Titanic Dataset: Who Most Likely Survived?

By (Sylar)Jiajian Guo, Qiqi Tiang, Lequn Yu, Scott McCoy, Tiam Moradi

QUESTROM SCHOOL OF BUSINESS

A Glimpse of Titanicers

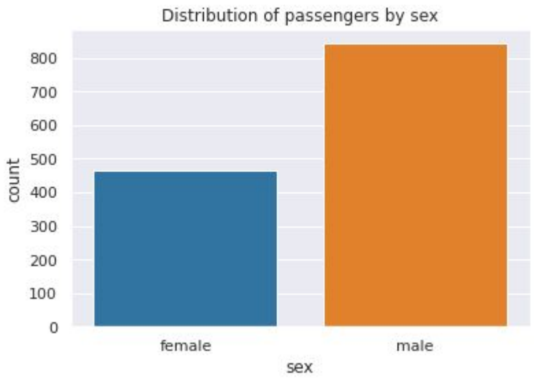
Titanic.head()

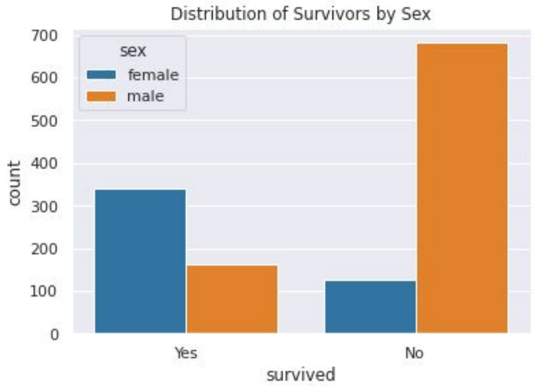
	pclass	survived	name	sex	age	sibsp	parch	ticket	fare	cabin	embarked	boat	body	home.dest
0	1	1	Allen, Miss. Elisabeth Walton	female	29.0000	0	0	24160	211.3375	B5	S	2	NaN	St Louis, MO
1	1	1	Allison, Master. Hudson Trevor	male	0.9167	1	2	113781	151.5500	C22 C26	S	11	NaN	Montreal, PQ / Chesterville, ON
2	1	0	Allison, Miss. Helen Loraine	female	2.0000	1	2	113781	151.5500	C22 C26	S	NaN	NaN	Montreal, PQ / Chesterville, ON
3	1	0	Allison, Mr. Hudson Joshua Creighton	male	30.0000	1	2	113781	151.5500	C22 C26	S	NaN	135.0	Montreal, PQ / Chesterville, ON
4	1	0	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	female	25.0000	1	2	113781	151.5500	C22 C26	S	NaN	NaN	Montreal, PQ / Chesterville, ON

Sex

	survived
sex	
female	0.727468
male	0.190985

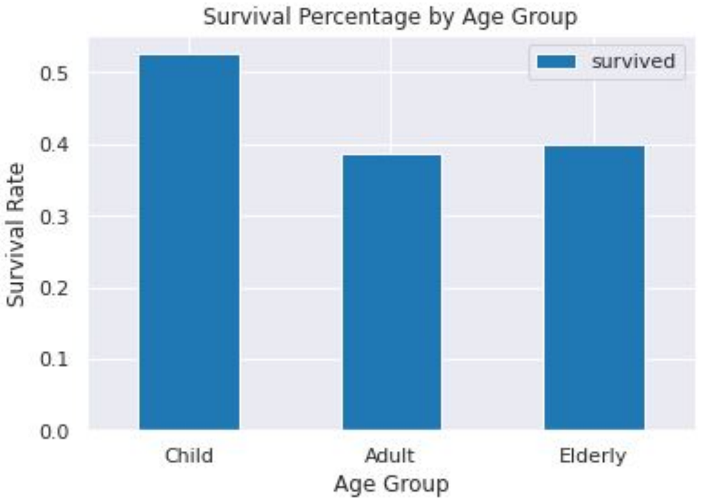
- The survival rate of female passengers is significant higher than that of male passengers.





Age

age_group	survived
Child	0.525974
Adult	0.386189
Elderly	0.400000

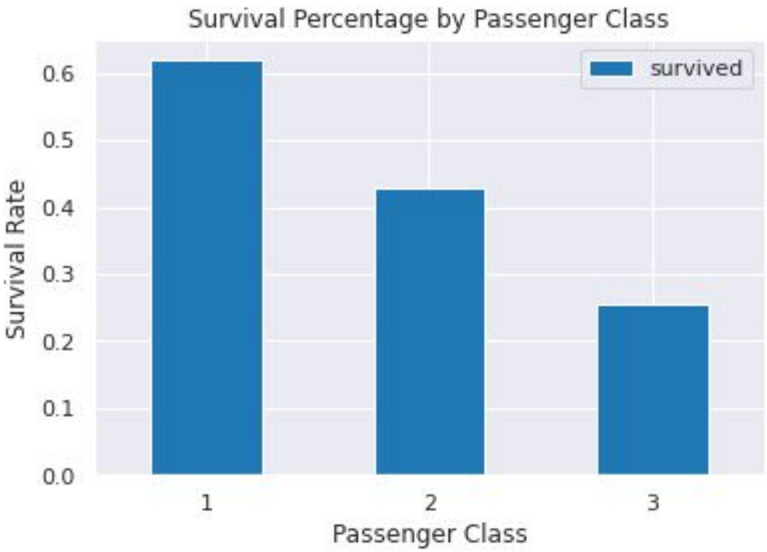


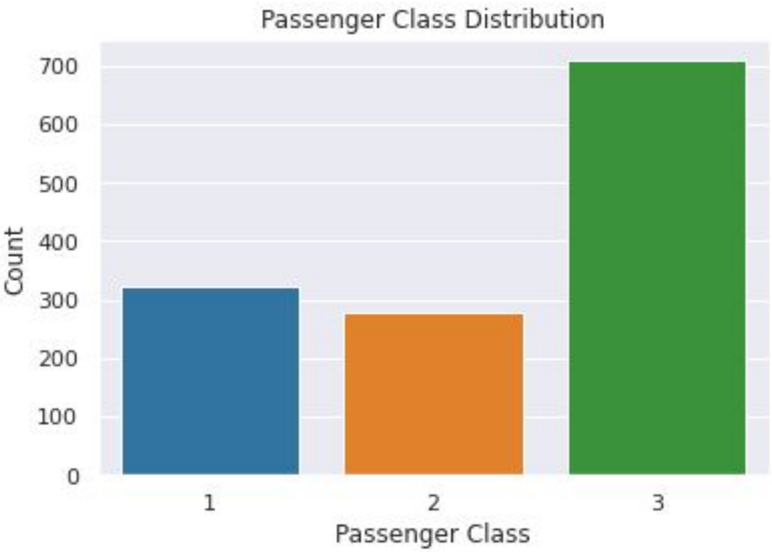
- Child: under 18
- Adult: 18-50
- Elderly: over 50

Child’s survival rate is the highest !

QUESTROM SCHOOL OF BUSINESS

Passenger Class



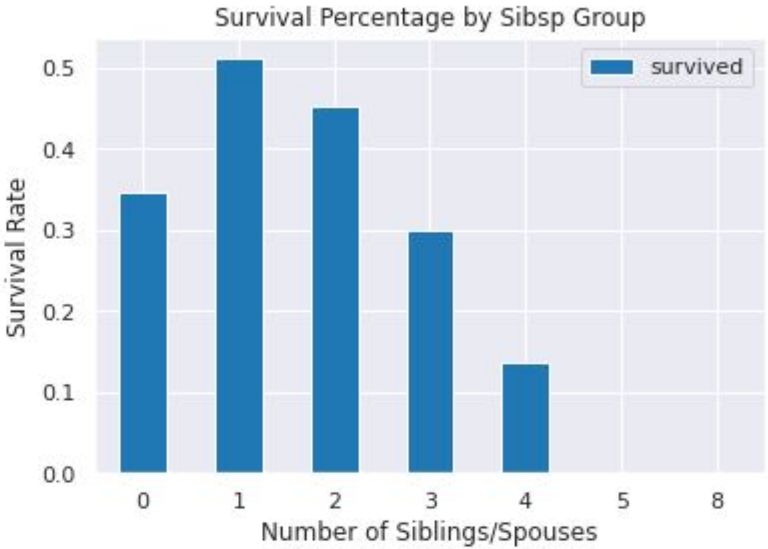


- The survival rate of the third class cabin is the lowest.
- The survival rate of the first class cabin is the highest.

Familial Relationships (sibsp / parch)

sibsp_group	survived
Alone	0.346801
1-2 companions	0.504155
more than 2 companions	0.157895

- Having 1 partner on Titanic had the highest survival rate



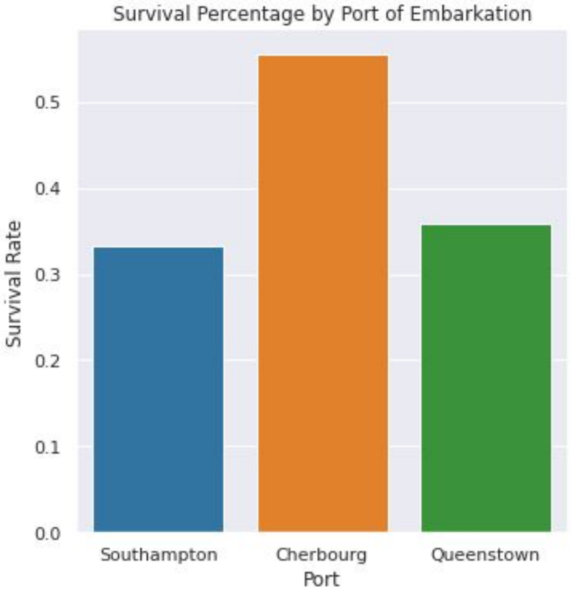
Fare

pclass	Average Fare	Survival Percentage
1	87.508992	0.619195
2	21.179196	0.429603
3	13.302889	0.255289

- Cheap: under 10 dollars
- Mid: between 10 and 30 dollars
- Expensive: above 30 dollars

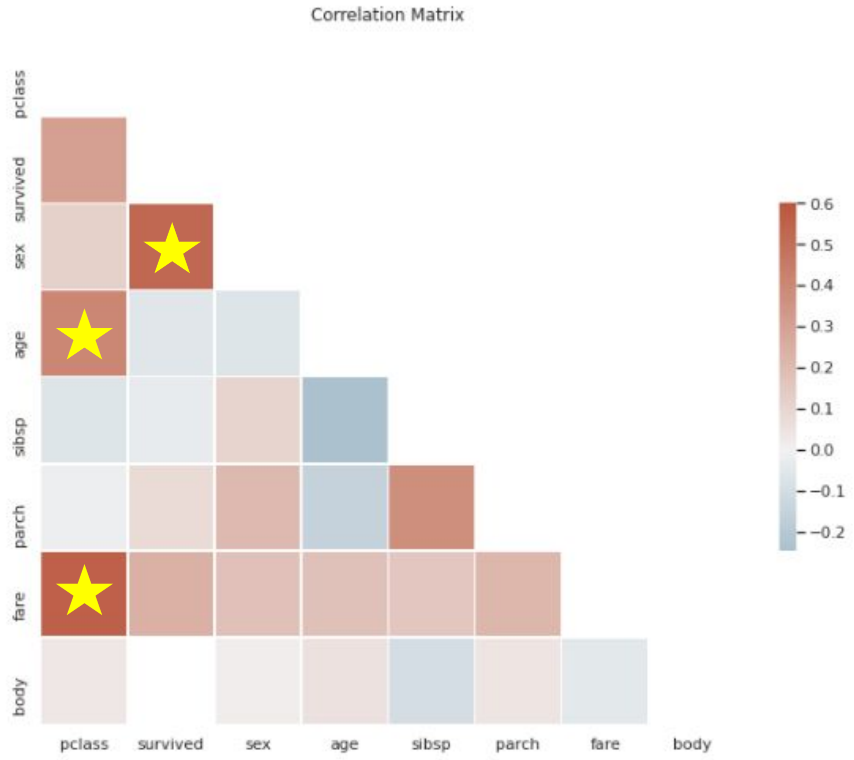
QUESTROM SCHOOL OF BUSINESS

Port Embarkation



embarked	pclass	Count
C	1	141
	2	28
	3	101
Q	1	3
	2	7
	3	113
S	1	177
	2	242
	3	495

Variables Correlation Matrix



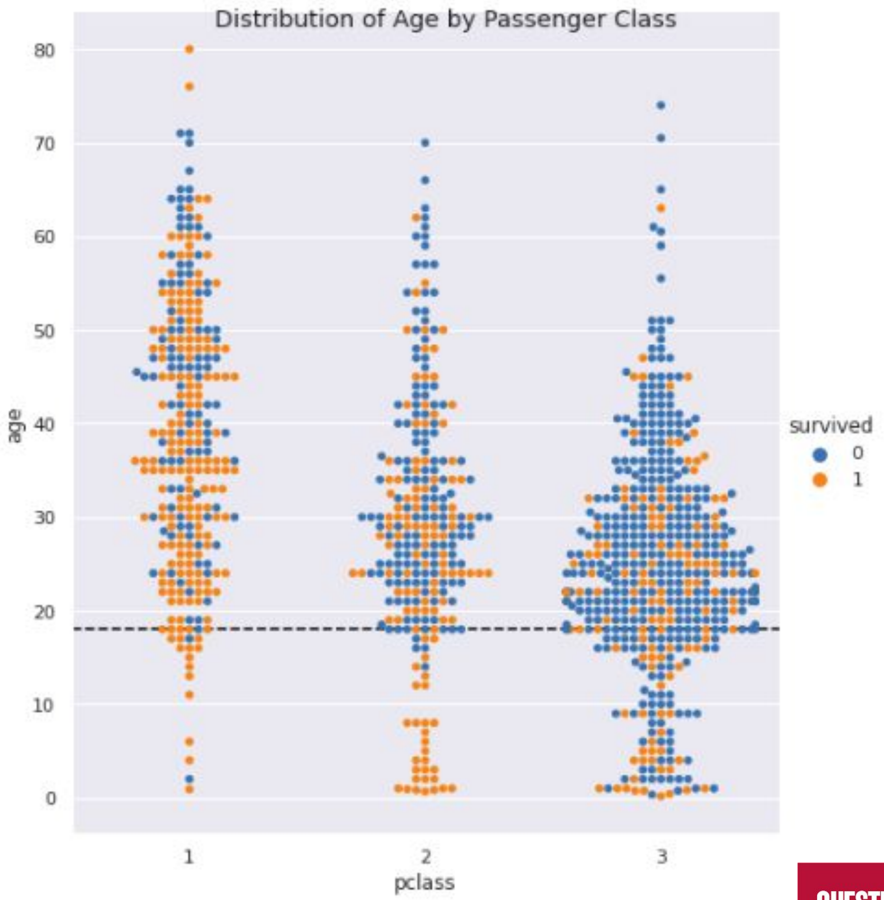
High Correlation

- Sex and Survived
- Age and Pclass
- Fare and Pclass

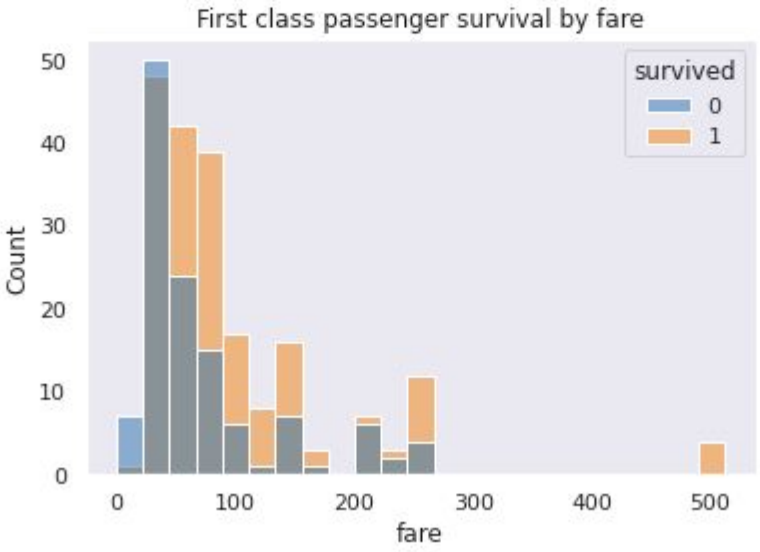
Age by Class

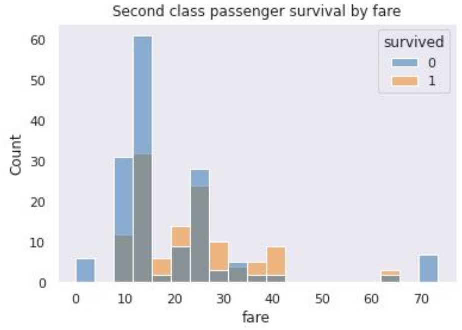
pclass	age
1	39.159918
2	29.506705
3	24.816367

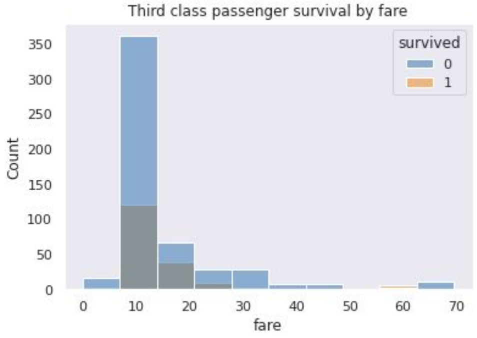
- First class passengers are on average almost 10 years older than second class passengers, and 15 years older than third class passengers.



Fare vs Class Survival Distribution

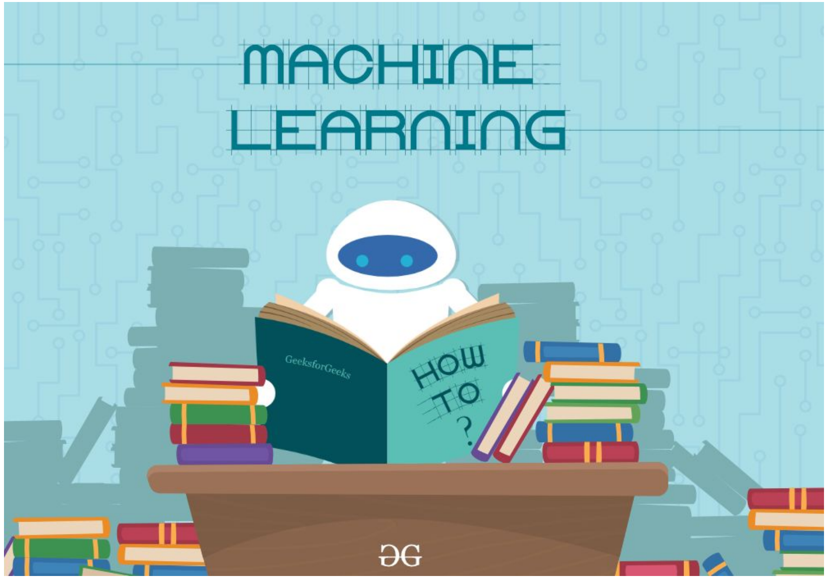






- Distribution differences can only be seen in the First Class

Who Survived: Machine Learning Classification



QUESTROM SCHOOL OF BUSINESS

Machine Learning: Preprocessing

```
# preprocessing for the Logistic Regression
if self.linear_model:
    print('We are turning categorical features into ohc and dropping some unhelpful columns...')
    df = pd.get_dummies(df,columns=['pclass','sex','embarked'])
    df['family_size'] = df['sibsp'] + df['parch']
    df.drop(['name','boat','home.dest','sex_male','body','cabin','ticket','sibsp','parch','pclass_3','embarked_S'],axis=1,inplace=True)
    # df.dropna(axis=0,inplace=True)
    df['age'].fillna(df['age'].median(),inplace=True)
    df['fare'].fillna(df['fare'].median(),inplace=True)

else:
# Preprocessing for the XGBoost Model
    df = pd.get_dummies(df,columns=['pclass','sex','embarked'])
    df['family_size'] = df['sibsp'] + df['parch']
    df['cabin'] = df[df.cabin.notnull()].cabin.apply(lambda cabin: cabin[0]).map({'A':1,'B':2,'C':3,'D':4,'E':5,'F':6,'G':7})
    df.drop(['name','boat','home.dest','sex_male','body','ticket','sibsp','parch','pclass_3','embarked_S'],axis=1,inplace=True)
    df.fillna(0,inplace=True)
```

Model Performance: Cross Validation Scores

- Logistic Regression

```
2 pd.DataFrame(model.cv_results_).sort_values('rank_test_score').iloc[0,:]
```

```
mean_fit_time                                                  0.00378466
std_fit_time                                                  0.000180305
mean_score_time                                               0.000973606
std_score_time                                                3.75194e-05
param_C                                                             0.365
param_max_iter                                                         95
param_penalty                                                          l1
param_random_state                                                    833
param_solver                                                    liblinear
params                  {'C': 0.365, 'max_iter': 95, 'penalty': 'l1', ...
split0_test_score                                                0.822335
split1_test_score                                                 0.77551
split2_test_score                                                 0.80102
split3_test_score                                                0.795918
split4_test_score                                                0.755102
mean_test_score                                                  0.789977
std_test_score                                                   0.022934
rank_test_score                                                         1
```

- XGBoost

```
2 pd.DataFrame(model_.cv_results_).sort_values('rank_test_score').iloc[0,:]
```

```
mean_fit_time                                                    0.193202
std_fit_time                                                   0.00786232
mean_score_time                                                0.00208974
std_score_time                                                 1.7916e-05
param_learning_rate                                                  0.02
param_loss                                                       deviance
param_max_depth                                                         5
param_n_estimators                                                    100
param_random_state                                                    833
params                 {'learning_rate': 0.02, 'loss': 'deviance', 'm...
split0_test_score                                                0.807107
split1_test_score                                                0.806122
split2_test_score                                                0.816327
split3_test_score                                                0.811224
split4_test_score                                                0.770408
mean_test_score                                                  0.802238
std_test_score                                                  0.0163167
rank_test_score                                                         1
```

Model Performance: Test Set

	accuracy	precision	recall	f1_score	roc_auc
0	0.79878	0.770642	0.672	0.717949	0.774424

- Logistic Regression

	accuracy	precision	recall	f1_score	roc_auc
0	0.820122	0.836735	0.656	0.735426	0.788591

- XGBoost

Examining Model Predictions

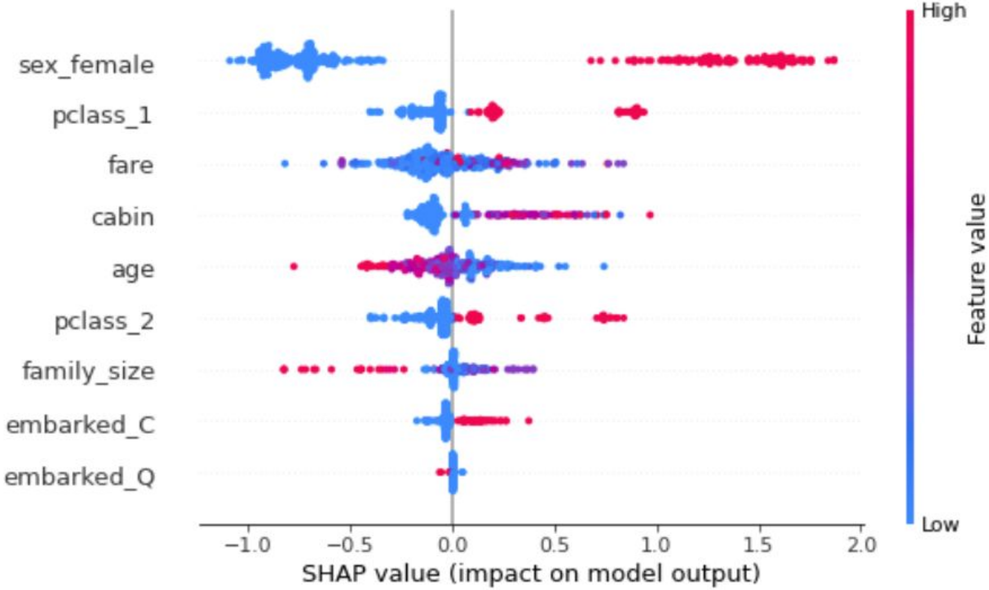
- Looking beyond performance metrics to better understand how the model made it's predictions

QUESTROM SCHOOL OF BUSINESS

Feature Importance:

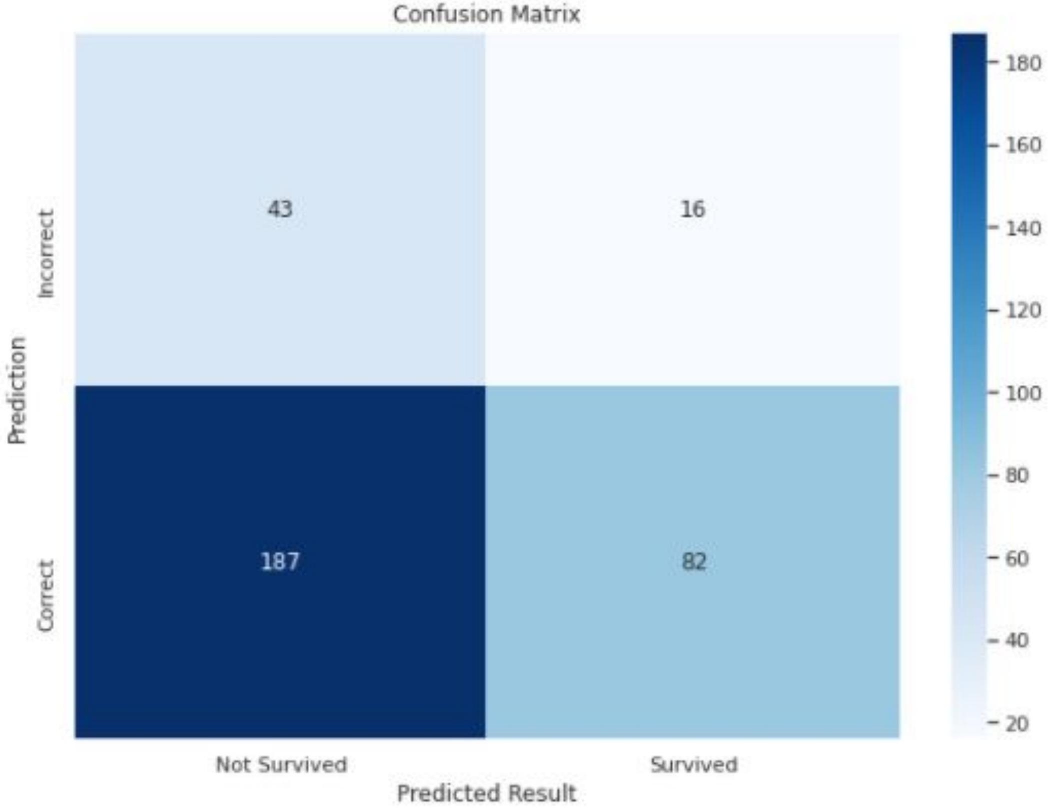
age	-0.030069
fare	0.003281
pclass_1	1.477765
pclass_2	0.668954
sex_female	2.452388
embarked_C	0.485291
embarked_Q	0.000000
family_size	-0.165658

Printing the SHAP Values of the



QUESTROM SCHOOL OF BUSINESS

Confusion Matrix



Correct Predictions: 269
Total Predictions: 328
Accuracy: 0.82

True Positives: 82
False Positives: 16
Precision: 0.837

True Positives: 82
False Negatives: 43
Recall: 0.656

True Negatives: 187
Total Negatives: 203
Specificity: 0.921

Dataframe with Predictions

	pclass	survived_x	name	sex	age	sibsp	parch	ticket	fare	cabin	embarked	boat	body	home.dest	survived_y	predicted	correct_pred
1	1	1	Allison, Master. Hudson Trevor	male	0.9167	1	2	113781	151.5500	C22 C26	S	11	NaN	Montreal, PQ / Chesterville, ON	1	0	False
3	1	0	Allison, Mr. Hudson Joshua Creighton	male	30.0000	1	2	113781	151.5500	C22 C26	S	NaN	135.0	Montreal, PQ / Chesterville, ON	0	0	True
7	1	0	Andrews, Mr. Thomas Jr	male	39.0000	0	0	112050	0.0000	A36	S	NaN	NaN	Belfast, NI	0	0	True
13	1	1	Barber, Miss. Ellen 'Nellie'	female	26.0000	0	0	19877	78.8500	NaN	S	6	NaN	NaN	1	1	True
15	1	0	Baumann, Mr. John D	male	NaN	0	0	PC 17318	25.9250	NaN	S	NaN	NaN	New York, NY	0	0	True
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1298	3	0	Wittevrongel, Mr. Camille	male	36.0000	0	0	345771	9.5000	NaN	S	NaN	NaN	NaN	0	0	True
1302	3	0	Yousif, Mr. Wazli	male	NaN	0	0	2647	7.2250	NaN	C	NaN	NaN	NaN	0	0	True
1304	3	0	Zabour, Miss. Hileni	female	14.5000	1	0	2665	14.4542	NaN	C	NaN	328.0	NaN	0	1	False
1307	3	0	Zakarian, Mr. Ortin	male	27.0000	0	0	2670	7.2250	NaN	C	NaN	NaN	NaN	0	0	True
1308	3	0	Zimmerman, Mr. Leo	male	29.0000	0	0	315082	7.8750	NaN	S	NaN	NaN	NaN	0	0	True

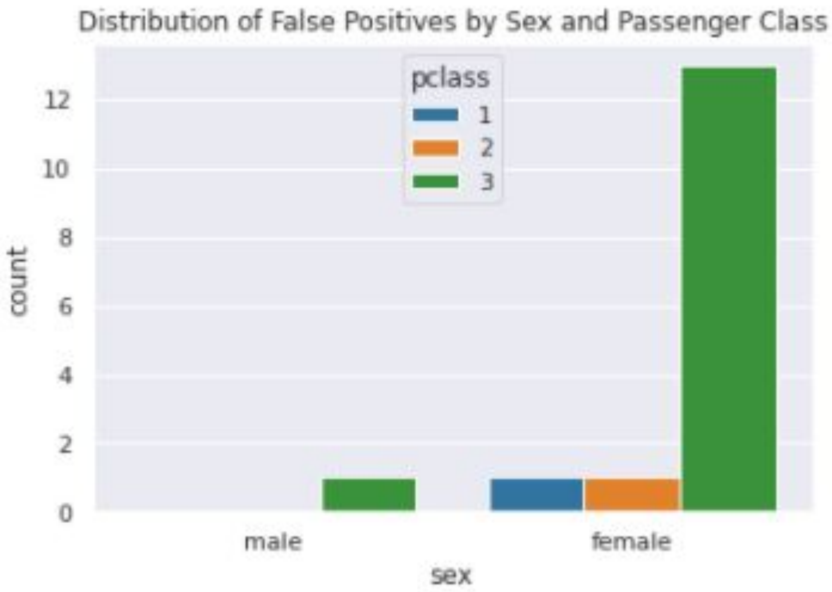
328 rows × 17 columns

Add column for:

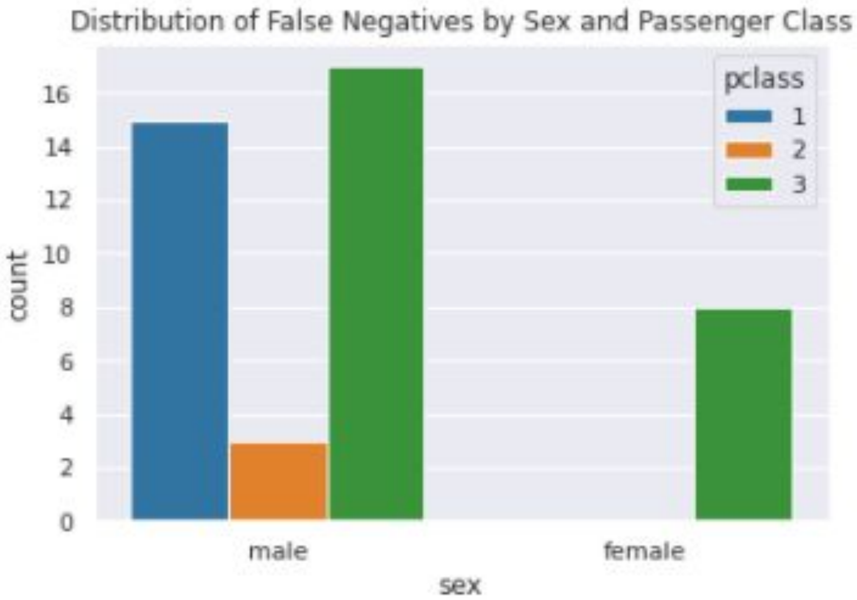
- Model’s predicted value
- Whether the prediction was correct

Visualizing Incorrect Predictions

- False Positives



- False Negatives



Conclusions

- Got a better understanding of passenger features and how they interact with each other to influence a passenger's chances of survival.
- Expanded and demystified our machine learning model to better explain its predictions.

QUESTROM SCHOOL OF BUSINESS

Q & A

QUESTROM SCHOOL OF BUSINESS

Old Presentation

Titanic Dataset: Who Most Likely Survived?

By (Sylar)Jiajian Guo, Qiqi Tiang, Lequn Yu, Scott McCoy, Tiam Moradi

Titanic Passenger Dataset

pclass	survived	name	sex	age	sibsp	parch	ticket	fare	cabin	embarked	boat	body	home_dest
3	0	Storey, Mr. Thomas	male	60.5	0	0	3701	NaN	None	S	None	261.0	None
1	0	Andrews, Mr. Thomas Jr	male	39.0	0	0	112050	0.0	A36	S	None	NaN	Belfast, NI
1	0	Chisholm, Mr. Roderick Robert Crispin	male	NaN	0	0	112051	0.0	None	S	None	NaN	Liverpool, England / Belfast
1	0	Fry, Mr. Richard	male	NaN	0	0	112058	0.0	B102	S	None	NaN	None
1	0	Harrison, Mr. William	male	40.0	0	0	112059	0.0	B94	S	None	110.0	None

	number_of_survivors	number_of_passengers	passenger_survival_percentage
0	500	1309	38.2

Titanic Passenger Dataset

pclass	survived	name	sex	age	sibsp	parch	ticket	fare	cabin	embarked	boat	body	home_dest
3	0	Storey, Mr. Thomas	male	60.5	0	0	3701	NaN	None	S	None	261.0	None
1	0	Andrews, Mr. Thomas Jr	male	39.0	0	0	112050	0.0	A36	S	None	NaN	Belfast, NI
1	0	Chisholm, Mr. Roderick Robert Crispin	male	NaN	0	0	112051	0.0	None	S	None	NaN	Liverpool, England / Belfast
1	0	Fry, Mr. Richard	male	NaN	0	0	112058	0.0	B102	S	None	NaN	None
1	0	Harrison, Mr. William	male	40.0	0	0	112059	0.0	B94	S	None	110.0	None

Features most associated with increased chance of survival:

-Sex
-Passenger Class

Sex and Survival

sex	Number_Passengers	Number_Survivors	Survival_Percentage
male	843	161	0.190985
female	466	339	0.727468

-Majority of passengers were male

-Majority of survivors were female

Age

age_group	Number_Passengers	Number_Survivors	Survival_Percentage
Child	154	81	0.525974
Elder	95	38	0.400000
Adult	797	308	0.386449
None	263	73	0.277567

```
(SELECT *,
CASE WHEN age > 0 AND age < 18 THEN  'Child'
WHEN  age >= 18 AND age  <= 50 THEN 'Adult'
WHEN age >50  THEN 'Elder'
ELSE NULL END AS age_group
FROM `ba775-team-6b.Project.passengers`
)
```

-Passengers under the age of 18 had the highest survival rate

-Very few passengers over the age of 50

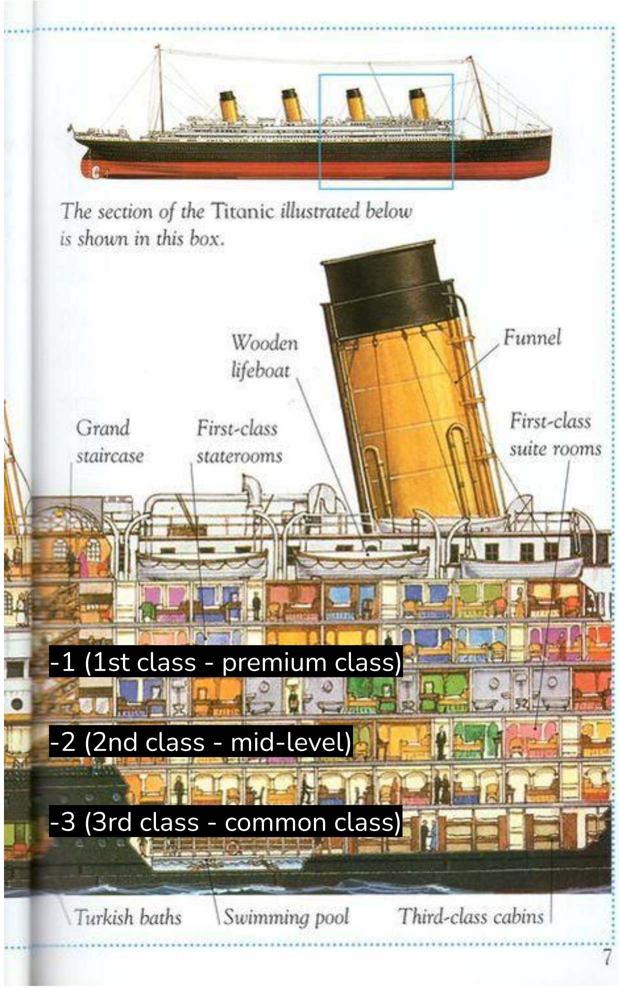
-263 null values in age column

Passenger Class

-Passengers who had better Class, had higher chances of survival

-Locations of different Class on Titanic affected this rate difference

pclass	Number_Passengers	Number_Survivors	Survival_Percentage
1	323	200	0.619195
2	277	119	0.429603
3	709	181	0.255289



Port of Embarkation

port_of_embarkation	Passengers	Survivors	Survival_Percentage
Cherbourg, France	270	150	0.555556
Queenstown, Ireland	123	44	0.357724
Southampton, England	914	304	0.332604

port_of_embarkation	pclass	Number_Passengers	pclass_percentage_by_port
Cherbourg, France	1	141	0.52
Cherbourg, France	2	28	0.10
Cherbourg, France	3	101	0.37
Queenstown, Ireland	1	3	0.02
Queenstown, Ireland	2	7	0.06
Queenstown, Ireland	3	113	0.92
Southampton, UK	1	177	0.19
Southampton, UK	2	242	0.27
Southampton, UK	3	495	0.54

-Higher survival rate Port had larger number of higher Class Passengers who had better locations, which affected chances of survival when scaping

Fare

-Passengers who had better Fare, had higher chances of survival

-Fare is highly correlated to Class

fare_group	Number_Passengers	Number_Survivors	Survival_Percentage
Expensive	350	202	0.577143
Mid	467	188	0.402570
Cheap	474	108	0.227848
None	18	2	0.111111

```
(SELECT *,
CASE WHEN fare > 0 AND fare < 10 THEN  'Cheap'
WHEN  fare >= 10 AND fare  < 30 THEN 'Mid'
WHEN fare >= 30  THEN 'Expensive'
ELSE NULL END AS fare_group
FROM `ba775-team-6b.Project.passengers`
)
```

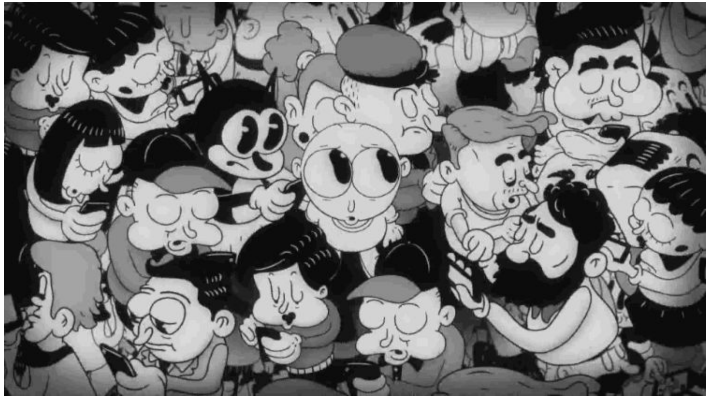
pclass	Number_Passengers	Number_Survivors	Survival_Percentage
1	323	200	0.619195
2	277	119	0.429603
3	709	181	0.255289

pclass	avg_fare
1	87.51
2	21.18
3	13.30

Number of Siblings/Spouses

SibSp	Number_Passengers	Number_Survivors	Survival_Percentage
0	891	309	0.346801
1	319	163	0.510972
2	42	19	0.452381
3	20	6	0.300000
4	22	3	0.136364
5	6	0	0.000000
8	9	0	0.000000

Imagine try to find all your family members on Titanic



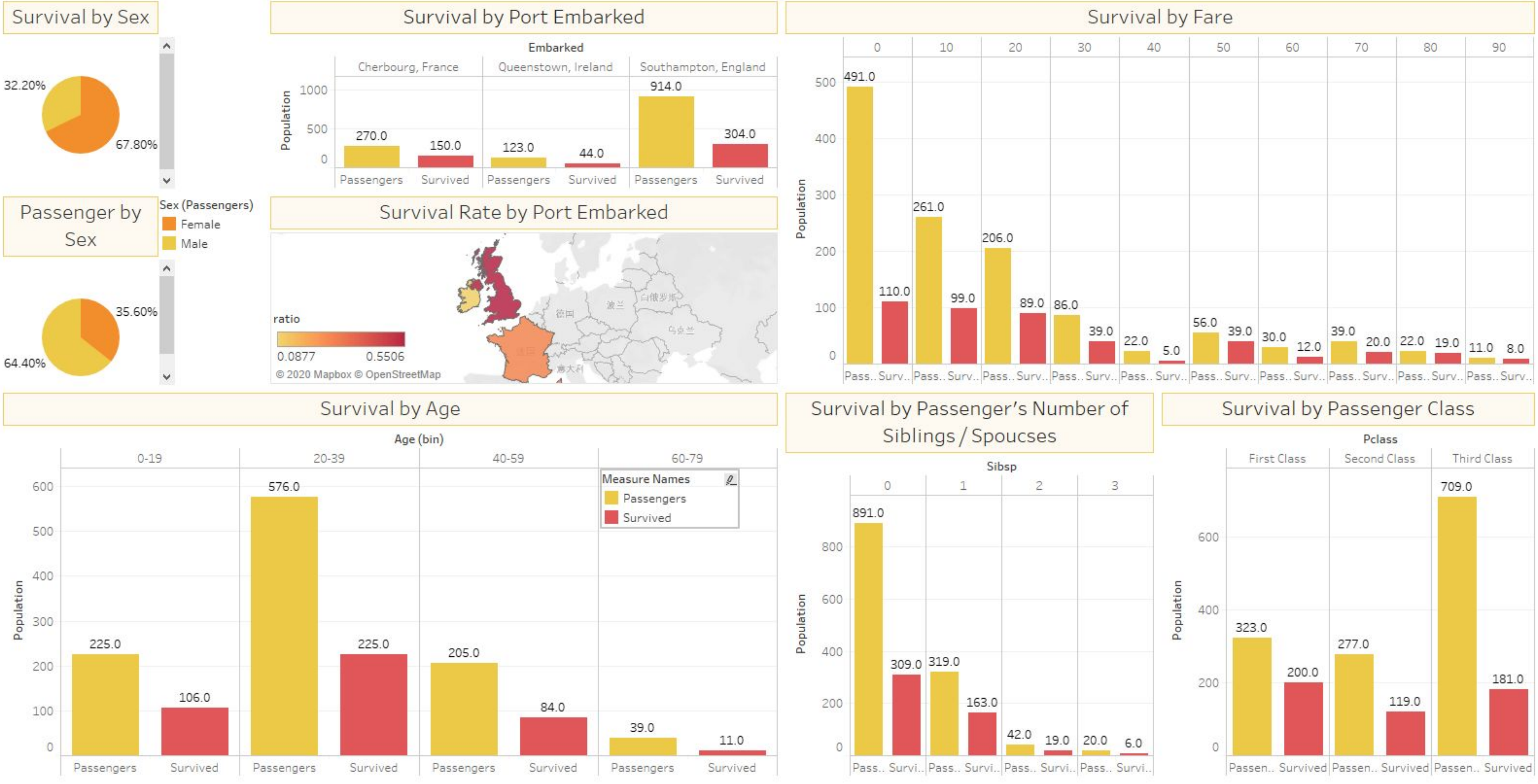
-Passengers who had one Siblings/Spouses, had higher chances of survival than zero or more

-Hypothesis of having one accompany is the sweet spot number under emergency natural disaster. Need more controlled experiments to test on.

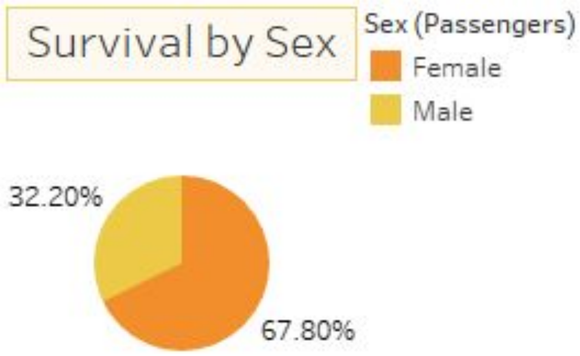
Titanic Dataset: Who Most Likely Survived?

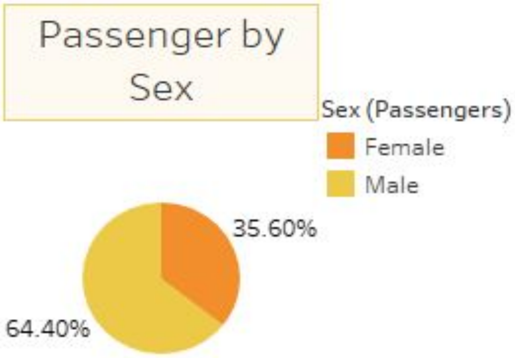
Our team attempted to predict whether passengers would survived the Titanic accident.Since our dataset has two discrete labels, survived and not survived, we are going to solving a binary classification problem.

By Cohort B Team 6: (Sylar)Jiajian Guo, Lequn Yu, Qiqi Tang, Scott McCoy, Tiam Moradi



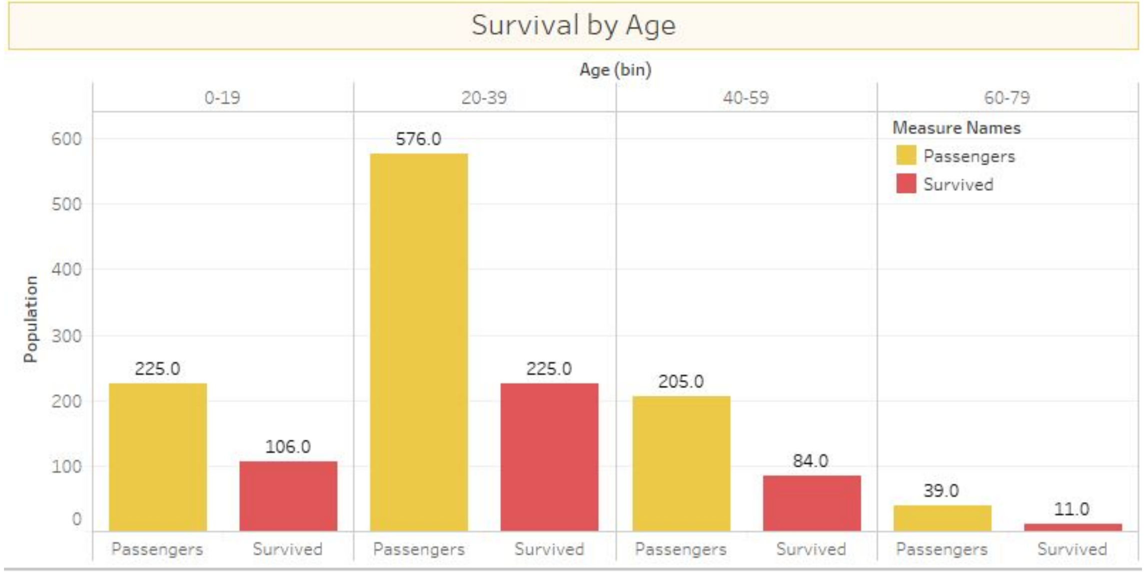
Survival by Sex





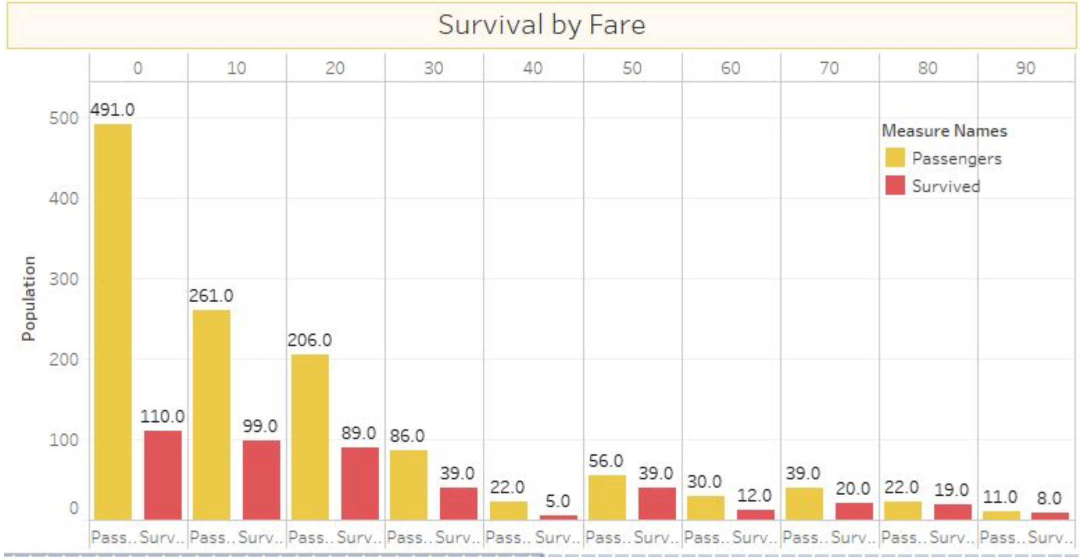
- The survival rate of female passengers is significant higher than that of male passengers.

Survival by Age



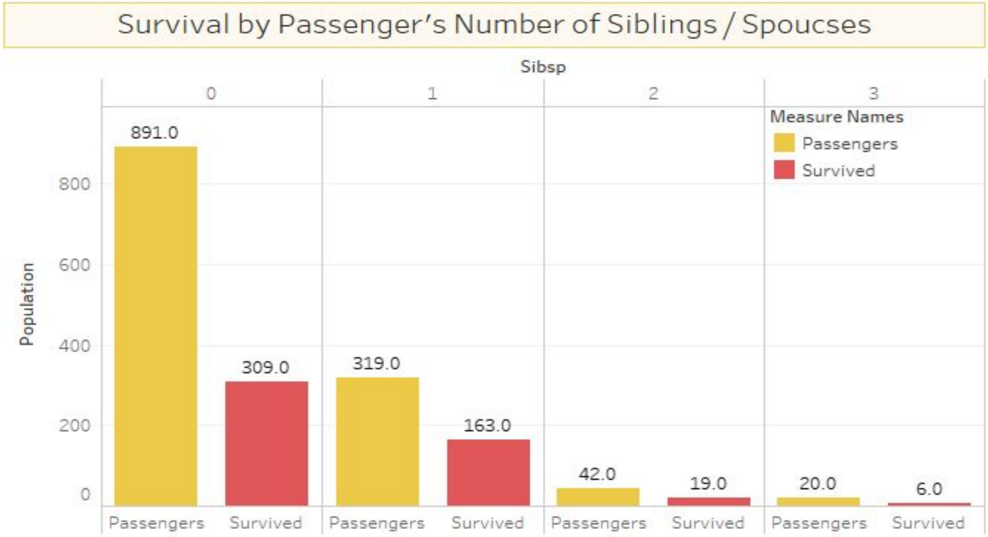
- The survival rate of the minor group (0-19) is the highest.
- The survival rate of the people over 60 is the lowest .

Survival by Fare



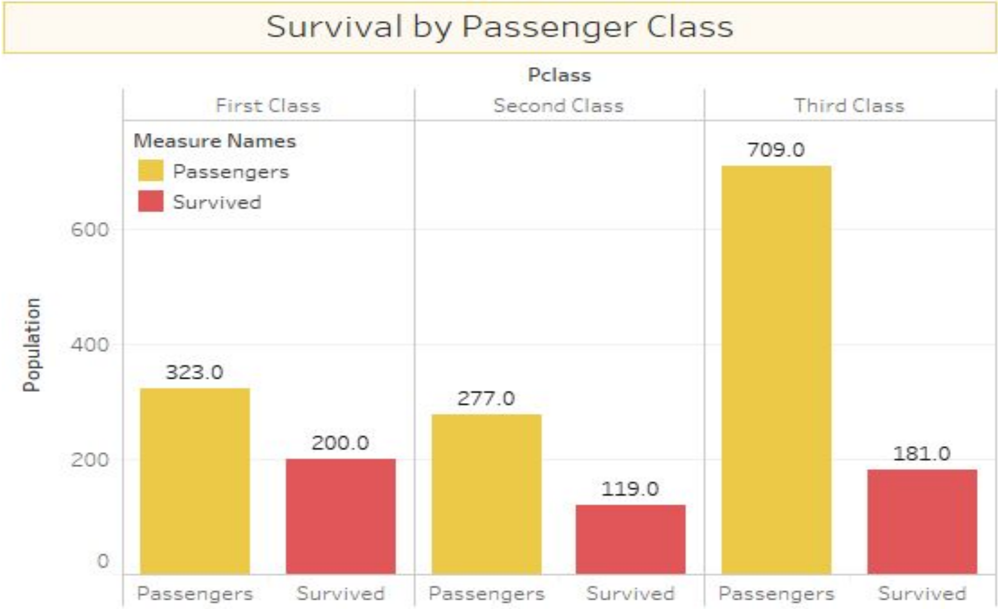
- Higher tickets' prices, Higher survival rate.

Survival by Number of Siblings/Spouses



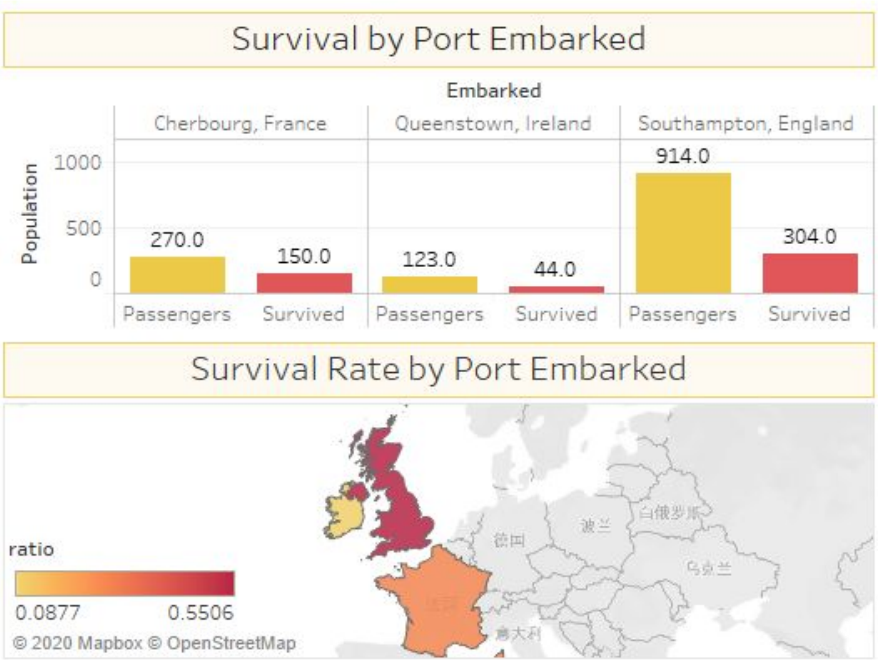
- Going out with 1 partner has the highest survival rate.

Survival by Passenger Class



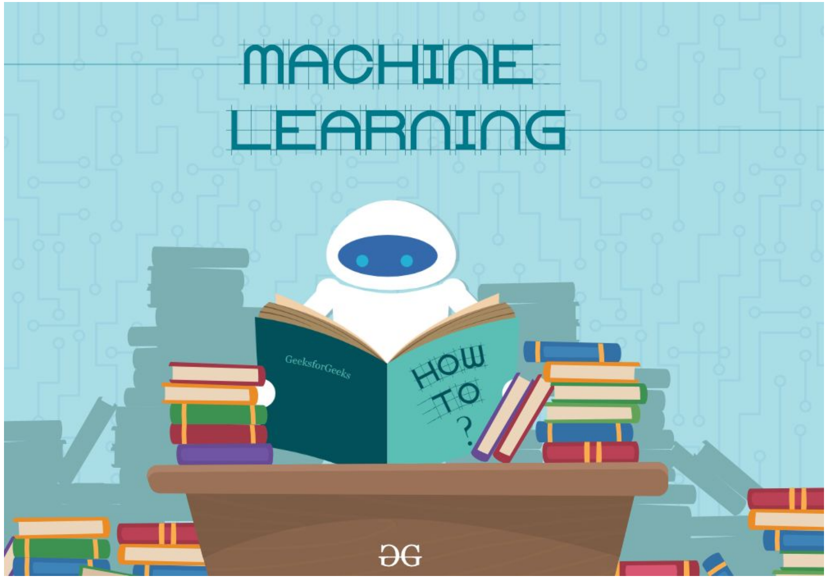
- The survival rate of the third class cabin is the lowest.
- The survival rate of the first class cabin is the highest.

Survival by Port Embarked



- The landing site may eventually be reflected in gender, cabin class and another aspect we talk about early.

Who Survived: Machine Learning Classification



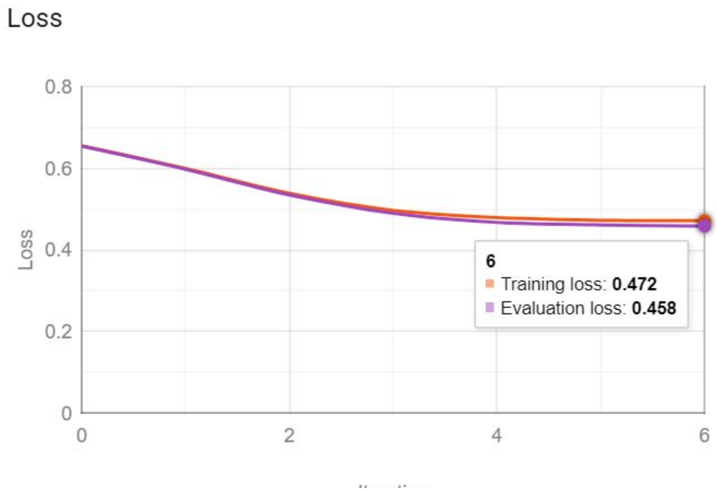
Preprocessing the Data

- Transformed categorical data into one hot encodings.
  - Sex
  - Embarked
  - PClass
- kept numerical features
  - SibSp
  - Parch
  - Fare
  - Age
    - Trained both original values vs scaled values
- Feature Engineering
  - Title
    - Ultimately removed because it hindered performance

Row	Survived	isMale	isFemale	Pclass1	Pclass2	Pclass3
1	0	0	1	0	0	1
2	1	0	1	0	0	1
3	1	0	1	0	0	1
4	0	0	1	0	0	1
5	0	0	1	0	0	1
6	0	0	1	0	0	1
7	1	0	1	0	0	1

Model Performance

- **Logistic Regression**
- Overall, our model is a good baseline score, but can be improved significantly.
- Here are our metrics on our test set.
    - **Accuracy: 79.5%**
    - **AUC ROC: 84.5%**
    - **Precision: 70.4%**
    - **Recall: 71.1%**
    - **F1-Score: 71.1%**
- Regularization prevents overfitting
    - Early Stopping
    - L1 Regularization



Feature Importance: Coefficients

- We can see that isMale, isFemale, Pclass1, Pclass3, and Embarked C have most predictive power.
  - Follows our exploratory analysis

	processed_input	weight
0	isMale	-1.189517
1	isFemale	1.189517
2	Pclass1	0.736785
3	Pclass2	0.000000
4	Pclass3	-0.641205
5	SibSp	-0.292447
6	Parch	0.000000
7	Age	-0.026890
8	Fare	0.000493
9	Embarked_S	-0.239489
10	Embarked_B	0.000000
11	Embarked_C	0.322280
12	__INTERCEPT__	1.015495

- Good amount of features were not impactful..
  - L1 regularization

Limitations

- Amount of data at our disposal.
  - Only 1300 samples and models need more to generalize well.
    - Can’t collect more data.
- Null and Missing values.
  - Certain models can handle these values
  - Imputing the feature, having too many imputed data points can confuse the model.

Conclusions

- We were able to explore and discover key finding that lead to certain passengers survive at a higher rate.
- Developed a machine learning model make predictions based on those features.

Q & A