SURVIVAL ANALYSES
ON THE TITANIC DATA SET

Maame Yaa Amanpene

December 2018

THE RESEARCH QUESTIONS

Were the survival and deaths on the titanic truly random or patterns exist?

Are there specific factors that strongly influenced whether a passenger perishes or survives?

Is it possible to develop a predictive model that sufficiently predicts the fate of passengers, based on their characteristics.

THE DATA

passenger_id	pclass	name	sex	age	sibsp	parch	ticket	fare	cabin	embarked	boat	body	home.dest	survived
1216	3	Smyth, Miss. Julia	female		0	0	335432	7.7333		Q	13			1
699	3	Cacic, Mr. Luka	male	38	0	0	315089	8.6625		S			Croatia	0
1267	3	Van Impe, Mrs. Jean Baptiste (Rosalie Paula Govaert)	female	30	1	1	345773	24.15		S				0
449	2	Hocking, Mrs. Elizabeth (Eliza Needs)	female	54	1	3	29105	23		S	4		Cornwall / Akron, OH	1
576	2	Veal, Mr. James	male	40	0	0	28221	13		S			Barre, Co Washington, VT	0
1083	3	Olsen, Mr. Henry Margido	male	28	0	0	C 4001	22.525		S		173		0
898	3	Johnson, Mr. William Cahoone Jr	male	19	0	0	LINE	0		S				0
560	2	Sinkkonen, Miss. Anna	female	30	0	0	250648	13		S	10		Finland / Washington, DC	1
1079	3	Ohman, Miss. Velin	female	22	0	0	347085	7.775		S	C			1
908	3	Jussila, Miss. Mari Aina	female	21	1	0	4137	9.825		S				0
313	1	Widener, Mr. Harry Elkins	male	27	0	2	113503	211.5	C82	C			Elkins Park, PA	0
43	1	Bucknell, Mrs. William Robert (Emma Eliza Ward)	female	60	0	0	11813	76.2917	D15	C	8		Philadelphia, PA	1
233	1	Potter, Mrs. Thomas Jr (Lily Alexenia Wilson)	female	56	0	1	11767	83.1583	C50	C	7		Mt Airy, Philadelphia, PA	1
446	2	Hocking, Miss. Ellen "Nellie"	female	20	2	1	29105	23		S	4		Cornwall / Akron, OH	1

Data summary: 1310 observations & 15 features

DATA SET FEATURES

survival - Survival (0 = No; 1 = Yes)
class - Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)
name - Name
sex – Sex (male / female)
age - Age
sibsp - Number of Siblings/Spouses Aboard
parch - Number of Parents/Children Aboard

ticket - Ticket Number
fare - Passenger Fare
cabin - Cabin
embarked - Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)
boat - Lifeboat (if survived)
body - Body number (if did not survive and body was recovered)

DATA WRANGLING: FEATURE SELECTION

RELEVANT FEATURES

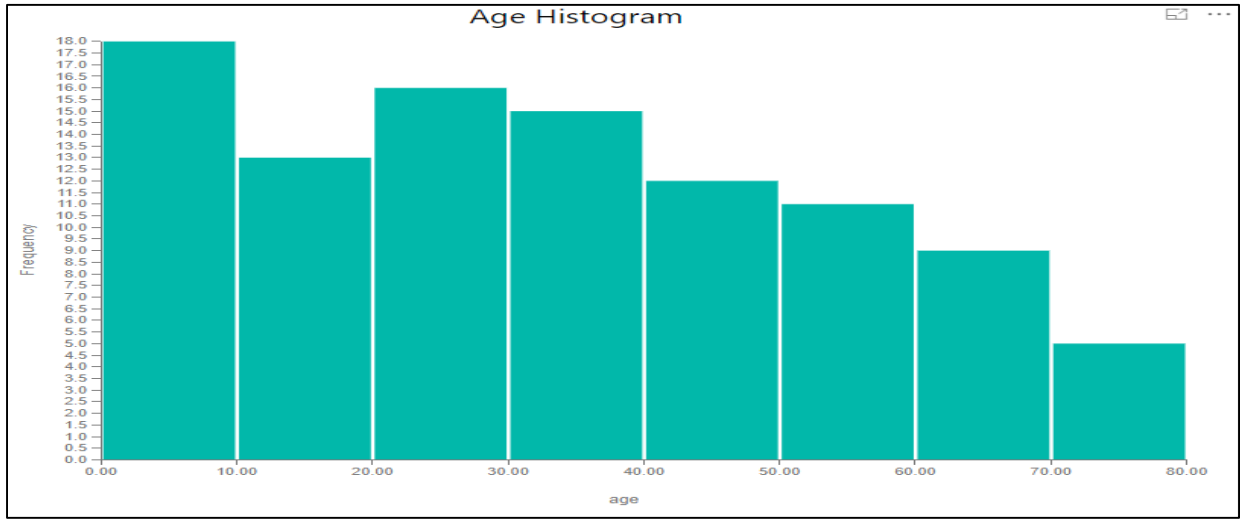
pclass	sex	age	sibsp	parch	fare	embarked	survived
3	female		0	0	7.7333	Q	1
3	male	38	0	0	8.6625	S	0
3	female	30	1	1	24.15	S	0
2	female	54	1	3	23	S	1
2	male	40	0	0	13	S	0
3	male	28	0	0	22.525	S	0
3	male	19	0	0	0	S	0
2	female	30	0	0	13	S	1
3	female	22	0	0	7.775	S	1
3	female	21	1	0	9.825	S	0
1	male	27	0	2	211.5	C	0
1	female	60	0	0	76.2917	C	1
1	female	56	0	1	83.1583	C	1
2	female	20	2	1	23	S	1
3	male	16	1	1	20.25	S	0
3	male	48	0	0	7.8542	S	0

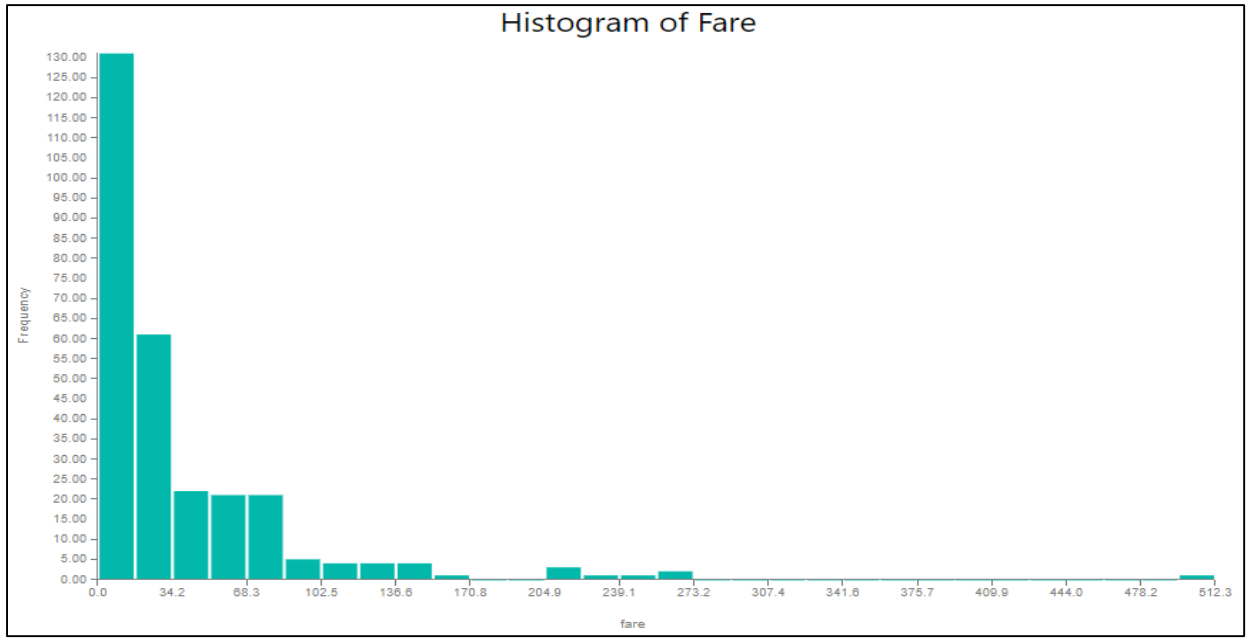
IRRELEVANT FEATURES

passenger_id	name	ticket	cabin	boat	body	home.dest
1216	Smyth, Miss. Julia	335432		13		
699	Cacic, Mr. Luka	315089				Croatia
1267	Van Impe, Mrs. Jean Baptiste (Rosalie Paula Govaert)	345773				
449	Hocking, Mrs. Elizabeth (Eliza Needs)	29105		4		Cornwall / Akron, OH
576	Veal, Mr. James	28221				Barre, Co Washington, VT
1083	Olsen, Mr. Henry Margido	C 4001			173	
898	Johnson, Mr. William Cahoone Jr	LINE				

DATA WRANGLING: HANDLING MISSING DATA

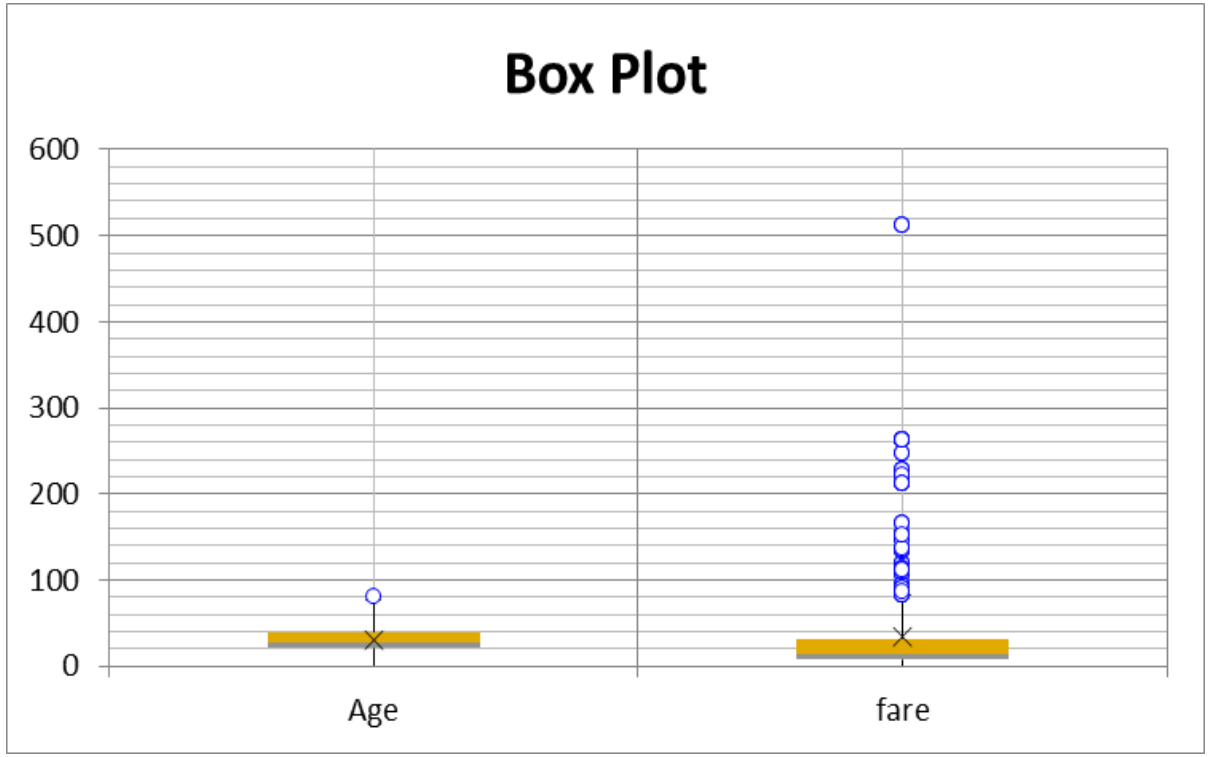
- Features with missing data
  - Age
  - Fare
- Age
  - Fairly even distribution
  - Mean, median and mode values likely to be similar.
- Fare
  - Distribution is skewed towards low fares
  - High fares mostly outliers





DATA WRANGLING: HANDLING MISSING DATA

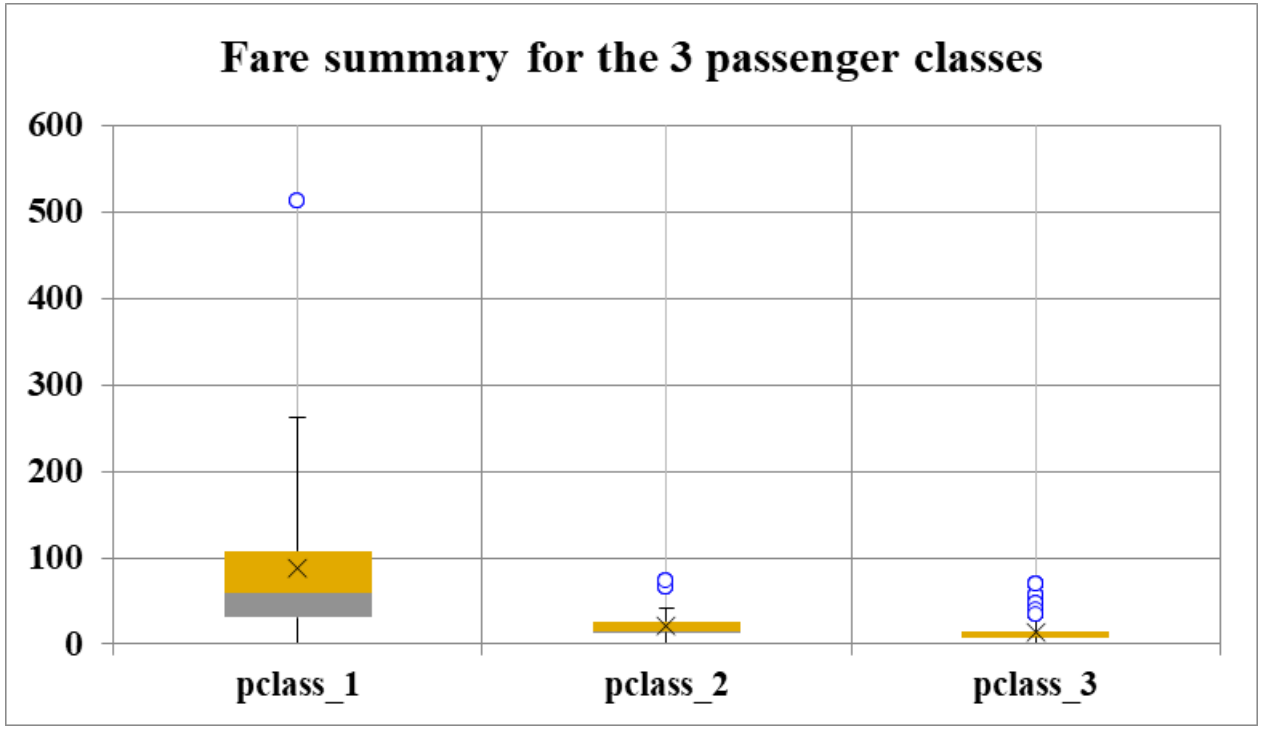
- Age
  - Box plot shows narrow value range.
  - There are only 80 outliers.
  - Mean, median and mode are close.
  - Use median (28.0) to fill missing data
- Fare
  - Several outliers (211) shifting mean.
  - Mean, median and mode are significantly different.
  - Consider fare sub-groups using pclass.



	Age	Fare
Mean	**29.9**	**33.3**
Mode	**24.0**	**8.1**
Median	**28.0**	**14.5**
Outliers	**80**	**211**

DATA WRANGLING: HANDLING MISSING DATA

- Fare
  - Box plot and summary statistics for fare, grouped by passenger classes
  - As suspected, mean, median and mode values are significantly different across the 3 passenger classes.
  - Therefore, missing fare entries will be filled using mean values for their specific pclass.



	pclass_1	pclass_2	pclass_3
Mean	**87.5**	**21.2**	**13.3**
Mode	**26.6**	**13.0**	**8.1**
Median	**60.0**	**15.0**	**8.1**

DATA WRANGLING: FEATURE TRANSFORMATION & ENGINEERING

- Recoded “sex” feature:
  - female =1
  - male = 0
- Created 3 dummy variables for “embarked” feature.
  - C = Cherbourg port
  - Q = Queenstown port
  - S = Southampton port
- Created 3 dummy variables for “pclass” feature.
  - pclass_1, pclass_2 and pclass_3

sex	sex_recoded
female	1
male	0
female	1
female	1
male	0

embarked	embarked_Q	embarked_S	embarked_C
Q	1	0	0
S	0	1	0
C	0	0	1
C	0	0	1
S	0	1	0
Q	1	0	0

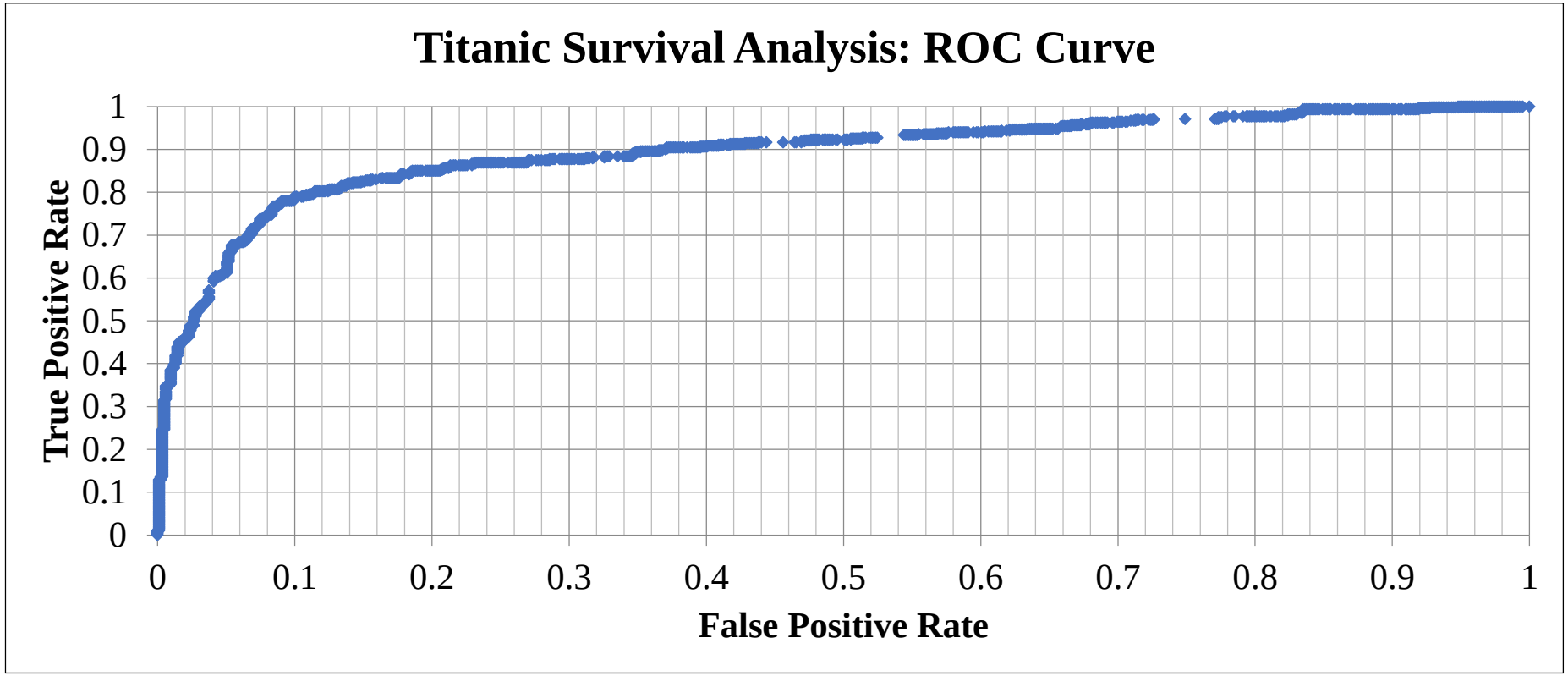
DATA WRANGLING: FEATURE STANDARDIZATION

- All the data features were standardized to improve predictive modeling.

FINAL MODEL FEATURES

pclass_1	pclass_2	pclass_3	Sex_ recoded	Age_ standardized	sibsp_ standardized	parch_ standardized	fare_ standardized	embarked_Q	embarked_S	embarked_C	survived
0	0	1	1	0.35	0	0	0.015094396	1	0	0	1
0	0	1	0	0.475	0	0	0.016908074	0	1	0	0
0	0	1	1	0.375	0.125	0.111111111	0.047137661	0	1	0	0
0	1	0	1	0.675	0.125	0.333333333	0.04489301	0	1	0	1
0	1	0	0	0.5	0	0	0.02537431	0	1	0	0
0	0	1	0	0.35	0	0	0.043965872	0	1	0	0
0	0	1	0	0.2375	0	0	0	0	1	0	0
0	1	0	1	0.375	0	0	0.02537431	0	1	0	1
0	0	1	1	0.275	0	0	0.015175789	0	1	0	1

MODELING RESULTS



- Model accuracy: 0.86
- AUC score: 0.894

Confusion Matrix		
	Yes-Observed	No-Observed
Yes-Predicted	370	73
No-Predicted	110	756

FUTURE ANALYSES

Remove outliers feature-by-feature.

Examine the correlation between “survival” and each feature / predictor.