Model evaluation is a critical task in data science. This is one of the few measures business stakeholders are interested in. Model performance heavily influences business impact of a data science project. Therefore, it is important take some time apart in an independent task in the process model.

So how are models evaluated? In supervised machine learning this is relatively straightforward since you can always create a ground truth and compare your results against ground truth.

So, we are either splitting data into training-, test- and validation-sets to assess model performance on the test set or we use cross validation. This all is explained in the following courser course https://www.coursera.org/learn/advanced-machine-learning-signal-processing/ Week 2.

In case we know what data set we can use as ground truth in supervised learning (classification and regression) we need to define a different measure for evaluation than in unsupervised learning (clustering). Since it depends on the type of model we create, the following none exhaustive lists can be used as a starting point for further research:

Classification:

- Confusion Matrix
- Accuracy
- Precision
- Recall
- Specificity
- True positive rate
- True negative rate
- False positive rate
- False negative rate
- F1-score
- Gain and Lift
- Kolomogorov Smirnov
- Area Under ROC
- Gini Coefficient
- Concordant – Discordant ratio

Regression:

- Root Mean Squared Error (RMSE)
- Mean Squared Error
- Mean Absolute Error (MAE)
- R-Squared
- Relative Squared Error
- Relative Absolute Error
- Sum of Differences

- ACF plot of residuals
- Histogram of residuals
- Residual plots against predictors
- Residual plots against fitted values

Clustering:

- Adjusted Rand index
- Mutual Information
- Homogeneity completeness
- V-measure
- Fowlkes-Mallows
- Silhouette Coefficient Calinski-Harabaz¶

References:
http://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation

Please choose at least one appropriate model performance measure, justify why you've used it and document how iterative changes in the feature creation task influence it.