Project 2: Comparison of Tree-based model and linear model on the Titanic dataset

Tse Justin Chung Heng [1] jchtse@connect.ust.hk
[1]: Department of Mathematics, HKUST

1. Introduction

In project 1, we have applied various model selection technique to optimize the logistic regression model. However, from other Kaggler's project, tree-based model seems to out perform logistic regression model in most case. Therefore, in this project, we will try to verify whether tree-based model have an advantage in this particular problem and we will compare this two kind of models and try to explain why tree-based model out perform logistic regression model for this problem.

2. Tree-Based model

Methodology

1. Decision Tree
2. Decision Tree with Bagging
3. Randomized Decision Trees (extra-trees)
4. AdaBoost
5. Gradient Boosting Classifier
6. Random Forest

3. Linear model

Methodology

1. Logistic Regression with L1 regularization
2. Logistic Regression with L2 regularization

3. Data processing and feature engineering

Compare to the feature of project 1, several modifications are made in order to incooperate more information in the dataset.
1. Family size is used instead of isAlone, hoping to add more information to the model.
2. More title is extracted from the Title variable.
3. One hot encoding is applied for nominal data.

	MLA Name	MLA Parameters	MLA Train Accuracy Mean	MLA Test Accuracy Mean	MLA Test Accuracy 3*STD	MLA Time
8	XGBClassifier	{'objective': 'binary:logistic', 'use_label_en...	0.89588	0.823134	0.03012	0.148902
3	GradientBoostingClassifier	{'ccp_alpha': 0.0, 'criterion': 'friedman_mse'...	0.872097	0.822388	0.033657	0.082205
4	RandomForestClassifier	{'bootstrap': True, 'ccp_alpha': 0.0, 'class_w...	0.899625	0.812313	0.053228	0.159975
1	BaggingClassifier	{'base_estimator': None, 'bootstrap': True, 'b...	0.895318	0.810448	0.053638	0.029325
2	ExtraTreesClassifier	{'bootstrap': False, 'ccp_alpha': 0.0, 'class_...	0.899813	0.809701	0.05298	0.142738
0	AdaBoostClassifier	{'algorithm': 'SAMME.R', 'base_estimator': Non...	0.829588	0.807836	0.051844	0.092847
7	DecisionTreeClassifier	{'ccp_alpha': 0.0, 'class_weight': None, 'crit...	0.899813	0.806343	0.044902	0.006781
6	RidgeClassifierCV	{'alphas': array([0.1, 1. , 10.]), 'class_w...	0.803371	0.781716	0.051844	0.008577
5	LogisticRegressionCV	{'Cs': 10, 'class_weight': None, 'cv': None, '...	0.80206	0.778731	0.052277	0.307422

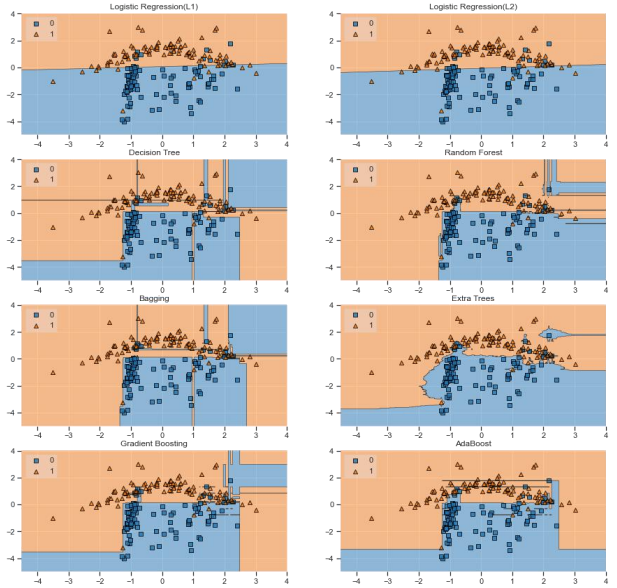
Comparison of tree-based model and linear model rank by test accuracy

4. Tree-based model v.s. linear model

For this dataset, since it has more than 3, it is hard to visualize the decision boundary. However, we believe that linear model is better at representing data that are linearly seperatable should be able to extend to higher dimension. This is possibily one of the reason why tree-based model perform better than linear model in this dataset.
Another reason that decision tree may better suit this problem is interpretibility.
On the other hand, linear model has lower variance than decision tree, it is true in this project as well, the varaince of decision tree, Logistic Regression with L1 regularization and Logistic Regression with L2 regularization are 0.86, 0.29 and 0.33 respectively.

5. Decision Tree vs ensemble method

Ensemble method such as bagging and random forest are designed to reduce the variance of the decision tree. The variance of decision tree, random forest and bagging are 0.86, 0.81 and 0.65 respectively. It agrees with the general believes that the bagging technique can reduce variance by training different decision tree based on differnt bootstrap samples and the random forest model can even further reduce variance by limiting the option of variables. An extra advantage of random forest over other method is that it is less vulnerable to correlated variables because of the reduction of correlation between trees.



Decsion boundary of each model

6. Conclusion

This project demonstrate that tree-based model, especially tree-based ensemble model is generally better model for this project in terms of accuracy.
In conclusion, Some mistake from project 1 has been fixed. Also, performace of the model is improved by using tree-based model and doing better feature engineering, we are able to improve the testing accuracy from 0.77 to 0.82.

7. References

1. https://medium.com/analytics-vidhya/decision-boundary-for-classifiers-an-introduction-cc67c6d3da0e
2. https://hackernoon.com/how-to-plot-a-decision-boundary-for-machine-learning-algorithms-in-python-3o1n3w07