**The Machine Learning Process**

The machine learning process consists of multiple steps from problem definition to model deployment. There are many definitions of the process.

1. **Define the purpose:** The purpose cannot be the prediction itself. Discuss with business team and define a goal. How are we going to use prediction?
2. **Obtain data:** Identify the various data sources and obtain necessary data.
3. **Explore and clean data:** Perform data exploration and cleaning, incl. outlier detection. A machine learning engineer spends 80% of his time in exploring and cleaning the data before running any machine learning algorithms. Partition the data into training and validation data.
4. **Determine Datamining Task:** What type of algorithm is needed? Regression/Classification? Once we have determined the datamining task, we can decide the actual algorithm to use.
5. **Apply Methods:** Apply the model decided in the earlier step.
6. **Evaluate Performance:** Carefully evaluate the impact of the model of our business use case.
7. **Model Deployment:** Deploy the model to for decision support.

In real-world scenarios, the process is not linear but has lot of twists and turns.

**Explanation vs Prediction**

Difference in model evaluation

* **Explanatory**: Emphasis is on goodness of fit measures like R2
* **Prediction:** We are concerned with its predictive abilities, on new unseen data.

**Data Partition**

* For simple projects, we split the data into two parts so that we can evaluate the performance on data that was not used in building process.
  + Training Data to build the model.
  + Validation Data to evaluate the performance of the model.
* For complex projects,
  + Training Data to build the model.
  + Validation Data to evaluate the performance of the model.
  + Revisit Step 1
  + Testing data: Final Validation Data to evaluate the performance of the model.
  + Predict/Classify using the final model on unseen data.

**Explanation vs Prediction**

Independent on the model’s purpose, the same basic analysis and principles apply, with different emphasis.

* Difference in model evaluation:
  + **Explanation**: Emphasis is on goodness of fit measures like R2
  + **Prediction**: We are concerned with the predictive abilities on new unseen data.
* Difference in feature selection:
  + **Explanatory**: Focus is on other factors that can explain the outcome, need to consider multicollinearity
  + **Prediction**: Emphasis is on features that support prediction and the timing of information
  + **Example**: In explaining final auction prices, the number of bidders may be critical in explaining the final price. However, if we are trying to predict the final auction price ahead of time, we do not know the number of bidders, and therefore, the number of bidders is not a useful variable of prediction.

**Prediction Accuracy Measures**

* For a continuous outcome variable, the error measures are based on the prediction errors.
* Notation:
  + be the actual true value of record i
  + be the predicted value of record i
  + the prediction error for the record i (the residual), defined as:

**Measures**

1. MAE: Mean Absolute Error
2. Average Error
3. MAPE: Mean Absolute Percentage Error
4. RMSE: Root mean squared error

The mean absolute error is calculated as below:

The average error is calculated as below:

The mean absolute percentage error can be calculated as follows:

RMSE (Root mean squared error):

The most commonly used measures are RMSE (Root mean squared error) and MAE(Mean Absolute Error)

**The Baseline – Naïve Rule:**

1. Whether $10,000 is good or bad prediction error depends!
2. We need a baseline to contrast the prediction accuracy against
3. Generally, the simplest model possible as the baseline; in terms of prediction models, we can use the sample mean.
   1. Predict sample mean of the outcome of all records
   2. Calculate the error
   3. Your model needs to improve significantly over the baseline
   4. R2=0 on the training set when the model predicts the sample mean

**Model Evaluation:**

1. Measure out of sample performance
   1. Data Partition: If you have sufficient data
   2. In a data sparse environment: Apply cross validation to estimate out of sample performance.
2. Risk machine learning bias
   1. Assess how different groups of people are affected. Make sure it does not change for different demographic groups.