Previous work (mainly focused on no-shows)

* <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4187098/>
* <https://www.hindawi.com/journals/cmmm/2020/9727096/>
* <https://www.sciencedirect.com/science/article/pii/S2666764921000175>

<https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8> - Great visualizations to find correlated features

For continuous predictors (independent variables), logistic regression assumes that predictors are linearly related to the log odds of the outcome (an assumption known as “linearity of the logit”). If this assumption is violated, logistic regression underestimates the strength of the association and rejects the association too easily, that is being not significant (not rejecting the null hypothesis) where it should be significant. The Box–Tidwell test can be performed to assess linearity in the log(odds) as required by logistic regression. If linearity is not observed, categorical scales for the continuous predictor can be examined on the basis of quartiles and logit graphs.

**Weights GridSearch**

There can be other combinations of weights which can perform equally good or may be better. In this part, will perform a grid search on different combinations of weights and will retain the one with a better performance score.

# define weight hyperparameter  
w = [{0:1000,1:100},{0:1000,1:10}, {0:1000,1:1.0},   
 {0:500,1:1.0}, {0:400,1:1.0}, {0:300,1:1.0}, {0:200,1:1.0},   
 {0:150,1:1.0}, {0:100,1:1.0}, {0:99,1:1.0}, {0:10,1:1.0},   
 {0:0.01,1:1.0}, {0:0.01,1:10}, {0:0.01,1:100},   
 {0:0.001,1:1.0}, {0:0.005,1:1.0}, {0:1.0,1:1.0},   
 {0:1.0,1:0.1}, {0:10,1:0.1}, {0:100,1:0.1},   
 {0:10,1:0.01}, {0:1.0,1:0.01}, {0:1.0,1:0.001}, {0:1.0,1:0.005},   
 {0:1.0,1:10}, {0:1.0,1:99}, {0:1.0,1:100}, {0:1.0,1:150},   
 {0:1.0,1:200}, {0:1.0,1:300},{0:1.0,1:400},{0:1.0,1:500},   
 {0:1.0,1:1000}, {0:10,1:1000},{0:100,1:1000} ]  
hyperparam\_grid = {"class\_weight": w }

https://towardsdatascience.com/weighted-logistic-regression-for-imbalanced-dataset-9a5cd88e68b

Logistic regression prefers categorical variables?? ->

<https://datascience.stackexchange.com/questions/109278/should-i-always-transform-data-to-normal-distribution>

ROC AUC vs PR AUC – PR AUC better when you care more about the positive class and ROC AUC is ideal when you care about both classes equally

[F1 Score vs ROC AUC vs Accuracy vs PR AUC: Which Evaluation Metric Should You Choose? - neptune.ai](https://neptune.ai/blog/f1-score-accuracy-roc-auc-pr-auc)

[IEEE-CIS Fraud Detection | Kaggle](https://www.kaggle.com/c/ieee-fraud-detection/overview)

NN info - <http://www.faqs.org/faqs/ai-faq/neural-nets/part1/preamble.html>

Why take log of feature: <https://datascience.stackexchange.com/questions/40089/what-is-the-reason-behind-taking-log-transformation-of-few-continuous-variables>

Noisy test curve on loss -

An unrepresentative validation dataset means that the validation dataset does not provide sufficient information to evaluate the ability of the model to generalize.

This may occur if the validation dataset has too few examples as compared to the training dataset.

This case can be identified by a learning curve for training loss that looks like a good fit (or other fits) and a learning curve for validation loss that shows noisy movements around the training loss.

Using logs

<https://stats.stackexchange.com/questions/298/in-linear-regression-when-is-it-appropriate-to-use-the-log-of-an-independent-va>