JULY 2021

TEXAS
The University of Texas at Austin

WHO SURVIVED IN THE TITANIC?

Was it a random group of people? Or were there any underlying factors which helped specific cohort of people to have a better chance at survival? We will find out!

GROUP 5: VISHU AGARWAL, ANKITA KUNDRA, BRUCE KIM, MARIO GONZALEZ, JESPER LI
MSBA, The University of Texas at Austin

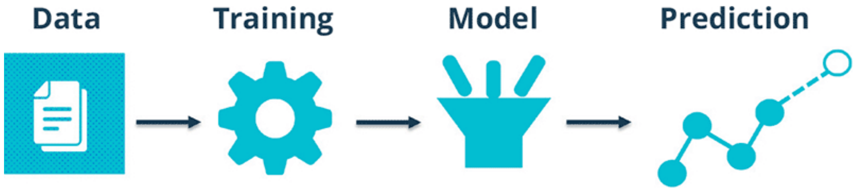
What are we trying to do?

- Everyone is aware of one of the deadliest peacetime maritime disasters, sinking of the RMS Titanic. Among the ~2200 passengers, only ~700 could survive

- At first, it would seem that the survival of passengers was just based on luck and nothing more. But what if there were some factors which played a vital role in determining the survival chances?

- **This is exactly what we will try to do in this project – determine the factors driving the survival rate of passengers on the Titanic!**

How will we do that?



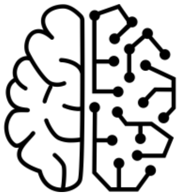
- We have sourced a dataset containing information about the passengers on the Titanic from Kaggle

- This dataset contains information like age, gender, fare, ticket class etc. We will analyze this information and build a model which can learn from the data to -
  - Provide insights into factors influential in determining the survival chances
  - Predict the survival chances of passengers in the test data

We will be using a two –fold approach for this analysis

Exploratory data analysis (EDA)

- Identify variables as categorical vs. numerical
- Visualize distribution of both variables
- Visualize survival rates among both types of variables
- Treat variables with missing values



Modelling

- Classification tree
- Random forest
- Boosting
- Logistic Regression
- KNN

Exploratory Data Analysis

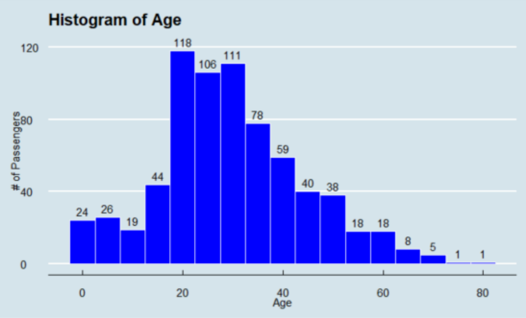
Variables (Predictors) present in the data have been identified as **Numerical** variables vs. **Categorical** variables

Numerical Variables	Description
Age	Age of passenger in years
SibSp	# of siblings/spouses aboard the Titanic
Parch	# of parents/children aboard the Titanic
Fare	Passenger fare price

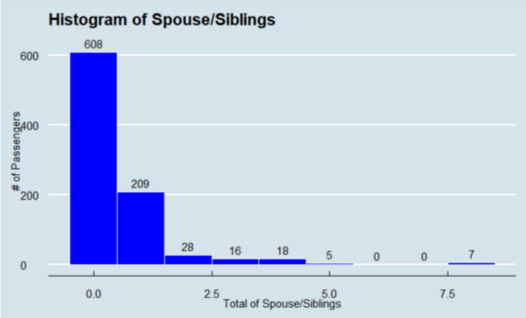
Categorical Variable	Description
Survived	Survival (0 = No; 1 = Yes)
Pclass	Ticket class (1 = 1st; 2 = 2nd; 3 = 3rd)
Sex	Gender (Male; Female)
Cabin	Cabin number
Embarked	Port of Embarkation

Visualizing Numerical variables

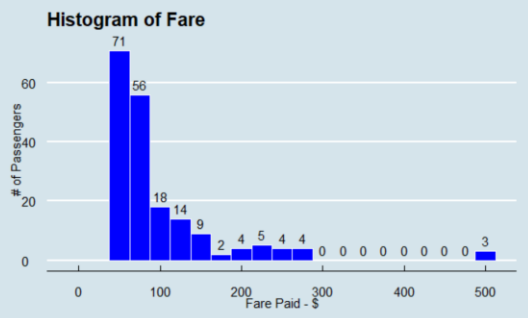
Age Distribution



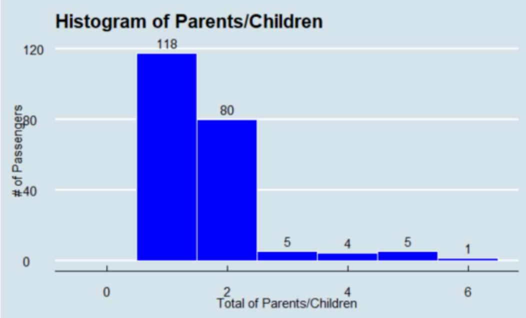
Siblings / Spouse Distribution



Fare Distribution

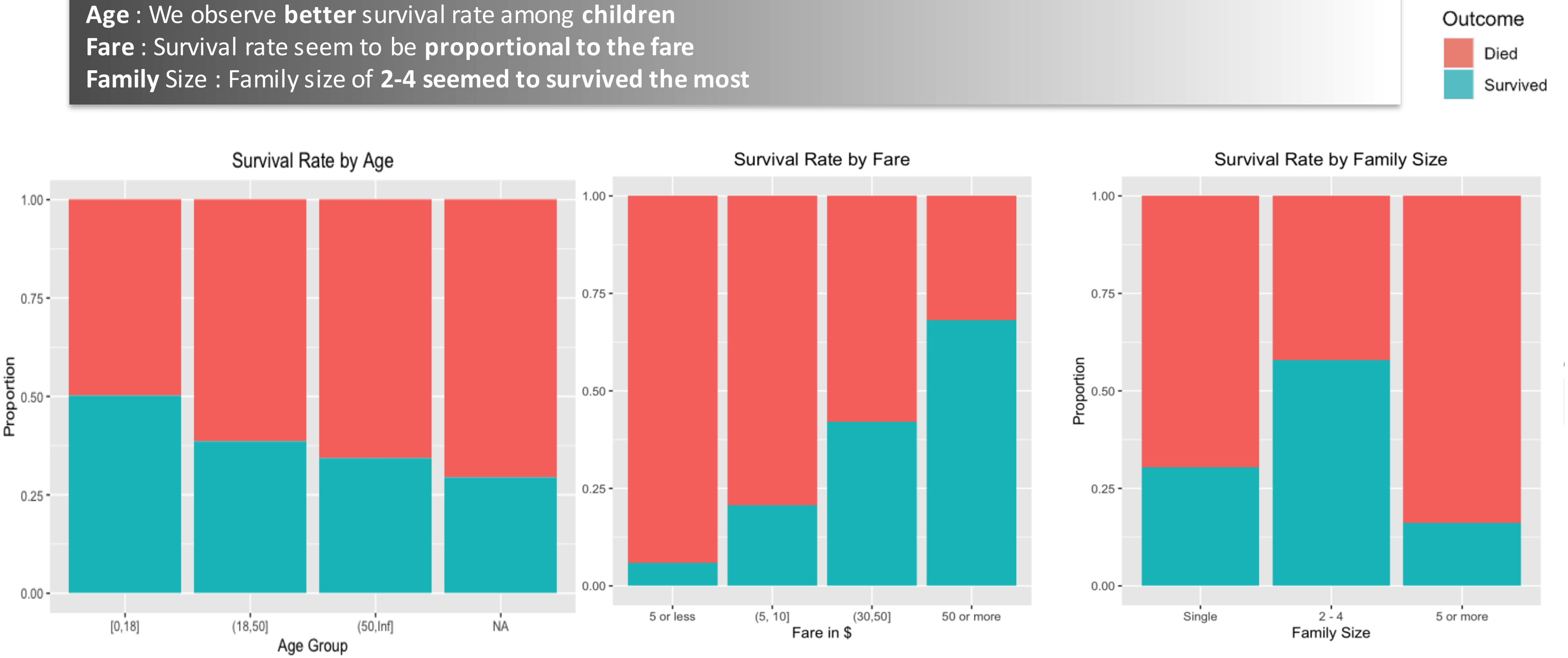


Parents / Children Distribution



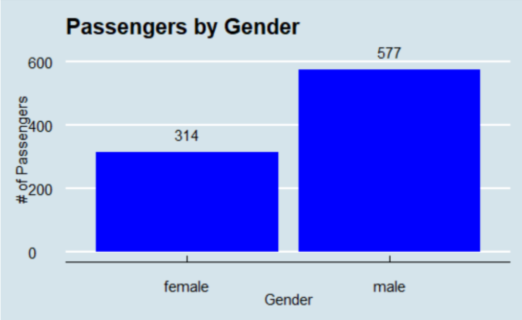
Survival rate by Numerical variables

Age : We observe **better** survival rate among **children**
Fare : Survival rate seem to be **proportional to the fare**
Family Size : Family size of **2-4 seemed to survived the most**

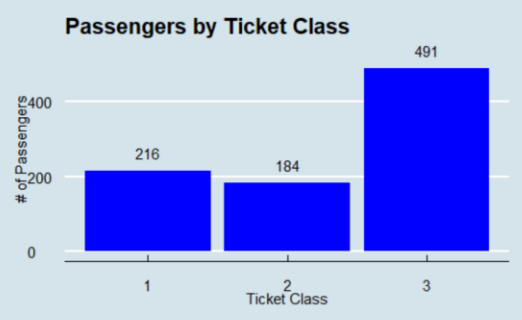


Visualizing Categorical variables

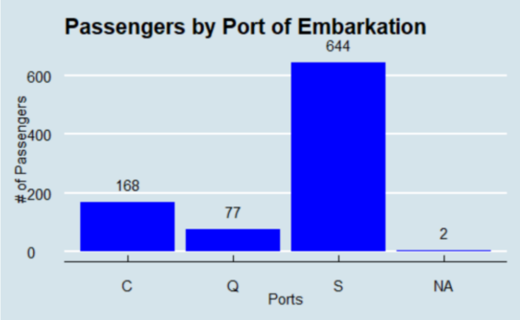
Gender Distribution



Ticket Class Distribution

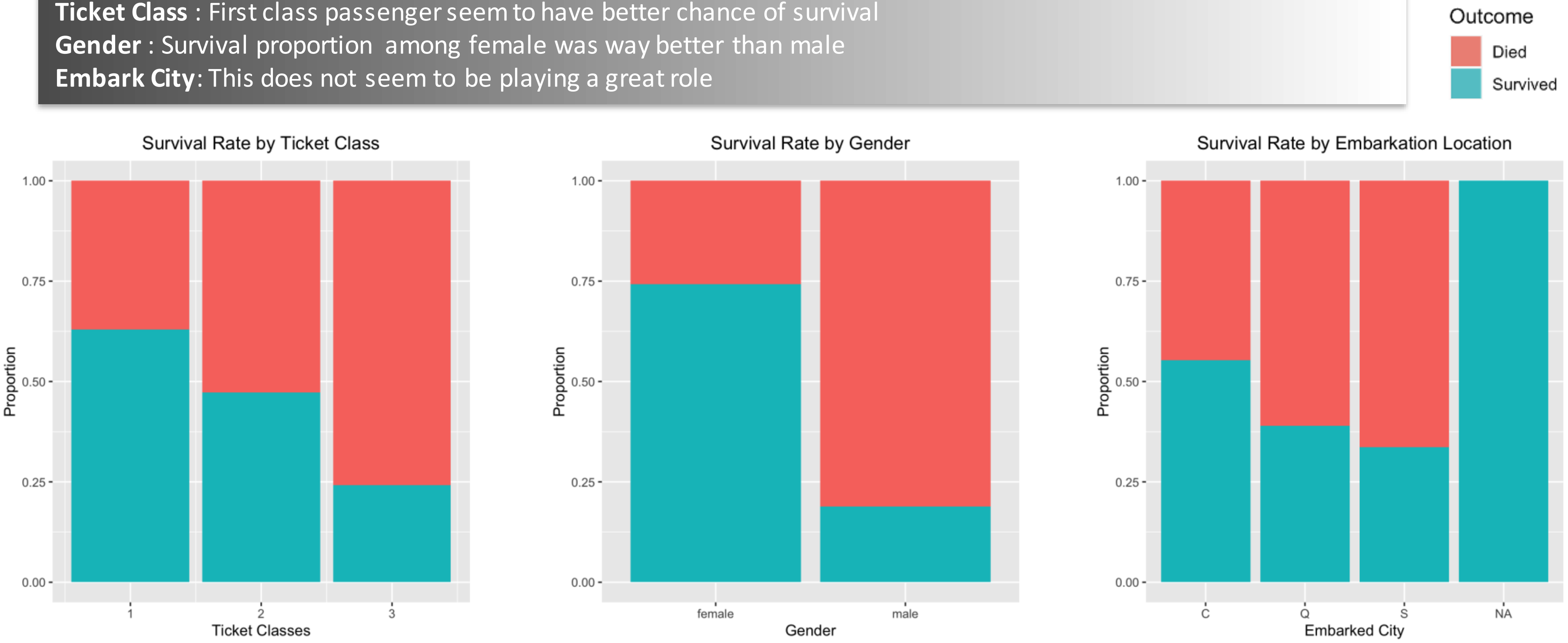


Embarkation Distribution



Survival rate by Categorical variables

Ticket Class : First class passenger seem to have better chance of survival
Gender : Survival proportion among female was way better than male
Embark City: This does not seem to be playing a great role



Treating variables with missing values

Predictors Age, Cabin and Embarked have missing values. Below is a quick summary of how we imputed values in Age and Cabin column. Embarked does not seem to have much effect on survival, so we did not treat it

Age

- Values missing for **177 (out of 891) records**
- Mean Age = 26.7 years
- Standard Deviation (S.D.) = 14.5
- Imputed value = Mean $\pm$ S.D.

Hence, we assigned random values between the mean $\pm$ S.D. to missing values in Age column

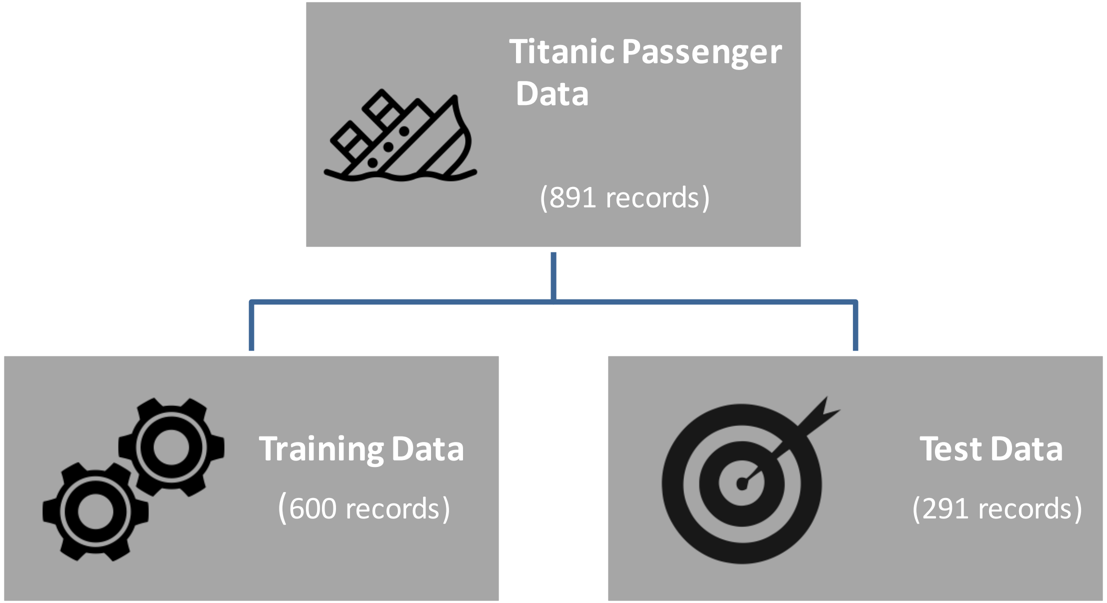
Cabin

- Values missing for **687 (out of 891) records**
- Contains cabin number for all passengers (example – A1, A2, C3, D13 ….)

Since >70% of records do not have Cabin number populated, we transformed this variable into a binary variable where 1 denotes that cabin information is present and vice versa

Modeling

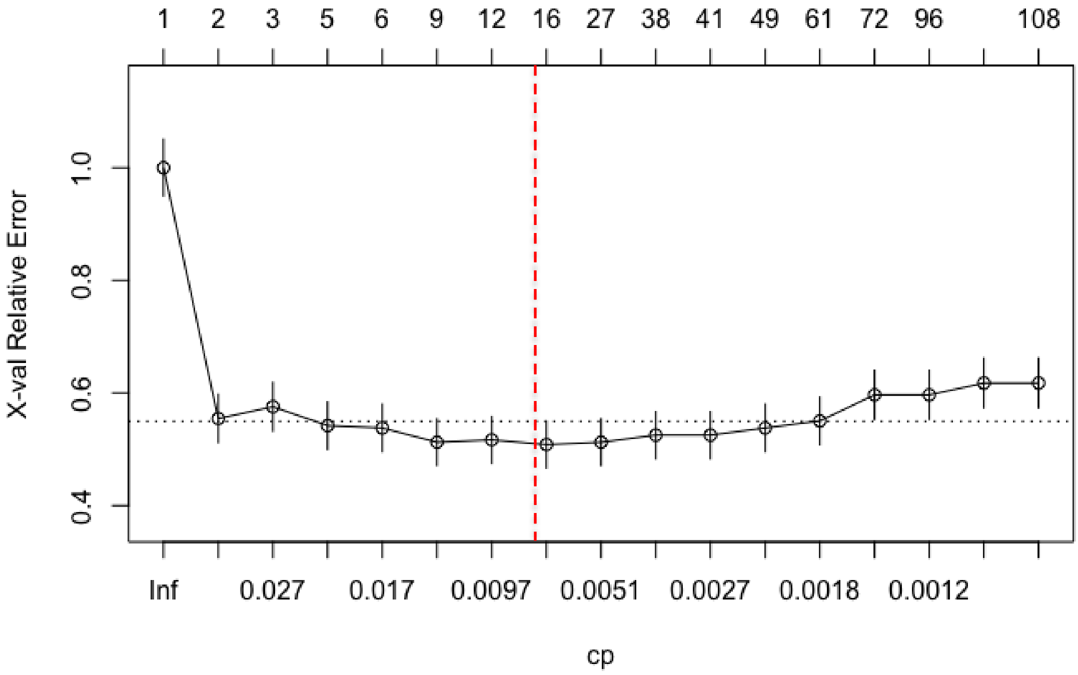
Preparing the data for modeling



Type	List of Variables included in model
Numerical	Age
	Fare
	Gender
Categorical	Class
	Siblings / Spouse
	Parents / Children

Library used : **rpart**

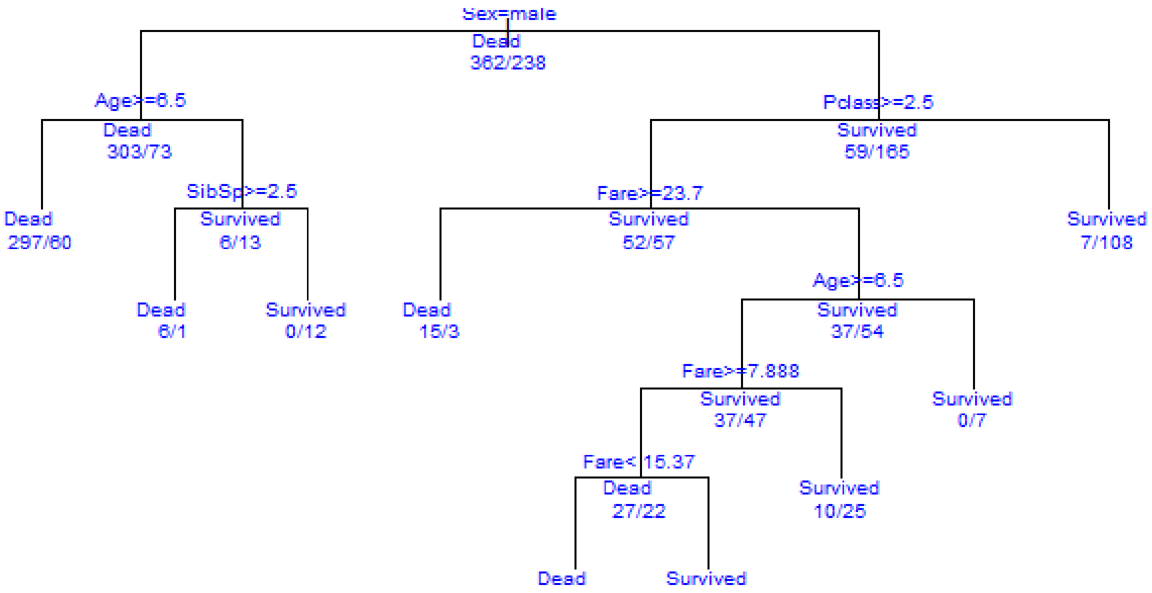
Classification Tree



- Aggressively build a complex tree with :
  - Minimum variables required to create a node (minsplit) - 4
  - Minimum complexity parameter (cp) – 0.0005
- This gives us a complex tree containing 108 nodes
- Plotting the errors vs. cp curve, we find that the minimum error corresponds to the tree with 15 nodes (cp = 0.006). Hence, we will perform pruning using this cp parameter

Library used : **rpart**

Classification Tree

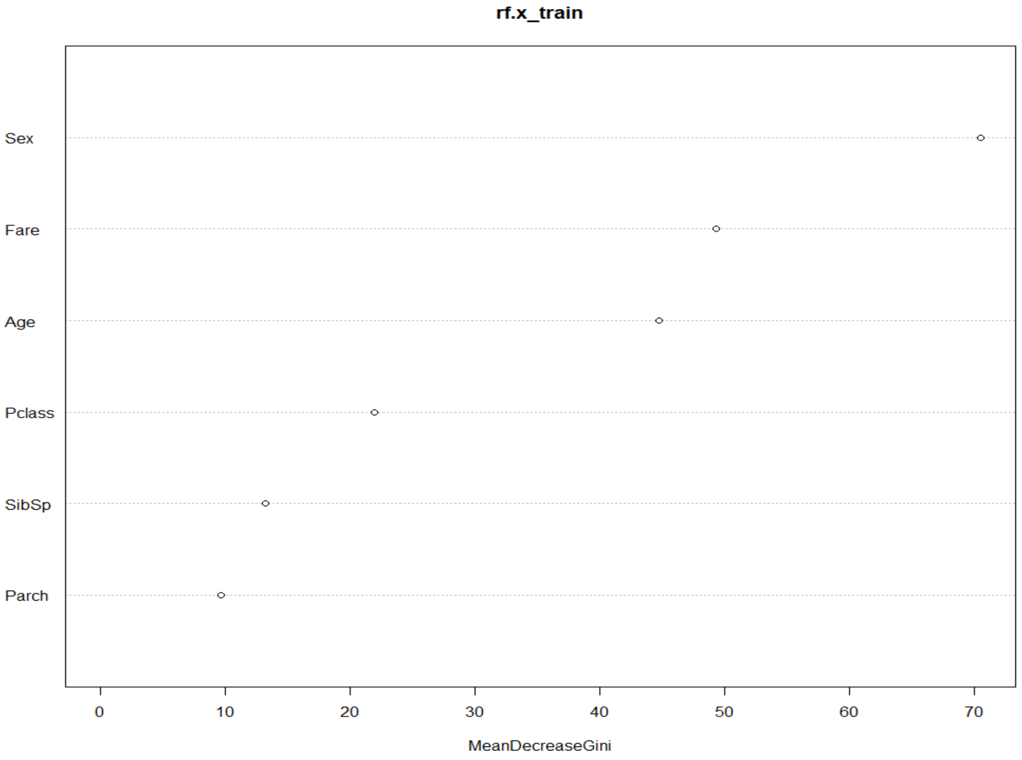


		Test Data	
		No	Yes
Predicted Data	No	177	35
	Yes	10	69
			Accuracy: 84.5%

The pruned tree correctly predicts that 177 people died and 69 people survived. The accuracy of this model is 84.5%

Library used : random forest

Random Forest

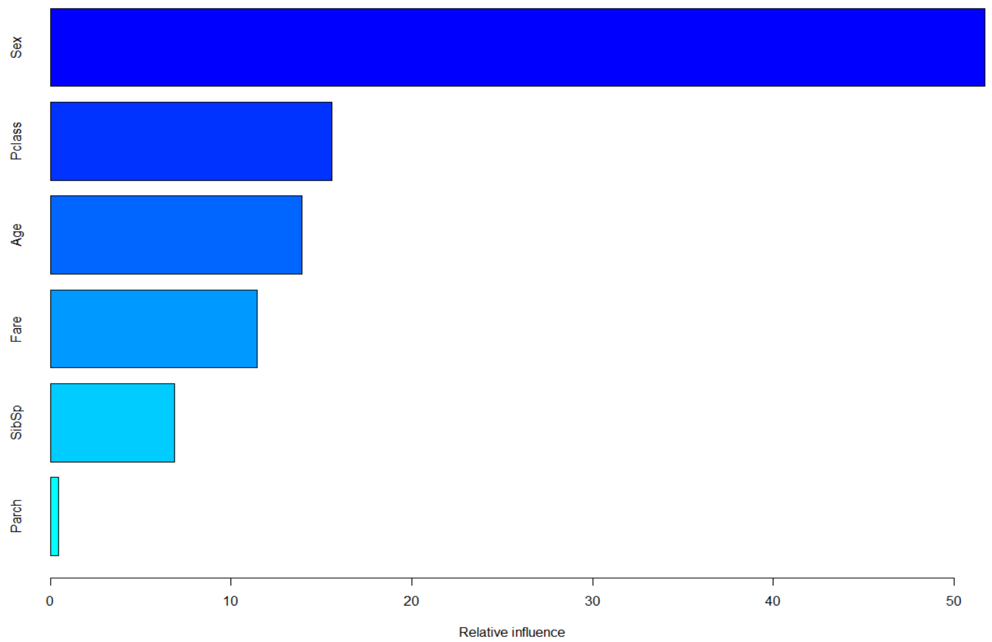


		Test Data	
		No	Yes
Predicted Data	No	174	39
	Yes	13	65
Accuracy: 82.1%			

- Number of variables chosen for each iteration is 2
- Model predicts that **Survival of the passengers is mostly influenced by Sex, Fare and Age**
- The random forest model correctly predicts that 176 people died and 68 people survived in the test data. The accuracy of this model is 82.1%

Library used : Boosting

Gradient Boosting



		Test Data	
		No	**Yes**
Predicted Data	**No**	180	43
	Yes	7	61

Accuracy: 82.8%

- Transformed the 'Survival' variable to 1s and 0s
- Number of trees = 1000 and shrinkage parameter = 0.01 was used to build the model
- Again, Gender has the highest influence on survival
- The boosting model correctly predicts that 180 people died and 61 people survived. The accuracy of this model is 82.8%

Library used : glm

Logistic Regression

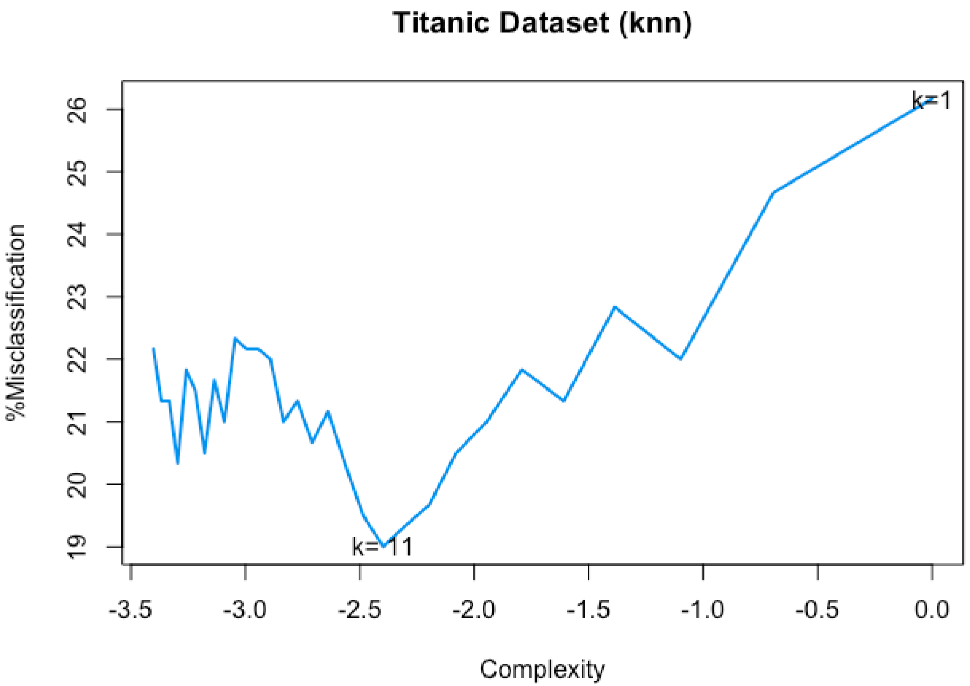
```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.967757   0.565618   3.479 0.000503 ***
Pclass      -1.083040   0.169695  -6.382 1.74e-10 ***
Sex          2.687319   0.239174  11.236  < 2e-16 ***
Age         -0.034188   0.008940  -3.824 0.000131 ***
SibSp       -0.405980   0.140352  -2.893 0.003821 **
Parch        0.008177   0.135383   0.060 0.951835
Fare         0.004005   0.002936   1.364 0.172594
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 805.96  on 599  degrees of freedom
Residual deviance: 537.86  on 593  degrees of freedom
AIC: 551.86
```

		Test Data	
		No	Yes
Predicted Data	No	178	40
	Yes	9	64
			Accuracy: 83.2%

- Categorical variables – Survival and Age were transformed to 1s and 0s
- The model summary suggests that the variables **Parch** and **Fare** are **not significant.** AIC reduces to 549.8 if we remove the Parch variable.
- The Logistic Regression model correctly predicts that 178 people died and 64 people survived in the test data. The accuracy of this model is 83.2%

Library used : kknn

k-Nearest Neighbors



		Test Data	
		No	Yes
Predicted Data	No	165	39
	Yes	22	65
			Accuracy: 79.1%

- The variables were scaled because KNN makes use of Euclidian Distance to calculate nearest neighbors
- The %Misclassification error is minimum for the test data when k value is 11
- The KNN model correctly predicts that 165 people died and 65 people survived in the test data. The accuracy of this model is 79.1%

Conclusion

- We get a similar story from all the models: Most influential variables:
  - **Sex**
  - **Fare**
  - **Age**
- We get the maximum accuracy of **84.5%** by using the classification tree model.

Model	Accuracy
Classification Tree	**84.5%**
Random Forest	82.1%
Gradient Boosting	82.8%
Logistic Regression	83.2%
KNN	79.1%