Problem

With the Titanic dataset, we would like to create a model that takes related variables into account to predict the survival possibility. These variables are personal information that associates with passengers. The model has to be as accurate as possible with the combination of highly suitable variables or interactions.

Significance

The problem is interesting because we might be able to predict if a person is able to survive the accident according to the sex, ticket class, port of embarkation, the number of family members, and age. For example, will a person that is an elderly woman, that has family members aboard the Titanic and bought a first-class ticket survive the tragedy? I assume this combination because I personally think rich, women, and elderly are the words that are associated with privilege during an incident.

Literature

- According to Predictors of death in severe COVID-19 patients at millennium COVID-19 care center in Ethiopia: a case-control study, a binary logistic regression model was used to predict death in severe COVID-19 Patients. The model integrates diabetes mellitus, fever, and shortness of breath as predictors of death. Diabetes and shortness of breath are the main predictors in this model.

- According to A systematic review of statistical models and outcomes of predicting fatal and serious injury crashes from driver crash and offense history data, binary logistic regression is also used to predict the fetal and serious injury of a car crash. Some of the models integrate crash history and offense history as predictors. People with a crash history and offense history are more likely to be predicted to be in fatal and serious injury crashes.

Data

I did find a few missing values. For missing values of age, I replace them with the median of Age. On the other hand, I replace the missing value for the Fare with the mean of the Fares. Other than that, I found that both test and train datasets have large amounts of missing value for the Cabin variable, so I remove that variable from both datasets. For missing values in the Embarked variable, I replace it with the most common value which is “S”.

After handling missing values, I remove some variables that have no impact on the regression model such as PassengerID, Name, Cabin, and Ticket.

Besides that, I changed the type of data for categorical data from integer and character to factor because those data were not converted to factor originally.

Lastly, I do think that family size might be an influential factor that will affect the accuracy of the model, so I create a family size variable by combining both SipSb and Parch.

Type of Models

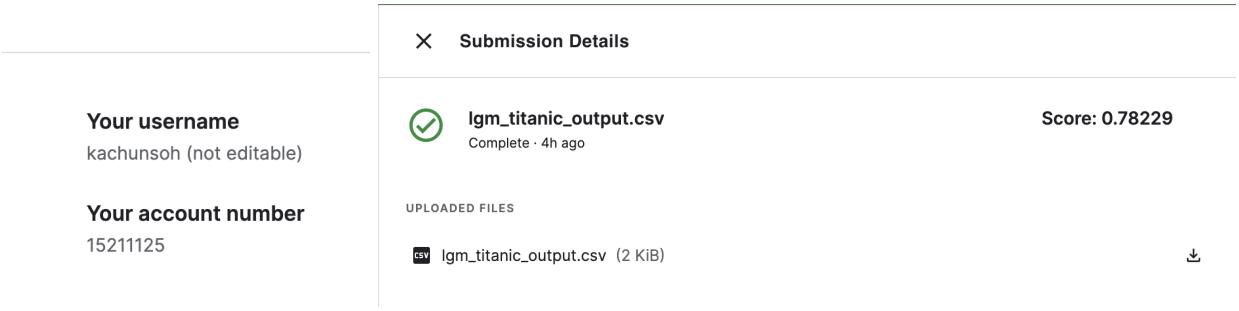
I built a binary classification model with logistic regression. In this problem, we have to predict the outcome of survival which is either survive or die. This model is considered as classifying or predicting whether that person will survive or die according to the predictors.

Formulation:

$\log(p/(1 - p)) = 4.166932 - 0.999507Pclass2 - 2.318650Pclass3 - 2.688409Sexmale - 0.036196Age - 0.237471\sqrt{family\_size} - 0.093867EmbarkedQ - 0.536743EmbarkedS$

Performance / Accuracy

The model performed on Kaggle did not lag too much behind. The accuracy rate on the train dataset is 0.8069585, while the accuracy for the test dataset on Kaggle is 0.78229. There's a 3.15% difference.



Limitations

My model requires no multicollinearity between independent variables. For example, the appearance of family size which is SipSb + Parch affects the SipSp variable and Parch variable as both variables do not influence the model significantly.

Besides that, my model overfitted the training dataset. The accuracy of the training dataset is higher than that of the test dataset. The model performs well with the training dataset but not with the test dataset. It does not predict accurate results for the test dataset.

Learning

Cleaning data is a crucial step for formulating a high-accuracy model. If missing data is not handled or data is not interpreted well. The accuracy of the model will never be good enough.

Other than that, there are many classification and prediction models that do better than logistic regression models. To find the best model is to test every model on the training dataset.

Do not assume variables that seem to be not related are useless. We should test the variables before we assume them.

I should factor family_size rather than leave it to be an integer. I ran out of submission but I will definitely do that next time.

Reference

- Leulseged, T. W., Maru, E. H., Hassen, I. S., Zewde, W. C., Chamiso, N. W., Abebe, D. S., Jagema, T. B., Banegyisa, A. B., Gezahegn, M. A., Tefera, O. S., Shiferaw, W. G., & Admasu, T. T. (2021). Predictors of death in severe COVID-19 patients at millennium COVID-19 care center in Ethiopia: a case-control study. *The Pan African medical journal*, *38*, 351. https://doi.org/10.11604/pamj.2021.38.351.28831
- Slikboer, R., Muir, S. D., Silva, S. S. M., & Meyer, D. (2020). A systematic review of statistical models and outcomes of predicting fatal and serious injury crashes from driver crash and offense history data. *Systematic reviews*, *9*(1), 220. https://doi.org/10.1186/s13643-020-01475-7