- Introduction

Titanic accidence is very famous in the world. It has caused many death.1502 in 2224 passengers or crew were killed. Many reasons may cause this death number, numbers, the lifeboat number, and their gender. In this report, I will try to analysis this data on how they will affect the death number.

- Titanic Dataset concludes 891 passengers' information and 418 test data, including their name, age, gender, and their cabin. Some of them are not important, some of them are missing and become useless. In this report, we only focus on some useful data to make the prediction as accurate as possible.
- Feature Extraction

Firstly we need to figure out the features we use to predict the survival. Though all the fields of data are listed above, we are not using all of them. Instead, we pick those which, we think, are crucial for prediction of the survival.

In this case, the fields that are determined to be features are “Sex”, “Age”, “Pclass”, “Cabin” and “Embarked”. “PassengerId” is the identification of a certain data vector, and “Survived” is the result we use to train and also the result we what to predict.

- Data Cleansing
  - Age

    The Age feature is missing 20% of its value. Since Age is an important feature to predict the survival, my first attempt is to fill all the missing data with random number from 5 to 70.

  2. Cabin

    The Cabin feature is missing 77% of its value. Apparently this feature cannot be utilized for analysis. Therefore I exclude Cabin from features.

  3. Embarked

    The Embarked feature is missing 0.2%, which is harmless if we fill with random value.

- Dataset Division

There are total 891 passengers (i.e. data vector). I divided the dataset into standard Training and Testing sets, and the scale is 7:3.

- First Model Result

Since this is a classic binary classification problem, I first use Logistic Regression as my model. I simply import the model from sklearn.linear_model, and the test result yields 59.19%.

- Second Model Result

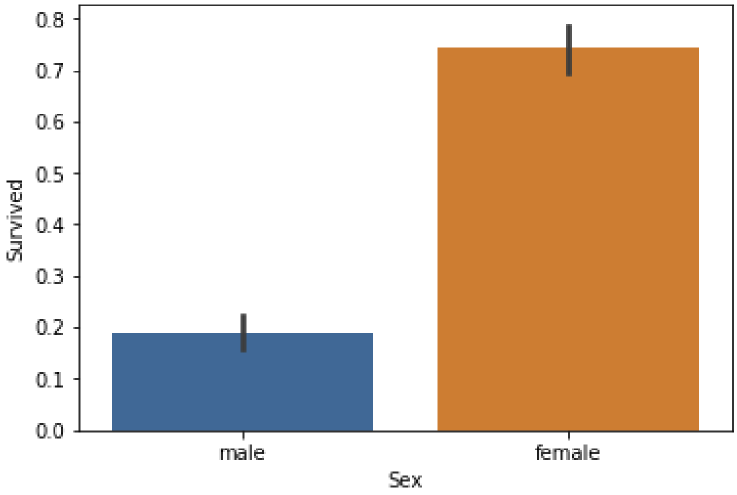
Perhaps Logistic Regression is not one of the best choices of our model, thus I switched to Support Vector Machine. I import SVC from sklearn.svm, and the result turns out to be much better, which is 76.73%.

- Reconsideration at Data Cleaning

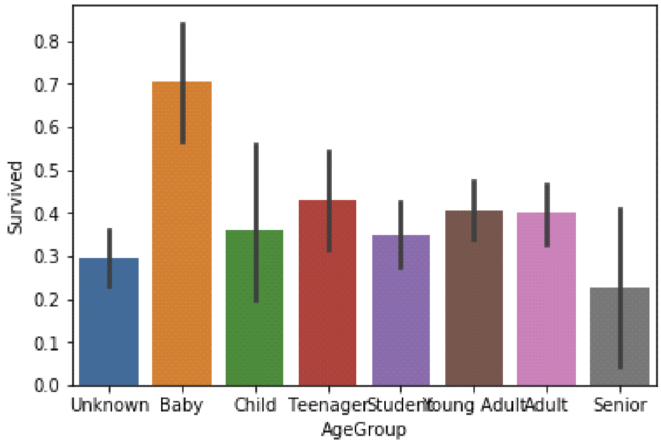
Although the accuracy is pretty high, it does not meet the strandard 90% expectation of a typical data scientist. Therefore I go back and reconsider the way I cleanse the data. Age is obviously an important feature, and I just random the missing value. Probably this way is not ideal. Thus I generate the missing value based on the distribution of the existing data. I test on SVM since our first run shows that SVM yields a better result, and the accuracy reached 83.73% this time.

- Inductive Bias

Inductive Bias is a certain factor we need to consider when we are training models. Analyzing the data and we can easily find out that female has a significantly higher survival rate than male.



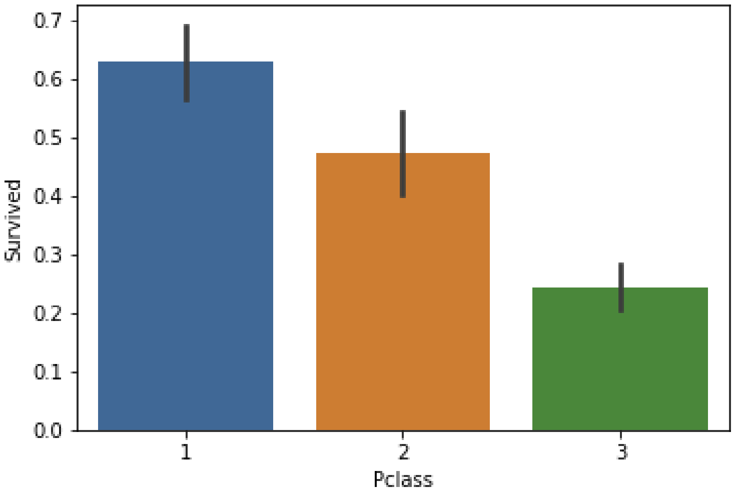
Moreover, if we divide the age evenly, we can see that people less than 18 has higher survival rate than people who are above.



We can easily predict all the above phenomenon since we have all heard the story: let women and children get on the lifeboat first. As a data scientist, this is clearly an Inductive bias if we want to generalize the model for future prediction. What people do might be different in a same voyage full of

people from a totally different culture or less civilized.

Also, survival ratio increases as pClass rises could be another inductive bias. This might caused by the design of Titanic, could be the position of those passengers on the boat or the lifeboats might be near first class area.



- Final Conclusion

In this research I mostly focusing on manipulating data, since the classification task is relatively easily: a simple, classic binary classification. Also the quantity of features is small, so I mainly focused on generating useful data for efficient training.

Inductive bias is a considerable factor in this model, thus the model cannot be generalized for prediction of events like this.

I tried to push the accuracy higher, but after trying several models I realized that it might not be the problems with data or model. This event was an accident after all, and there were so many factors to determine a person live or not in that situation. A locked door, a slippery deck or many other factors could result in the death of a person. Therefore I consider this dataset a nice one to practice, but not one for generalization or further investigation.

Acknowledgements

This report was done by Jim Cao, and his partner.

References

[1] Kaggle.com. “Titanic: Machine Learning from Disaster.” Kaggle, 28 Sept. 2012, www.kaggle.com/c/titanic/data.