PREDICTIVE ANALYTICS

Northeastern University

ALY6020, FALL 2020

ASSIGNMENT 3

NUID: 001084592

CRN: 71717

SUBMITTED BY: VISHRUTA MAKHIJA

EMAIL ID: makhija.v@northeastern.edu

SUBMITTED TO: PROF. JUSTIN GROSZ

NORTHEASTERN UNIVERSITY, COLLEGE OF PROFESSIONAL STUDIES

PREDICTIVE ANALYTICS

This week we are working on the Titanic Dataset. The goal of this week's assignment is to implement predictive model to understand the features that determined the survival of a person on the Titanic.

DATA CLEANING

(Treated Missing value, Outliers and Categorical Data)

Initially we checked the data for missing values and observed that there were no missing values in the dataset. We then plotted a catplot for Sibling/Spouse Aboard and Parent/Children Aboard. From the catplot, we observed that people with a greater number of siblings or spouses have less probability of survival. Similar to the Spouse/Sibling feature, the Parent/Children feature has the same trend, that is, when the number of parent or children increase, the probability of survival decreases. Since these two features have similar trends, we have combined the two features into one, that is, Family_cnt making the dataset cleaner. After making Family_cnt column, we deleted the Spouse/Sibling Aboard feature and Parent/Children Aboard feature from our dataset. Then we converted the categorical variables such as Sex to numerical. We then created boxplots for the continuous data columns namely Family_cnt, Age and Fare. Outlier removal was done by taking the quantile range of less than 1% and more than 99% to remove the extreme outliers. First we removed outliers in the Fare column, followed by Family_cnt, after which we were left with 871 datapoints. The boxplots can be seen in the Appendix. Also, the Name column didn't provide any valuable insights, hence removed the name column as well.

We then plotted bar graphs to explore the data further. To check if the data was imbalanced, we calculated the percentage of the responses in the Survived column and also plotted a graph

representing the same. We did not observe extreme imbalance in the data, 61% of the rows had value as 0 and 38% of the rows had value as 1.

PREDICTIVE MODELS

After cleaning the data, we then created dummy variables for our categorical variables such as Pclass and Family_cnt and then combined them with the continuous variables having 871 data points and 13 columns. Before running the predictive models, we divided the dataset into train and test, with 30% of the data classified as testing and 70% as train dataset. We then performed standard scaling on the continuous variables Fare and Age. The concept behind Standard Scaler is that it transforms data such that the distribution has a mean value of 0 and standard deviation of 1. We then fitted our three predictive models namely Decision Tree, Gradient Boosting and Logistic Regression. First, we have used the DecisionTreeClassifier function for prediction using decision tree and fitted the model on our training dataset and then predicted it on the test dataset. Similarly, we used the GradientBoostingClassifier function to predict using Gradient Boost Model, followed by Logistic model using LogisticRegression function. We also calculated different performance metrices to compare the different predictive models.

We then summarized the above calculated performance metrices in the form of a table as shown in Figure 1, for easy understanding and comparison.

	Model	Train Score	Test Score	Recall	Precision
1	Gradient Boosting	89.655172	80.916031	64.150943	85.000000
2	Logistic Regression	81.444992	80.916031	68.867925	81.111111
0	Decision Tree	98.686371	76.717557	60.377358	77.108434

Figure 1: Performance Metrices of different Predictive models

From the performance metrices, we observed that train score is highest in case of Decision Tree having a value of 98%, followed by Gradient Boost Model with 89% and Logistic Model having the lowest accuracy score of 81% on train set. Whereas for the test dataset, Gradient Boosting and Logistic have same accuracy of 81% but the accuracy of decision tree falls drastically to 77%, which implies that the Decision Tree model is highly overfitted and will not perform well if new data is introduced. Gradient Boost model is also overfitted as compared to the Logistic model. The Recall values for all the three models is low whereas the precision value can be considered as fine. According to the table in Figure 1, we can say that the Gradient Boost model and the Logistic model are equally efficient at predicting the survival of a person.

We performed Hyperparameter tuning for the Gradient Boost model and Logistic model to improve the accuracy and reduce the overfitting. We used GridSearchCV function for hyperparameter tunning of the predictive models. GridSearchCV is hyperparameter tuning method that is performed to identify the optimum values for a given model. So, we performed GridSearchCV on both the logistic and Gradient boost model. The performance metrices for the hyperparameter tunned models can be seen in Figure 2.

	Model	Train Score	Test Score	Recall	Precision	Score Tunned Train	Score Tunned Test	Recall Tunned	Precision tunned
1	Gradient Boosting	89.655172	80.916031	64.150943	85.000000	85.714286	82.442748	66.981132	86.585366
2	Logistic Regression	81.444992	80.916031	68.867925	81.111111	82.266010	81.679389	68.867925	82.954545
0	Decision Tree	98.686371	77.099237	61.320755	77.380952	0.000000	0.000000	0.000000	0.000000

Figure 2: Performance Metrices of different Predictive models with Hyperparameter Tunning.

In Figure 2, we have mentioned all the performance metrices for easy comparison. We can see that the accuracy of Gradient Boost model increased to 86% and 82% for both train as well as test dataset respectively after hyperparameter tunning whereas the accuracy of the Logistic model

remains almost the same. Moreover, the Gradient Boost model is no longer overfitted after hyperparameter tunning. From the above observations we can conclude that Gradient boost model is the best model for survival prediction having an accuracy of 82%.

CONCLUSION

From exploratory data analysis, we observed that Sex is an important determinant whether a person dies or survives. We plotted a bar graph (attached in the appendix), where we can clearly see that the survival rate of males is comparatively less than the females. Also, the class is a huge factor in determination of the survival. Pclass 1 and Pclass 2 seem to have almost same number of people, but the survival rate of Pclass 1 is higher than Pclass 2 and Pclass 3. Moreover, Pclass 3 has the highest number of people in comparison to the other classes and almost thrice the number of people in Pclass 3 die to the number of people who survive. This shows that the Pclass 1 and Pclass 2 are given the priority to be rescued, which implies that they are categorized as the elite or privileged class on the ship, whereas Pclass 3 is considered the lowest class on the ship. When combined together, the Pclass feature and the Gender, it is evident that males have lower survival rate even in Class 1. So, based on this fact, we can conclude that Females from Pclass 1 have a very high probability of survival as compared to any other person on the ship and also the females are given priority over males irrespective of the Class they belong to. In addition to Pclass and Gender, Family count also plays an important role in determining the survival. People having family members between 2 to 4 have higher probability of surviving than those who are alone or single or have a family size of more than 4. People who are alone on the ship tend to have the highest number of non survivors on the ship.

REFERENCES

StandardScaler?, C., Vu, T., & Petraglia, R. (2020). Can anyone explain me StandardScaler?. Retrieved 13 October 2020, from https://stackoverflow.com/questions/40758562/can-anyone-explain-me-standardscaler

An Introduction to Grid Search CV | What is Grid Search. (2020). Retrieved 13 October 2020, from https://www.mygreatlearning.com/blog/gridsearchcv/

APPENDIX

DATA CLEANING

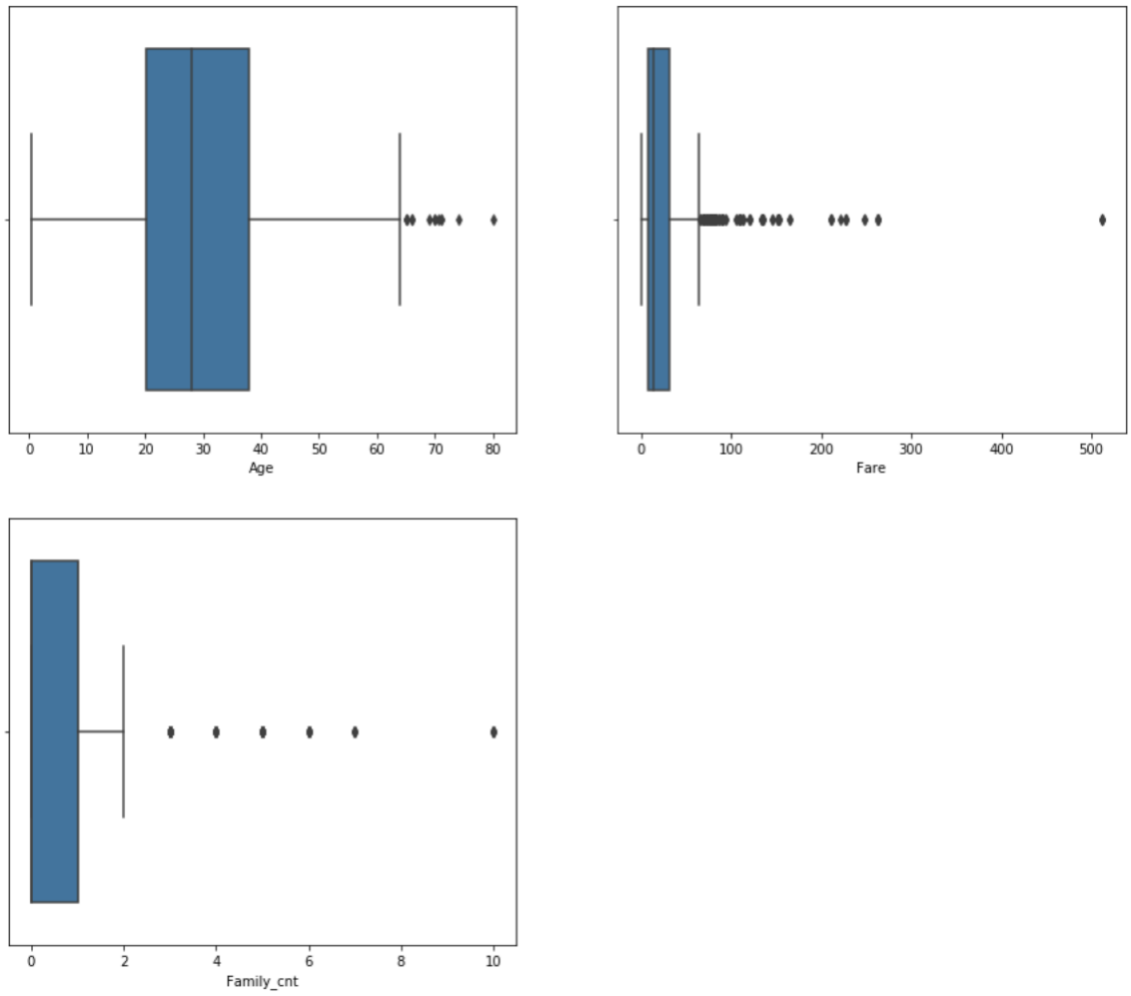

Figure 1: Boxplot visualization for outliers in continuous data.

```
(871, 6)
<matplotlib.axes._subplots.AxesSubplot at 0x7ff40b542650>
```

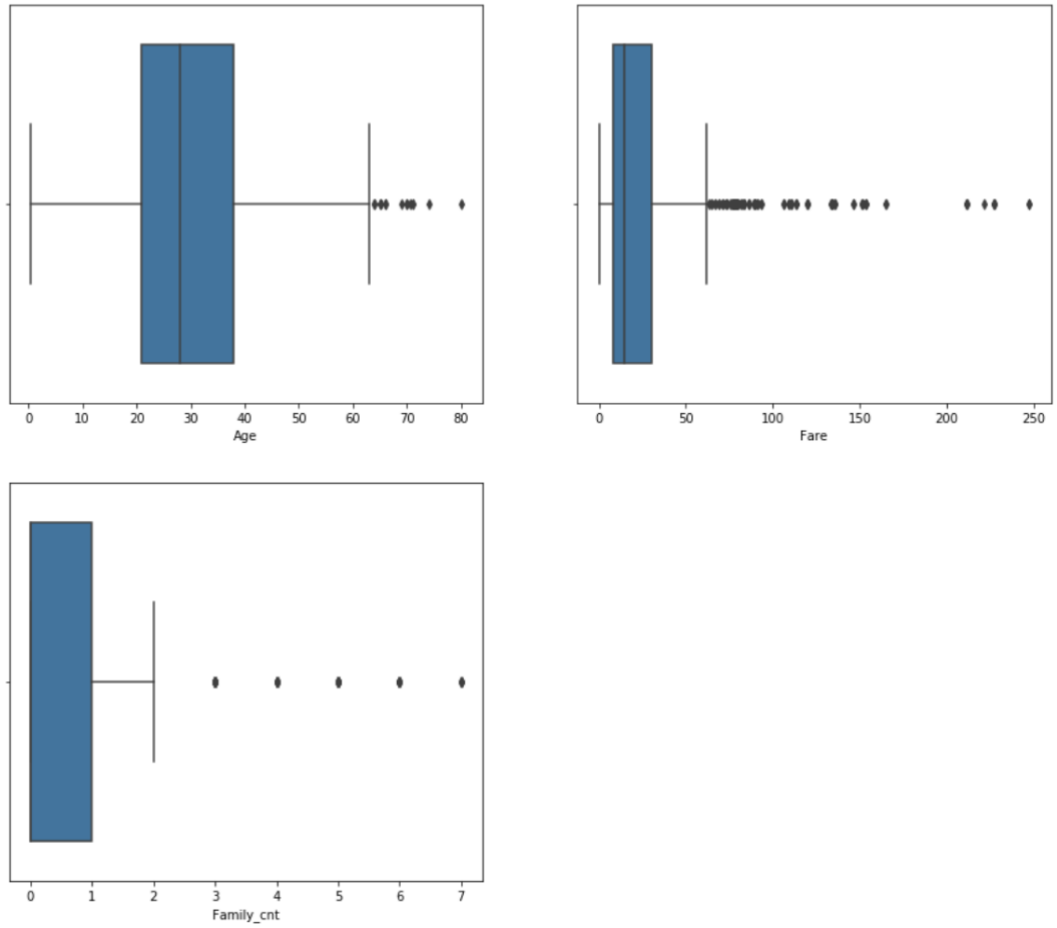

Figure 2: Final Data shape and Boxplot visualization for outliers in continuous data after treating all the Outliers.

EXPLORATORY DATA ANALYSIS

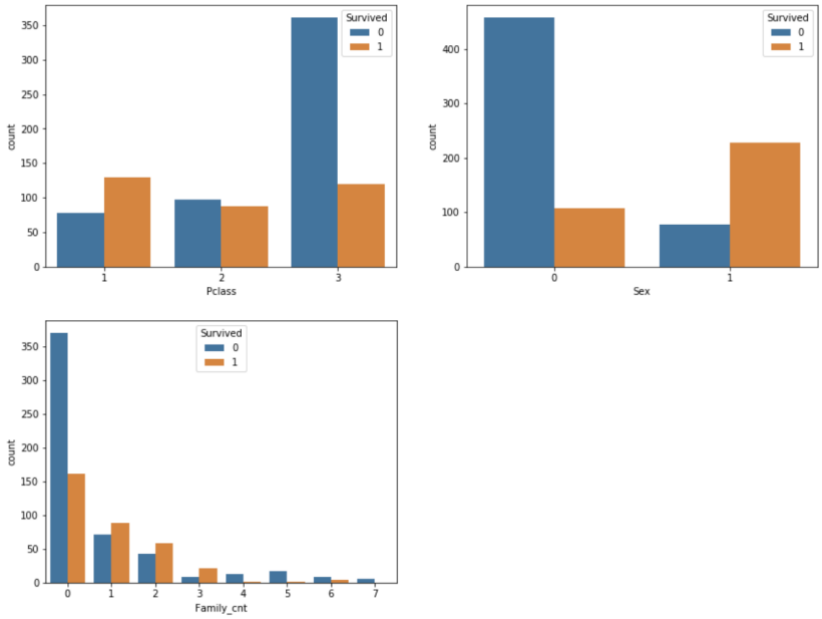

Figure 3: Bar graphs representing Response with respect to other variables.

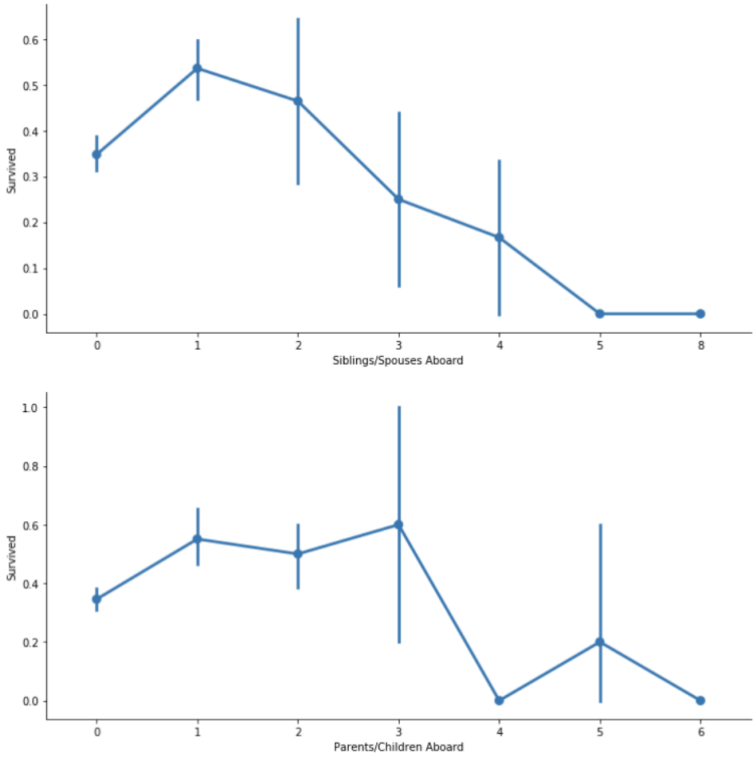

Figure 4: Plots representing the similar trend of Spouse/Sibling Aboard and Parent/Children Aboard Feature.

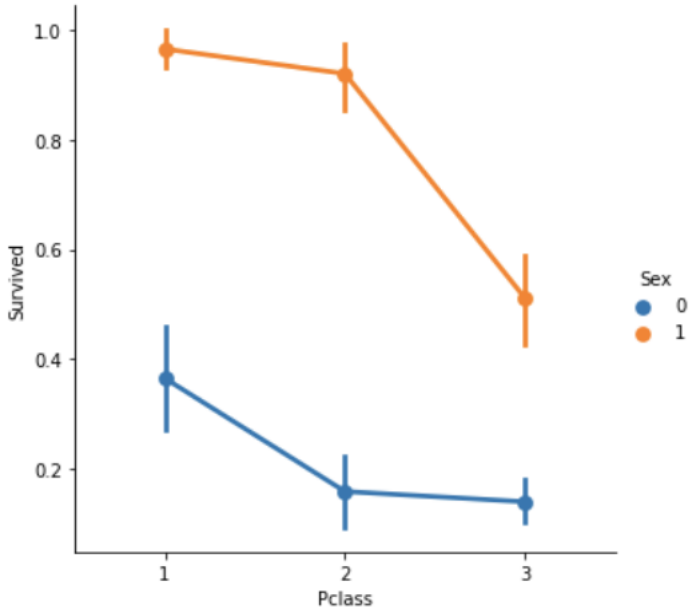

Figure 5: Graphical representation of Survival with respect to Pclass for both males and females.