Kaggle

Kaggle

- data set: training set (build models), test set (submit predictions)
- data analysis and visualization
  - separate variables to numeric/categorical features
  - analyze missingness distribution and choose proper imputation method
  - For numeric features, analyze feature distribution and choose a method to deal with outliner (box-cox transformation/ medium replacement)
  - For categorical features, analyze the counts of each level
  - correlation matrix of variables (choose the right model)
- feature engineering to boost the accuracy of machine learning models
  - add new features based on data analysis and domain knowledge
- machine learning models
  - split to training/validation data set
  - build models to training set and choose the final model based on the performance of validation set
  - for every model, use grid search/bayesian optimization to choose the best combination of tuning parameters based on RMSE (regression) / AUC(classification)
- ensemble/stacking multiple models and submit the final results
  - Ensemble: average the prediction results of multiple models
  - Stacking: build another model (input: prediction results of multiple models, output: target)

Kaggle

- team work and effective commutation
- time management
- efforts (multiple submission every day)
- effective methods and codes