Data 622 Machine Learning and Big Data_HW3

By: Jiaxin Zheng

I. Three Articles:

1. Predicting property prices with machine learning algorithms including support vector machine (SVM), random forest (RF) and gradient boosting machine (GBM). The methods were applied to the data of 40,000 housing transactions in period of over 18 years in Hong Kong and compares the results of these algorithms. In terms of predictive power, random forest (RF) and gradient boosting machine (GBM) have achieved better performance when compared to SVM. And the study has found that SVM is still useful in data fitting, it can produce reasonably accurate predictions within a tight time.

   https://www.tandfonline.com/doi/full/10.1080/09599916.2020.1832558

2. In these articles, they report that in their property price dataset, decision tree algorithm performed better than SVM. In their experience, they gathered 20 participants and set a minimum power of 80% using a G power calculator for predicting future property prices. As result, when it comes to predicting property prices, decision tree algorithms perform better than support vector machine method.

   https://pubs.aip.org/aip/acp/article-abstract/3267/1/020098/3349583/Improved-prediction-accuracy-of-house-price-using?redirectedFrom=fulltext

3. In this article, they evaluate the performance of various machine learning methods in predicting both the selling price and the sale velocity of properties. The machine learning they used in this study were random forest, decision tree, K-nearest neighbor, support vector regression and multilayer perceptron. After preprocessing the data set used comprises 560,000 distinct data from the Swedish housing market, the results demonstrate that random forest is best performance than other machine learning methods in predicting property selling prices.

   https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1783631&dswid=8418

II. Introduction:

This assignment is focused on using the Support Vector Machine (SVM) algorithm to examine the dataset from homework #2 and compare the results with the results from previous work.

The Bank Marketing dataset collect information from a marketing campaign of a Portuguese bank. The goal is to predict whether a client will subscribe to a term deposit. This is a binary classification analysis with both categorical variables (e.g., job, education, marital, housing loan) and numeric variables (e.g., age, balance, campaign, pdays, previous). You can download the dataset from https://archive.ics.uci.edu/dataset/222/bank+marketing.

The dataset has already been prepared for this experiment in the last assessment. The data preparation process consists of loading the data, cleaning format variables properly, and handle 'unknown' values.

III. SVM Algorithm

1. SVM Linear:

Hypothesis: The linear SVM with the default setting is better model than in previous assignments.

In the baseline linear SVM, it achieved solid accuracy, but low recall. The model achieved an accuracy of 0.8979095, precision of 0.6495536, recall of 0.2753075, F1-score of 0.386711, and an AUC of 0.9040328. The low recall implies that the model struggles to identify all positive cases. It underfits when the relationship between predictors and the outcome is nonlinear. But the default linear SVM is little better than default decision tree.

2. SVM Linear Tuned:

Hypothesis: The tuned linear SVM improved linear SVM performance.

The tuned linear SVM improved performance slightly after adjusting the parameter cost. In this model, I use a grid search accompanied by cross-validation was conducted utilizing tune.svm. I got an accuracy of 0.8926004, precision of 0.6524823, recall of 0.1740776, F1-score of 0.274832, and an AUC of 0.8970666. However, since the data structure remains nonlinear, the model did not increase that much.

3. **SVM Radial:**

   Hypothesis: Radial SVM is expected to capture complex non-linear relationships better.

   The radial SVM may not capture specific non-linear patterns in the data, but in this default radial SVM achieved accuracy of 0.8977989, precision of 0.6481069, recall of 0.2753075, F1-score of 0.3864542, and an AUC of 0.9040328. The result is slightly lower than SVM Linear and SVM Linear Tuned. While the SVM radial kernel helps to transform data into a high- dimensional space, using default settings fails to capture specific non-linear patterns.

4. **SVM Radial Tuned:**

Hypothesis: Tuned SVM Radial will improve radial SVM performance.

In this experiment, I expect that after tuning parameters of the SVM Radial, the model should significantly improve. In this model achieved accuracy of 0.8977989, precision of 0.6481069, recall of 0.2753075, F1-score of 0.3864542, and an AUC of 0.9040328. It is same as SVM Radial default settings.

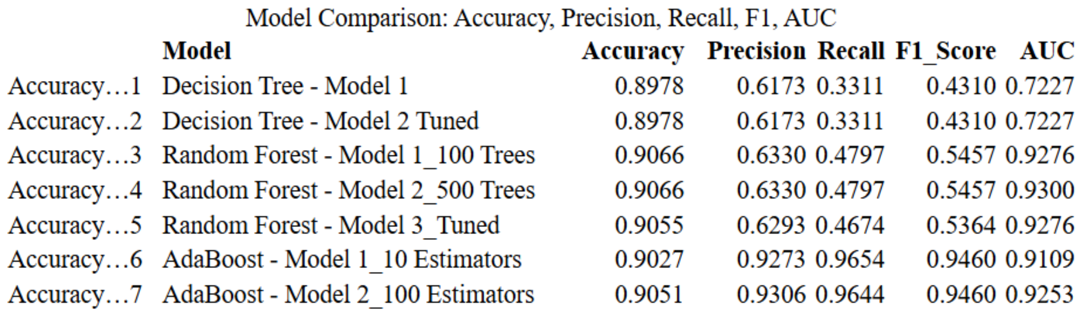
5. **SVM Polynomial:**

Hypothesis: SVM Polynomial model will provide better performance.

The SVM Polynomial was able to capture some of the non-linear interactions between features. However, it's overall slightly better than the linear SVM, but below the tuned SVM Radial Tuned. The model achieved accuracy of 0.8937064, precision of 0.6643836, recall of 0.1835383, F1-score of 0.2896205, and an AUC of 0.8829763.

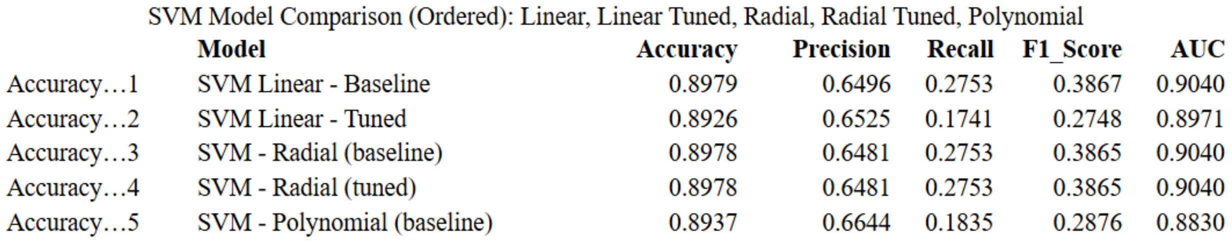
IV. Conclusion:

Compared with previous assignment's model. The AdaBoost and Random Forest clearly outperformed the other algorithms. AdaBoost demonstrated the best overall classification performance in recall and F-1 score. SVM models are stable and interpretable in decision-making, underperformed in recall and F1 metrics.



Model Comparison: Accuracy, Precision, Recall, F1, AUC

	Model	Accuracy	Precision	Recall	F1_Score	AUC
Accuracy…1	Decision Tree - Model 1	0.8978	0.6173	0.3311	0.4310	0.7227
Accuracy…2	Decision Tree - Model 2 Tuned	0.8978	0.6173	0.3311	0.4310	0.7227
Accuracy…3	Random Forest - Model 1_100 Trees	0.9066	0.6330	0.4797	0.5457	0.9276
Accuracy…4	Random Forest - Model 2_500 Trees	0.9066	0.6330	0.4797	0.5457	0.9300
Accuracy…5	Random Forest - Model 3_Tuned	0.9055	0.6293	0.4674	0.5364	0.9276
Accuracy…6	AdaBoost - Model 1_10 Estimators	0.9027	0.9273	0.9654	0.9460	0.9109
Accuracy…7	AdaBoost - Model 2_100 Estimators	0.9051	0.9306	0.9644	0.9460	0.9253



SVM Model Comparison (Ordered): Linear, Linear Tuned, Radial, Radial Tuned, Polynomial

	Model	Accuracy	Precision	Recall	F1_Score	AUC
Accuracy…1	SVM Linear - Baseline	0.8979	0.6496	0.2753	0.3867	0.9040
Accuracy…2	SVM Linear - Tuned	0.8926	0.6525	0.1741	0.2748	0.8971
Accuracy…3	SVM - Radial (baseline)	0.8978	0.6481	0.2753	0.3865	0.9040
Accuracy…4	SVM - Radial (tuned)	0.8978	0.6481	0.2753	0.3865	0.9040
Accuracy…5	SVM - Polynomial (baseline)	0.8937	0.6644	0.1835	0.2876	0.8830