Stacked Ensembles Algorithm after Feature Engineering for Data Analytics

Brief by Tianlei SHI

Why should you care? – **Stacked Ensembles** is a supervised ensemble **machine learning algorithm** with a relative high performance and generalization, and it is also called Super Learning or Stacked Regression. The Stacked Ensembles algorithm produces an optimal prediction model, which is a combination of a collection of prediction algorithms that after the processing of Bagging, or Bosting method, and this makes Stacked Ensembles algorithm generally better predictive in **Data Analytics** tasks. Stacked Ensembles algorithm is often used to solve problems related to "Big Data" due to the superior performance it has shown when combined with appropriate **Feature Engineering**.

Approach – The primary technologies used in this data analytic task is feature engineering and stacked ensembles algorithm.

Feature engineering performs a series of processes on the raw data to produce more efficient features for later analysis. We conducted the following feature engineering in this task: (1) **Feature selection** by dropping irrelevant features according to the covariance matrix. (2) **Data Preprocessing** by using One-hot Encoding to categorical features, converting temporal features to time period of numerical features, and performing normalization. (3) **Value Imputation** by detecting and removing outliers, and conducting missing value imputation using KNN. (4) **Feature construction** by using interaction features.

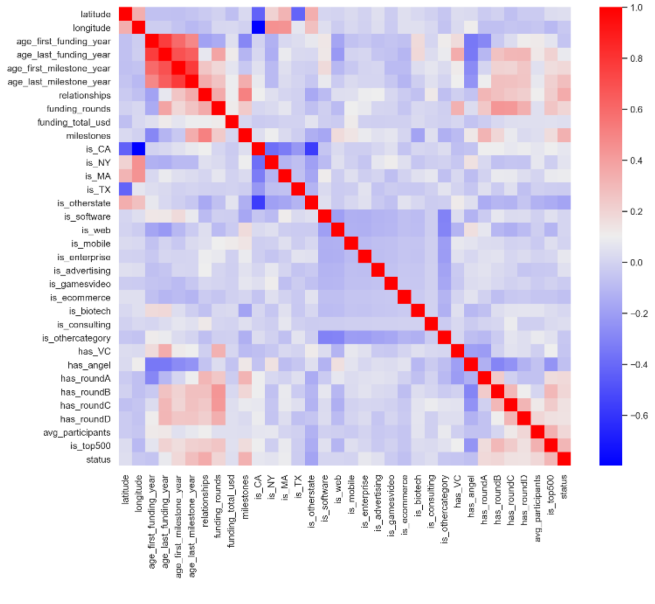

Figure 1: Covariance matrix shows correlations between features.

Stacked ensembles algorithm choose several machine learning algorithms as base algorithms, and train base algorithms and then collect prediction results by using k-fold cross-validation to obtain "level-one" data. Finally, stacked ensembles algorithms select a learning rule as metalearning algorithm, and train this metalearning algorithm on level-one data to produce prediction model.

Results - Stacked ensembles algorithm's competitors are other effective machine learning algorithms. Typical algorithms only consider single model, and ensemble learning algorithms such as Bagging and Boosting considers multiple models. Moreover, the idea of stacked ensembles algorithm is to integrate the results of ensemble learning. Therefore, because of the different principles and implementation of algorithms, they also present different performance.

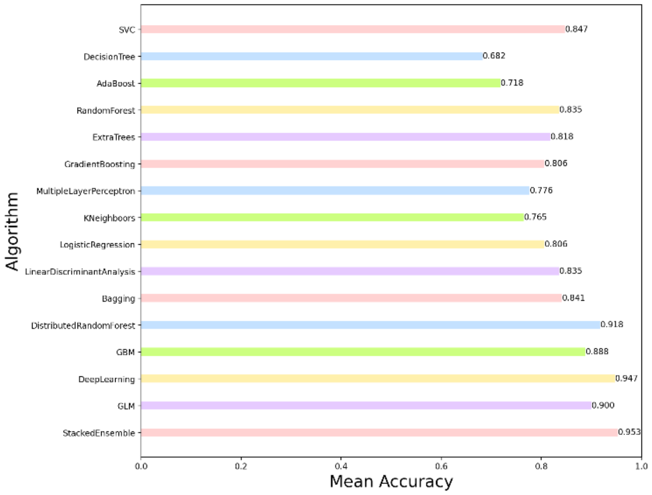

Figure 2: Performance comparison between different algorithms, and stacked ensembles algorithm finally obtain 0.95 public score and 1.0 private score in the Kaggle competition, better than other algorithms.

Pros/Cons

- Pro: Relatively good bias-variance tradeoff can be acquired, and optimal performance is almost always achieved compare to base algorithms.
- Con: Possibly get prediction results have not much improvement or worse compare to base algorithms, and Large computational quantity.
- Contributor thoughts: The algorithm integrates strong "perspectives" of base algorithms for better performance. However, weakness may be ignored, resulting in degraded performance. Thus, create accurate base models with different "perspectives" by randomized search may helpful.

Conclusion - Stacked ensembles is a powerful algorithm that can perform well in a variety of tasks. It always achieves optimal performance and better bias-variance tradeoff by integrating the advantage of base algorithms.