DS-GA 1001 Capstone Project: Debunking the Myth Regarding Ghost Month and Drowning

Group Name: ASE
Suei-Wen Chen, swc435@nyu.edu; Annika Kufrovich, amk9634@nyu.edu;
Elaine Zhang, qz691@nyu.edu
Date: Dec 22, 2021

Introduction and Data Description

In Taiwan and some parts of Asia, the seventh month of the Chinese calendar is known as the Ghost Month. During this month, ghosts are believed to enter our world from the underworld. On the fifteenth day of the month is the Ghost Festival, when sacrificial offerings are given to the ghosts and various traditions are observed. It is believed that people should not go swimming in open waters during the month because ghosts need to drown living people for reincarnation. It is also believed that people should avoid fishing because ghosts may take the shape of a fish to drown the fishers. This project is an attempt to debunk these traditional beliefs by analyzing the data on drowning rescue records from the Taiwanese fire departments[1]. The reprocessed dataset is stored in the file "DrowningData.csv", and contains 5588 records of rescuing drowning people from 2014 to 2020. Each entry has the following attributes:

- *City_or_County*: the municipality where the incident happened
- *Year*, *Month*, *Day*: the date of the incident (in CE)
- *CC_Year*, *CC_Month*, *CC_Day*: the date of the incident in Chinese Calendar[2]
- *Hour*, *Minute*: the time of the incident
- *Types_of_waters*: Ditch, Dock, Fish Pond, Lake, Offshore (<1km), Offshore (>1km), Others, Pond, Reservoir, River, Swimming Pool
- *Drowning_reasons*: Capsizing, Fishing, Floating Corpse, Other, Playing, Saving Others, Slipping, Snorkeling, Suicide, Traffic Accident, Work
- *Drowning_results*: Dead, Missing, Rescued
- *Gender*: Male, Female, Unknown
- *Age*: an integer value or Unknown
- *Swimming_skills*: Yes, No, Unknown

We aim to answer the following three questions using this dataset:

1. Are drownings more likely to happen or more fatal during the ghost month compared to other months?
2. What reasons for drowning (for example, fishing) are more prevalent in the ghost month than other months?
3. How well can a classification algorithm built with this data predict whether a certain drowning individual will be rescued?

Throughout this paper, the significance level is set to 0.005.

---

[1] The dataset used in this project can be found here: https://data.gov.tw/dataset/7065
[2] The dates in the Chinese Calendar are obtained using the Microsoft Excel function *TEXT(CEDateString, "[$-130000]yyyy-mm-dd")*. The file "DataConversion.py" then relabel the intercalary months (September in 2014, June in 2016 and April in 2020) as the 13th month of the years.

Question 1

Given the belief that drownings are more likely or more fatal during the ghost month, are drowning incidents actually more likely or more fatal during the ghost month, as compared to other months?

To answer this question we first performed a chi-squared test across all Chinese Calendar months. As a note, we re-coded incidents listed as "dead" or "missing" to the single label of "dead or missing" when performing the chi square test. We believe it's not too drastic to assume that people listed as missing are more likely dead than not and there are so few of these observations that it does not drastically affect our comparisons (difference in p value of < 0.00007 after recode). The chi-squared test, after re-coding, had a p-value of 0.00477, so we are 99.5% confident that the rate of fatality differs in at least 1 month. Table 1 shows the counts along with columns and row sums, and Table 2 shows the row-wise (monthly) percentage of drowning fatalities.

Table 1

Month	Dead/Miss	Rescued	Total
1	143	81	224
2	208	90	298
3	228	162	390
4	255	143	398
5	318	203	521
6	387	265	652
7 (ghost)	424	289	713
8	386	187	573
9	345	180	525
10	274	169	443
11	191	118	309
12	172	91	263
13	160	119	279
Total	3491	2097	5588

Table 2

Month	Dead/Miss	Rescued
1	63.84%	36.16%
2	69.79%	30.21%
3	58.46%	41.54%
4	64.07%	35.93%
5	61.04%	38.96%
6	59.36%	40.64%
7 (ghost)	59.47%	40.53%
8	67.36%	32.63%
9	65.71%	34.29%
10	61.85%	38.15%
11	61.81%	38.19%
12	65.39%	34.61%
13	57.35%	42.65%
Overall	62.47%	37.52%

Now that we have confirmed there is some difference across months in the rate of fatalities, a closer look at the table shows that the ghost month is one of the less fatal months. It is of note that the second and eighth months have the highest overall fatality rates of any month but that difference is not significant upon closer inspection with a t-test ($p > .2$). As for overall incidents, the ghost month clearly has the highest amount. However, this could be explained by ghost month occurring during the summer, when more people go to the waters. Comparing the

total incidents in the ghost month to total incidents in the sixth month (nearest summer month with closest incident rate) using a chi-squared test, shows that it is not a significant difference ($p > .09$), at least within season.

Question 2

Given the drowning data records in Taiwan, we are also interested in whether certain reasons for drowning are more prevalent in the ghost month as compared to other months. In other words, we want to test the difference between the proportions of each drowning reason in the ghost month versus other months.

To test our hypothesis, we chose to use Welch's t-test because it is reasonable to dichotomize and reduce the drowning reasons data to sample means (which equate to proportions), and, as shown in question 1, we have much less drownings in the ghost month than in all the other months combined, which may introduce unequal variances. We performed a test for each of the 11 drowning reasons as listed in Table 3 to see whether the reason is significant. In the end, we found that no reasons for drowning are significantly more prevalent in the ghost month at the 99.5% confidence level (as listed in *Table 3*). As a note, using the Bonferroni correction, the significance level was re-calculated to 0.005/11=0.00045, which is lower than any of the p-values that we got.

#	Drowning reason	Number of total data points	P-value of independent t-test
1	Work	227	0.354592
2	Suicide	1301	0.783954
3	Floating Corpse	1255	0.763625
4	Capsizing	170	0.999551
5	Slipping	536	0.12742
6	Traffic Accident	148	0.832147
7	Playing	580	0.121389
8	Snorkeling	131	0.331844
9	Saving Others	38	0.704759
10	Fishing	304	0.007326
11	Other	898	0.80895

Table 3: p-values of hypothesis testing from different drowning reasons

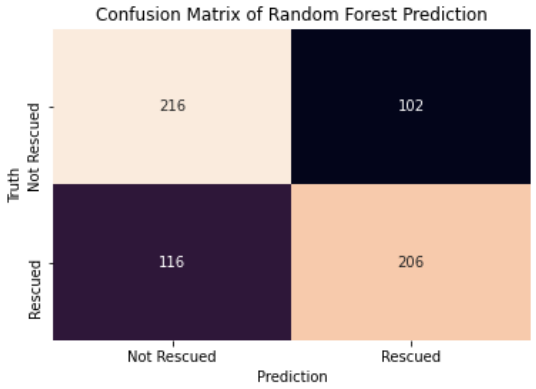
Question 3

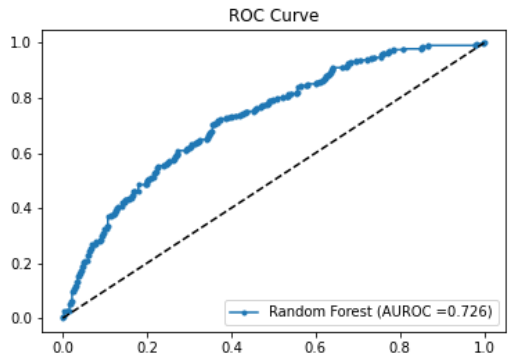
The goal is to build a classification algorithm that predicts whether a given drowning individual, under a given set of circumstances, will be rescued or not. Because our dataset contains both numerical and categorical variables, we chose to use a random forest as the classification algorithm using the following predictors: *CC_Month*, *Hour*, *Gender*, *Age*, *Swimming_skills*, *Types_of_waters*, and *Drowning_reasons*. To prepare our data, entries with unknown gender or age, 'Others' in *Types_of_waters* and 'Others' or 'Floating Corpse' (which is necessarily dead) in *Drowning_reasons* were dropped, leaving 3196 rows. Next, categorical predictors were converted as follows:

- *Drowning_results*: Dead = 0, Missing = 0, Rescued = 1
- *Gender*: Male = 1, Female = 0
- *Swimming_skills*: No = -1, Unknown = 0, Yes = 1
- *Types_of_waters* and *Drowning_reasons*: labelled in alphabetical order

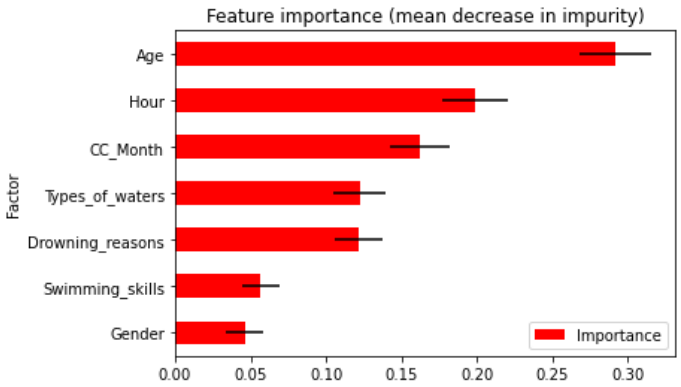
We kept entries with unknown Swimming_skills because they represent a large portion of our dataset (n = 2368). In order to keep things fair, we encoded those entries with 0 so that they would lie in the middle of "yes" and "no." For *Types_of_waters* and *Drowning_reasons*, we choose not to use one-hot encoding because it introduces issues of sparsity and dependence among columns.

With the data fully prepped, the dataset was split such that 80% of our data was used for training and the remaining 20% for testing. A grid search was performed to choose the parameters *max_features*, *min_samples_leaf* and *max_depth*, with 5-fold cross validation on the training set. Next, the whole training set was used to build a random forest with the chosen parameters {'max_depth': 13, 'max_features': 2, 'min_samples_leaf': 1, 'n_estimators': 500} and random_state=10. The accuracy of our algorithm on the training set is 86% while that on the test set is 65.9%, and the ROC AUC score is 0.726. The confusion matrix and the ROC curve are given below:





As far as reliability, if we regard "Not rescued" as positive, we see that the algorithm has sensitivity 65% and specificity 67%. When making predictions, the model outperforms random guessing (with an accuracy rate of 322/(318+322)=50.3%) by around 15.6%, which is better but not outstanding. Lastly, we inspected how much one average each feature affects the decision-making process of the trees in the random forest, using the mean decrease in Gini impurity (MDI):



The black bars represent the standard deviation of the MDI's. From the bar graph we see that the two most predictive factors as to whether a drowning person will be rescued, *Age* and *Hour*, account for around 48% of the total importance, whereas *CC_Month* only accounts for around 14.7%. Owing to the ensemble nature of random forests, however, how these factors affect the prediction process cannot be readily interpreted without further analyses.

It would be unwise to dismiss features with low importance in the model, such as swimming skills and gender, as insignificant predictors of rescue. This is because we dropped a considerable amount of data with *Drowning_reaons*='Floating Corpse', which necessarily means being dead. Dropping entries with unknown *Swimming_skills* (which leaves 1057 rows) and performing a Welch's t-test with the alternative hypothesis that those who know how to swim are more likely to be rescued, we found that the p-value is 0.0029, which is significant. Likewise, dropping entries with unknown *Gender* (leaving 5518 rows) and performing a Welch's t-test shows that women are more likely to be rescued than men, with a p-value of 2.7e-12, which is, again, significant.

Conclusion

Overall, it does not appear that the drownings in the ghost month are significantly more likely or more fatal. Rather, reasons such as swimming skills and age significantly correlate to these outcomes. Additionally, no reasons for drowning were more prevalent in the ghost month than in other months. Given our results, we do not believe there is any relationship between drownings and the ghost month, but there is room for further research on this topic.

Some of the limitations of our analysis include that not all drownings may be reported or recorded, potentially biasing our data. Our data is also entirely observational, and while there are a decent number of variables, we cannot comfortably determine causality for any of our questions without controlled intervention. We had quite a bit of missing data across several columns, making it difficult to fully utilize our data in creating the classification algorithm. This is especially the case for swimming skills, as it is a significant factor but does not account for much of the reduction of impurity in the random forest. Additionally, time series analysis methods, such as decomposition, might have been useful for our first two questions, so that seasonality and other dependency issues could be better measured and accounted for. Other than time series methods, some measure of temperature might also be useful to account for seasonal increases in swimmers. Finally, by virtue of the data, we were limited to drownings in

Taiwan. It might be useful to see if our findings are consistent across other regions in Asia where the ghost month is recognized.

Lastly, we think future work on this question, or using this data, might consider some other confounding variables beyond those mentioned above. For example, tides and weather conditions might be predictive of fatalities. In addition, while we dropped exact locations of the incidents from our dataset, they might be useful in determining where drowning incidents are most common. Finally, other latent factors such as the presence of dangerous currents might be inferable from these variables and usable in future analyses.