To Get Results Advertise Your Wants In the Daily Globe

The Boston Daily Globe.

Advertise Your April 19th Sales In The Daily Globe

VOL. LXXXI NO. 107. BOSTON, TUESDAY EVENING, APRIL 16, 1912—TWENTY PAGES. PRICE TWO CENTS.

EVENING EDITION—7:30 O'CLOCK

ALL DROWNED BUT 868

About 1232 Lost Lives in the Titanic's Plunge, Greatest Sea Disaster for Years.

EXCITING EVENTS BEFORE TITANIC'S FINAL PLUNGE

SCENE WHEN LINER TITANIC WAS PUTTING OFF BOATLOADS OF WOMEN AND CHILDREN

Virginian and Parisian Found None Alive.

Women and Children Safe But Few Notable Men.

Carpathia Has Survivors—On Way to New York

Only Partial List of Them Received Owing to Interruption.

BAY STATE PEOPLE SAVED

[illegible]

CHARLES M. HAYS SAVED.

President of Grand Trunk Railroad One of Few Men Rescued.

[illegible]

THROUGH FIELD OF ICE.

SOME WHO MAY BE SAVED

[illegible]

NEARLY ALL MEN LOST.

All the Boats Launched Have Been Accounted For by Carpathia.

[illegible]

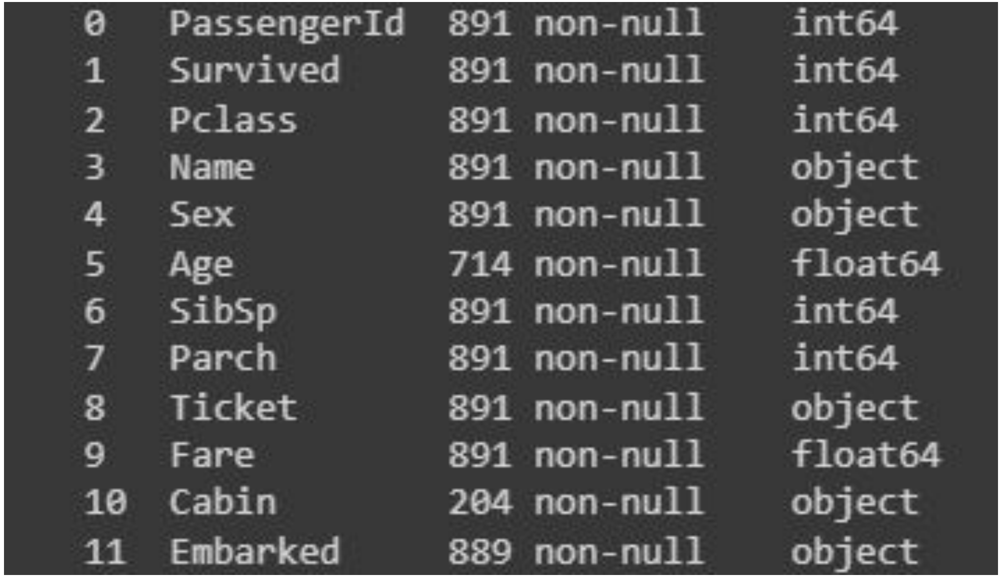
PREDICTING SURVIVAL ON THE TITANIC

GABRIELLA RIVERA
VIVIAN DO
JOEL DAY

OVERVIEW

Data Overview

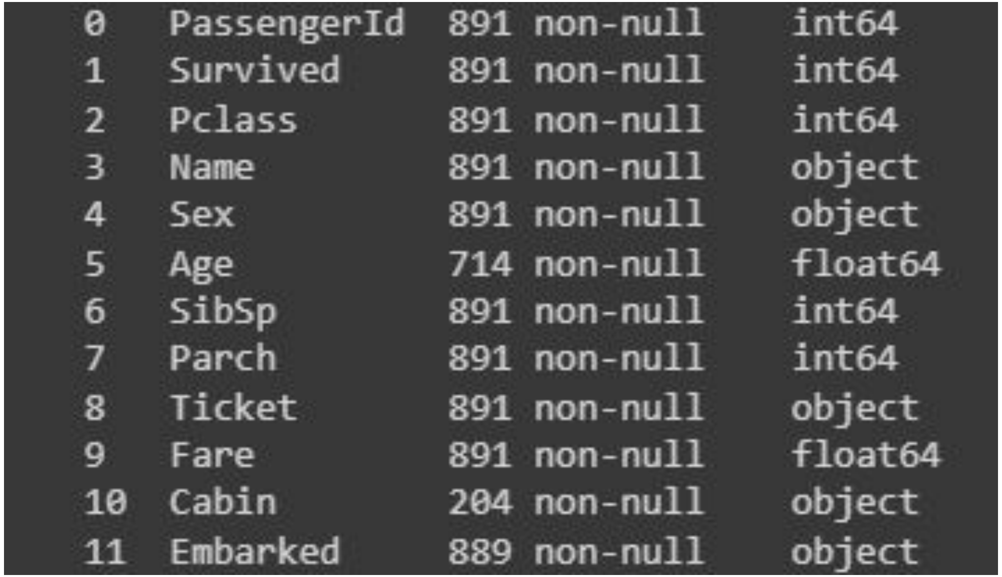
- ‘Titanic - Machine Learning from Disaster’ Kaggle competition
- Train.csv
- 891 Entries
- 12 Features



```
 0   PassengerId  891 non-null    int64
 1   Survived     891 non-null    int64
 2   Pclass       891 non-null    int64
 3   Name         891 non-null    object
 4   Sex          891 non-null    object
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64
 7   Parch        891 non-null    int64
 8   Ticket       891 non-null    object
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object
 11  Embarked     889 non-null    object
```

Data Overview

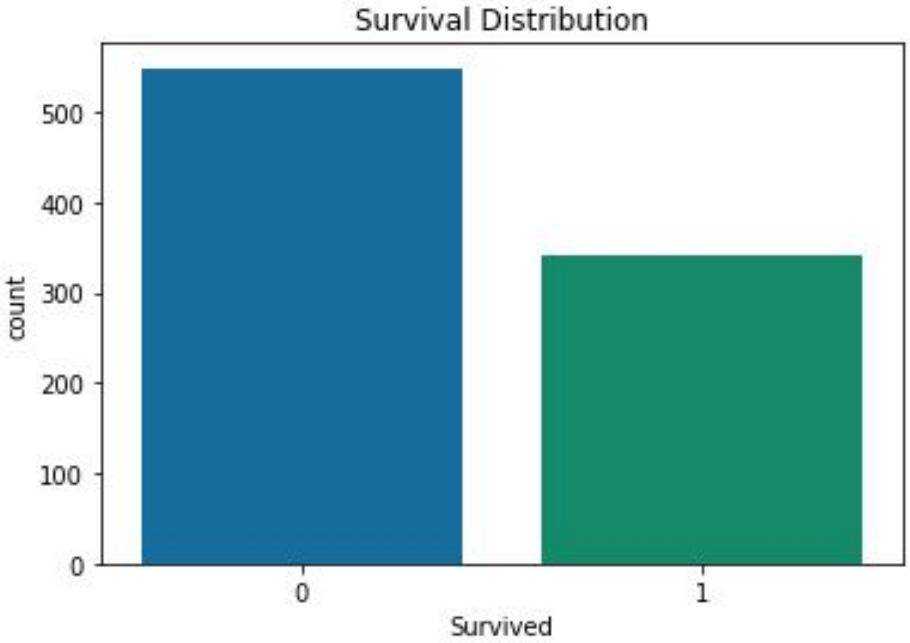
- Three features had null values. Cabin=687, Age=177, and Embarked=2.



```
 0   PassengerId  891 non-null    int64
 1   Survived     891 non-null    int64
 2   Pclass       891 non-null    int64
 3   Name         891 non-null    object
 4   Sex          891 non-null    object
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64
 7   Parch        891 non-null    int64
 8   Ticket       891 non-null    object
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object
 11  Embarked     889 non-null    object
```

Data Overview

- 342 Survivors
- 549 Deaths



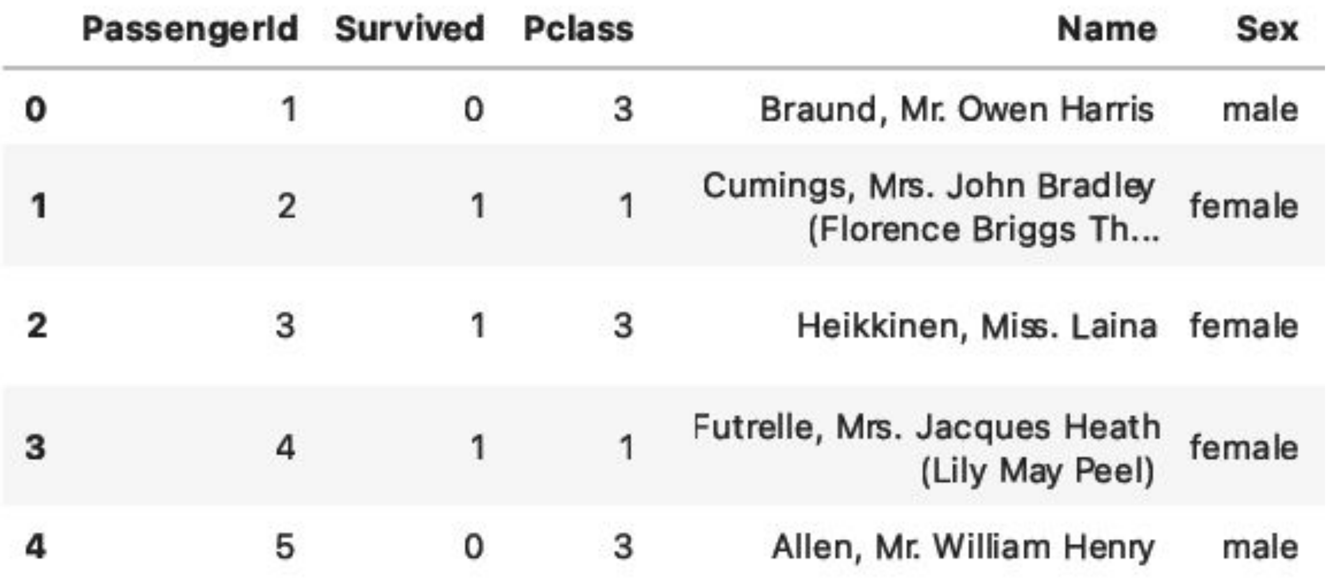
DATA PREPARATION

Partitioning the Dataset

- Each observation was randomly assigned to a training set (67%) and a test set (33%)
- Training set
  - Used to build the model
  - Rebalanced
- Test set
  - Model validation and evaluation

Title Extracted from Name

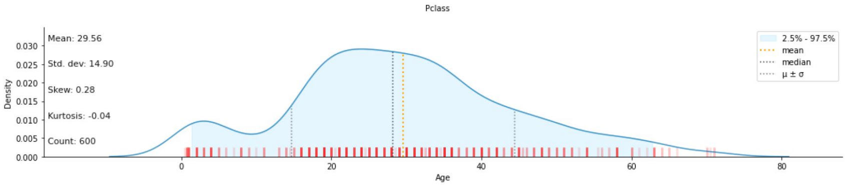
- Indication of marital status and occupation

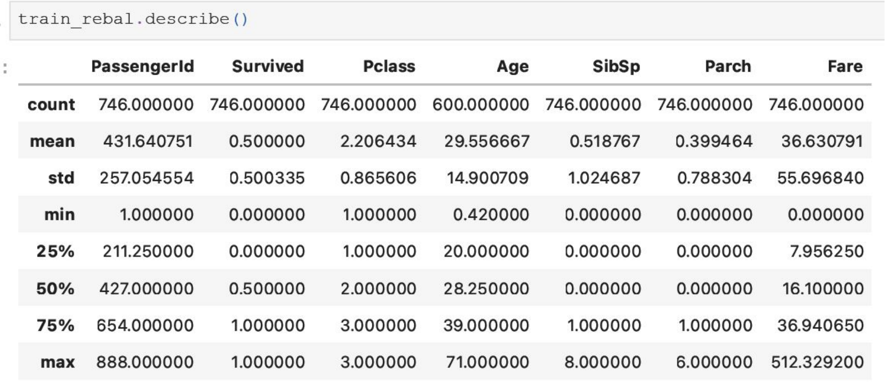


	PassengerId	Survived	Pclass	Name	Sex
0	1	0	3	Braund, Mr. Owen Harris	male
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female
2	3	1	3	Heikkinen, Miss. Laina	female
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female
4	5	0	3	Allen, Mr. William Henry	male

Age and Family Predictors

- ❖ Age Predictor:
  - ➢ Missing Data Imputation with Median Value
    - ■ Positively skewed = mean is greater than median
  - ➢ Transformation
    - ■ Numerical to categorical
      - ● Baby/Toddle = 0-3 years old
      - ● Child = 4-17 years old
      - ● Adult = 18-63 years old
      - ● Elderly = 64-99 years old
- ❖ Fam Predictor:
  - ➢ Combined SibSp and Pach variables

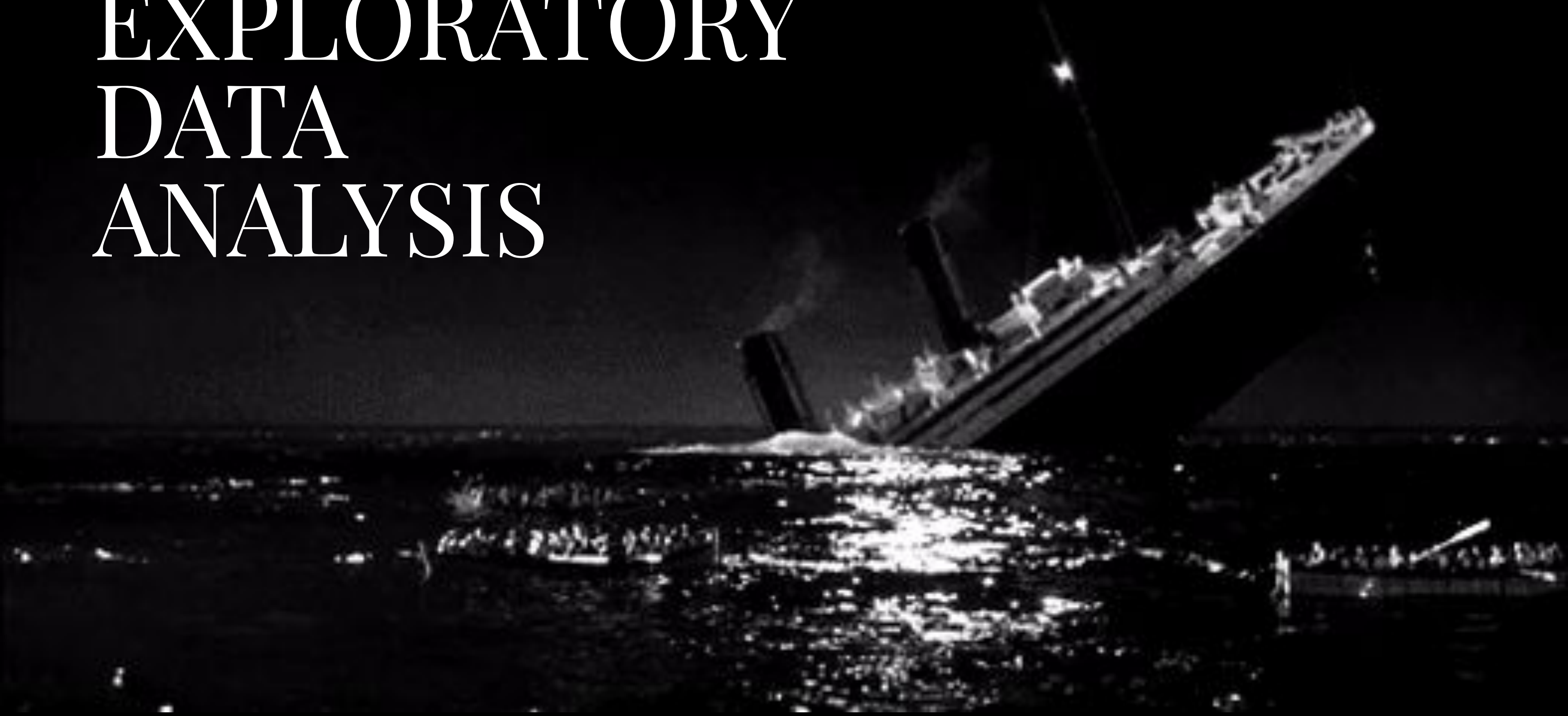




```
train_rebal.describe()
```

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	746.000000	746.000000	746.000000	600.000000	746.000000	746.000000	746.000000
mean	431.640751	0.500000	2.206434	29.556667	0.518767	0.399464	36.630791
std	257.054554	0.500335	0.865606	14.900709	1.024687	0.788304	55.696840
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	211.250000	0.000000	1.000000	20.000000	0.000000	0.000000	7.956250
50%	427.000000	0.500000	2.000000	28.250000	0.000000	0.000000	16.100000
75%	654.000000	1.000000	3.000000	39.000000	1.000000	1.000000	36.940650
max	888.000000	1.000000	3.000000	71.000000	8.000000	6.000000	512.329200

EXPLORATORY DATA ANALYSIS



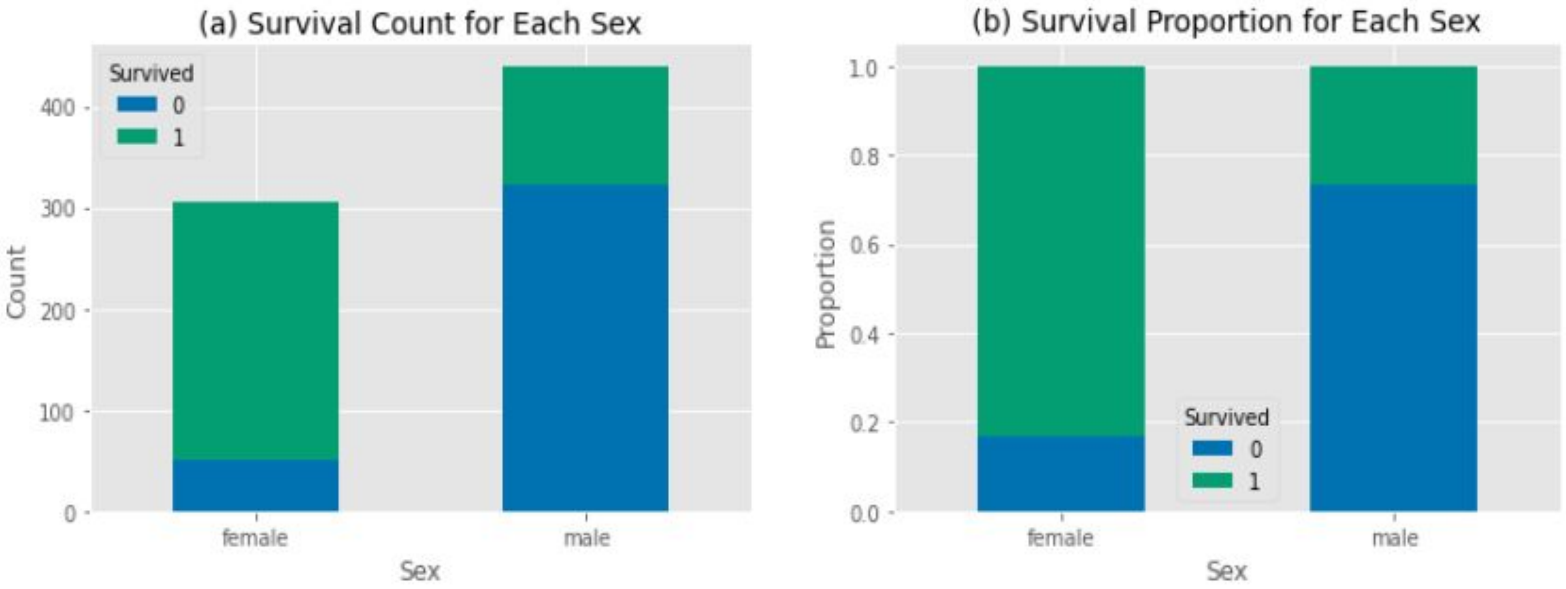
Title

- Four main titles: Mr (53%), Miss (25%), Mrs (18%), Master (5%)
  - Occupational titles (e.g. Reverend, Doctor) were all male
  - The captain of the Titanic did not survive
- Young boys and women were most likely to survive

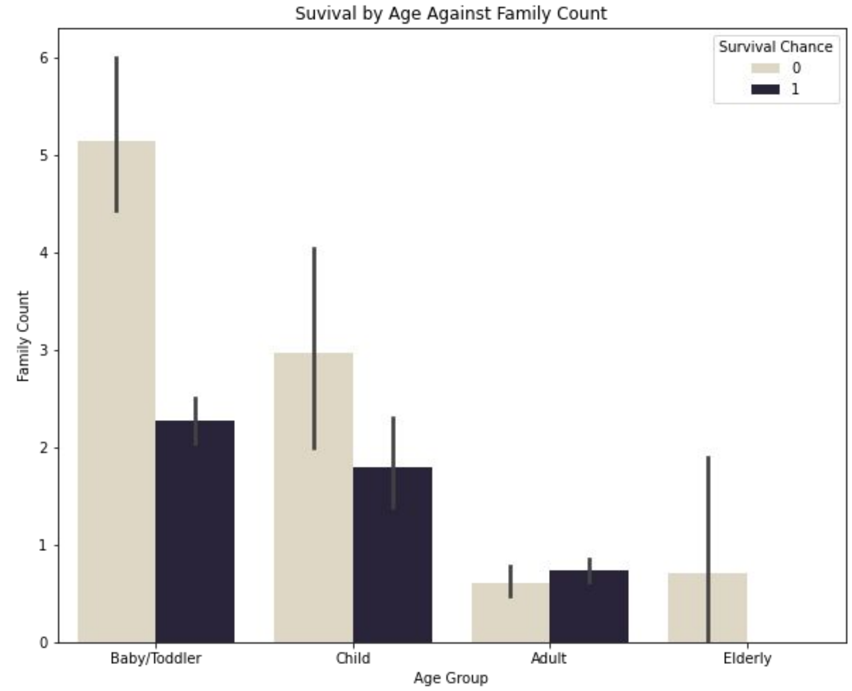
Title	Survivor Count	Likelihood of Survival
Master	25	71.34%
Mr	85	21.41%
Miss	151	82.51%
Mrs	112	85.50%

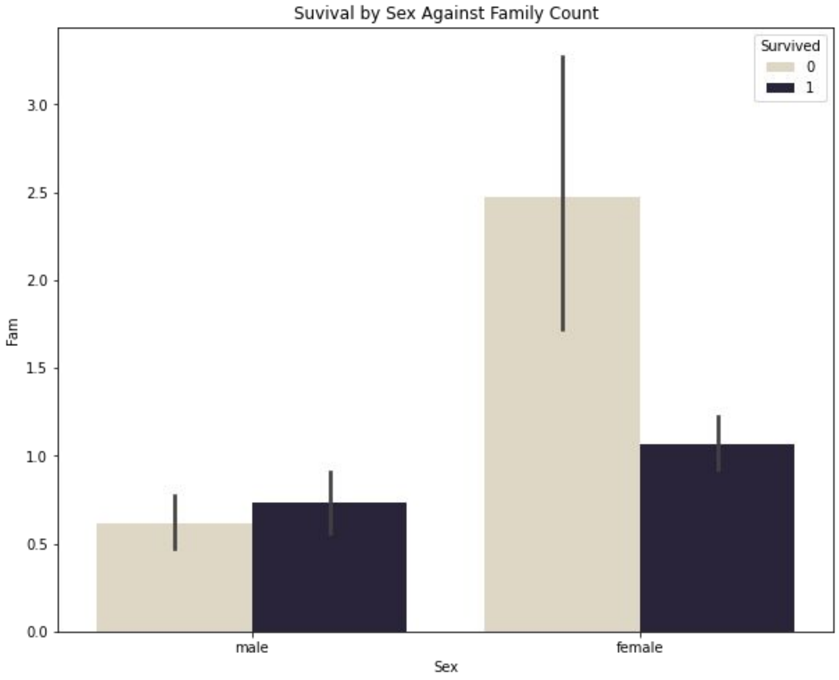
Sex

- ❖ 58% males and 42% females
- ❖ Girls and women were more likely to survive (84%) compared to males (25%)



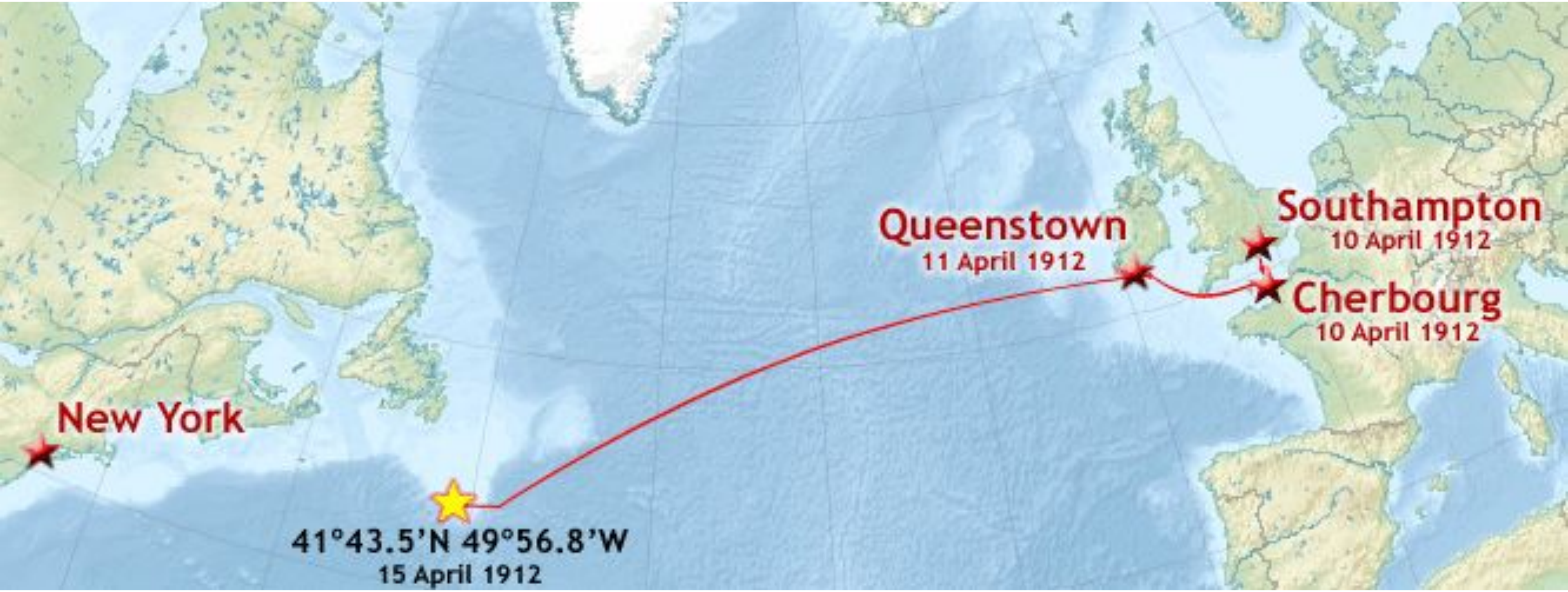
Age, Sex, and Family Predictors





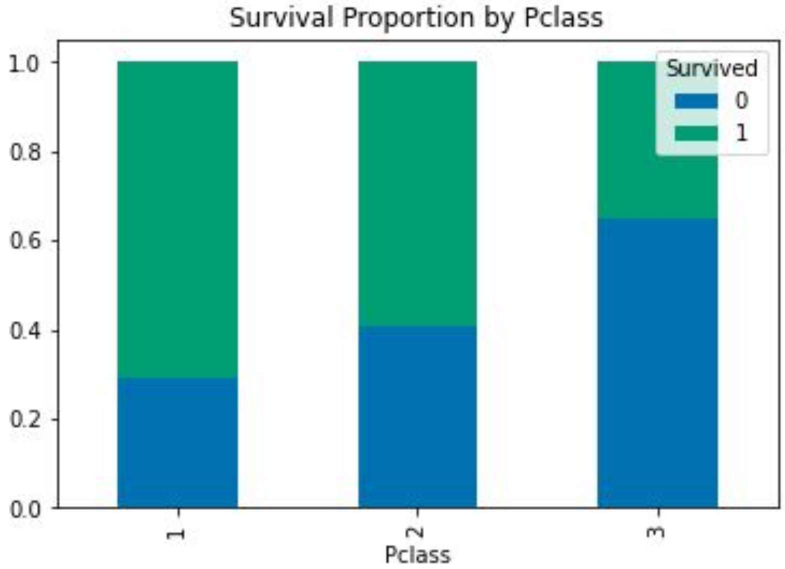
Embarked

- Three ports of embarkation: Port of Cherbourg (C), Port of Queenstown (Q) and Port of Southhampton (S)
  - Most passengers boarded at Southhampton (72%)
- Proportionately more passengers survived who boarded in Cherbourg (67%) and Queenstown (53%)
- Passengers who boarded in Southhampton had a survival rate of 45%


Southampton
10 April 1912
Queenstown
11 April 1912
Cherbourg
10 April 1912
New York
41°43.5'N 49°56.8'W
15 April 1912

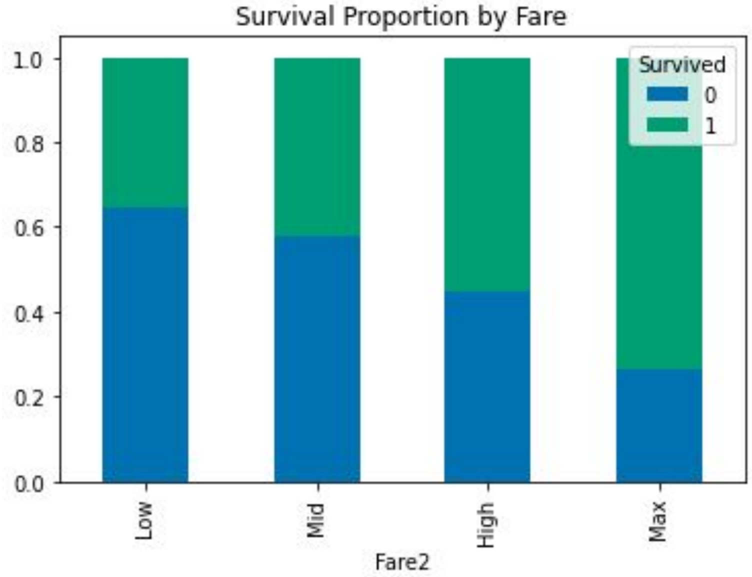
Pclass

- Ordinal
- Three ordered categories: 1, 2, & 3



Fare

- ❖ Continuous
- ❖ Divided into 4 bins based on fare amount
  - ➢ Low, Mid, High, and Max



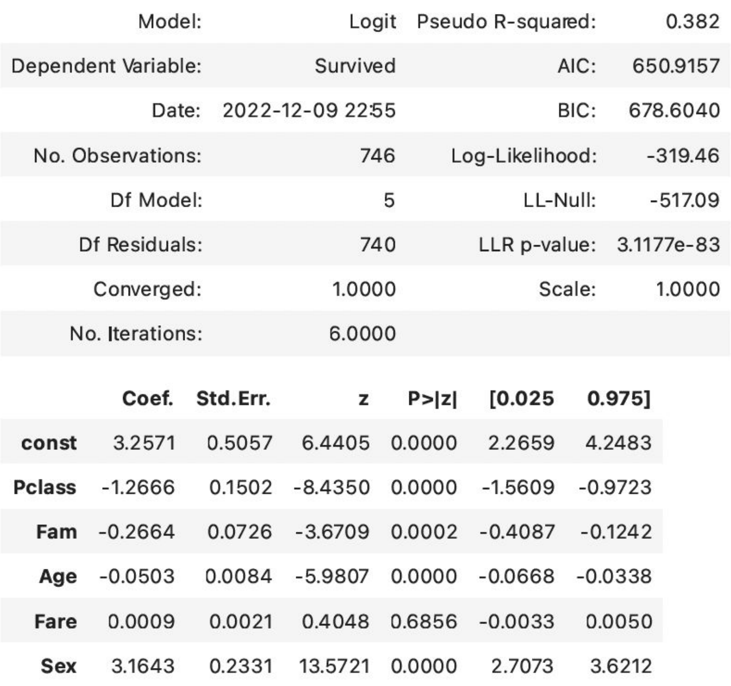
Fare & Ticket

- Ticket, was removed due to being highly correlated with Fare
- Chi-squared Test
- P-value 6.05e-73

MODELING

Logistic Regression

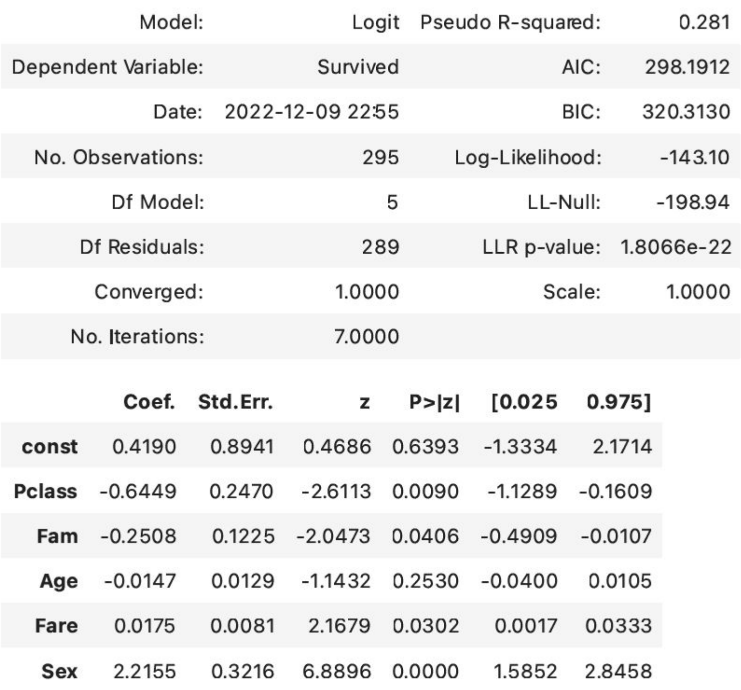
Titanic Training Dataset



Model:	Logit	Pseudo R-squared:	0.382
Dependent Variable:	Survived	AIC:	650.9157
Date:	2022-12-09 22:55	BIC:	678.6040
No. Observations:	746	Log-Likelihood:	-319.46
Df Model:	5	LL-Null:	-517.09
Df Residuals:	740	LLR p-value:	3.1177e-83
Converged:	1.0000	Scale:	1.0000
No. Iterations:	6.0000		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
const	3.2571	0.5057	6.4405	0.0000	2.2659	4.2483
Pclass	-1.2666	0.1502	-8.4350	0.0000	-1.5609	-0.9723
Fam	-0.2664	0.0726	-3.6709	0.0002	-0.4087	-0.1242
Age	-0.0503	0.0084	-5.9807	0.0000	-0.0668	-0.0338
Fare	0.0009	0.0021	0.4048	0.6856	-0.0033	0.0050
Sex	3.1643	0.2331	13.5721	0.0000	2.7073	3.6212

Titanic Test Dataset

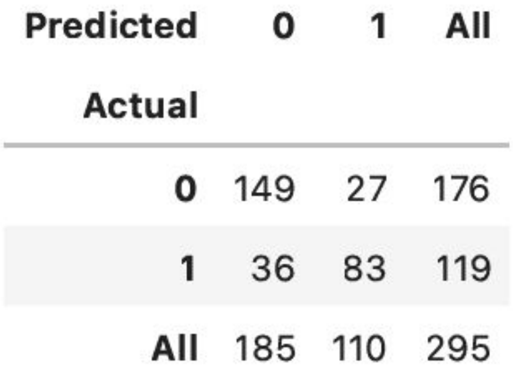


Model:	Logit	Pseudo R-squared:	0.281
Dependent Variable:	Survived	AIC:	298.1912
Date:	2022-12-09 22:55	BIC:	320.3130
No. Observations:	295	Log-Likelihood:	-143.10
Df Model:	5	LL-Null:	-198.94
Df Residuals:	289	LLR p-value:	1.8066e-22
Converged:	1.0000	Scale:	1.0000
No. Iterations:	7.0000		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
const	0.4190	0.8941	0.4686	0.6393	-1.3334	2.1714
Pclass	-0.6449	0.2470	-2.6113	0.0090	-1.1289	-0.1609
Fam	-0.2508	0.1225	-2.0473	0.0406	-0.4909	-0.0107
Age	-0.0147	0.0129	-1.1432	0.2530	-0.0400	0.0105
Fare	0.0175	0.0081	2.1679	0.0302	0.0017	0.0333
Sex	2.2155	0.3216	6.8896	0.0000	1.5852	2.8458

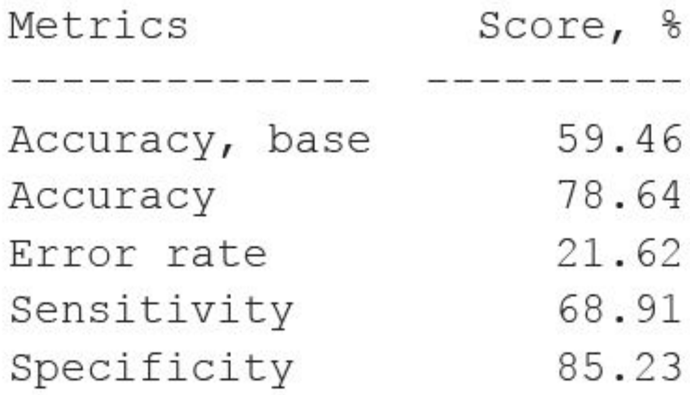
Logistic Regression

Contingency Table



Predicted Actual	0	1	All
0	149	27	176
1	36	83	119
All	185	110	295

Evaluation Metrics Summary:



```
Metrics                Score, %
---------------   ----------
Accuracy, base         59.46
Accuracy               78.64
Error rate             21.62
Sensitivity            68.91
Specificity            85.23
```

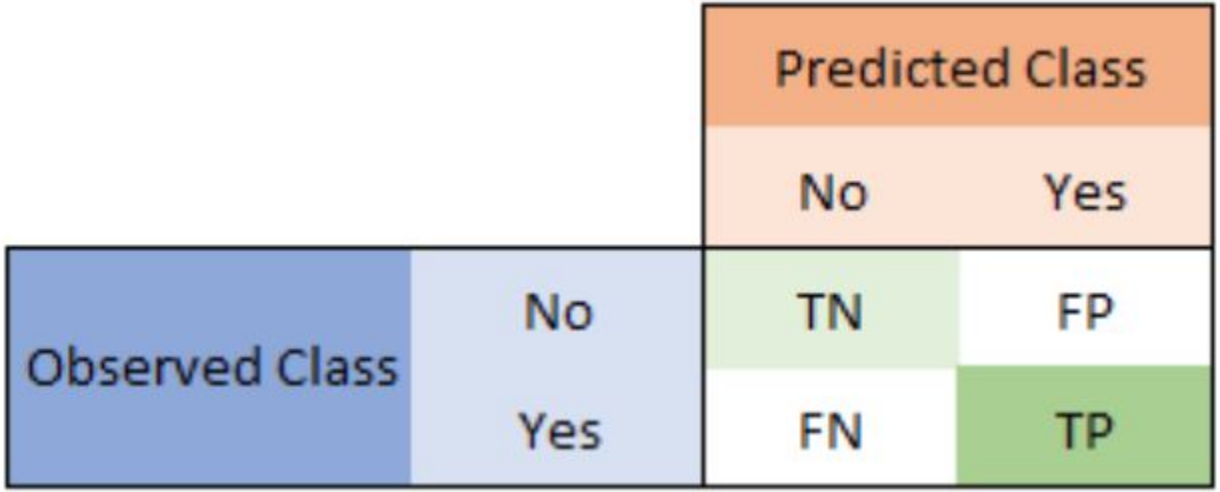
Naive Bayes, Random Forest, CART

- Categorical Dummy variables
  - Sex
  - Embarked
  - Age
  - Pclass
- Continuous
  - Fare
  - Fam

MODEL EVALUATION

Confusion Matrix

- ❖ Summarize predicted results compared to actual distribution
- ❖ Model predicts: Did the passenger survive?



		Predicted Class	
		No	Yes
Observed Class	No	TN	FP
	Yes	FN	TP

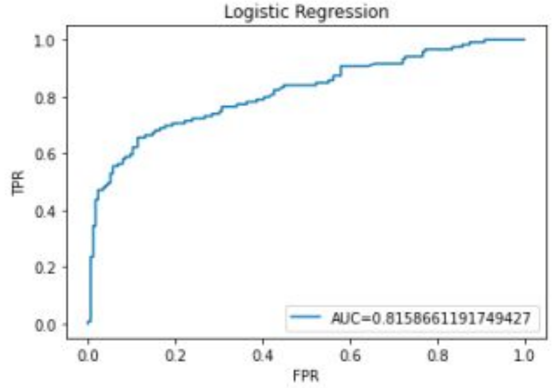
Evaluation Metrics

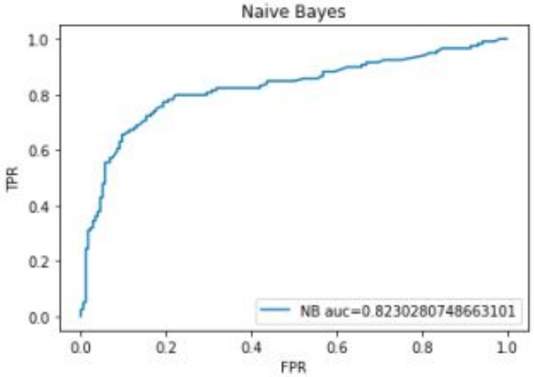
- Accuracy: Proportion of correct predictions
- Error Rate: Proportion of incorrect predictions (1-Accuracy)
- Sensitivity: Proportion of survivors correctly identified
- Specificity: Proportion of deaths correctly identified

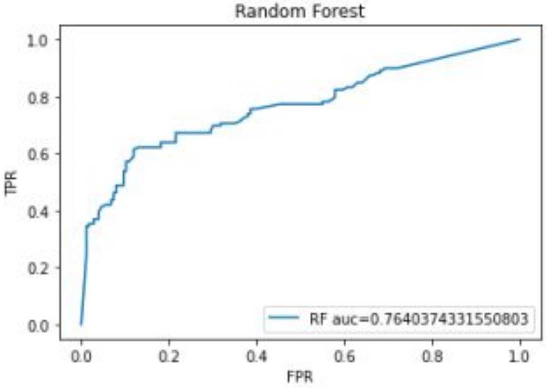
Model	Accuracy	Error Rate	Sensitivity	Specificity
Logistic Regression	78.64%	21.36%	68.91%	85.23%
CART	76.95%	23.05%	62.18%	86.93%
Random Forest	75.59%	24.41%	62.18%%	84.09%
Naïve Bayes	80.34%	19.66%	64.71%	90.91%

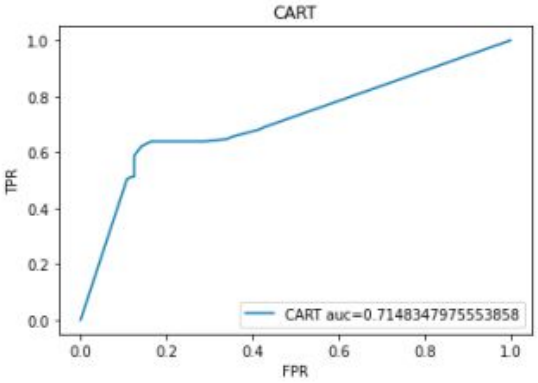
ROC

- ❖ Naïve Bayes = .823
- ❖ Logistic Regression = .816
- ❖ Random Forest =.764
- ❖ CART = .714









CONCLUSION

Given a passenger’s age, sex, family size, boarding class, and fare price, the Naive Bayes model would correctly identify the **survival status of that passenger 79.66% of the time.** Additionally, the Naive Bayes model is able to identify **63.87% of all survivors** and **90.34% of all deaths.**

References:

Brownlee, J. (2018, August 30). How to Use ROC Curves and Precision-Recall Curves for Classification in Python. *MachineLearningMastery.Com*. https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-classifiction-in-python/

History.com Editors. (2009, November 9). Titanic. History.com. Retrieved December 4, 2022, from https://www.history.com/topics/early-20th-century-us/titanic

Kaggle. *Titanic - Machine Learning from Disaster.* Retrieved November 7, 2022, from https://www.kaggle.com/competitions/titanic/overview

Larose, C. D. & Larose, D. T. (2019). *Data science Using Python and R*. John Wiley & Sons,Inc.

Riding, A. (1998, April 26). *Why 'titanic' conquered the world*. The New York Times. Retrieved December 9, 2022, from https://www.nytimes.com/1998/04/26/movies/why-titanic-conquered-the-world.html

The Shipyard. (2022, May 25). *13 maritime disasters more tragic than the Titanic*. The Shipyard. Retrieved December 4, 2022, from https://www.theshipyardblog.com/13-maritime-disasters-more-tragic-than-the-titanic/