<https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8>

[Recursive Feature Elimination (RFE)](http://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html) is based on the idea to repeatedly construct a model and choose either the best or worst performing feature, setting the feature aside and then repeating the process with the rest of the features. This process is applied until all features in the dataset are exhausted. The goal of RFE is to select features by recursively considering smaller and smaller sets of features.

ROC

—-------------------------------------------------------------------

<https://www.analyticsvidhya.com/blog/2021/09/guide-for-building-an-end-to-end-logistic-regression-model/>

Hyperparameter tuning\

—-------------------------------------------------------------------------

<https://www.simplilearn.com/tutorials/machine-learning-tutorial/logistic-regression-in-python>

Detailed into logistic regression

—----------------------------------------------------------------------------

<https://www.justintodata.com/logistic-regression-for-machine-learning-tutorial/>

Detailed background of the logistic regression

—----------------------------------------------------------------------

### 2.5 Model improvement

There are multiple methods that can be used to improve your logistic regression model.

#### 2.5.1 Data preprocessing

The greatest improvements are usually achieved with a proper data cleaning process. Logistic regression uses a linear model, so it suffers from the same issues that [linear regression does](https://www.kdnuggets.com/2020/04/data-transformation-standardization-normalization.html). To properly prepare the data for logistic regression modeling, you need to:

1. Remove outliers. Outliers will skew your model to perform less well.
2. Remove multicollinearity. Logistic regression assumes that the predictor variables (features) are not correlated with one another. Check their pairwise correlation and from the analysis, remove those variables which are highly correlated.
3. Assert linear assumption. If your independent variables do not have a linear relationship with your predictor variable, you need to log transform them to reshape polynomial relationships into linear.
4. Assert normal distribution. The model assumes that the independent variables follow a Gaussian distribution. Transform your variables with log transform or BoxCox if they are not normally distributed.

Logistic regression has additional assumptions and needs for cleaning:

1. Binary output variable. Transform your output variable into 0 or 1.
2. Failure to converge. The maximum likelihood estimation model (the ‘maths’) behind logistic regression assumes that no single variable will perfectly predict class membership. In the event that you have a feature that perfectly predicts the target class, the algorithm will try to assign it infinite weights (because it is so important) and thus will fail to converge to a solution. If you have a perfect predictor, simply remove it from the feature set... or just don’t model your data at all. At the end of the day, you do not need a machine learning model if you have a perfect predictor.

#### 2.5.2 Feature scaling

Feature values can be comparably different by orders of magnitude. For instance, loan size is in the tens of thousands ($50,000), while “number of months late” is in single digits (0, 1, 2, …).

Features of different scales convert slower (or not at all) with gradient descent.

[Normalize and standardize](https://www.kdnuggets.com/2020/04/data-transformation-standardization-normalization.html) your features to speed up and improve model training.

#### 2.5.3 Regularization

Regularization is particularly useful in settings with multiple features (or independent variables). Regularization takes a complex model (with multiple predictors) and sets their weights to zero (L1 regularization). This effectively removes a predictor from the linear equation or lowers its weights towards zero (L2 regularization), making the feature less impactful on the final logistic regression equation.

—----------------------------------------------------------------