The Titanic Data Analysis

Tie Ma

2024-02-26

In the framework of this analysis, linear regression, logistic regression, and random forest models have been selected to evaluate and contrast their effectiveness on the training dataset for predicting passenger survival within the test dataset. The dataset consists of 11 variables, 8 of which have been used for this study.

Given the constraints of the variable set, the necessity for extensive dimensionality reduction methods such as Principal Component Analysis (PCA) and lasso or double lasso regression is not prioritized, as their performance may not be significantly better than that of simple linear regression and logistic regression. Both models are better suited to situations where there are a larger number of variables with existing multicollinearity. Therefore, I have chosen the linear regression model along with the logistic regression model.

To better grasp complex and nonlinear relationships that linear regression and logistic regression cannot capture, I'm turning to the Random Forest model. This model allows us to dive deeper into the data's patterns, going beyond what simpler models can achieve.

The examination of the Titanic dataset through both linear and logistic regression models suggests a pattern of determinants affecting survival outcomes. Passenger class has a negative effect on survival, with higher classes associated with a decreasing chance of survival. Gender plays a significant role, with females having a higher probability of survival. Age negatively correlates with survival, suggesting that younger passengers are more likely to survive. The presence of siblings or spouses aboard shows a slight negative impact on survival chances. Furthermore, individuals who embarked from Southampton are slightly less likely to survive, primarily because the majority of passengers boarded the Titanic at Southampton.

The prediction performance of linear regression, logistic regression and radom forest model on the test test show prediction scales with marginal differences. The Random Forest model exhibited the highest prediction score of 0.77751, followed by logistic regression with a score of 0.77272, and linear regression trailing with a score of 0.77033.

Although the differences are marginal, they are significant, with the Random Forest model demonstrating a superior advantage of 0.00479 over logistic regression and 0.00718 over linear regression. It demonstrates the Random Forest model's capability in handling complex and non-linear relationships between variables where linear models fall short. Because linear and logistic regression models are both constructed based on the assumption of linearity between the dependent and independent variables, an assumption that may not hold true in real-world situations. In the context of the Titanic dataset, the assumption of linearity between variables such as socio-economic status, age, and family connections may interact in non-linear ways. For instance, the benefit of a higher socio-economic status on survival likely varies more significantly for adults than for children, who were more likely to be prioritized for lifeboat sport regardless of social class.

This report comprises three parts. In the first part, I perform data cleaning and evaluation. The second part involves constructing and evaluating the performance of linear and logistic regression models. Finally, in the third part, I will construct and optimize a random forest model.

Part One: Data Cleaning and Evaluation

The Titanic training data set including following 11 variable:

- `PassengerId`: The unique Identification number.
- `Survived`: Indicates if a passenger survived(1) or not (0).
- `Pclass` passenger class (1 for first-class, 2 for second-class, 3 for third-class).
- `Sex`: The gender of the passenger (male or female).
- `Age`: The age of a passenger.
- `SibSp`: The number of siblings or spouses aboard.
- `Parch`: The number of parents or children aboard.
- `Fare`: The ticket fare.
- `Embarked`: port of embarkation (C for Cherbourg, Q for Queenstown, S for Southampton).
- `Cabin`: The cabin number.
- `Ticket`: The ticket number.

To optimize the model's training efficacy, names, ticket numbers, and cabin numbers were omitted due to their insufficiency in providing meaningful patterns for the regression model. Specifically, cabin numbers were excluded because they are missing in 687 instances, indicating that 77% of the training data lack cabin numbers

First, I checked for any NA (missing) values in the training dataset.

```
## PassengerId    Survived      Pclass         Sex         Age       SibSp
##           0           0           0           0         177           0
##       Parch        Fare    Embarked
##           0           0           2
```

The training data is missing 177 Age values and 2 Embarked values. Considering the size of the training dataset is only 881 rows, removing all the missing Age variables would further reduce the available data for the model. Therefore, I decided to use the average age to fill the missing 177 values. However, I do need to check the average age between the people who survived and those who did not to avoid introducing bias into the training dataset.

The average age by Survived as following.

```
##   Survived      Age
## 1        0 30.62618
## 2        1 28.34369
```

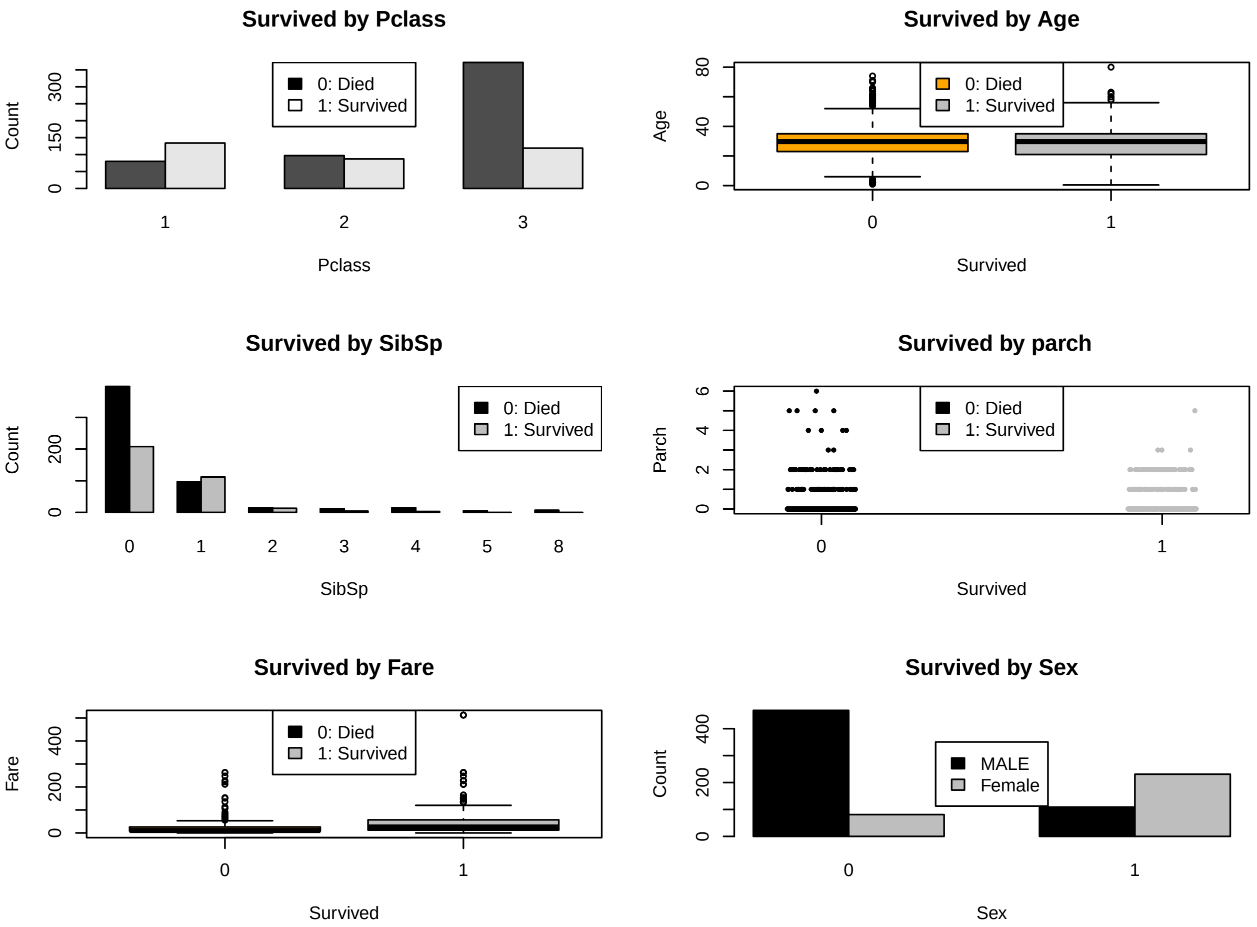
The average age of the entire data set

```
## [1] 29.69912
```

Replacing missing values with the average age of 29.69 in the entire training dataset, which lies between the mean ages of survivors (28.343) and non-survivors (30.62), will introduce a bias towards the non-survivor group in the analysis of the correlation between age and survival outcomes. However, compared to the disadvantage of deleting all missing age values and thereby losing a critical variable in predicting survival, a minor bias is relatively acceptable

Furthermore, after filling in all the missing age values, I removed the 2 missing rows of the Embarked value from the training dataset and transformed the 'Sex' column into a dummy variable.

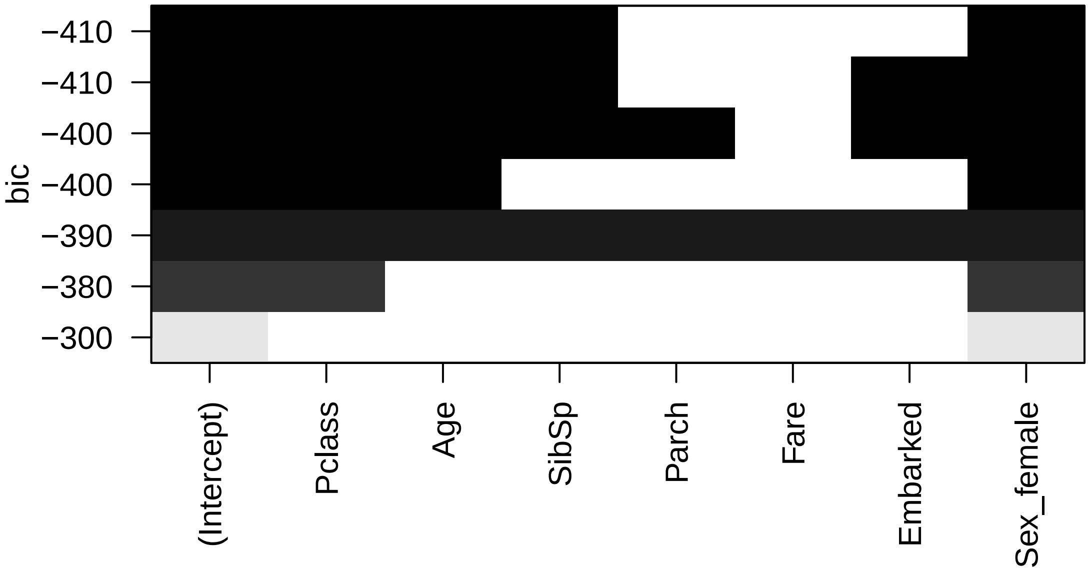
For a better understanding of the possible patterns in the data and to determine potential variables for the linear regression and logit models, I examined graphs comparing the impact of different variables on passenger survival. From these graphs, we can observe that Passenger Class (Pclass), SibSp, and gender significantly influence the survival of passengers. Furthermore, the fare, representing the ticket price, shows that passengers who paid higher fares were more likely to survive. This observation aligns with the conclusion drawn from the Passenger Class graphs.Furthermore, the rest of the variables do not provide a visible impact on survival according to the graph.



Part two: Data Evaluation

In order to select the optimal variables for constructing the linear regression model without risking over fitting, I will use the best subset selection and Ridge regression to explore all possible linear combinations of variables ’

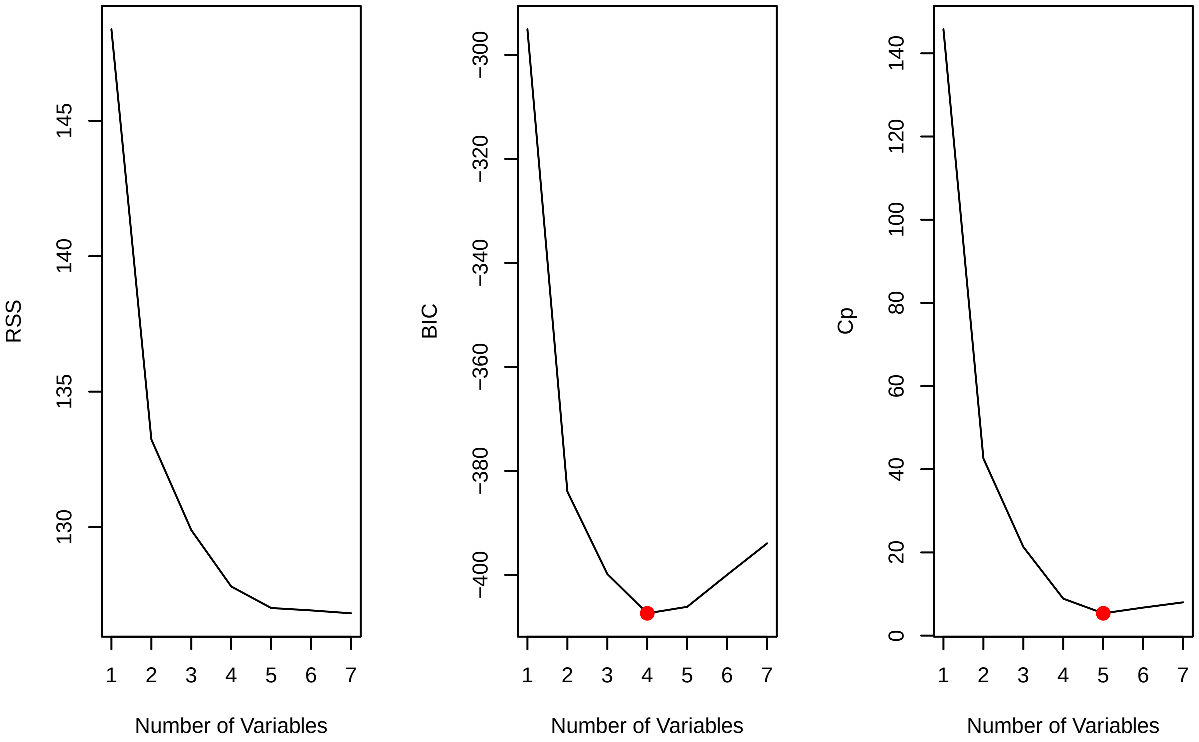
First, I will use the best subset selection to search all possible combinations and compare their BIC values.



The Bayesian Information Criterion (BIC) evaluates a model's performance by considering how well the model explains the data and penalizing models for their number of parameters, which helps prevent the problem of overfitting. A lower BIC value suggests that the model is more effective at explaining the data while using fewer parameters, striking a balance between a good fit and avoiding overfitting.

From the graph, we can see that the combination of Pclass, Age, SibSp, and sex_female has the lowest BIC, indicating that this model is likely the most efficient among the tested combinations in terms of balancing model complexity and goodness of fit.

Next, we determine the number of variables.



Mallows' Cp (Cp) assesses a function similar to that of the BIC; it measures a model's performance by evaluating the trade-off between the model's complexity and its ability to fit the data closely. Similar to BIC, a lower Cp value suggests that the model best balances fitting the data and complexity. The Cp graph suggests that the optimal number of variables is 5.

Part Two: Linear Regression and Logistic Regression

According to the best subset selection, I should choose the two group with the following variables.

- The first group includes:
  - Pclass, Age, SibSp, Sex_female
- The second group includes:
  - Pclass, Age, SibSp, Embarked, Sex_female

Next, I checked for heteroscedasticity and multicollinearity in the training dataset for both vairable in both linear regression and logistic regression models.

For The heteroscedasticity test for linear regression, I am using BP test.

```
##
##  studentized Breusch-Pagan test
##
## data:  data_test_1_lm
## BP = 5.4777, df = 4, p-value = 0.2417
```

```
##
```

```
##  studentized Breusch-Pagan test
##
## data:  data_test_2_lm
## BP = 5.132, df = 5, p-value = 0.4
```

Given that the p-values for the variable sets under the linear regression model exceed the 0.05 , we fail to reject the null hypothesis. We conclude that there is not enough statistically significant evidence to suggest heteroskedasticity exists in the dataset.

For the multicollinearity test for linear regression, I am using the VIF test.

```
##     Pclass        Age      SibSp Sex_female
##   1.153247   1.192362   1.067780   1.047674
```

```
##     Pclass        Age      SibSp   Embarked Sex_female
##   1.178775   1.193368   1.073569   1.042342   1.057325
```

```
##     Pclass Sex_female        Age      SibSp
##   1.296395   1.144130   1.277105   1.126283
```

```
##     Pclass Sex_female        Age      SibSp   Embarked
##   1.301521   1.137749   1.275564   1.129368   1.015937
```

All the VIF values for the variable are between 1 and 1.6. This indicates that the training data show moderate correlation which will not impact the model outcome.

Next, we will proceed with the implementation of Ridge regression

```
##              Variable           OLS         Ridge
## Pclass         Pclass -0.1688667659 -0.1564689820
## Age               Age -0.0058566464 -0.0053474253
## SibSp           SibSp -0.0411820740 -0.0388136461
## Parch           Parch -0.0168729693 -0.0138132743
## Fare             Fare  0.0002829035  0.0003878972
## Embarked     Embarked -0.0353433928 -0.0358988675
## Sex_female Sex_female  0.5059113514  0.4799392450
```

Ridge regression applies a penalty to the size of coefficients to shrink them toward zero, particularly targeting less important predictors. notably, the coefficients for the variables "Sex_female" and "Pclass" remain relatively high in both OLS and Ridge regression, suggesting a strong relationship with the dependent variable. The coefficients for SibSp, Embarked, and Parch follow this trend. However, Ridge regression only shrinks the coefficients of two variables, "Age" and "Fare," towards zero. Therefore, it does not provide any new model suggestions.

Model Comparison and Kaggle Scale

In this step, the training dataset was divided into two segments: 90% of the data was allocated for training, while the remaining 10% served as the test set. The training dataset was used to train the models, which were then employed to predict the output for the test set. By comparing the predicted outcomes between predicted and actual values were utilized to compute the Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) to evaluate the prediction ability of the model.

MSE and RMSE impose greater penalties for larger prediction errors, emphasizing the importance of minimizing such errors. Conversely, MAE focuses on the overall performance of the model, providing an average error measure without specific emphasis on large errors. By evaluating these metrics, we could gain insights into the predictive performance of both linear and logistic regression models.

Additionally, the Kaggle prediction scale was integrated into the analysis to facilitate comparison and interpretation of the model's performance within a standardized framework.

```
##           Test Model_1_lm Model_1_glm Model_2_lm Model_2_glm
## 1          MSE  0.1223255   0.1179534  0.1241875   0.1191073
## 2         RMSE  0.3497506   0.3434435  0.3524024   0.3451192
## 3          MAE  0.2756727   0.2625819  0.2783762   0.2646737
## 4 Kaggle Scale  0.7679400   0.7538500  0.7703300   0.7727200
```

The logistic regression model with variable set two (Pclass + Sex_female + Age + SibSp + Embarked) has the highest kaggle prediction scale with the smallest MSE, RMSE and MAE.

Part Three: Random Forest Model

In construction optimal Random Forest model, it is essenrtial to identify teh best value for both `ntree()` and `mtry().ntree()` defines the number of trees in the forest, influencing the model's stability, while `mtry()` dictates the number of variables considered at each decision point, balancing model bias and variance for better robustness.

To determine the optimal `mtry()` value, I am using 10-fold cross-validation to examine the performance for values ranging from 1 to 10 and select the one with the highest accuracy. This approach divides the dataset into 10 segments, using 9 of these segments to train the data and one to evaluate the performance of different `mtry()` values.

Compared to `mtry()`, the setting for `ntree()` is less critical. It is sufficient to select a sufficiently high number to stabilize the error rate. Thus, I have chosen to set ntree() to 10,000.

```
## Random Forest
##
## 889 samples
##   9 predictor
##   2 classes: '0', '1'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 800, 800, 801, 800, 800, 800, ...
## Resampling results across tuning parameters:
```

```
##
##   mtry  Accuracy   Kappa
##    1    0.8087717  0.5741640
##    2    0.8245020  0.6157622
##    3    0.8267748  0.6228800
##    4    0.8324055  0.6353924
##    5    0.8256512  0.6236363
##    6    0.8200332  0.6114304
##    7    0.8189224  0.6100784
##    8    0.8177988  0.6072135
##    9    0.8188968  0.6106127
##   10    0.8188968  0.6101722
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was mtry = 4.
```

According to the output, the `mtry()` = 4 has the highest accuracy with the highest Kappa scale.

Then I constructed my Random Forest with 10,000 trees `ntree()`, set `mtry()` equal to 3, and `nodesize()` to 1 for classifying the output.

```
##
## Call:
##  randomForest(formula = Survived ~ ., data = TTD, ntree = 10000,      mtry = 4, nodes
##                Type of random forest: classification
##                      Number of trees: 10000
## No. of variables tried at each split: 4
##
##         OOB estimate of  error rate: 17.66%
## Confusion matrix:
##     0   1 class.error
## 0 492  57   0.1038251
## 1 100 240   0.2941176
```

The result suggest the Out of Bag error rate (OBB) at 17.5% and a prediction error bias towards survivors, with a prediction error of 10.4% for non-survivors and 29.4% for survivors. Moreover, with the prediction score is 0.77751 which is the higher than logistic model and linear regression.

The Out of Bag error rate (OBB) at 17.55%, means that 17.5% of the prediction made by the Random Forest on the out of bag samples are incorrect.

The confusion matrix shows that the Random Forest model accurately identified 492 non-survivors and 241 survivors. Prediction errors occurred in 57 cases incorrectly predicted as survivors and in 99 cases incorrectly predicted as non-survivors. The prediction error rates are approximately 10.38% for non-survivors and 29.4% for survivors. The discrepancy in prediction error rates may stem from replacing all missing values with the dataset's average

age, particularly since the average age for the training dataset is higher than that of the survivors.

Conclusion

In this assignment, I used three different models—linear regression, logistic regression, and random forest models—to train on the Titanic dataset and predict passenger survival in the test set. The results suggest that the random forest model delivers the best performance with a prediction score of 0.77751, enhancing accuracy by 0.00718 compared to the best linear regression and logistic regression, which both have a variable set at 0.77033. Comparing logistic regression to linear regression, logistic regression performs slightly better in terms of MAE, RMSE, MSE and prediction scare with 0.00239

This research demonstrates that random forest models are more suited for predicting survival on the Titanic dataset due to their ability to capture non-linear and complex relationships between variables.

However, due to the author's limited knowledge and training in statistics, a bias towards non-survival outcomes related to age has been introduced to both the training and test dataset by using the average age to replace missing age values. This may potentially decrease the accuracy of model predictions due to bias.