonOutcomes`	(numeric) An numeric specifying the require number of non-outcomes per outcome
`sampleSeed`	(numeric) A seed to use when splitting the data for reproducibility (if not set a random number will be generated)

Details

Returns an object of class `sampleSettings` that specifies the sampling function that will be called and the settings

Value

An object of class `sampleSettings`

`createSplineSettings` *Create the settings for adding a spline for continuous variables*

Description

Create the settings for adding a spline for continuous variables

Usage

```
createSplineSettings(continousCovariateId, knots, analysisId = 683)
```

Arguments

`continousCovariateId`	The covariateId to apply splines to
`knots`	Either number of knots of vector of split values
`analysisId`	The analysisId to use for the spline covariates

Details

Returns an object of class `featureEngineeringSettings` that specifies the sampling function that will be called and the settings

Value

An object of class `featureEngineeringSettings`

`createStratifiedImputationSettings` *Create the settings for adding a spline for continuous variables*

Description

Create the settings for adding a spline for continuous variables

Usage

```
createStratifiedImputationSettings(covariateId, ageSplits = NULL)
```

Arguments

`covariateId`	The covariateId that needs imputed values
`ageSplits`	A vector of age splits in years to create age groups

Details

Returns an object of class `featureEngineeringSettings` that specifies how to do stratified imputation

Value

An object of class `featureEngineeringSettings`

`createStudyPopulation` *Create a study population*

Description

Create a study population

Usage

```
createStudyPopulation(
  plpData,
  outcomeId,
  populationSettings,
  population = NULL
)
```

Arguments

`plpData`	An object of type `plpData` as generated using `getplpData`.
`outcomeId`	The ID of the outcome.
`populationSettings`	An object of class populationSettings created using `createPopulationSettings`
`population`	If specified, this population will be used as the starting point instead of the cohorts in the `plpData` object.

Details

Create a study population by enforcing certain inclusion and exclusion criteria, defining a risk window, and determining which outcomes fall inside the risk window.

Value

A data frame specifying the study population. This data frame will have the following columns:

rowId A unique identifier for an exposure

subjectId The person ID of the subject

cohortStartdate The index date

outcomeCount The number of outcomes observed during the risk window

timeAtRisk The number of days in the risk window

survivalTime The number of days until either the outcome or the end of the risk window

createStudyPopulationSettings

create the study population settings

Description

create the study population settings

Usage

```
createStudyPopulationSettings(
  binary = T,
  includeAllOutcomes = T,
  firstExposureOnly = FALSE,
  washoutPeriod = 0,
  removeSubjectsWithPriorOutcome = TRUE,
  priorOutcomeLookback = 99999,
  requireTimeAtRisk = T,
  minTimeAtRisk = 364,
  riskWindowStart = 1,
  startAnchor = "cohort start",
  riskWindowEnd = 365,
  endAnchor = "cohort start",
  restrictTarToCohortEnd = F
)
```

Arguments

`binary`	Forces the outcomeCount to be 0 or 1 (use for binary prediction problems)
`includeAllOutcomes`	(binary) indicating whether to include people with outcomes who are not observed for the whole at risk period
`firstExposureOnly`	Should only the first exposure per subject be included? Note that this is typically done in the `createStudyPopulation` function,
`washoutPeriod`	The mininum required continuous observation time prior to index date for a person to be included in the cohort.
`removeSubjectsWithPriorOutcome`	Remove subjects that have the outcome prior to the risk window start?
`priorOutcomeLookback`	How many days should we look back when identifying prior outcomes?
`requireTimeAtRisk`	Should subject without time at risk be removed?
`minTimeAtRisk`	The minimum number of days at risk required to be included
`riskWindowStart`	The start of the risk window (in days) relative to the index date (+ days of exposure if the `addExposureDaysToStart` parameter is specified).
`startAnchor`	The anchor point for the start of the risk window. Can be "cohort start" or "cohort end".

`riskWindowEnd`	The end of the risk window (in days) relative to the index data (+ days of exposure if the `addExposureDaysToEnd` parameter is specified).
`endAnchor`	The anchor point for the end of the risk window. Can be "cohort start" or "cohort end".
`restrictTarToCohortEnd`	If using a survival model and you want the time-at-risk to end at the cohort end date set this to T

Details

Takes as input the inputs to create study population

Value

A list containing all the settings required for creating the study population

---

`createTempModelLoc` *Create a temporary model location*

Description

Create a temporary model location

Usage

```
createTempModelLoc()
```

`createUnivariateFeatureSelection` *Create the settings for defining any feature selection that will be done*

Description

Create the settings for defining any feature selection that will be done

Usage

```
createUnivariateFeatureSelection(k = 100)
```

Arguments

`k`	This function returns the K features most associated (univariately) to the outcome

Details

Returns an object of class `featureEngineeringSettings` that specifies the sampling function that will be called and the settings

Value

An object of class `featureEngineeringSettings`

createValidationDesign

createValidationDesign - Define the validation design for external validation

Description

createValidationDesign - Define the validation design for external validation

Usage

```
createValidationDesign(
  targetId,
  outcomeId,
  populationSettings,
  restrictPlpDataSettings,
  plpModelList,
  recalibrate = NULL,
  runCovariateSummary = TRUE
)
```

Arguments

`targetId`	The targetId of the target cohort to validate on
`outcomeId`	The outcomeId of the outcome cohort to validate on
`populationSettings`	A list of population restriction settings created by `createPopulationSettings`
`restrictPlpDataSettings`	A list of plpData restriction settings created by `createRestrictPlpDataSettings`
`plpModelList`	A list of plpModels objects created by `runPlp` or a path to such objects
`recalibrate`	A vector of characters specifying the recalibration method to apply,
`runCovariateSummary`	whether to run the covariate summary for the validation data

createValidationSettings

createValidationSettings define optional settings for performing external validation

Description

This function creates the settings required by externalValidatePlp

Usage

```
createValidationSettings(recalibrate = NULL, runCovariateSummary = T)
```

Arguments

`recalibrate`	A vector of characters specifying the recalibration method to apply
`runCovariateSummary`	Whether to run the covariate summary for the validation data

Details

Users need to specify whether they want to sample or recalibate when performing external validation

Value

A setting object of class `validationSettings` containing a list of settings for externalValidatePlp

---

`diagnoseMultiplePlp` *Run a list of predictions diagnoses*

Description

Run a list of predictions diagnoses

Usage

```
diagnoseMultiplePlp(
  databaseDetails = createDatabaseDetails(),
 modelDesignList = list(createModelDesign(targetId = 1, outcomeId = 2, modelSettings =
   setLassoLogisticRegression()), createModelDesign(targetId = 1, outcomeId = 3,
    modelSettings = setLassoLogisticRegression())),
  cohortDefinitions = NULL,
  logSettings = createLogSettings(verbosity = "DEBUG", timeStamp = T, logName =
    "diagnosePlp Log"),
  saveDirectory = getwd()
)
```

Arguments

`databaseDetails`	The database settings created using `createDatabaseDetails()`
`modelDesignList`	A list of model designs created using `createModelDesign()`
`cohortDefinitions`	A list of cohort definitions for the target and outcome cohorts
`logSettings`	The setting spexcifying the logging for the analyses created using `createLogSettings()`
`saveDirectory`	Name of the folder where all the outputs will written to.

Details

This function will run all specified prediction design diagnoses as defined using .

Value

A data frame with the following columns:

`analysisId`	The unique identifier for a set of analysis choices.
`targetId`	The ID of the target cohort populations.
`outcomeId`	The ID of the outcomeId.
`dataLocation`	The location where the plpData was saved
`the settings ids`	The ids for all other settings used for model development.

`diagnosePlp` *diagnostic - Investigates the prediction problem settings - use before training a model*

Description

This function runs a set of prediction diagnoses to help pick a suitable T, O, TAR and determine whether the prediction problem is worth executing.

Usage

```
diagnosePlp(
  plpData = NULL,
  outcomeId,
  analysisId,
  populationSettings,
  splitSettings = createDefaultSplitSetting(),
  sampleSettings = createSampleSettings(),
  saveDirectory = NULL,
  featureEngineeringSettings = createFeatureEngineeringSettings(),
  modelSettings = setLassoLogisticRegression(),
  logSettings = createLogSettings(verbosity = "DEBUG", timeStamp = T, logName =
    "diagnosePlp Log"),
  preprocessSettings = createPreprocessSettings()
)
```

Arguments

`plpData`	An object of type `plpData` - the patient level prediction data extracted from the CDM. Can also include an initial population as plpData$popualtion.
`outcomeId`	(integer) The ID of the outcome.
`analysisId`	(integer) Identifier for the analysis. It is used to create, e.g., the result folder. Default is a timestamp.
`populationSettings`	An object of type `populationSettings` created using `createStudyPopulationSettings` that specifies how the data class labels are defined and addition any exclusions to apply to the plpData cohort
`splitSettings`	An object of type `splitSettings` that specifies how to split the data into train/validation/test. The default settings can be created using `createDefaultSplitSetting`.

`sampleSettings`	An object of type `sampleSettings` that specifies any under/over sampling to be done. The default is none.
`saveDirectory`	The path to the directory where the results will be saved (if NULL uses working directory)
`featureEngineeringSettings`	An object of `featureEngineeringSettings` specifying any feature engineering to be learned (using the train data)
`modelSettings`	An object of class `modelSettings` created using one of the function: • setLassoLogisticRegression() A lasso logistic regression model • setGradientBoostingMachine() A gradient boosting machine • setAdaBoost() An ada boost model • setRandomForest() A random forest model • setDecisionTree() A decision tree model • setKNN() A KNN model
`logSettings`	An object of `logSettings` created using `createLogSettings` specifying how the logging is done
`preprocessSettings`	An object of `preprocessSettings`. This setting specifies the minimum fraction of target population who must have a covariate for it to be included in the model training and whether to normalise the covariates before training

Details

Users can define set of Ts, Os, databases and population settings. A list of data.frames containing details such as follow-up time distribution, time-to-event information, characteriszation details, time from last prior event, observation time distribution.

Value

An object containing the model or location where the model is save, the data selection settings, the preprocessing and training settings as well as various performance measures obtained by the model.

`distribution`	list for each O of a data.frame containing: i) Time to observation end distribution, ii) Time from observation start distribution, iii) Time to event distribution and iv) Time from last prior event to index distribution (only for patients in T who have O before index)
`incident`	list for each O of incidence of O in T during TAR
`characterization`	list for each O of Characterization of T, TnO, Tn~O

Examples

```
## Not run:
#********* EXAMPLE 1 *********

## End(Not run)
```

`diagnosticOddsRatio` *Calculate the diagnostic odds ratio*

Description

Calculate the diagnostic odds ratio

Usage

```
diagnosticOddsRatio(TP, TN, FN, FP)
```

Arguments

`TP`	Number of true positives
`TN`	Number of true negatives
`FN`	Number of false negatives
`FP`	Number of false positives

Details

Calculate the diagnostic odds ratio

Value

diagnosticOddsRatio value

`evaluatePlp` *evaluatePlp*

Description

Evaluates the performance of the patient level prediction model

Usage

```
evaluatePlp(prediction, typeColumn = "evaluationType")
```

Arguments

`prediction`	The patient level prediction model's prediction
`typeColumn`	The column name in the prediction object that is used to stratify the evaluation

Details

The function calculates various metrics to measure the performance of the model

Value

A list containing the performance values

externalValidateDbPlp *externalValidateDbPlp - Validate a model on new databases*

Description

This function extracts data using a user specified connection and cdm_schema, applied the model and then calcualtes the performance

Usage

```
externalValidateDbPlp(
  plpModel,
  validationDatabaseDetails = createDatabaseDetails(),
  validationRestrictPlpDataSettings = createRestrictPlpDataSettings(),
  settings = createValidationSettings(recalibrate = "weakRecalibration"),
  logSettings = createLogSettings(verbosity = "INFO", logName = "validatePLP"),
  outputFolder = getwd()
)
```

Arguments

`plpModel`	The model object returned by runPlp() containing the trained model
`validationDatabaseDetails`	A list of objects of class `databaseDetails` created using `createDatabaseDetails`
`validationRestrictPlpDataSettings`	A list of population restriction settings created by `createRestrictPlpDataSettings()`
`settings`	A settings object of class `validationSettings` created using `createValidationSettings`
`logSettings`	An object of `logSettings` created using `createLogSettings` specifying how the logging is done
`outputFolder`	The directory to save the validation results to (subfolders are created per database in validationDatabaseDetails)

Details

Users need to input a trained model (the output of runPlp()) and new database connections. The function will return a list of length equal to the number of cdm_schemas input with the performance on the new data

Value

A list containing the performance for each validation_schema

extractDatabaseToCsv *Exports all the results from a database into csv files*

Description

Exports all the results from a database into csv files

Usage

```
extractDatabaseToCsv(
  conn = NULL,
  connectionDetails,
  databaseSchemaSettings = createDatabaseSchemaSettings(resultSchema = "main"),
  csvFolder,
  minCellCount = 5,
  sensitiveColumns = getPlpSensitiveColumns(),
  fileAppend = NULL
)
```

Arguments

`conn`	The connection to the database with the results
`connectionDetails`	The connectionDetails for the result database
`databaseSchemaSettings`	The result database schema settings
`csvFolder`	Location to save the csv files
`minCellCount`	The min value to show in cells that are sensitive (values less than this value will be replaced with -1)
`sensitiveColumns`	A named list (name of table columns belong to) with a list of columns to apply the minCellCount to.
`fileAppend`	If set to a string this will be appended to the start of the csv file names

Details

Extracts the results from a database into a set of csv files

f1Score *Calculate the f1Score*

Description

Calculate the f1Score

Usage

```
f1Score(TP, TN, FN, FP)
```

Arguments

`TP`	Number of true positives
`TN`	Number of true negatives
`FN`	Number of false negatives
`FP`	Number of false positives

Details

Calculate the f1Score

Value

f1Score value

`falseDiscoveryRate` *Calculate the falseDiscoveryRate*

Description

Calculate the falseDiscoveryRate

Usage

```
falseDiscoveryRate(TP, TN, FN, FP)
```

Arguments

`TP`	Number of true positives
`TN`	Number of true negatives
`FN`	Number of false negatives
`FP`	Number of false positives

Details

Calculate the falseDiscoveryRate

Value

falseDiscoveryRate value

`falseNegativeRate` *Calculate the falseNegativeRate*

Description

Calculate the falseNegativeRate

Usage

```
falseNegativeRate(TP, TN, FN, FP)
```

Arguments

`TP`	Number of true positives
`TN`	Number of true negatives
`FN`	Number of false negatives
`FP`	Number of false positives

Details

Calculate the falseNegativeRate

Value

falseNegativeRate value

`falseOmissionRate` *Calculate the falseOmissionRate*

Description

Calculate the falseOmissionRate

Usage

```
falseOmissionRate(TP, TN, FN, FP)
```

Arguments

`TP`	Number of true positives
`TN`	Number of true negatives
`FN`	Number of false negatives
`FP`	Number of false positives

Details

Calculate the falseOmissionRate

Value

falseOmissionRate value

falsePositiveRate *Calculate the falsePositiveRate*

Description

Calculate the falsePositiveRate

Usage

```
falsePositiveRate(TP, TN, FN, FP)
```

Arguments

`TP`	Number of true positives
`TN`	Number of true negatives
`FN`	Number of false negatives
`FP`	Number of false positives

Details

Calculate the falsePositiveRate

Value

falsePositiveRate value

fitPlp *fitPlp*

Description

Train various models using a default parameter gird search or user specified parameters

Usage

```
fitPlp(trainData, modelSettings, search = "grid", analysisId, analysisPath)
```

Arguments

`trainData`	An object of type `TrainData` created using `splitData` data extracted from the CDM.
`modelSettings`	An object of class `modelSettings` created using one of the function: • setLassoLogisticRegression() A lasso logistic regression model • setGradientBoostingMachine() A gradient boosting machine • setRandomForest() A random forest model • setKNN() A KNN model
`search`	The search strategy for the hyper-parameter selection (currently not used)
`analysisId`	The id of the analysis
`analysisPath`	The path of the analysis

Details

The user can define the machine learning model to train (regularised logistic regression, random forest, gradient boosting machine, neural network and)

Value

An object of class `plpModel` containing:

`model`	The trained prediction model
`preprocessing`	The preprocessing required when applying the model
`prediction`	The cohort data.frame with the predicted risk column added
`modelDesign`	A list specifiying the modelDesign settings used to fit the model
`trainDetails`	The model meta data
`covariateImportance`	The covariate importance for the model

`getCalibrationSummary` *Get a sparse summary of the calibration*

Description

Get a sparse summary of the calibration

Usage

```
getCalibrationSummary(
  prediction,
  predictionType,
  typeColumn = "evaluation",
  numberOfStrata = 100,
  truncateFraction = 0.05
)
```

Arguments

`prediction`	A prediction object as generated using the `predict` functions.
`predictionType`	The type of prediction (binary or survival)
`typeColumn`	A column that is used to stratify the results
`numberOfStrata`	The number of strata in the plot.
`truncateFraction`	This fraction of probability values will be ignored when plotting, to avoid the x-axis scale being dominated by a few outliers.

Details

Generates a sparse summary showing the predicted probabilities and the observed fractions. Predictions are stratefied into equally sized bins of predicted probabilities.

Value

A dataframe with the calibration summary

getCohortCovariateData *Extracts covariates based on cohorts*

Description

Extracts covariates based on cohorts

Usage

```
getCohortCovariateData(
  connection,
  oracleTempSchema = NULL,
  cdmDatabaseSchema,
  cdmVersion = "5",
  cohortTable = "#cohort_person",
  rowIdField = "row_id",
  aggregated,
  cohortIds,
  covariateSettings
)
```

Arguments

`connection`	The database connection
`oracleTempSchema`	The temp schema if using oracle
`cdmDatabaseSchema`	The schema of the OMOP CDM data
`cdmVersion`	version of the OMOP CDM data
`cohortTable`	the table name that contains the target population cohort
`rowIdField`	string representing the unique identifier in the target population cohort
`aggregated`	whether the covariate should be aggregated
`cohortIds`	cohort id for the target cohort
`covariateSettings`	settings for the covariate cohorts and time periods

Details

The user specifies a cohort and time period and then a covariate is constructed whether they are in the cohort during the time periods relative to target population cohort index

Value

The models will now be in the package

`getDemographicSummary` *Get a calibration per age/gender groups*

Description

Get a calibration per age/gender groups

Usage

```
getDemographicSummary(prediction, predictionType, typeColumn = "evaluation")
```

Arguments

`prediction`	A prediction object
`predictionType`	The type of prediction (binary or survival)
`typeColumn`	A column that is used to stratify the results

Details

Generates a data.frame with the calibration per each 5 year age group and gender group

Value

A dataframe with the calibration summary

`getPlpData` *Get the patient level prediction data from the server*

Description

This function executes a large set of SQL statements against the database in OMOP CDM format to extract the data needed to perform the analysis.

Usage

```
getPlpData(databaseDetails, covariateSettings, restrictPlpDataSettings)
```

Arguments

`databaseDetails`	The cdm database details created using `createDatabaseDetails()`
`covariateSettings`	An object of type `covariateSettings` as created using the `createCovariateSettings` function in the `FeatureExtraction` package.
`restrictPlpDataSettings`	Extra settings to apply to the target population while extracting data. Created using `createRestrictPlpDataSettings()`.

Details

Based on the arguments, the at risk cohort data is retrieved, as well as outcomes occurring in these subjects. The at risk cohort is identified through user-defined cohorts in a cohort table either inside the CDM instance or in a separate schema. Similarly, outcomes are identified through user-defined cohorts in a cohort table either inside the CDM instance or in a separate schema. Covariates are automatically extracted from the appropriate tables within the CDM. If you wish to exclude concepts from covariates you will need to manually add the concept_ids and descendants to the `excludedCovariateConceptIds` of the `covariateSettings` argument.

Value

Returns an object of type `plpData`, containing information on the cohorts, their outcomes, and baseline covariates. Information about multiple outcomes can be captured at once for efficiency reasons. This object is a list with the following components:

outcomes A data frame listing the outcomes per person, including the time to event, and the outcome id. Outcomes are not yet filtered based on risk window, since this is done at a later stage.

cohorts A data frame listing the persons in each cohort, listing their exposure status as well as the time to the end of the observation period and time to the end of the cohort (usually the end of the exposure era).

covariates An ffdf object listing the baseline covariates per person in the two cohorts. This is done using a sparse representation: covariates with a value of 0 are omitted to save space.

covariateRef An ffdf object describing the covariates that have been extracted.

metaData A list of objects with information on how the cohortMethodData object was constructed.

The generic `()` and `summary()` functions have been implemented for this object.

---

getPredictionDistribution *Calculates the prediction distribution*

Description

Calculates the prediction distribution

Usage

```
getPredictionDistribution(
  prediction,
  predictionType,
  typeColumn = "evaluation"
)
```

Arguments

`prediction`	A prediction object
`predictionType`	The type of prediction (binary or survival)
`typeColumn`	A column that is used to stratify the results

Details

Calculates the quantiles from a predition object

Value

The 0.00, 0.1, 0.25, 0.5, 0.75, 0.9, 1.00 quantile pf the prediction, the mean and standard deviation per class

---

`getPredictionDistribution_binary` *Calculates the prediction distribution*

Description

Calculates the prediction distribution

Usage

```
getPredictionDistribution_binary(prediction, evalColumn, ...)
```

Arguments

`prediction`	A prediction object
`evalColumn`	A column that is used to stratify the results
`...`	Other inputs

Details

Calculates the quantiles from a predition object

Value

The 0.00, 0.1, 0.25, 0.5, 0.75, 0.9, 1.00 quantile pf the prediction, the mean and standard deviation per class

---

`getThresholdSummary` *Calculate all measures for sparse ROC*

Description

Calculate all measures for sparse ROC

Usage

```
getThresholdSummary(prediction, predictionType, typeColumn = "evaluation")
```

Arguments

`prediction`	A prediction object
`predictionType`	The type of prediction (binary or survival)
`typeColumn`	A column that is used to stratify the results

Details

Calculates the TP, FP, TN, FN, TPR, FPR, accuracy, PPF, FOR and Fmeasure from a prediction object

Value

A data.frame with all the measures

`getThresholdSummary_binary` *Calculate all measures for sparse ROC when prediction is bianry classification*

Description

Calculate all measures for sparse ROC when prediction is bianry classification

Usage

```
getThresholdSummary_binary(prediction, evalColumn, ...)
```

Arguments

`prediction`	A prediction object
`evalColumn`	A column that is used to stratify the results
`...`	Other inputs

Details

Calculates the TP, FP, TN, FN, TPR, FPR, accuracy, PPF, FOR and Fmeasure from a prediction object

Value

A data.frame with all the measures

ici — *Calculate the Integrated Calibration Information from Austin and Steyerberg https://onlinelibrary.wiley.com/doi/full/10.1002/sim.8281*

Description

Calculate the Integrated Calibration Information from Austin and Steyerberg https://onlinelibrary.wiley.com/doi/full/10.1

Usage

```
ici(prediction)
```

Arguments

`prediction`	the prediction object found in the plpResult object

Details

Calculate the Integrated Calibration Information

Value

Integrated Calibration Information

insertCsvToDatabase — *Function to insert results into a database from csvs*

Description

This function converts a folder with csv results into plp objects and loads them into a plp result database

Usage

```
insertCsvToDatabase(
  csvFolder,
  connectionDetails,
  databaseSchemaSettings,
  modelSaveLocation,
  csvTableAppend = ""
)
```

Arguments

`csvFolder`	The location to the csv folder with the plp results
`connectionDetails`	A connection details for the plp results database that the csv results will be inserted into
`databaseSchemaSettings`	A object created by `createDatabaseSchemaSettings` with all the settings specifying the result tables to insert the csv results into
`modelSaveLocation`	The location to save any models from the csv folder - this should be the same location you picked when inserting other models into the database
`csvTableAppend`	A string that appends the csv file names

Details

The user needs to have plp csv results in a single folder and an existing plp result database

Value

Returns a data.frame indicating whether the results were inported into the database

---

`insertModelDesignInDatabase`
Insert a model design into a PLP result schema database

Description

This function inserts a model design and all the settings into the result schema

Usage

```
insertModelDesignInDatabase(
  object,
  conn,
  databaseSchemaSettings,
  cohortDefinitions
)
```

Arguments

`object`	An object of class modelDesign, runPlp or externalValidatePlp
`conn`	A connection to a database created by using the function `connect` in the `DatabaseConnector` package.
`databaseSchemaSettings`	A object created by `createDatabaseSchemaSettings` with all the settings specifying the result tables
`cohortDefinitions`	A set of one or more cohorts extracted using ROhdsiWebApi::exportCohortDefinitionSet()

Details

This function can be used to upload a model design into a database

Value

Returns NULL but uploads the model design into the database schema specified in databaseSchemaSettings

---

`insertResultsToSqlite` *Create sqlite database with the results*

Description

This function create an sqlite database with the PLP result schema and inserts all results

Usage

```
insertResultsToSqlite(
  resultLocation,
  cohortDefinitions,
  databaseList = NULL,
  sqliteLocation = file.path(resultLocation, "sqlite")
)
```

Arguments

`resultLocation`	(string) location of directory where the main package results were saved
`cohortDefinitions`	A set of one or more cohorts extracted using ROhdsiWebApi::exportCohortDefinitionSet()
`databaseList`	A list created by `createDatabaseList` to specify the databases
`sqliteLocation`	(string) location of directory where the sqlite database will be saved

Details

This function can be used upload PatientLevelPrediction results into an sqlite database

Value

Returns the location of the sqlite database file

`listAppend` *join two lists*

Description

join two lists

Usage

```
listAppend(a, b)
```

Arguments

`a`	A list
`b`	Another list

Details

This function joins two lists

`listCartesian` *Cartesian product*

Description

Computes the Cartesian product of all the combinations of elements in a list

Usage

```
listCartesian(allList)
```

Arguments

`allList`	a list of lists

Value

A list with all possible combinations from the input list of lists

---

`loadPlpAnalysesJson` *Load the multiple prediction json settings from a file*

Description

Load the multiple prediction json settings from a file

Usage

```
loadPlpAnalysesJson(jsonFileLocation)
```

Arguments

`jsonFileLocation`	The location of the file 'predictionAnalysisList.json' with the modelDesignList

Details

This function interprets a json with the multiple prediction settings and creates a list that can be combined with connection settings to run a multiple prediction study

Examples

```
## Not run:
modelDesignList <- loadPlpAnalysesJson('location of json settings')$analysis

## End(Not run)
```

`loadPlpData` *Load the cohort data from a folder*

Description

`loadPlpData` loads an object of type plpData from a folder in the file system.

Usage

```
loadPlpData(file, readOnly = TRUE)
```

Arguments

`file`	The name of the folder containing the data.
`readOnly`	If true, the data is opened read only.

Details

The data will be written to a set of files in the folder specified by the user.

Value

An object of class plpData.

Examples

```
# todo
```

`loadPlpModel` *loads the plp model*

Description

loads the plp model

Usage

```
loadPlpModel(dirPath)
```

Arguments

`dirPath`	The location of the model

Details

Loads a plp model that was saved using `savePlpModel()`

`loadPlpResult` *Loads the evalaution dataframe*

Description

Loads the evalaution dataframe

Usage

```
loadPlpResult(dirPath)
```

Arguments

`dirPath`	The directory where the evaluation was saved

Details

Loads the evaluation

`loadPlpShareable` *Loads the plp result saved as json/csv files for transparent sharing*

Description

Loads the plp result saved as json/csv files for transparent sharing

Usage

```
loadPlpShareable(loadDirectory)
```

Arguments

`loadDirectory`	The directory with the results as json/csv files

Details

Load the main results from json/csv files into a runPlp object

`loadPrediction` *Loads the prediciton dataframe to csv*

Description

Loads the prediciton dataframe to csv

Usage

```
loadPrediction(fileLocation)
```

Arguments

`fileLocation`	The location with the saved prediction

Details

Loads the prediciton RDS file

MapIds *Map covariate and row Ids so they start from 1*

Description

this functions takes covariate data and a cohort/population and remaps the covariate and row ids, restricts to pop and saves/creates mapping

Usage

```
MapIds(covariateData, cohort = NULL, mapping = NULL)
```

Arguments

`covariateData`	a covariateData object
`cohort`	if specified rowIds restricted to the ones in cohort
`mapping`	A pre defined mapping to use

migrateDataModel *Migrate Data model*

Description

Migrate data from current state to next state

It is strongly advised that you have a backup of all data (either sqlite files, a backup database (in the case you are using a postgres backend) or have kept the csv/zip files from your data generation.

Usage

```
migrateDataModel(connectionDetails, databaseSchema, tablePrefix = "")
```

Arguments

`connectionDetails`	DatabaseConnector connection details object
`databaseSchema`	String schema where database schema lives
`tablePrefix`	(Optional) Use if a table prefix is used before table names (e.g. "cd_")

modelBasedConcordance

Calculate the model-based concordance, which is a calculation of the expected discrimination performance of a model under the assumption the model predicts the "TRUE" outcome as detailed in van Klaveren et al. https://pubmed.ncbi.nlm.nih.gov/27251001/

Description

Calculate the model-based concordance, which is a calculation of the expected discrimination performance of a model under the assumption the model predicts the "TRUE" outcome as detailed in van Klaveren et al. https://pubmed.ncbi.nlm.nih.gov/27251001/

Usage

```
modelBasedConcordance(prediction)
```

Arguments

`prediction`	the prediction object found in the plpResult object

Details

Calculate the model-based concordance

Value

model-based concordance value

negativeLikelihoodRatio

Calculate the negativeLikelihoodRatio

Description

Calculate the negativeLikelihoodRatio

Usage

```
negativeLikelihoodRatio(TP, TN, FN, FP)
```

Arguments

`TP`	Number of true positives
`TN`	Number of true negatives
`FN`	Number of false negatives
`FP`	Number of false positives

Details

Calculate the negativeLikelihoodRatio

Value

negativeLikelihoodRatio value

---

negativePredictiveValue *Calculate the negativePredictiveValue*

Description

Calculate the negativePredictiveValue

Usage

```
negativePredictiveValue(TP, TN, FN, FP)
```

Arguments

`TP`	Number of true positives
`TN`	Number of true negatives
`FN`	Number of false negatives
`FP`	Number of false positives

Details

Calculate the negativePredictiveValue

Value

negativePredictiveValue value

---

outcomeSurvivalPlot *Plot the outcome incidence over time*

Description

Plot the outcome incidence over time

Usage

```
outcomeSurvivalPlot(
  plpData,
  outcomeId,
 populationSettings = createStudyPopulationSettings(binary = T, includeAllOutcomes = T,
   firstExposureOnly = FALSE, washoutPeriod = 0, removeSubjectsWithPriorOutcome = TRUE,
   priorOutcomeLookback = 99999, requireTimeAtRisk = F, riskWindowStart = 1, startAnchor
     = "cohort start", riskWindowEnd = 3650, endAnchor = "cohort start"),
  riskTable = T,
  confInt = T,
  yLabel = "Fraction of those who are outcome free in target population"
)
```

Arguments

`plpData`	The plpData object returned by running getPlpData()
`outcomeId`	The cohort id corresponding to the outcome
`populationSettings`	The population settings created using `createStudyPopulationSettings`
`riskTable`	(binary) Whether to include a table at the bottom of the plot showing the number of people at risk over time
`confInt`	(binary) Whether to include a confidence interval
`yLabel`	(string) The label for the y-axis

Details

This creates a survival plot that can be used to pick a suitable time-at-risk period

Value

TRUE if it ran

`pfi` *pfi*

Description

Calculate the permutation feature importance for a PLP model.

Usage

```
pfi(
  plpResult,
  population,
  plpData,
  repeats = 1,
  covariates = NULL,
  cores = NULL,
  log = NULL,
  logthreshold = "INFO"
)
```

Arguments

`plpResult`	An object of type `runPlp`
`population`	The population created using createStudyPopulation() who will have their risks predicted
`plpData`	An object of type `plpData` - the patient level prediction data extracted from the CDM.
`repeats`	The number of times to permute each covariate
`covariates`	A vector of covariates to calculate the pfi for. If NULL it uses all covariates included in the model.
`cores`	Number of cores to use when running this (it runs in parallel)
`log`	A location to save the log for running pfi
`logthreshold`	The log threshold (e.g., INFO, TRACE, ...)

Details

The function permutes the each covariate/features <repeats> times and calculates the mean AUC change caused by the permutation.

Value

A dataframe with the covariateIds and the pfi (change in AUC caused by permuting the covariate) value

---

`plotDemographicSummary`
Plot the Observed vs. expected incidence, by age and gender

Description

Plot the Observed vs. expected incidence, by age and gender

Usage

```
plotDemographicSummary(
  plpResult,
  typeColumn = "evaluation",
  saveLocation = NULL,
  fileName = "roc.png"
)
```

Arguments

`plpResult`	A plp result object as generated using the `runPlp` function.
`typeColumn`	The name of the column specifying the evaluation type
`saveLocation`	Directory to save plot (if NULL plot is not saved)
`fileName`	Name of the file to save to plot, for example 'plot.png'. See the function ggsave in the ggplot2 package for supported file formats.

Details

Create a plot showing the Observed vs. expected incidence, by age and gender #'

Value

A ggplot object. Use the `ggsave` function to save to file in a different format.

`plotF1Measure` *Plot the F1 measure efficiency frontier using the sparse thresholdSummary data frame*

Description

Plot the F1 measure efficiency frontier using the sparse thresholdSummary data frame

Usage

```
plotF1Measure(
  plpResult,
  typeColumn = "evaluation",
  saveLocation = NULL,
  fileName = "roc.png"
)
```

Arguments

`plpResult`	A plp result object as generated using the `runPlp` function.
`typeColumn`	The name of the column specifying the evaluation type
`saveLocation`	Directory to save plot (if NULL plot is not saved)
`fileName`	Name of the file to save to plot, for example 'plot.png'. See the function ggsave in the ggplot2 package for supported file formats.

Details

Create a plot showing the F1 measure efficiency frontier using the sparse thresholdSummary data frame

Value

A ggplot object. Use the `ggsave` function to save to file in a different format.

`plotGeneralizability` *Plot the train/test generalizability diagnostic*

Description

Plot the train/test generalizability diagnostic

Usage

```
plotGeneralizability(
  covariateSummary,
  saveLocation = NULL,
  fileName = "Generalizability.png"
)
```

Arguments

`covariateSummary`	A prediction object as generated using the `runPlp` function.
`saveLocation`	Directory to save plot (if NULL plot is not saved)
`fileName`	Name of the file to save to plot, for example 'plot.png'. See the function ggsave in the ggplot2 package for supported file formats.

Details

Create a plot showing the train/test generalizability diagnostic #'

Value

A ggplot object. Use the `ggsave` function to save to file in a different format.

`plotLearningCurve` *plotLearningCurve*

Description

Create a plot of the learning curve using the object returned from `createLearningCurve`.

Usage

```
plotLearningCurve(
  learningCurve,
  metric = "AUROC",
  abscissa = "events",
  plotTitle = "Learning Curve",
  plotSubtitle = NULL,
  fileName = NULL
)
```

Arguments

`learningCurve` An object returned by `createLearningCurve` function.

`metric` Specifies the metric to be plotted:

- `'AUROC'` - use the area under the Receiver Operating Characteristic curve
- `'AUPRC'` - use the area under the Precision-Recall curve
- `'sBrier'` - use the scaled Brier score

`abscissa` Specify the abscissa metric to be plotted:

- `'events'` - use number of events
- `'observations'` - use number of observations

`plotTitle` Title of the learning curve plot.

`plotSubtitle` Subtitle of the learning curve plot.

`fileName` Filename of plot to be saved, for example `'plot.png'`. See the function ggsave in the ggplot2 package for supported file formats.

Value

A ggplot object. Use the ggsave function to save to file in a different format.

Examples

```
## Not run:
# create learning curve object
learningCurve <- createLearningCurve(population,
                                     plpData,
                                     modelSettings)
# plot the learning curve
plotLearningCurve(learningCurve)

## End(Not run)
```

---

plotPlp *Plot all the PatientLevelPrediction plots*

Description

Plot all the PatientLevelPrediction plots

Usage

```
plotPlp(plpResult, saveLocation = NULL, typeColumn = "evaluation")
```

Arguments

plpResult	Object returned by the runPlp() function
saveLocation	Name of the directory where the plots should be saved (NULL means no saving)
typeColumn	The name of the column specifying the evaluation type (to stratify the plots)

Details

Create a directory with all the plots

Value

TRUE if it ran

plotPrecisionRecall *Plot the precision-recall curve using the sparse thresholdSummary data frame*

Description

Plot the precision-recall curve using the sparse thresholdSummary data frame

Usage

```
plotPrecisionRecall(
  plpResult,
  typeColumn = "evaluation",
  saveLocation = NULL,
  fileName = "roc.png"
)
```

Arguments

plpResult	A plp result object as generated using the runPlp function.
typeColumn	The name of the column specifying the evaluation type
saveLocation	Directory to save plot (if NULL plot is not saved)
fileName	Name of the file to save to plot, for example 'plot.png'. See the function ggsave in the ggplot2 package for supported file formats.

Details

Create a plot showing the precision-recall curve using the sparse thresholdSummary data frame

Value

A ggplot object. Use the ggsave function to save to file in a different format.

plotPredictedPDF *Plot the Predicted probability density function, showing prediction overlap between true and false cases*

Description

Plot the Predicted probability density function, showing prediction overlap between true and false cases

Usage

```
plotPredictedPDF(
  plpResult,
  typeColumn = "evaluation",
  saveLocation = NULL,
  fileName = "PredictedPDF.png"
)
```

Arguments

`plpResult`	A plp result object as generated using the `runPlp` function.
`typeColumn`	The name of the column specifying the evaluation type
`saveLocation`	Directory to save plot (if NULL plot is not saved)
`fileName`	Name of the file to save to plot, for example 'plot.png'. See the function ggsave in the ggplot2 package for supported file formats.

Details

Create a plot showing the predicted probability density function, showing prediction overlap between true and false cases

Value

A ggplot object. Use the `ggsave` function to save to file in a different format.

`plotPredictionDistribution` *Plot the side-by-side boxplots of prediction distribution, by class#'*

Description

Plot the side-by-side boxplots of prediction distribution, by class#'

Usage

```
plotPredictionDistribution(
  plpResult,
  typeColumn = "evaluation",
  saveLocation = NULL,
  fileName = "PredictionDistribution.png"
)
```

Arguments

`plpResult`	A plp result object as generated using the `runPlp` function.
`typeColumn`	The name of the column specifying the evaluation type
`saveLocation`	Directory to save plot (if NULL plot is not saved)
`fileName`	Name of the file to save to plot, for example 'plot.png'. See the function ggsave in the ggplot2 package for supported file formats.

Details

Create a plot showing the side-by-side boxplots of prediction distribution, by class #'

Value

A ggplot object. Use the `ggsave` function to save to file in a different format.

`plotPreferencePDF` *Plot the preference score probability density function, showing prediction overlap between true and false cases #'*

Description

Plot the preference score probability density function, showing prediction overlap between true and false cases #'

Usage

```
plotPreferencePDF(
  plpResult,
  typeColumn = "evaluation",
  saveLocation = NULL,
  fileName = "plotPreferencePDF.png"
)
```

Arguments

`plpResult`	A plp result object as generated using the `runPlp` function.
`typeColumn`	The name of the column specifying the evaluation type
`saveLocation`	Directory to save plot (if NULL plot is not saved)
`fileName`	Name of the file to save to plot, for example 'plot.png'. See the function ggsave in the ggplot2 package for supported file formats.

Details

Create a plot showing the preference score probability density function, showing prediction overlap between true and false cases #'

Value

A ggplot object. Use the `ggsave` function to save to file in a different format.

`plotSmoothCalibration` *Plot the smooth calibration as detailed in Calster et al. "A calibration heirarchy for risk models was defined: from utopia to empirical data" (2016)*

Description

Plot the smooth calibration as detailed in Calster et al. "A calibration heirarchy for risk models was defined: from utopia to empirical data" (2016)

Usage

```
plotSmoothCalibration(
  plpResult,
  smooth = "loess",
  span = 0.75,
  nKnots = 5,
  scatter = FALSE,
  bins = 20,
  sample = TRUE,
  typeColumn = "evaluation",
  saveLocation = NULL,
  fileName = "smoothCalibration.pdf"
)
```

Arguments

`plpResult`	The result of running `runPlp` function. An object containing the model or location where the model is save, the data selection settings, the preprocessing and training settings as well as various performance measures obtained by the model.
`smooth`	options: 'loess' or 'rcs'
`span`	This specifies the width of span used for loess. This will allow for faster computing and lower memory usage.
`nKnots`	The number of knots to be used by the rcs evaluation. Default is 5
`scatter`	plot the decile calibrations as points on the graph. Default is False
`bins`	The number of bins for the histogram. Default is 20.
`sample`	If using loess then by default 20,000 patients will be sampled to save time
`typeColumn`	The name of the column specifying the evaluation type
`saveLocation`	Directory to save plot (if NULL plot is not saved)
`fileName`	Name of the file to save to plot, for example 'plot.png'. See the function ggsave in the ggplot2 package for supported file formats.

Details

Create a plot showing the smoothed calibration #'

Value

A ggplot object.

`plotSparseCalibration` *Plot the calibration*

Description

Plot the calibration

Usage

```
plotSparseCalibration(
  plpResult,
  typeColumn = "evaluation",
  saveLocation = NULL,
  fileName = "roc.png"
)
```

Arguments

`plpResult`	A plp result object as generated using the `runPlp` function.
`typeColumn`	The name of the column specifying the evaluation type
`saveLocation`	Directory to save plot (if NULL plot is not saved)
`fileName`	Name of the file to save to plot, for example 'plot.png'. See the function ggsave in the ggplot2 package for supported file formats.

Details

Create a plot showing the calibration #'

Value

A ggplot object. Use the `ggsave` function to save to file in a different format.

`plotSparseCalibration2` *Plot the conventional calibration*

Description

Plot the conventional calibration

Usage

```
plotSparseCalibration2(
  plpResult,
  typeColumn = "evaluation",
  saveLocation = NULL,
  fileName = "roc.png"
)
```

Arguments

`plpResult`	A plp result object as generated using the `runPlp` function.
`typeColumn`	The name of the column specifying the evaluation type
`saveLocation`	Directory to save plot (if NULL plot is not saved)
`fileName`	Name of the file to save to plot, for example 'plot.png'. See the function ggsave in the ggplot2 package for supported file formats.

Details

Create a plot showing the calibration #'

Value

A ggplot object. Use the `ggsave` function to save to file in a different format.

`plotSparseRoc` *Plot the ROC curve using the sparse thresholdSummary data frame*

Description

Plot the ROC curve using the sparse thresholdSummary data frame

Usage

```
plotSparseRoc(
  plpResult,
  typeColumn = "evaluation",
  saveLocation = NULL,
  fileName = "roc.png"
)
```

Arguments

`plpResult`	A plp result object as generated using the `runPlp` function.
`typeColumn`	The name of the column specifying the evaluation type
`saveLocation`	Directory to save plot (if NULL plot is not saved)
`fileName`	Name of the file to save to plot, for example 'plot.png'. See the function ggsave in the ggplot2 package for supported file formats.

Details

Create a plot showing the Receiver Operator Characteristics (ROC) curve.

Value

A ggplot object. Use the `ggsave` function to save to file in a different format.

plotVariableScatterplot
Plot the variable importance scatterplot

Description

Plot the variable importance scatterplot

Usage

```
plotVariableScatterplot(
  covariateSummary,
  saveLocation = NULL,
  fileName = "VariableScatterplot.png"
)
```

Arguments

`covariateSummary`	A prediction object as generated using the `runPlp` function.
`saveLocation`	Directory to save plot (if NULL plot is not saved)
`fileName`	Name of the file to save to plot, for example 'plot.png'. See the function ggsave in the ggplot2 package for supported file formats.

Details

Create a plot showing the variable importance scatterplot #'

Value

A ggplot object. Use the `ggsave` function to save to file in a different format.

plpDataSimulationProfile
A simulation profile

Description

A simulation profile

Usage

```
data(plpDataSimulationProfile)
```

Format

A data frame containing the following elements:

covariatePrevalence prevalence of all covariates

outcomeModels regression model parameters to simulate outcomes

metaData settings used to simulate the profile

covariateRef covariateIds and covariateNames

timePrevalence time window

exclusionPrevalence prevalence of exclusion of covariates

`positiveLikelihoodRatio` *Calculate the positiveLikelihoodRatio*

Description

Calculate the positiveLikelihoodRatio

Usage

```
positiveLikelihoodRatio(TP, TN, FN, FP)
```

Arguments

`TP`	Number of true positives
`TN`	Number of true negatives
`FN`	Number of false negatives
`FP`	Number of false positives

Details

Calculate the positiveLikelihoodRatio

Value

positiveLikelihoodRatio value

`positivePredictiveValue` *Calculate the positivePredictiveValue*

Description

Calculate the positivePredictiveValue

Usage

```
positivePredictiveValue(TP, TN, FN, FP)
```

Arguments

`TP`	Number of true positives
`TN`	Number of true negatives
`FN`	Number of false negatives
`FP`	Number of false positives

Details

Calculate the positivePredictiveValue

Value

positivePredictiveValue value

`predictCyclops` *Create predictive probabilities*

Description

Create predictive probabilities

Usage

```
predictCyclops(plpModel, data, cohort)
```

Arguments

`plpModel`	An object of type `predictiveModel` as generated using `fitPlp`.
`data`	The new plpData containing the covariateData for the new population
`cohort`	The cohort to calculate the prediction for

Details

Generates predictions for the population specified in plpData given the model.

Value

The value column in the result data.frame is: logistic: probabilities of the outcome, poisson: Poisson rate (per day) of the outome, survival: hazard rate (per day) of the outcome.

`predictPlp` *predictPlp*

Description

Predict the risk of the outcome using the input plpModel for the input plpData

Usage

```
predictPlp(plpModel, plpData, population, timepoint)
```

Arguments

`plpModel`	An object of type `plpModel` - a patient level prediction model
`plpData`	An object of type `plpData` - the patient level prediction data extracted from the CDM.
`population`	The population created using createStudyPopulation() who will have their risks predicted or a cohort without the outcome known
`timepoint`	The timepoint to predict risk (survival models only)

Details

The function applied the trained model on the plpData to make predictions

Value

A dataframe containing the prediction for each person in the population with an attribute metaData containing prediction details.

`preprocessData` *A function that wraps around FeatureExtraction::tidyCovariateData to normalise the data and remove rare or redundant features*

Description

A function that wraps around FeatureExtraction::tidyCovariateData to normalise the data and remove rare or redundant features

Usage

```
preprocessData(covariateData, preprocessSettings)
```

Arguments

`covariateData`	The covariate part of the training data created by `splitData` after being sampled and having any required feature engineering
`preprocessSettings`	The settings for the preprocessing created by `createPreprocessSettings`

Details

Returns an object of class `covariateData` that has been processed

Value

The data processed

`recalibratePlp` *recalibratePlp*

Description

Train various models using a default parameter gird search or user specified parameters

Usage

```
recalibratePlp(
  prediction,
  analysisId,
  typeColumn = "evaluationType",
  method = c("recalibrationInTheLarge", "weakRecalibration")
)
```

Arguments

`prediction`	A prediction dataframe
`analysisId`	The model analysisId
`typeColumn`	The column name where the strata types are specified
`method`	Method used to recalibrate ('recalibrationInTheLarge' or 'weakRecalibration')

Details

The user can define the machine learning model to train (regularised logistic regression, random forest, gradient boosting machine, neural network and)

Value

An object of class `runPlp` that is recalibrated on the new data

`recalibratePlpRefit` *recalibratePlpRefit*

Description

Train various models using a default parameter gird search or user specified parameters

Usage

```
recalibratePlpRefit(plpModel, newPopulation, newData)
```

Arguments

`plpModel`	The trained plpModel (runPlp$model)
`newPopulation`	The population created using createStudyPopulation() who will have their risks predicted
`newData`	An object of type `plpData` - the patient level prediction data extracted from the CDM.

Details

The user can define the machine learning model to train (regularised logistic regression, random forest, gradient boosting machine, neural network and)

Value

An object of class `runPlp` that is recalibrated on the new data

`runMultiplePlp` *Run a list of predictions analyses*

Description

Run a list of predictions analyses

Usage

```
runMultiplePlp(
  databaseDetails = createDatabaseDetails(),
 modelDesignList = list(createModelDesign(targetId = 1, outcomeId = 2, modelSettings =
   setLassoLogisticRegression()), createModelDesign(targetId = 1, outcomeId = 3,
    modelSettings = setLassoLogisticRegression())),
  onlyFetchData = F,
  cohortDefinitions = NULL,
  logSettings = createLogSettings(verbosity = "DEBUG", timeStamp = T, logName =
    "runPlp Log"),
  saveDirectory = getwd(),
  sqliteLocation = file.path(saveDirectory, "sqlite")
)
```

Arguments

`databaseDetails`	The database settings created using `createDatabaseDetails()`
`modelDesignList`	A list of model designs created using `createModelDesign()`
`onlyFetchData`	Only fetches and saves the data object to the output folder without running the analysis.
`cohortDefinitions`	A list of cohort definitions for the target and outcome cohorts
`logSettings`	The setting specifying the logging for the analyses created using `createLogSettings()`
`saveDirectory`	Name of the folder where all the outputs will written to.
`sqliteLocation`	(optional) The location of the sqlite database with the results

Details

This function will run all specified predictions as defined using .

Value

A data frame with the following columns:

`analysisId`	The unique identifier for a set of analysis choices.
`targetId`	The ID of the target cohort populations.
`outcomeId`	The ID of the outcomeId.
`dataLocation`	The location where the plpData was saved
`the settings ids`	The ids for all other settings used for model development.

`runPlp` *runPlp - Develop and internally evaluate a model using specified settings*

Description

This provides a general framework for training patient level prediction models. The user can select various default feature selection methods or incorporate their own, The user can also select from a range of default classifiers or incorporate their own. There are three types of evaluations for the model patient (randomly splits people into train/validation sets) or year (randomly splits data into train/validation sets based on index year - older in training, newer in validation) or both (same as year spliting but checks there are no overlaps in patients within training set and validaiton set - any overlaps are removed from validation set)

Usage

```
runPlp(
  plpData,
  outcomeId = plpData$metaData$call$outcomeIds[1],
  analysisId = paste(Sys.Date(), plpData$metaData$call$outcomeIds[1], sep = "-"),
  analysisName = "Study details",
  populationSettings = createStudyPopulationSettings(),
  splitSettings = createDefaultSplitSetting(type = "stratified", testFraction = 0.25,
    trainFraction = 0.75, splitSeed = 123, nfold = 3),
  sampleSettings = createSampleSettings(type = "none"),
  featureEngineeringSettings = createFeatureEngineeringSettings(type = "none"),
  preprocessSettings = createPreprocessSettings(minFraction = 0.001, normalize = T),
  modelSettings = setLassoLogisticRegression(),
  logSettings = createLogSettings(verbosity = "DEBUG", timeStamp = T, logName =
    "runPlp Log"),
  executeSettings = createDefaultExecuteSettings(),
  saveDirectory = getwd()
)
```

Arguments

`plpData`	An object of type `plpData` - the patient level prediction data extracted from the CDM. Can also include an initial population as plpData$popualtion.
`outcomeId`	(integer) The ID of the outcome.
`analysisId`	(integer) Identifier for the analysis. It is used to create, e.g., the result folder. Default is a timestamp.
`analysisName`	(character) Name for the analysis
`populationSettings`	An object of type `populationSettings` created using `createStudyPopulationSettings` that specifies how the data class labels are defined and addition any exclusions to apply to the plpData cohort
`splitSettings`	An object of type `splitSettings` that specifies how to split the data into train/validation/test. The default settings can be created using `createDefaultSplitSetting`.
`sampleSettings`	An object of type `sampleSettings` that specifies any under/over sampling to be done. The default is none.
`featureEngineeringSettings`	An object of `featureEngineeringSettings` specifying any feature engineering to be learned (using the train data)
`preprocessSettings`	An object of `preprocessSettings`. This setting specifies the minimum fraction of target population who must have a covariate for it to be included in the model training and whether to normalise the covariates before training
`modelSettings`	An object of class `modelSettings` created using one of the function: • setLassoLogisticRegression() A lasso logistic regression model • setGradientBoostingMachine() A gradient boosting machine • setAdaBoost() An ada boost model • setRandomForest() A random forest model • setDecisionTree() A decision tree model • setKNN() A KNN model
`logSettings`	An object of `logSettings` created using `createLogSettings` specifying how the logging is done
`executeSettings`	An object of `executeSettings` specifying which parts of the analysis to run
`saveDirectory`	The path to the directory where the results will be saved (if NULL uses working directory)

Details

This function takes as input the plpData extracted from an OMOP CDM database and follows the specified settings to develop and internally validate a model for the specified outcomeId.

Value

An object containing the following:

- model The developed model of class `plpModel`
- executionSummary A list containing the hardward details, R package details and execution time
- performanceEvaluation Various internal performance metrics in sparse format
- prediction The plpData cohort table with the predicted risks added as a column (named value)
- covariateSummary A characterization of the features for patients with and without the outcome during the time at risk
- analysisRef A list with details about the analysis

Examples

```
## Not run:
#******** EXAMPLE 1 *********
#load plpData:
plpData <- loadPlpData(file.path('C:','User','home','data'))

# specify the outcome to predict (the plpData can have multiple outcomes)
outcomeId <- 2042

# specify a unique identifier for the analysis
analysisId <- 'lrModel'

# create population settings (this defines the labels in the data)
#create study population to develop model on
#require minimum of 365 days observation prior to at risk start
#no prior outcome and person must be observed for 365 after index (minTimeAtRisk)
#with risk window from 0 to 365 days after index
populationSettings <- createStudyPopulationSettings(plpData,
                                    firstExposureOnly = FALSE,
                                    washoutPeriod = 365,
                                    removeSubjectsWithPriorOutcome = TRUE,
                                    priorOutcomeLookback = 99999,
                                    requireTimeAtRisk = TRUE,
                                    minTimeAtRisk=365,
                                    riskWindowStart = 0,
                                    addExposureDaysToStart = FALSE,
                                    riskWindowEnd = 365,
                                    addExposureDaysToEnd = FALSE)

# create the split setting by specifying how you want to
# partition the data into development (train/validation) and evaluation (test or CV)
splitSettings <- createDefaultSplitSetting(testFraction = 0.25,
                                           trainFraction = 0.75,
                                           splitSeed = sample(100000,1),
                                           nfold=3,
                                           type = 'stratified')
```

```
# create the settings specifying any under/over sampling
# in this example we do not do any
sampleSettings <- createSampleSettings(type = 'none')

# specify any feature engineering that will be applied to the train data
# in this example we do not do any
featureEngineeringSettings <- createFeatureEngineeringSettings(type = 'none')

# specify whether to use normalization and removal of rare features
# preprocessSettings <- ...


#lasso logistic regression predicting outcome 200 in cohorts 10
#using no feature selection with a time split evaluation with 30% in test set
#70% in train set where the model hyper-parameters are selected using 3-fold cross validation:
#and results are saved to file.path('C:','User','home')
modelSettingsLR <- setLassoLogisticRegression()

# specify how you want the logging for the analysis
# generally this is saved in a file with the results
# but you can define the level of logging
logSettings <- createLogSettings(verbosity = 'DEBUG',
                                 timeStamp = T,
                                 logName = 'runPlp LR Log')

# specify what parts of the analysis to run:
# in this example we run everything
executeSettings <- createExecuteSettings(runSplitData = T,
                                         runSampleData = T,
                                         runfeatureEngineering = T,
                                         runProcessData = T,
                                         runModelDevelopment = T,
                                         runCovariateSummary = T)

lrModel <- runPlp(plpData = plpData,
                  outcomeId = outcomeId,
                  analysisId = analysisId,
                  populationSettings = populationSettings,
                  splitSettings = splitSettings,
                  sampleSettings = sampleSettings,
                  featureEngineeringSettings = featureEngineeringSettings,
                  preprocessSettings = preprocessSettings,
                  modelSettings = modelSettingsLR,
                  logSettings = logSettings
                  executeSettings = executeSettings,
                  saveDirectory = saveDirectory
                  )

#******** EXAMPLE 2 *********
# Gradient boosting machine with a grid search to select hyper parameters
# using the test/train/folds created for the lasso logistic regression above
modelSettingsGBM <- gradientBoostingMachine.set(rsampRate=c(0.5,0.9,1),csampRate=1,
                           ntrees=c(10,100), bal=c(F,T),
                           max_depth=c(4,5), learn_rate=c(0.1,0.01))
```

```
analysisId <- 'gbmModel'

gbmModel <- runPlp(plpData = plpData,
                  outcomeId = outcomeId,
                  analysisId = analysisId,
                  populationSettings = populationSettings,
                  splitSettings = splitSettings,
                  sampleSettings = sampleSettings,
                  featureEngineeringSettings = featureEngineeringSettings,
                  preprocessSettings = preprocessSettings,
                  modelSettings = modelSettingsGBM,
                  logSettings = logSettings
                  executeSettings = executeSettings,
                  saveDirectory = saveDirectory
                  )

## End(Not run)
```

---

savePlpAnalysesJson *Save the modelDesignList to a json file*

Description

Save the modelDesignList to a json file

Usage

```
savePlpAnalysesJson(
  modelDesignList = list(createModelDesign(targetId = 1, outcomeId = 2, modelSettings =
    setLassoLogisticRegression()), createModelDesign(targetId = 1, outcomeId = 3,
     modelSettings = setLassoLogisticRegression())),
  cohortDefinitions = NULL,
  saveDirectory = NULL
)
```

Arguments

modelDesignList	A list of modelDesigns created using `createModelDesign()`
cohortDefinitions	A list of the cohortDefinitions (generally extracted from ATLAS)
saveDirectory	The directory to save the modelDesignList settings

Details

This function creates a json file with the modelDesignList saved

Examples

```
## Not run:
savePlpAnalysesJson(
modelDesignList = list(
createModelDesign(targetId = 1, outcomeId = 2, modelSettings = setLassoLogisticRegression()),
createModelDesign(targetId = 1, outcomeId = 3, modelSettings = setLassoLogisticRegression())
),
saveDirectory = 'C:/bestModels'
)

## End(Not run)
```

---

savePlpData *Save the cohort data to folder*

Description

savePlpData saves an object of type plpData to folder.

Usage

```
savePlpData(plpData, file, envir = NULL, overwrite = F)
```

Arguments

plpData	An object of type plpData as generated using getPlpData.
file	The name of the folder where the data will be written. The folder should not yet exist.
envir	The environment for to evaluate variables when saving
overwrite	Whether to force overwrite an existing file

Details

The data will be written to a set of files in the folder specified by the user.

Examples

```
# todo
```

`savePlpModel` *Saves the plp model*

Description

Saves the plp model

Usage

```
savePlpModel(plpModel, dirPath)
```

Arguments

`plpModel`	A trained classifier returned by running `runPlp()$model`
`dirPath`	A location to save the model to

Details

Saves the plp model to a user specificed folder

`savePlpResult` *Saves the result from runPlp into the location directory*

Description

Saves the result from runPlp into the location directory

Usage

```
savePlpResult(result, dirPath)
```

Arguments

`result`	The result of running runPlp()
`dirPath`	The directory to save the csv

Details

Saves the result from runPlp into the location directory

`savePlpShareable` *Save the plp result as json files and csv files for transparent sharing*

Description

Save the plp result as json files and csv files for transparent sharing

Usage

```
savePlpShareable(result, saveDirectory, minCellCount = 10)
```

Arguments

`result`	An object of class runPlp with development or validation results
`saveDirectory`	The directory the save the results as csv files
`minCellCount`	Minimum cell count for the covariateSummary and certain evaluation results

Details

Saves the main results json/csv files (these files can be read by the shiny app)

`savePrediction` *Saves the prediction dataframe to RDS*

Description

Saves the prediction dataframe to RDS

Usage

```
savePrediction(prediction, dirPath, fileName = "prediction.rds")
```

Arguments

`prediction`	The prediciton data.frame
`dirPath`	The directory to save the prediction RDS
`fileName`	The name of the RDS file that will be saved in dirPath

Details

Saves the prediction data frame returned by predict.R to an RDS file and returns the fileLocation where the prediction is saved

sensitivity — *Calculate the sensitivity*

Description

Calculate the sensitivity

Usage

```
sensitivity(TP, TN, FN, FP)
```

Arguments

`TP`	Number of true positives
`TN`	Number of true negatives
`FN`	Number of false negatives
`FP`	Number of false positives

Details

Calculate the sensitivity

Value

sensitivity value

setAdaBoost — *Create setting for AdaBoost with python DecisionTreeClassifier base estimator*

Description

Create setting for AdaBoost with python DecisionTreeClassifier base estimator

Usage

```
setAdaBoost(
  nEstimators = list(10, 50, 200),
  learningRate = list(1, 0.5, 0.1),
  algorithm = list("SAMME.R"),
  seed = sample(1e+06, 1)
)
```

Arguments

`nEstimators`	(list) The maximum number of estimators at which boosting is terminated. In case of perfect fit, the learning procedure is stopped early.
`learningRate`	(list) Weight applied to each classifier at each boosting iteration. A higher learning rate increases the contribution of each classifier. There is a trade-off between the learningRate and nEstimators parameters There is a trade-off between learningRate and nEstimators.
`algorithm`	(list) If 'SAMME.R' then use the SAMME.R real boosting algorithm. base_estimator must support calculation of class probabilities. If 'SAMME' then use the SAMME discrete boosting algorithm. The SAMME.R algorithm typically converges faster than SAMME, achieving a lower test error with fewer boosting iterations.
`seed`	A seed for the model

Examples

```
## Not run:
model.adaBoost <- setAdaBoost(nEstimators = list(10,50,200), learningRate = list(1, 0.5, 0.1),
                              algorithm = list('SAMME.R'), seed = sample(1000000,1)
                              )

## End(Not run)
```

`setCoxModel` *Create setting for lasso Cox model*

Description

Create setting for lasso Cox model

Usage

```
setCoxModel(
  variance = 0.01,
  seed = NULL,
  includeCovariateIds = c(),
  noShrinkage = c(),
  threads = -1,
  upperLimit = 20,
  lowerLimit = 0.01,
  tolerance = 2e-07,
  maxIterations = 3000
)
```

Arguments

`variance`	Numeric: prior distribution starting variance
`seed`	An option to add a seed when training the model
`includeCovariateIds`	a set of covariate IDS to limit the analysis to

`noShrinkage`	a set of covariates whcih are to be forced to be included in the final model. default is the intercept
`threads`	An option to set number of threads when training model
`upperLimit`	Numeric: Upper prior variance limit for grid-search
`lowerLimit`	Numeric: Lower prior variance limit for grid-search
`tolerance`	Numeric: maximum relative change in convergence criterion from successive iterations to achieve convergence
`maxIterations`	Integer: maximum iterations of Cyclops to attempt before returning a failed-to-converge error

Examples

```
model.lr <- setCoxModel()
```

`setDecisionTree` *Create setting for the scikit-learn 1.0.1 DecisionTree with python*

Description

Create setting for the scikit-learn 1.0.1 DecisionTree with python

Usage

```
setDecisionTree(
  criterion = list("gini"),
  splitter = list("best"),
  maxDepth = list(as.integer(4), as.integer(10), NULL),
  minSamplesSplit = list(2, 10),
  minSamplesLeaf = list(10, 50),
  minWeightFractionLeaf = list(0),
  maxFeatures = list(100, "sqrt", NULL),
  maxLeafNodes = list(NULL),
  minImpurityDecrease = list(10^-7),
  classWeight = list(NULL),
  seed = sample(1e+06, 1)
)
```

Arguments

`criterion`	The function to measure the quality of a split. Supported criteria are “gini” for the Gini impurity and “entropy” for the information gain.
`splitter`	The strategy used to choose the split at each node. Supported strategies are “best” to choose the best split and “random” to choose the best random split.
`maxDepth`	(list) The maximum depth of the tree. If NULL, then nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples.
`minSamplesSplit`	The minimum number of samples required to split an internal node

`minSamplesLeaf`	The minimum number of samples required to be at a leaf node. A split point at any depth will only be considered if it leaves at least minSamplesLeaf training samples in each of the left and right branches. This may have the effect of smoothing the model, especially in regression.
`minWeightFractionLeaf`	The minimum weighted fraction of the sum total of weights (of all the input samples) required to be at a leaf node. Samples have equal weight when sampleWeight is not provided.
`maxFeatures`	(list) The number of features to consider when looking for the best split (int/'sqrt'/NULL)
`maxLeafNodes`	(list) Grow a tree with max_leaf_nodes in best-first fashion. Best nodes are defined as relative reduction in impurity. If None then unlimited number of leaf nodes. (int/NULL)
`minImpurityDecrease`	Threshold for early stopping in tree growth. A node will split if its impurity is above the threshold, otherwise it is a leaf.
`classWeight`	(list) Weights associated with classes 'balance' or NULL
`seed`	The random state seed

Examples

```
## Not run:
model.decisionTree <- setDecisionTree(maxDepth=10,minSamplesLeaf=10, seed=NULL )

## End(Not run)
```

---

setGradientBoostingMachine

Create setting for gradient boosting machine model using gbm_xgboost implementation

---

Description

Create setting for gradient boosting machine model using gbm_xgboost implementation

Usage

```
setGradientBoostingMachine(
  ntrees = c(100, 300),
  nthread = 20,
  earlyStopRound = 25,
  maxDepth = c(4, 6, 8),
  minChildWeight = 1,
  learnRate = c(0.05, 0.1, 0.3),
  scalePosWeight = 1,
  lambda = 1,
  alpha = 0,
  seed = sample(1e+07, 1)
)
```

Arguments

`ntrees`	The number of trees to build
`nthread`	The number of computer threads to use (how many cores do you have?)
`earlyStopRound`	If the performance does not increase over earlyStopRound number of trees then training stops (this prevents overfitting)
`maxDepth`	Maximum depth of each tree - a large value will lead to slow model training
`minChildWeight`	Minimum sum of of instance weight in a child node - larger values are more conservative
`learnRate`	The boosting learn rate
`scalePosWeight`	Controls weight of positive class in loss - useful for imbalanced classes
`lambda`	L2 regularization on weights - larger is more conservative
`alpha`	L1 regularization on weights - larger is more conservative
`seed`	An option to add a seed when training the final model

Examples

```
model.gbm <- setGradientBoostingMachine(ntrees=c(10,100), nthread=20,
                        maxDepth=c(4,6), learnRate=c(0.1,0.3))
```

---

setIterativeHardThresholding
Create setting for lasso logistic regression

Description

Create setting for lasso logistic regression

Usage

```
setIterativeHardThresholding(
  K = 10,
  penalty = "bic",
  seed = sample(1e+05, 1),
  exclude = c(),
  forceIntercept = F,
  fitBestSubset = FALSE,
  initialRidgeVariance = 10000,
  tolerance = 1e-08,
  maxIterations = 10000,
  threshold = 1e-06,
  delta = 0
)
```

Arguments

`K`	The maximum number of non-zero predictors
`penalty`	Specifies the IHT penalty; possible values are ‘BIC‘ or ‘AIC‘ or a numeric value
`seed`	An option to add a seed when training the model
`exclude`	A vector of numbers or covariateId names to exclude from prior
`forceIntercept`	Logical: Force intercept coefficient into regularization
`fitBestSubset`	Logical: Fit final subset with no regularization
`initialRidgeVariance`	integer
`tolerance`	numeric
`maxIterations`	integer
`threshold`	numeric
`delta`	numeric

Examples

```
model.lr <- setLassoLogisticRegression()
```

`setKNN` *Create setting for knn model*

Description

Create setting for knn model

Usage

```
setKNN(k = 1000, indexFolder = file.path(getwd(), "knn"), threads = 1)
```

Arguments

`k`	The number of neighbors to consider
`indexFolder`	The directory where the results and intermediate steps are output
`threads`	The number of threads to use when applying big knn

Examples

```
## Not run:
model.knn <- setKNN(k=10000)

## End(Not run)
```

setLassoLogisticRegression

Create setting for lasso logistic regression

Description

Create setting for lasso logistic regression

Usage

```
setLassoLogisticRegression(
  variance = 0.01,
  seed = NULL,
  includeCovariateIds = c(),
  noShrinkage = c(0),
  threads = -1,
  forceIntercept = F,
  upperLimit = 20,
  lowerLimit = 0.01,
  tolerance = 2e-06,
  maxIterations = 3000,
  priorCoefs = NULL
)
```

Arguments

`variance`	Numeric: prior distribution starting variance
`seed`	An option to add a seed when training the model
`includeCovariateIds`	a set of covariate IDS to limit the analysis to
`noShrinkage`	a set of covariates whcih are to be forced to be included in the final model. default is the intercept
`threads`	An option to set number of threads when training model
`forceIntercept`	Logical: Force intercept coefficient into prior
`upperLimit`	Numeric: Upper prior variance limit for grid-search
`lowerLimit`	Numeric: Lower prior variance limit for grid-search
`tolerance`	Numeric: maximum relative change in convergence criterion from successive iterations to achieve convergence
`maxIterations`	Integer: maximum iterations of Cyclops to attempt before returning a failed-to-converge error
`priorCoefs`	Use coefficients from a previous model as starting points for model fit (transfer learning)

Examples

```
model.lr <- setLassoLogisticRegression()
```

setLightGBM *Create setting for gradient boosting machine model using lightGBM (https://github.com/microsoft/LightGBM/tree/master/R-package).*

Description

Create setting for gradient boosting machine model using lightGBM (https://github.com/microsoft/LightGBM/tree/maste
package).

Usage

```
setLightGBM(
  nthread = 20,
  earlyStopRound = 25,
  numIterations = c(100),
  numLeaves = c(31),
  maxDepth = c(5, 10),
  minDataInLeaf = c(20),
  learningRate = c(0.05, 0.1, 0.3),
  lambdaL1 = c(0),
  lambdaL2 = c(0),
  scalePosWeight = 1,
  isUnbalance = FALSE,
  seed = sample(1e+07, 1)
)
```

Arguments

`nthread`	The number of computer threads to use (how many cores do you have?)
`earlyStopRound`	If the performance does not increase over earlyStopRound number of trees then training stops (this prevents overfitting)
`numIterations`	Number of boosting iterations.
`numLeaves`	This hyperparameter sets the maximum number of leaves. Increasing this parameter can lead to higher model complexity and potential overfitting.
`maxDepth`	This hyperparameter sets the maximum depth . Increasing this parameter can also lead to higher model complexity and potential overfitting.
`minDataInLeaf`	This hyperparameter sets the minimum number of data points that must be present in a leaf node. Increasing this parameter can help to reduce overfitting
`learningRate`	This hyperparameter controls the step size at each iteration of the gradient descent algorithm. Lower values can lead to slower convergence but may result in better performance.
`lambdaL1`	This hyperparameter controls L1 regularization, which can help to reduce overfitting by encouraging sparse models.
`lambdaL2`	This hyperparameter controls L2 regularization, which can also help to reduce overfitting by discouraging large weights in the model.
`scalePosWeight`	Controls weight of positive class in loss - useful for imbalanced classes

`isUnbalance`	This parameter cannot be used at the same time with scalePosWeight, choose only one of them. While enabling this should increase the overall performance metric of your model, it will also result in poor estimates of the individual class probabilities.
`seed`	An option to add a seed when training the final model

Examples

```
model.lightgbm <- setLightGBM(
    numLeaves = c(20, 31, 50), maxDepth = c(-1, 5, 10),
    minDataInLeaf = c(10, 20, 30), learningRate = c(0.05, 0.1, 0.3)
)
```

---

setMLP *Create setting for neural network model with python*

Description

Create setting for neural network model with python

Usage

```
setMLP(
  hiddenLayerSizes = list(c(100), c(20)),
  activation = list("relu"),
  solver = list("adam"),
  alpha = list(0.3, 0.01, 1e-04, 1e-06),
  batchSize = list("auto"),
  learningRate = list("constant"),
  learningRateInit = list(0.001),
  powerT = list(0.5),
  maxIter = list(200, 100),
  shuffle = list(TRUE),
  tol = list(1e-04),
  warmStart = list(TRUE),
  momentum = list(0.9),
  nesterovsMomentum = list(TRUE),
  earlyStopping = list(FALSE),
  validationFraction = list(0.1),
  beta1 = list(0.9),
  beta2 = list(0.999),
  epsilon = list(1e-08),
  nIterNoChange = list(10),
  seed = sample(1e+05, 1)
)
```

Arguments

`hiddenLayerSizes`	(list of vectors) The ith element represents the number of neurons in the ith hidden layer.

`activation`	(list) Activation function for the hidden layer. • "identity": no-op activation, useful to implement linear bottleneck, returns f(x) = x • "logistic": the logistic sigmoid function, returns f(x) = 1 / (1 + exp(-x)). • "tanh": the hyperbolic tan function, returns f(x) = tanh(x). • "relu": the rectified linear unit function, returns f(x) = max(0, x)
`solver`	(list) The solver for weight optimization. ('lbfgs', 'sgd', 'adam')
`alpha`	(list) L2 penalty (regularization term) parameter.
`batchSize`	(list) Size of minibatches for stochastic optimizers. If the solver is 'lbfgs', the classifier will not use minibatch. When set to "auto", batchSize=min(200, n_samples).
`learningRate`	(list) Only used when solver='sgd' Learning rate schedule for weight updates. 'constant', 'invscaling', 'adaptive', default='constant'
`learningRateInit`	(list) Only used when solver='sgd' or 'adam'. The initial learning rate used. It controls the step-size in updating the weights.
`powerT`	(list) Only used when solver='sgd'. The exponent for inverse scaling learning rate. It is used in updating effective learning rate when the learning_rate is set to 'invscaling'.
`maxIter`	(list) Maximum number of iterations. The solver iterates until convergence (determined by 'tol') or this number of iterations. For stochastic solvers ('sgd', 'adam'), note that this determines the number of epochs (how many times each data point will be used), not the number of gradient steps.
`shuffle`	(list) boolean: Whether to shuffle samples in each iteration. Only used when solver='sgd' or 'adam'.
`tol`	(list) Tolerance for the optimization. When the loss or score is not improving by at least tol for nIterNoChange consecutive iterations, unless learning_rate is set to 'adaptive', convergence is considered to be reached and training stops.
`warmStart`	(list) When set to True, reuse the solution of the previous call to fit as initialization, otherwise, just erase the previous solution.
`momentum`	(list) Momentum for gradient descent update. Should be between 0 and 1. Only used when solver='sgd'.
`nesterovsMomentum`	(list) Whether to use Nesterov's momentum. Only used when solver='sgd' and momentum > 0.
`earlyStopping`	(list) boolean Whether to use early stopping to terminate training when validation score is not improving. If set to true, it will automatically set aside 10 percent of training data as validation and terminate training when validation score is not improving by at least tol for n_iter_no_change consecutive epochs.
`validationFraction`	(list) The proportion of training data to set aside as validation set for early stopping. Must be between 0 and 1. Only used if earlyStopping is True.
`beta1`	(list) Exponential decay rate for estimates of first moment vector in adam, should be in 0 to 1.
`beta2`	(list) Exponential decay rate for estimates of second moment vector in adam, should be in 0 to 1.
`epsilon`	(list) Value for numerical stability in adam.

`nIterNoChange`	(list) Maximum number of epochs to not meet tol improvement. Only effective when solver='sgd' or 'adam'.
`seed`	A seed for the model

Examples

```
## Not run:
model.mlp <- setMLP()

## End(Not run)
```

setNaiveBayes *Create setting for naive bayes model with python*

Description

Create setting for naive bayes model with python

Usage

```
setNaiveBayes()
```

Examples

```
## Not run:
model.nb <- setNaiveBayes()

## End(Not run)
```

setPythonEnvironment *Use the virtual environment created using configurePython()*

Description

Use the virtual environment created using configurePython()

Usage

```
setPythonEnvironment(envname = "PLP", envtype = NULL)
```

Arguments

`envname`	A string for the name of the virtual environment (default is 'PLP')
`envtype`	An option for specifying the environment as'conda' or 'python'. If NULL then the default is 'conda' for windows users and 'python' for non-windows users

Details

This function sets PatientLevelPrediction to use a virtual environment

setRandomForest — *Create setting for random forest model with python (very fast)*

Description

Create setting for random forest model with python (very fast)

Usage

```
setRandomForest(
  ntrees = list(100, 500),
  criterion = list("gini"),
  maxDepth = list(4, 10, 17),
  minSamplesSplit = list(2, 5),
  minSamplesLeaf = list(1, 10),
  minWeightFractionLeaf = list(0),
  mtries = list("sqrt", "log2"),
  maxLeafNodes = list(NULL),
  minImpurityDecrease = list(0),
  bootstrap = list(TRUE),
  maxSamples = list(NULL, 0.9),
  oobScore = list(FALSE),
  nJobs = list(NULL),
  classWeight = list(NULL),
  seed = sample(1e+05, 1)
)
```

Arguments

`ntrees`	(list) The number of trees to build
`criterion`	(list) The function to measure the quality of a split. Supported criteria are “gini” for the Gini impurity and “entropy” for the information gain. Note: this parameter is tree-specific.
`maxDepth`	(list) The maximum depth of the tree. If NULL, then nodes are expanded until all leaves are pure or until all leaves contain less than minSamplesSplit samples.
`minSamplesSplit`	(list) The minimum number of samples required to split an internal node
`minSamplesLeaf`	(list) The minimum number of samples required to be at a leaf node. A split point at any depth will only be considered if it leaves at least minSamplesLeaf training samples in each of the left and right branches. This may have the effect of smoothing the model, especially in regression.
`minWeightFractionLeaf`	(list) The minimum weighted fraction of the sum total of weights (of all the input samples) required to be at a leaf node. Samples have equal weight when sampleWeight is not provided.
`mtries`	(list) The number of features to consider when looking for the best split: • int then consider max_features features at each split. • float then max_features is a fraction and round(max_features * n_features) features are considered at each split

- 'sqrt' then max_features=sqrt(n_features)
- 'log2' then max_features=log2(n_features)
- NULL then max_features=n_features

`maxLeafNodes`	(list) Grow trees with max_leaf_nodes in best-first fashion. Best nodes are defined as relative reduction in impurity. If None then unlimited number of leaf nodes.
`minImpurityDecrease`	(list) A node will be split if this split induces a decrease of the impurity greater than or equal to this value.
`bootstrap`	(list) Whether bootstrap samples are used when building trees. If False, the whole dataset is used to build each tree.
`maxSamples`	(list) If bootstrap is True, the number of samples to draw from X to train each base estimator.
`oobScore`	(list) Whether to use out-of-bag samples to estimate the generalization score. Only available if bootstrap=True.
`nJobs`	The number of jobs to run in parallel.
`classWeight`	(list) Weights associated with classes. If not given, all classes are supposed to have weight one. NULL, "balanced", "balanced_subsample"
`seed`	A seed when training the final model

Examples

```
## Not run:
model.rf <- setRandomForest(mtries=list('auto',5,20),  ntrees=c(10,100),
                            maxDepth=c(5,20))

## End(Not run)
```

---

setSVM   *Create setting for the python sklearn SVM (SVC function)*

Description

Create setting for the python sklearn SVM (SVC function)

Usage

```
setSVM(
  C = list(1, 0.9, 2, 0.1),
  kernel = list("rbf"),
  degree = list(1, 3, 5),
  gamma = list("scale", 1e-04, 3e-05, 0.001, 0.01, 0.25),
  coef0 = list(0),
  shrinking = list(TRUE),
  tol = list(0.001),
  classWeight = list(NULL),
  cacheSize = 500,
  seed = sample(1e+05, 1)
)
```

Arguments

`C`	(list) Regularization parameter. The strength of the regularization is inversely proportional to C. Must be strictly positive. The penalty is a squared l2 penalty.
`kernel`	(list) Specifies the kernel type to be used in the algorithm. one of 'linear', 'poly', 'rbf', 'sigmoid', 'precomputed'. If none is given 'rbf' will be used.
`degree`	(list) degree of kernel function is significant only in poly, rbf, sigmoid
`gamma`	(list) kernel coefficient for rbf and poly, by default 1/n_features will be taken. 'scale', 'auto' or float, default='scale'
`coef0`	(list) independent term in kernel function. It is only significant in poly/sigmoid.
`shrinking`	(list) whether to use the shrinking heuristic.
`tol`	(list) Tolerance for stopping criterion.
`classWeight`	(list) Class weight based on imbalance either 'balanced' or NULL
`cacheSize`	Specify the size of the kernel cache (in MB).
`seed`	A seed for the model

Examples

```
## Not run:
model.svm <- setSVM(kernel='rbf', seed = NULL)

## End(Not run)
```

`simulatePlpData` *Generate simulated data*

Description

`simulateplpData` creates a plpData object with simulated data.

Usage

```
simulatePlpData(plpDataSimulationProfile, n = 10000)
```

Arguments

`plpDataSimulationProfile`	An object of type `plpDataSimulationProfile` as generated using the `createplpDataSimulationProfile` function.
`n`	The size of the population to be generated.

Details

This function generates simulated data that is in many ways similar to the original data on which the simulation profile is based. The contains same outcome, comparator, and outcome concept IDs, and the covariates and their 1st order statistics should be comparable.

Value

An object of type `plpData`.

`sklearnFromJson` *Loads sklearn python model from json*

Description

Loads sklearn python model from json

Usage

```
sklearnFromJson(path)
```

Arguments

`path`	path to the model json file

`sklearnToJson` *Saves sklearn python model object to json in path*

Description

Saves sklearn python model object to json in path

Usage

```
sklearnToJson(model, path)
```

Arguments

`model`	a fitted sklearn python model object
`path`	path to the saved model file

`specificity` *Calculate the specificity*

Description

Calculate the specificity

Usage

```
specificity(TP, TN, FN, FP)
```

Arguments

`TP`	Number of true positives
`TN`	Number of true negatives
`FN`	Number of false negatives
`FP`	Number of false positives

Details

Calculate the specificity

Value

specificity value

`splitData` *Split the plpData into test/train sets using a splitting settings of class* `splitSettings`

Description

Split the plpData into test/train sets using a splitting settings of class `splitSettings`

Usage

```
splitData(
  plpData = plpData,
  population = population,
  splitSettings = splitSettings
)
```

Arguments

`plpData`	An object of type `plpData` - the patient level prediction data extracted from the CDM.
`population`	The population created using `createStudyPopulation` that define who will be used to develop the model
`splitSettings`	An object of type `splitSettings` specifying the split - the default can be created using `createDefaultSplitSetting`

Details

Returns a list containing the training data (Train) and optionally the test data (Test). Train is an Andromeda object containing

- covariates: a table (rowId, covariateId, covariateValue) containing the covariates for each data point in the train data
- covariateRef: a table with the covariate information
- labels: a table (rowId, outcomeCount, ...) for each data point in the train data (outcomeCount is the class label)
- folds: a table (rowId, index) specifying which training fold each data point is in.

Test is an Andromeda object containing

- covariates: a table (rowId, covariateId, covariateValue) containing the covariates for each data point in the test data
- covariateRef: a table with the covariate information
- labels: a table (rowId, outcomeCount, ...) for each data point in the test data (outcomeCount is the class label)

Value

An object of class `splitSettings`

`toSparseM` *Convert the plpData in COO format into a sparse R matrix*

Description

Converts the standard plpData to a sparse matrix

Usage

```
toSparseM(plpData, cohort = NULL, map = NULL)
```

Arguments

`plpData`	An object of type `plpData` with covariate in coo format - the patient level prediction data extracted from the CDM.
`cohort`	If specified the plpData is restricted to the rowIds in the cohort (otherwise plpData$labels is used)
`map`	A covariate map (telling us the column number for covariates)

Details

This function converts the covariate file from ffdf in COO format into a sparse matrix from the package Matrix

Value

Returns a list, containing the data as a sparse matrix, the plpData covariateRef and a data.frame named map that tells us what covariate corresponds to each column This object is a list with the following components:

data A sparse matrix with the rows corresponding to each person in the plpData and the columns corresponding to the covariates.

covariateRef The plpData covariateRef.

map A data.frame containing the data column ids and the corresponding covariateId from covariateRef.

Examples

```
#TODO
```

`validateExternal` *externalValidatePlp - Validate model performance on new data*

Description

externalValidatePlp - Validate model performance on new data

Usage

```
validateExternal(
  validationDesignList,
  databaseDetails,
  logSettings,
  outputFolder
)
```

Arguments

`validationDesignList`	A list of objects created with `createValidationDesign`
`databaseDetails`	A list of objects of class `databaseDetails` created using `createDatabaseDetails`
`logSettings`	An object of `logSettings` created using `createLogSettings`
`outputFolder`	The directory to save the validation results to (subfolders are created per database in validationDatabaseDetails)

`validateMultiplePlp` *externally validate the multiple plp models across new datasets*

Description

This function loads all the models in a multiple plp analysis folder and validates the models on new data

Usage

```
validateMultiplePlp(
  analysesLocation,
  validationDatabaseDetails,
  validationRestrictPlpDataSettings = createRestrictPlpDataSettings(),
  recalibrate = NULL,
  cohortDefinitions = NULL,
  saveDirectory = NULL
)
```

Arguments

`analysesLocation`	The location where the multiple plp analyses are
`validationDatabaseDetails`	A single or list of validation database settings created using `createDatabaseDetails()`
`validationRestrictPlpDataSettings`	The settings specifying the extra restriction settings when extracting the data created using `createRestrictPlpDataSettings()`.
`recalibrate`	A vector of recalibration methods (currently supports 'RecalibrationintheLarge' and/or 'weakRecalibration')
`cohortDefinitions`	A list of cohortDefinitions
`saveDirectory`	The location to save to validation results

Details

Users need to input a location where the results of the multiple plp analyses are found and the connection and database settings for the new data

---

`viewDatabaseResultPlp` *open a local shiny app for viewing the result of a PLP analyses from a database*

Description

open a local shiny app for viewing the result of a PLP analyses from a database

Usage

```
viewDatabaseResultPlp(
  mySchema,
  myServer,
  myUser,
  myPassword,
  myDbms,
  myPort = NULL,
  myTableAppend
)
```

Arguments

`mySchema`	Database result schema containing the result tables
`myServer`	server with the result database
`myUser`	Username for the connection to the result database
`myPassword`	Password for the connection to the result database
`myDbms`	database management system for the result database
`myPort`	Port for the connection to the result database
`myTableAppend`	A string appended to the results tables (optional)

Details

Opens a shiny app for viewing the results of the models from a database

---

`viewMultiplePlp` *open a local shiny app for viewing the result of a multiple PLP analyses*

Description

open a local shiny app for viewing the result of a multiple PLP analyses

Usage

```
viewMultiplePlp(analysesLocation)
```

Arguments

`analysesLocation`	The directory containing the results (with the analysis_x folders)

Details

Opens a shiny app for viewing the results of the models from various T,O, Tar and settings settings.

`viewPlp` *viewPlp - Interactively view the performance and model settings*

Description

This is a shiny app for viewing interactive plots of the performance and the settings

Usage

```
viewPlp(runPlp, validatePlp = NULL, diagnosePlp = NULL)
```

Arguments

`runPlp`	The output of runPlp() (an object of class ’runPlp’)
`validatePlp`	The output of externalValidatePlp (on object of class ’validatePlp’)
`diagnosePlp`	The output of diagnosePlp()

Details

Either the result of runPlp and view the plots

Value

Opens a shiny app for interactively viewing the results

Index