How to improve performance of regression model

General guideline for developing ML model

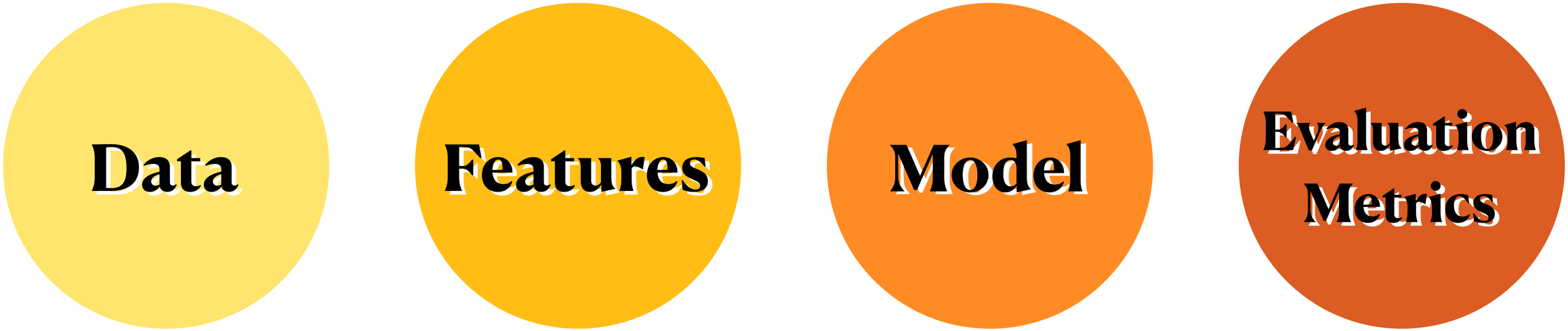
AI 05 Hyungjin Kim

Contents

Flow of the presentation

follow the process of developing ML model

(From collecting data to evaluating model)



Collecting Data

The more data we collect, the greater the performance of model

Let data tell itself

Add more data

to get more information from data itself

Dealing with Features (EDA)

Handling missing values & outliers

Handling missing values & outliers

Missing values - imputing with mean, median, or mode for **continuous** values
using methods like KNN imputation for **categorical** values

Outliers - deleting observations, transforming or binning values

making **raw data** better in terms of reducing bias and errors

Dealing with Features (EDA)

Feature Engineering & Feature Selection

Feature Engineering

help to **extract more information** from existing features

related to **hypothesis**

Variable **Transformation**
(e.g. Data normalization)

Feature **Creation**

Feature Selection

find the features that **explain data better** than others

Domain Knowledge

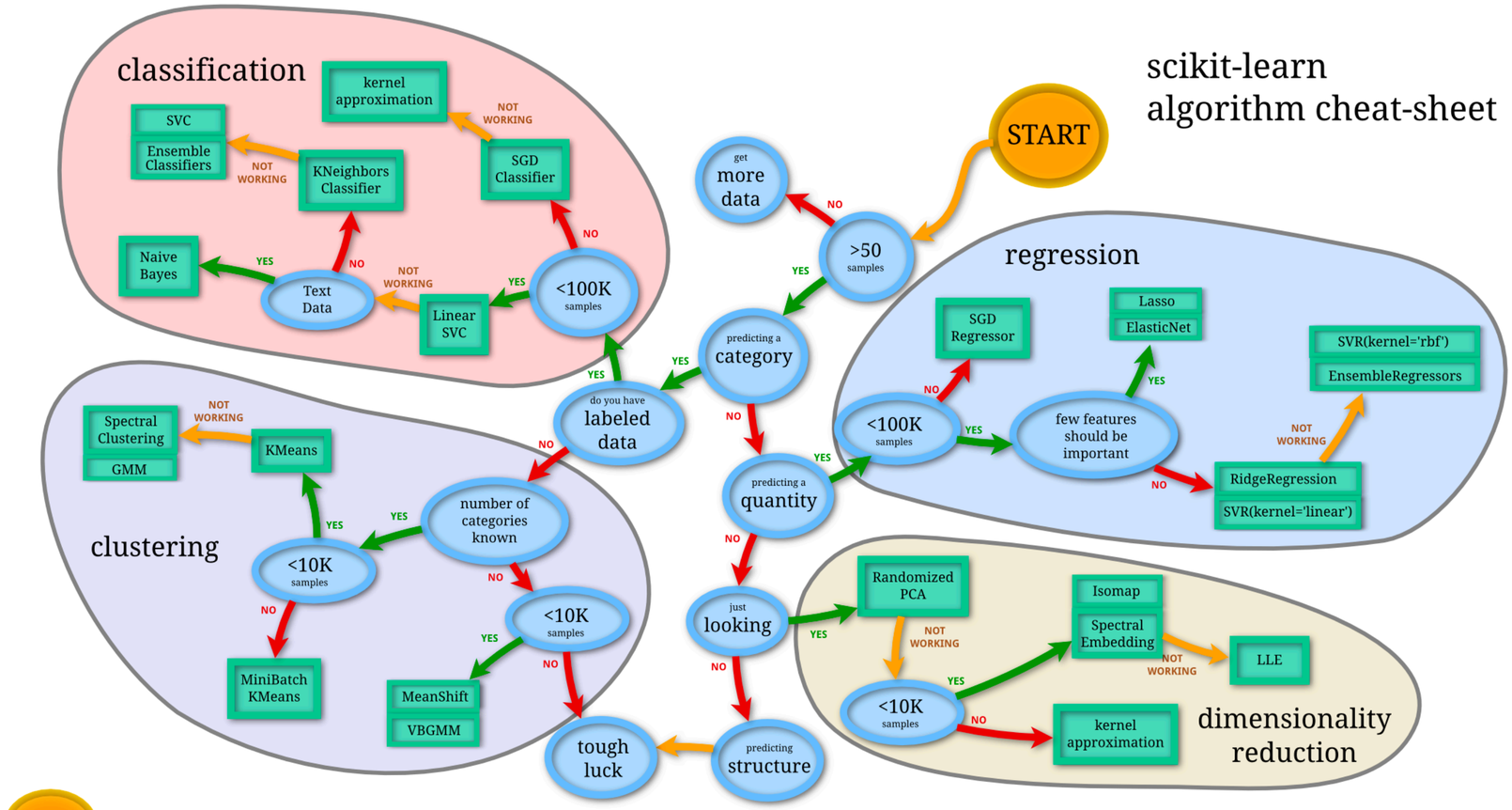
Visualization

Statistical Methods
(e.g. PCA)

+ Regularization

Developing Model

Model Selection



Developing Model

Hyperparameter Tuning

sklearn.linear_model.LinearRegression

class sklearn.linear_model.LinearRegression(*, *fit_intercept=True*, *normalize='deprecated'*, *copy_X=True*, *n_jobs=None*, *positive=False*) [source]

sklearn.linear_model.ElasticNet

class sklearn.linear_model.ElasticNet(*alpha=1.0*, *, *l1_ratio=0.5*, *fit_intercept=True*, *normalize='deprecated'*, *precompute=False*, *max_iter=1000*, *copy_X=True*, *tol=0.0001*, *warm_start=False*, *positive=False*, *random_state=None*, *selection='cyclic'*) [source]

Evaluating Performance

Choose proper evaluation metrics

Regression

MAE (Mean Absolute Error)

MSE (Mean Squared Error)

R-Squared (R^2)

.
.
.

Classification

Accuracy

Precision / Recall

F1 Score

.
.
.

Takeaway

What we should keep in mind

Prepare **data** as much as possible

Use methods that are helpful for **reducing errors**

e.g. Feature Engineering, Feature Selection, Handling missing values & outliers

Choose proper **model** & **evaluation metric**

There is no optimal solution for all

Be cautious of ‘Overfitting’

References

https://towardsdatascience.com/how-to-improve-the-accuracy-of-a-regression-model-3517accf8604

https://stackoverflow.com/questions/47577168/how-can-i-increase-the-accuracy-of-my-linear-regression-modelmachine-learning

https://www.analyticsvidhya.com/blog/2015/12/improve-machine-learning-results/

https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/

https://evidentlyai.com/blog/evidently-016-regression-model-performance

https://nycdatascience.com/blog/student-works/improving-model-accuracy-kaggle-competition/

https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.ElasticNet.html#sklearn.linear_model.ElasticNet

Thank You

for listening 💛