Alice Tang Final Project Code: https://rpubs.com/aliciatang/874677

Problem and Significance

This Kaggle competition wanted participants to take a deeper look into the tragic shipwreck of 1912— the Titanic. Though luck was involved to survive this disaster, it is argued that certain factors contributed to a passenger's survival. Participants were tasked with creating a predictive model that would determine which types of people were more likely to survive by using passenger data. This problem is worth investigating since it is interesting to see how different factors affected the survival rate.

Descriptive Statistics and Cleaning the Data

As we can see in the the descriptive statistics, there are missing values in *Age*, *Fare, Embarked*, and *Cabin*. This is where data cleaning is incredibly important, and has to be done before creating the predictive model. Because there are 1014 missing values in *Cabin* it would be impractical to fill them all in, so I decided to drop the *Cabin* variable. For *Embarked,* I chose to use the mode for the 2 missing values— in this case the mode was "S". Whereas for *Age* and *Fare*, I decided to use the median to fill in missing values.

Next, I had to convert the columns (*Pclass, Sex, Embarked*) into categorical variables by using the factor() function and assigned it back in the data. I did this because the variables were integers, but needed to be categorical variables. I did not do *Survived* with the other variables because there were three categories in *Survived*. I do *Survived* after the dataset is split, so the N/A (missing values) do not appear as apart of the *Survived* category.

Data Visualization

From the graphs we are able to see how few people survived, and how more females survived than males. The graphs also allow up to infer the survival rate was higher for 1st class passengers. Lastly, we can see that younger passengers (though there are outliers) were much more likely to survive.

Type of Model and Implementation in R

Though we did not go over the random forest model in class, I stumbled upon this model when doing my research on this problem. I thought the random forest model was fitting because when looking at the Titanic data, there were a few outliers. I discovered that random forest models are robust to outliers and get averaged out, so I was curious to try it over a logistic regression method.

To use this model, I installed the package "randomForest". I then specifically chose the columns I wanted to build my predictive model out of. I chose PClass, Sex, Age, SibSp, Parch, Fare, and Embarked to use in my formula. My Survived formula was "Survived ~ Pclass + Sex + Age + SibSp + Parch + Fare + Embarked". After building these set of relationships, I applied it

and ran a prediction using the predict() function on the test dataset. Before submitting, I had to create a dataframe as well to write it out as a CSV for Kaggle.

Related Peer Reviewed Journals

Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer's Disease

In an Alzheimer's Disease study, they decided to use Random Forest (RF) since it produced the most optimal accuracy in the neurological diseases field. Similar to the reasons I chose RF for the Titanic dataset, these researchers used RF because it showed advantages over the other models because of its robust nature and ability to handle outliers. The researchers explained that they used RF because it is considered "more stable" (Sarica et al.) with outliers (which must occur frequently in their research) and with "high dimensional parameter spaces" (Sarica et al.) than other models.

Outlier Detection in OpenStreetMap Data using Random Forest Algorithim and Variable Contributions

OpenStreetMap (OSM) is an online source that allows volunteers to contribute data by creating "digitized geography objects with user assigned tags" (Wen and Rinner.). After experimenting with RF models on OSM data, they found that RF's outlier detection could lower search times for irregular objects, and make data patterns easier to interpret in OSM. RF was found to be beneficial and was used in similar situations as the others. RF models have proven to be a useful tool for its flexibility.

Performance and Limitations

My model only scored a 0.77272. I can think of a few reasons as to why and the limitations of my model. To obtain a better predictive model, the data could have been cleaned better. Using the median made accuracy go down because the median is different for genders and different for Pclasses. To improve this model, I could build a regression model to predict the missing values instead. I could have also cross-validated my data, which would have trained the model on all of the data. Fixing these things would make my model more accurate.

Key Learning Points

In this Kaggle Titanic Competition, I learned how essential data cleaning was to creating a better model. The majority of my time was spent on preparing the data and filling in missing values to create the best predictive model— unexpected to me. I also learned that there are many

great resources available to help and practice provided by R studio and Kaggle, which made the experience much less daunting.

Works Cited

Sarica, Alessia, et al. “Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer's Disease: A Systematic Review.” *Frontiers in Aging Neuroscience*, Frontiers, 6 Oct. 2017, https://www.frontiersin.org/articles/10.3389/fnagi.2017.00329/full.

Wen, Richard, and Claus Rinner. *Outlier Detection in OpenStreetMap Data Using the Random Forest Algorithm and Variable Contributions*. Sept. 2016, https://www.researchgate.net/publication/317099129_Outlier_Detection_in_OpenStreetMap_Data_using_the_Random_Forest_Algorithm_and_Variable_Contributions.