HARVARD EXTENSION SCHOOL

EXT CSCI E-106 Model Data Class Group Project Template

Author One Author Two Author Three Author Four Author Five Author Six

01 November 2024

Abstract

This is the location for your abstract.
It must consist of two paragraphs.

Contents

Classify whether a passenger on board the maiden voyage of the RMS Titanic in 1912 survived given their age, sex and class. Sample-Data-Titanic-Survival.csv to be used in the Final Project

Variable	Description
pclass	**Passanger Class, could be 1st, 2nd or 3rd**
survived	*Survival Status: 0=No, 1=Yes*
name	*Name of the Passanger*
Sex	*Sex*
sibsp	*Number of Siblings or Spouses aboard*
parch	*Number of Parents or Children aboard*
ticket	*Ticket Number*
fare	*Passenger Fare*
cabin	*Cabin number, "C85" would mean the cabin is on deck C and is numbered 85.*
embarked	*Port of Embarkation: C=Cherburg, S=Southampton, Q=Queenstown*
boat	*Lifeboat ID, if passanger survived*
body	*Body number (if passanger did not survive and body was recovered*
home.dest	*The intended home destination of the passanger*

Instructions:

0. Join a team with your fellow students with appropriate size (Four Students total)
1. Load and Review the dataset named "Titanic_Survival_Data.csv"
2. Create the train data set which contains 70% of the data and use set.seed (1023). The remaining 30% will be your test data set.
3. Investigate the data and combine the level of categorical variables if needed and drop variables as needed. For example, you can drop id, Latitude, Longitude, etc.
4. Build approprite model to predict the probability of survival.
5. Create scatter plots and a correlation matrix for the train data set. Interpret the possible relationship between the response.
6. Build the best models by using the appropriate selection method. Compare the performance of the best logistic linear models.
7. Make sure that model assumption(s) are checked for the final model. Apply remedy measures (transformation, etc.) that helps satisfy the assumptions.
8. Investigate unequal variances and multicollinearity.
9. Build an alternative to your model based on one of the following approaches as applicable to predict the probability of survival: logistic regression, classification Tree, NN, or SVM. Check the applicable model assumptions. Explore using a negative binomial regression and a Poisson regression.
10. Use the test data set to assess the model performances from above.
11. Based on the performances on both train and test data sets, determine your primary (champion) model and the other model which would be your benchmark model.
12. Create a model development document that describes the model following this template, input the name of the authors, Harvard IDs, the name of the Group, all of your code and calculations, etc..:

Due Date: December 18th, 2024 at 11:59 pm EST

Notes No typographical errors, grammar mistakes, or misspelled words, use English language All tables need to be numbered and describe their content in the body of the document All figures/graphs need to be numbered and describe their content All results must be accurate and clearly explained for a casual reviewer to fully understand their purpose and impact Submit both the RMD markdown file and PDF with the sections with appropriate explanations. A more formal document in Word can be used in place of the pdf file but must include all appropriate explanations.

Executive Summary

This section will describe the model usage, your conclusions and any regulatory and internal requirements. In a real world scneario, this section is for senior management who do not need to know the details. They need to know high level (the purpose of the model, limitations of the model and any issues).

I. Introduction (5 points)

This section needs to introduce the reader to the problem to be resolved, the purpose, and the scope of the statistical testing applied. What you are doing with your prediction? What is the purpose of the model? What methods were trained on the data, how large is the test sample, and how did you build the model?

I. Description of the data and quality (15 points)

Here you need to review your data, the statistical test applied to understand the predictors and the response and how are they correlated. Extensive graph analysis is recommended. Is the data continuous, or categorical, do any transformation needed? Do you need dummies?

III. Model Development Process (15 points)

Build an appropriate model to predict probability of survival. And of course, create the train data set which contains 70% of the data and use set.seed (1023). The remaining 30% will be your test data set. Investigate the data and combine the level of categorical variables if needed and drop variables. For example, you can passenger name, cabin, etc..

IV. Model Performance Testing (15 points)

Use the test data set to assess the model performances. Here, build the best model by using appropriate selection method. You may compare the performance of the best two logistic or other classification model selected. Apply remedy measures as applicable (transformation, etc.) that helps satisfy the assumptions of your particular model. Deeply investigate unequal variances and multicollinearity if warranted.

V. Challenger Models (15 points)

Build an alternative model based on one of the following approaches to predict survival as applicable:logistic regression, decision tree, NN, or SVM, Poisson regression or negative binomial. Check the applicable model assumptions. Apply in-sample and out-of-sample testing, back testing and review the comparative goodness of fit of the candidate models. Describe step by step your procedure to get to the best model and why you believe it is fit for purpose.

VI. Model Limitation and Assumptions (15 points)

Based on the performances on both train and test data sets, determine your primary (champion) model and the other model which would be your benchmark model. Validate your models using the test sample. Do the residuals look normal? Does it matter given your technique? How is the prediction performance using Pseudo R^2, SSE, RMSE? Benchmark the model against alternatives. How good is the relative fit? Are there any serious violations of the logistic model assumptions? Has the model had issues or limitations that the user must know? (Which assumptions are needed to support the Champion model?)

VII. Ongoing Model Monitoring Plan (5 points)

How would you picture the model needing to be monitored, which quantitative thresholds and triggers would you set to decide when the model needs to be replaced? What are the assumptions that the model must comply with for its continuous use?

VIII. Conclusion (5 points)

Summarize your results here. What is the best model for the data and why?

Bibliography (7 points)

Please include all references, articles and papers in this section.

Appendix (3 points)

Please add any additional supporting graphs, plots and data analysis.