MATH 4995 Mini-Project 1: Predict Survival on the Titanic

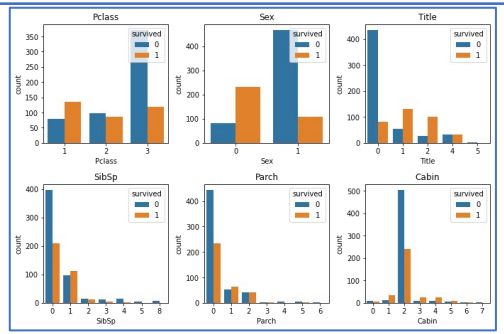
Ngai Nok Yiu (20510180), Cheung Hang Yee (20514796)

1. Introduction

The Titanic survival problem is a binary classification problem. Given two type of features: passengers' personal information(e.g. age,sex) and their trip information(e.g. Ticket number, Cabin number), we need to predict whether they will survive the accident or not. We test two classification models - logistic regression and random forest classifier and compare their performance.

2. Feature Engineering

To fit the logistic regression model, we need to encode the qualitative features to numeric one using mapping function. Although name of each passenger is not important, the title in the name reflect their social status which may affect the survival rate. We therefore extracted the title and build a new feature "Title". We use the mean and median to fill the empty values in "Age" and "Fare". We only keep the letter in "Cabin".

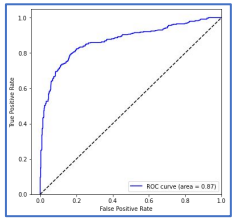


Count Plot of features

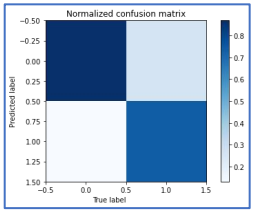
3. Logistic Regression

We fit the training data to a logistic regression model and perform 10-fold cross validation, the result average accuracy is 0.81, with sd 0.03. The ROC score is 0.87. The Kaggle score is 0.76794.

Model: $$\log\left(\frac{P(y="1"|x)}{P(y="0"|x)}\right) = W^T \cdot x$$



ROC curve



Confusion Matrix

4. Random Forest

We fit the data to a random Forest: model and test it with 100 n_estimators. The average accuracy is 0.84, with sd 0.03. The ROC score is 0.92. The Kaggle score is 0.78947.

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b(x')$$

5. Analysis

One reason why logistic regression performs slightly worse than random forest may be because the linearity of the features is not strong. Besides, most of the features are not continuous values which might not work well with logistic regression.

Logistic Regression cannot handle missing values unlike random forest which is immune to it as its underlyings are decision trees. So Logistic Regression require data pre-processing, which is not convenience.

Overfitting occur in both model as the training set accuracy score is higher than the Kaggle score.

Random Forest has a good Performance on imbalanced datasets than Logistic Regression.

6. Conclusion

Although the difference of two model are very small, We think random forest model is better for this Titanic survival problem. First, using Random Forest do not need to pre-process the data which saves time.

Second, Logistic Regression cannot handle missing values, which is not suitable for Titanic survival problem. Some of the data in Titanic survival problem are missing and we use mean, mode and median to replace them when using Logistic Regression. It will affect the accuracy and may not be general enough for new data.

7. Contribution

Logistic Regression: Ngai Nok Yiu

Random Forest: Cheung Hang Yee