Copyright (c) 2022 Mayur Khandetod

Permission is hereby granted, free of charge, to any person obtaining a copy  
of this software and associated documentation files (the "Software"), to deal  
in the Software without restriction, including without limitation the rights  
to use, copy, modify, merge, publish, distribute, sublicense, and/or sell  
copies of the Software, and to permit persons to whom the Software is  
furnished to do so, subject to the following conditions:

The above copyright notice and this permission notice shall be included in all  
copies or substantial portions of the Software.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR  
IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY,  
FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE  
AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER  
LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM,  
OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE  
SOFTWARE.

**Abstract:**

Aim: To develop a prediction model that can determine whether a patient has lung cancer based on a set of risk indicators.

Description of data set:

This dataset contains a List of Risk Factors for Lung Cancer. Based on the input values listed below, this dataset is used to forecast if a patient is likely to get lung cancer. Each row of data in the table contains pertinent information about the patient.

Description of fields:

* Gender: 1(male), 0(female)
* Age: Age of the patient
* Smoking: YES=2 , NO=1.
* Yellow fingers: YES=2 , NO=1.
* Anxiety: YES=2 , NO=1.
* Peer\_pressure: YES=2 , NO=1.
* Chronic Disease: YES=2 , NO=1.
* Fatigue: YES=2 , NO=1.
* Allergy: YES=2 , NO=1.
* Wheezing: YES=2 , NO=1.
* Alcohol: YES=2 , NO=1.
* Coughing: YES=2 , NO=1.
* Shortness of Breath: YES=2 , NO=1.
* Swallowing Difficulty: YES=2 , NO=1.
* Chest pain: YES=2 , NO=1.
* Lung Cancer: YES=2 , NO=1.

Imported required libraries and train, test dataset. Set a minimum memory size and a run time. Start the H2O server on a random port. Import dataset using H2O. Split the data in training and test sets. Set the features and target. Set up AutoML. Train model using H2O. Display AML leaderboard. Find best model. Plot variable importance. Plot standardized coefficient magnitudes. Find best model accuracy. Plot partial plot for best model. Plot correlation matrix & heat map. Compute p-value. Performed regularization.

**Questions/ Answers:**

\* Is the relationship significant?

Ans: As shown in the table below, p-values for all the independent variables is greater than 0.05, hence the relationship is statistically insignificant.

\* Are any model assumptions violated?

Ans:

Model assumptions for classification are as follows:

* It assumes that there is minimal or no multicollinearity among the independent variables.
* It usually requires a large sample size to predict properly.

No model assumptions are violated because no multicollinearity is observed in the model, sample size is large (approx. 55,000).

\* Is there any multicollinearity in the model?

Ans: After observing the correlation matrix, it is concluded that no predictor variable is related to any other predictor variable because the absolute value of correlation coefficient is less than 0.7 for all the predictor variables. Hence, there is no multicollinearity observed in the model.

\* In the multivariate models are predictor variables independent of all the other predictor variables?

Ans: As per the correlation matrix, all predictor variables are independent of all other predictor variables as the absolute value is less than 0.7 for all the predictor variables.

\* In in multivariate models rank the most significant predictor variables and exclude insignificant ones from the model.

Ans:

As per the variable importance plot, AGE is the most significant predictor variable.

As per the correlation matrix, all predictor variables are independent of all other predictor variables as the absolute value is less than 0.7 for all the predictor variables. Hence, all predictors are independent of each other & hence important for lung cancer prediction.

\* Does the model make sense?

Ans: No, the model does not make any sense because variable importance graph shows ‘AGE’ as the most significant variable whereas standardized coefficient magnitude shows ‘COUGHING’ as the most significant variable. We can perform regularization to make the model better.

\* Does regularization help?

Yes, regularization does help.

With model regularization, the difference in error values for train and test data is relatively small, whereas the difference in error values for unregularized data is much larger. As a result, regularization aids in overcoming overfitting.

Error values for train and test unregularized data:

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Error values for train and test regularized data:

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application, chat or text message

Description automatically generated

\* Which independent variables are significant?

Ans: As none of the independent variables have p-value less than 0.05, all independent variables are insignificant.

\* Which hyperparameters are important?

The hyperparameters which are available for DRF model are number\_of\_trees, number\_of\_internal\_trees, model\_size\_in\_bytes, min\_depth, max\_depth, mean\_depth, min\_leaves, max\_leaves and mean\_leaves.

Firstly, tuning without defining any hyperparameter gives accuracy as 0.5019. Then, after defining ntrees, sample\_rate, max\_depth as hyperparameters for DRF model increases accuracy to 0.5027. Moreover, further increasing ntrees and max\_depth to 100, 25 respectively increases accuracy to 0.5062. Further altering hyperparameters ntrees and max\_depth to 150, 30 respectively does not increases the accuracy.

\* Coding professionalism?

Ans:

* https://github.com/nikbearbrown/AI\_Research\_Group/blob/main/Kaggle\_Datasets/AutoML/VAI\_Churn\_H2O\_automl.ipynb
* <https://www.kdnuggets.com/2021/02/machine-learning-assumptions.html>
* <https://www.simplilearn.com/tutorials/machine-learning-tutorial/regularization-in-machine-learning>
* <https://h2o-release.s3.amazonaws.com/h2o/rel-zipf/6/docs-website/h2o-docs/performance-and-prediction.html>
* <https://pro.arcgis.com/en/pro-app/2.8/tool-reference/spatial-statistics/what-is-a-z-score-what-is-a-p-value.htm>
* <https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/glm.html>
* <https://github.com/Avkash/mldl/blob/master/orgs/h2o/guide/algo/grid/h2o_grid_drf_houseprice_python.md>
* https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/upliftdrf.html

Conclusion:

Build a predictive model using AutoML to predict whether a patient has lung cancer based on a set of risk factors. Computed pvalues for all the independent variables is greater than 0.05, the relationship is statistically insignificant. Observed that no model assumptions are violated. As per the correlation matrix, no multicollinearity observed in the model. Hence, all predictor variables are independent of all other predictor variables. Observed that model does not make any sense. With model regularization, the difference in error values for train and test data is relatively small, whereas the difference in error values for unregularized data is much larger. As a result, regularization aids in overcoming overfitting and performed Hyperparameter Tuning.