The Titanic Data Analysis

Tie Ma

2024-02-24

In this assignment, I have chosen to employ linear regression, logistic regression, and regression tree models to delve into the relatively compact Titanic dataset, which consists of 11 variables, albeit only 8 are applicable. The dataset's constrained size somewhat limits the effectiveness of principal components or lasso/double lasso models, which excel when tasked with condensing variable sets and boosting model efficacy due to their prowess in reducing dimensionality and selecting relevant features.

My preference for the linear regression model stems from it being the mode I have been trained in and studied for a long time. Logistic regression is selected due to its superior performance with binary outputs, particularly for predicting survival on the Titanic. For machine learning and to capture the non-linear relationships among the variables, I chose the regression tree model.

According to my analysis, performance metrics from Kaggle indicate a close match between linear and logistic regression at 0.77033, with the regression tree model slightly outperforming them at 0.777051, enhancing accuracy by a mere 0.00718.

The regression tree model has the best Kaggle score because it can capture non-linear relationships. Linear and logistic regression models are both constructed based on the assumption of linearity between the dependent and independent variables. Such an assumption may not hold true in real-world scenarios. In contrast, the regression tree does not assume linear dependence. It searches for the best path through the data that leads to the most similar outcomes within a leaf at the end of the path.

The first part includes data cleaning, checking, and variable selection using the best subset and Ridge regression, and compares the performance and Kaggle scores for two different variable sets using logistic regression and linear regression. The outcomes are identical due to the small, marginally better performance difference between logistic regression and linear regression, which does not significantly impact the final outcome. However, the best-performing model is logistic regression with the following variables: Pclass + Sex_female + Age + SibSp + Embarked.

The second part of this assignment involves using a regression tree model. I first used 10-fold cross-validation to determine the optimal settings for mtry, and then generated the regression tree model.

Part one: The data evaluation process.

The Taitanic training data set including following 11 variable:

- `PassengerId`: A unique numerical ID given to each passenger.
- `Survived`: Indicates if a passenger survived (1) or not (0).
- `Pclass`: Denotes the passenger class (1 for first-class, 2 for second-class, 3 for third-class), which is a proxy for socio-economic status.
- `Sex`: The gender of the passenger (male or female).
- `Age`: The age of the passenger in years.
- `SibSp`: The number of siblings or spouses aboard with the passenger.
- `Parch`: The number of parents or children aboard with the passenger.
- `Fare`: The ticket fare paid by the passenger.
- `Embarked`: The port of embarkation (C for Cherbourg, Q for Queenstown, S for Southampton).
- `Cabin`: The cabin number where the passenger stayed.

- `Ticket`: The ticket number for the passenger's voyage.

I removed the names, ticket numbers, and cabin numbers during the data import process because they are not important for training the dataset, and having only a few cabin numbers does not aid in training the models. First, I checked whether there are any NA (missing) values in the training data set.

```
## 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
```

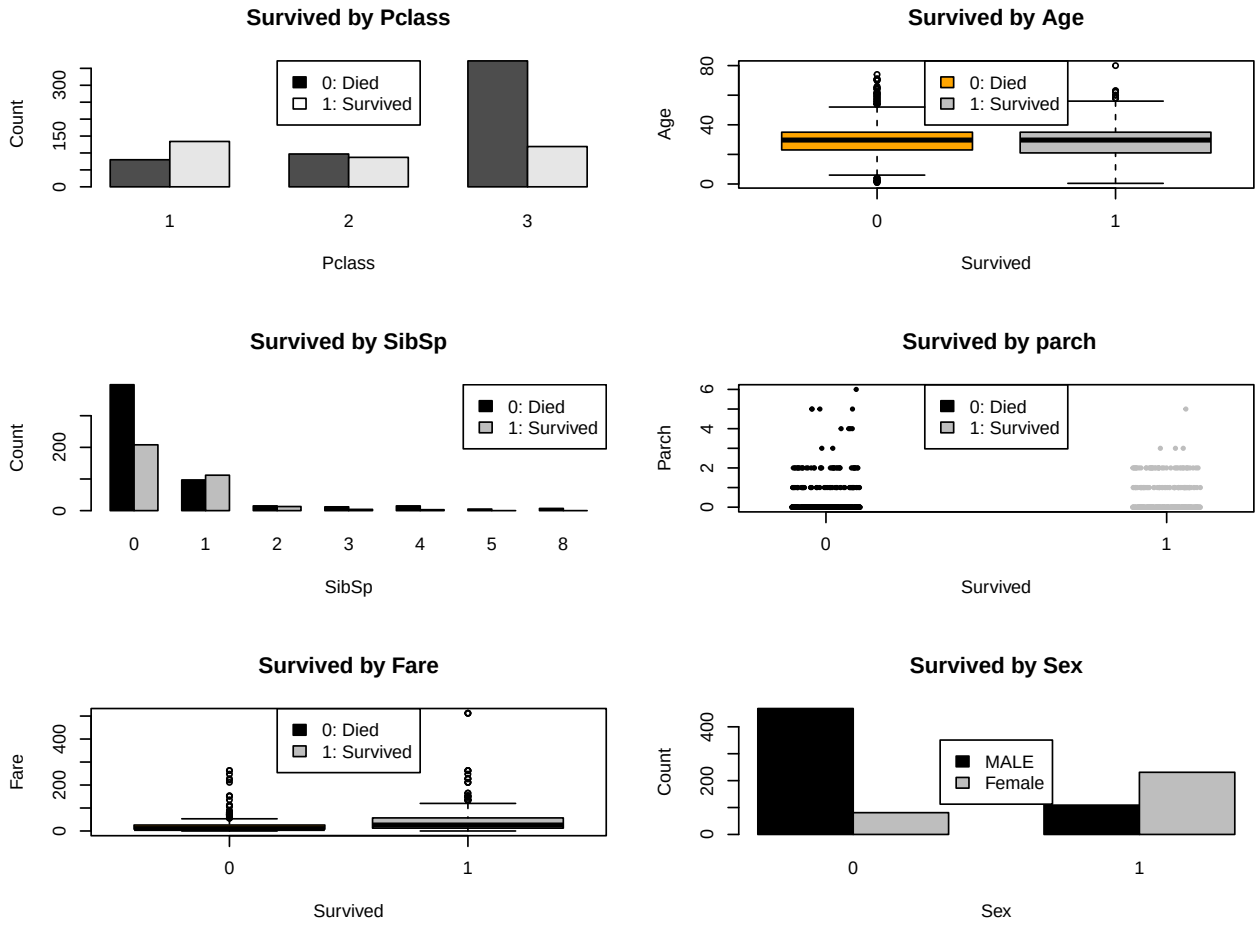
The average age of the entire training data set are 29.69 which is right in between the average age for people who survived and died. Therefore using the average age will not introduce significant bias toward the training data.

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. This transformation was necessary because machine learning models require numerical input, and converting 'Sex' into a dummy variable. .

Step two: Check The Data.

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 analysis.



For better perform of the linear regression, I check the heteroscedasticity and the multicollinearity of the training data set.

For The heteroscedasticity test, I am using BP test.

```
##
##  studentized Breusch-Pagan test
##
## data:  data_test_1
## BP = 9.8992, df = 7, p-value = 0.1944
```

Since the p value is 0.19 which is greater than 0.05, we fail to reject the null hypothesis. We state that there is not enough statisticallty significant evidence of heteroskedasticity exist in the data set.

For the multicollinearity, I am using the VIF test.

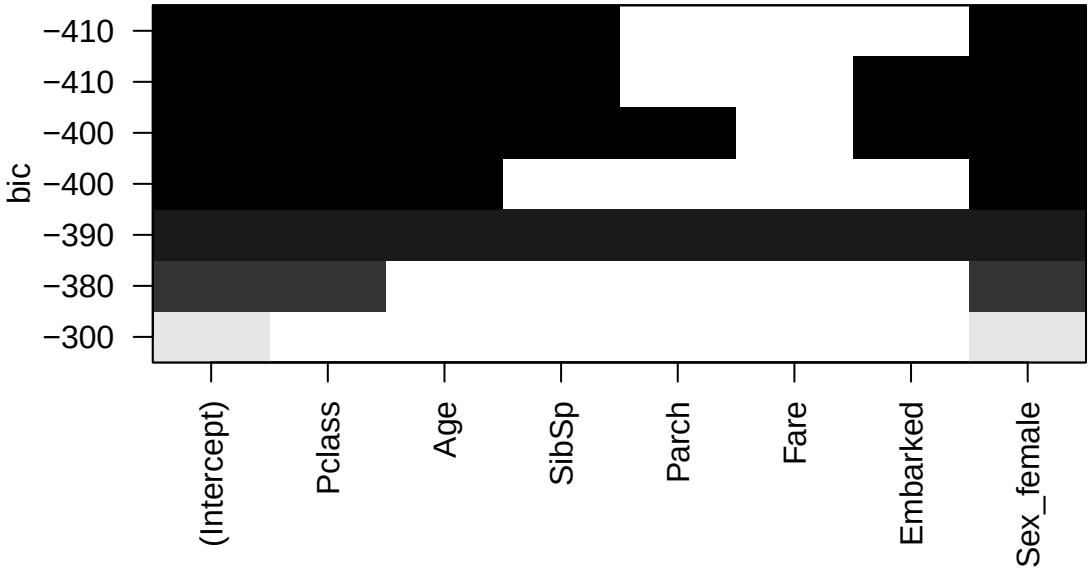
```
##     Pclass        Age      SibSp      Parch       Fare   Embarked Sex_female
##   1.664130   1.203003   1.281804   1.323066   1.645957   1.081159   1.109890
```

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.

Step Three: Choose The Variable

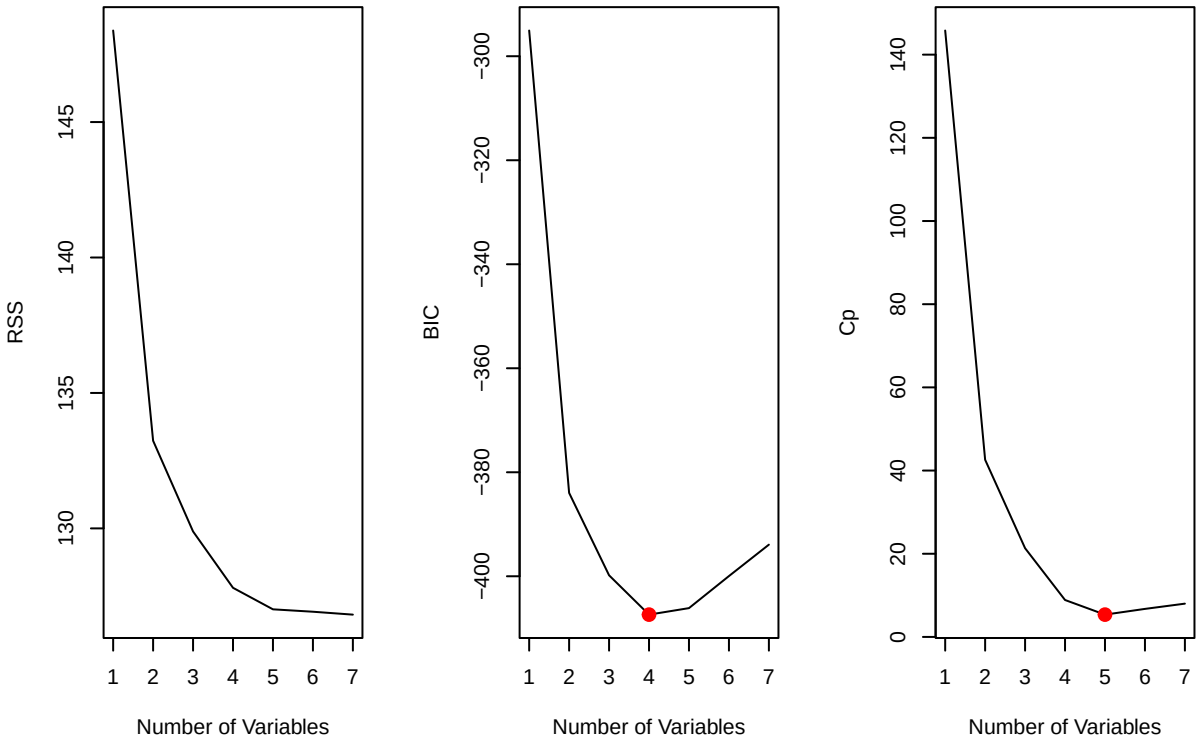
In order to select the optimal variable for constructing the linear regression model without risking over fitting, I will use the best subset selection and Ridge regression to explore all possible linear combinations of variables

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



From the graph, we can see that the combination of Pclass, Age, SibSp, and sex_female has the lowest BIC, which means 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.



The BIC graphy suggest with the four variables will reach the lowest BIC same as the early grpahy point out. As CP(AIC) graph suggest that the optimal variables number is 5.

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"

Now, let's proform the 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
```

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.

Step Four: Model Comparison and Kaggle Scale

In this step, I divided the training set into two parts: 90% of the data will be used for training, and the remaining 10% will serve as the test set. This division will allow us to calculate the Mean Squared Error, Root Mean Squared Error, and Mean Absolute Error to evaluate the performance of two sets of linear regression and two sets of logistic regression models. And then I combine the Kaggle prediction scale together with the table.

```
##           Test Model_1_lm Model_1_glm Model_2_lm Model_2_glm
## 1          MSE  0.1664768   0.1653768  0.1631783   0.1633131
## 2         RMSE  0.4080156   0.4066654  0.4039534   0.4041201
## 3          MAE  0.3172978   0.3086485  0.3146302   0.3061866
## 4 Kaggle Scale  0.7679400   0.7679400  0.7703300   0.7703300
```

The comparison scales between variable set one and variable set two are close, with only around a 0.01 difference in magnitude. Upon a detailed examination of the performance, we can observe that the logistic regression model performs better on a small scale. However, a detailed line-by-line analysis reveals that variable set two, when paired with logistic regression models, exhibits the best performance among all the models.

Comparing the Kaggle scale for predictions, we notice that the logistic regression model and the linear regression model for the same variable set have identical scales. This is because I forced R to assign a value of 1 when the model's output is greater than 0.5, and similarly, to assign a value of 0 when the prediction is lower than 0.5. As a result, even though the logistic regression model has better performance, the improvement is not significant enough to increase the probability of survival to the extent that it would change the outcome from death to survival compared to the linear regression model

Part 2: Regression Tree Model

I transfer the survived variable into the factor and using 10 ford validation to determine the optimal number of mtry I should use.

```
## 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.
```

Then I constructed my regression tree with 10,000 trees (ntree), set ntry equal to 3, and nodesize to 1 for classifying the output.

```
##
## Call:
##  randomForest(formula = Survived ~ ., data = TTD, ntree = 10000,      mtry = 4, nodesize = 1, importa
##                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
```

For the regression tree model, the prediction score is 0.777051, which outperforms the best linear regression and logistic regression models by 0.00718, approximately 0.718%.

Conclstion:

In this assignment, I used three three different models—linear regression, logistic regression, and regression tree models—to train on the Titanic dataset and predict passenger survival in the test set. The results suggest that the regression tree model delivers the best performance with a prediction score of 0.777051, 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, and MSE. However, this marginal improvement is not sufficient to alter the outcome when predictions greater than 0.5 are classified as 1 and those smaller than 0.5 as 0. Ultimately, the prediction scores are the same between logistic regression and linear regression.