Titanic ML from dataset

Table of contents

Project Brief

The sinking of the Titanic is one of the most infamous shipwrecks in history.

On April 15, 1912, during her maiden voyage, the widely considered “unsinkable” RMS Titanic sank after colliding with an iceberg. Unfortunately, there weren’t enough lifeboats for everyone onboard, resulting in the death of 1502 out of 2224 passengers and crew. While there was some element of luck involved in surviving, it seems some groups of people were more likely to survive than others.

In this challenge, we ask you to build a predictive model that answers the question: “**what sorts of people were more likely to survive?**” using passenger data (ie name, age, gender, socio-economic class, etc).

Labels

- **PassengerId**: The index values of the passengers who traveled in the Titanic ship
- **Survived:** The survival of the passengers, a categorical data of 0 (No) and 1(yes)

  Pclass: The class at which the passengers are aboard, a categorical data of 1 (Upper), 2 (Middle), and 3(Lower) is in ascending order
- **Name:** The names of the passengers aboard
- **Sex:** The genders of the passengers aboard, categorical data of Male and Female
- **Age:** The age of the passengers aboard
- **SibSp:** The siblings and spouses of the passenger aboard, categorical data
- **Parch:** The children and parents of the passenger aboard, categorical data
- **Ticket:** The passenger ticket number
- **Fare:** The fare of the tickets of the passengers
- **Cabin:** The cabin of the passengers aboard
- **Embarked:** The boarding point, C = Cherbourg; Q = Queenstown; S = Southampton

Dataset

```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
dtype: int64
```

Categorical variables - survived , Pclass, sex, parch, carbin, embarked, name

Numerical variables - passengerId, age, fare

Missing values - age, cabin, embarked

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Execution Order

```
ge of missing values in 'Cabin' column: 77.10437710437711
```

Transformation

	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked	Last_Name
0	0	3	male	22.0	1	0	7.2500	S	Braund
1	1	1	female	38.0	1	0	71.2833	C	Cumings
2	1	3	female	26.0	0	0	7.9250	S	Heikkinen
3	1	1	female	35.0	1	0	53.1000	S	Futrelle
4	0	3	male	35.0	0	0	8.0500	S	Allen
...	...	...	...	...	...	...	...	...	...
886	0	2	male	27.0	0	0	13.0000	S	Montvila
887	1	1	female	19.0	0	0	30.0000	S	Graham
888	0	3	female	28.0	1	2	23.4500	S	Johnston
889	1	1	male	26.0	0	0	30.0000	C	Behr
890	0	3	male	32.0	0	0	7.7500	Q	Dooley

```
test_titanic
[394] ✓ 0.0s
```

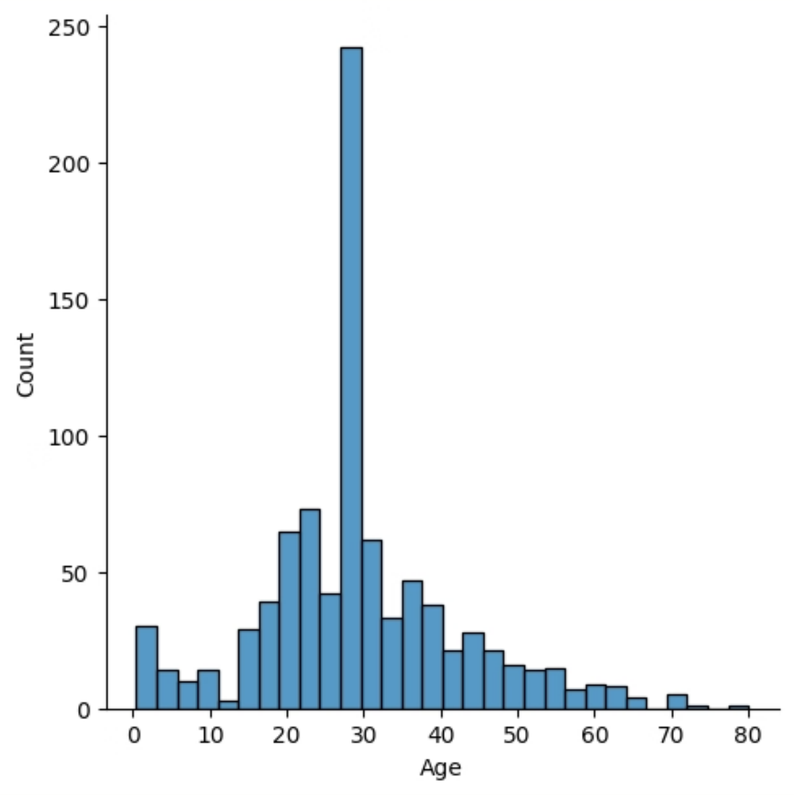
	Survived	Pclass	Age	SibSp	Parch	Fare	last_name_encoded	Embarked_C	Embarked_Q	Embarked_S	Embarked_nan	Sex_female	Sex_male
0	0	3	22.0	1	0	7.2500	73	0.0	0.0	1.0	0.0	0.0	1.0
1	1	1	38.0	1	0	71.2833	136	1.0	0.0	0.0	0.0	1.0	0.0
2	1	3	26.0	0	0	7.9250	251	0.0	0.0	1.0	0.0	1.0	0.0
3	1	1	35.0	1	0	53.1000	198	0.0	0.0	1.0	0.0	1.0	0.0
4	0	3	35.0	0	0	8.0500	11	0.0	0.0	1.0	0.0	0.0	1.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
886	0	2	27.0	0	0	13.0000	406	0.0	0.0	1.0	0.0	0.0	1.0
887	1	1	19.0	0	0	30.0000	221	0.0	0.0	1.0	0.0	1.0	0.0
888	0	3	0.0	1	2	23.4500	293	0.0	0.0	1.0	0.0	1.0	0.0
889	1	1	26.0	0	0	30.0000	52	1.0	0.0	0.0	0.0	0.0	1.0
890	0	3	32.0	0	0	7.7500	159	0.0	1.0	0.0	0.0	0.0	1.0

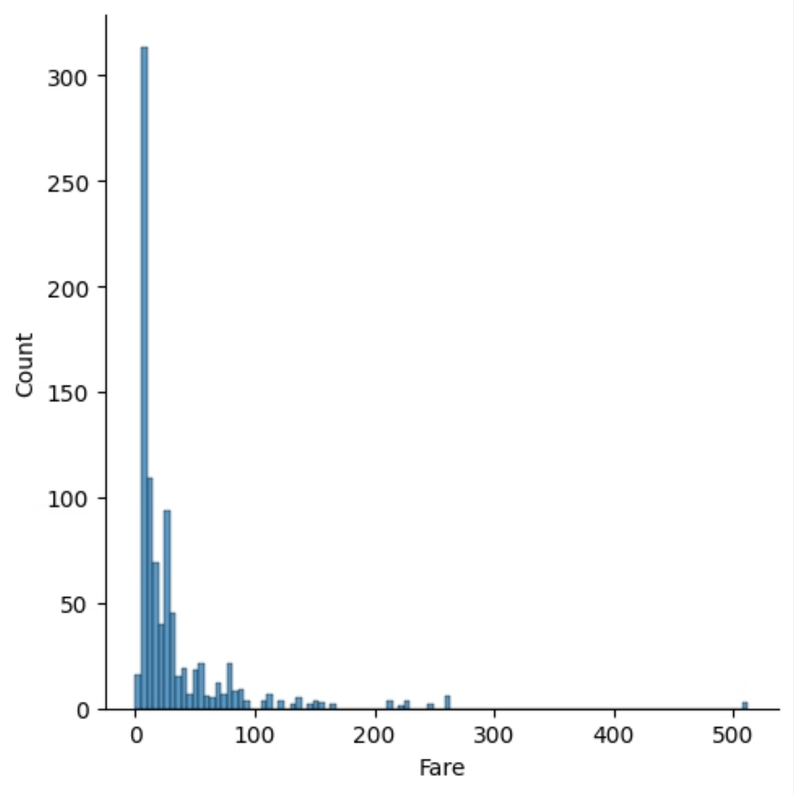
891 rows × 13 columns

Survived	Pclass	SibSp	Parch	Embarked_C	Embarked_Q	Embarked_S	Embarked_nan	Sex_female	Sex_male	last_name_encoded_scaled	Age_bucket	Fare_bucket
0	3	1	0	0.0	0.0	1.0	0.0	0.0	1.0	0.109610	2.0	0.0
1	1	1	0	1.0	0.0	0.0	0.0	1.0	0.0	0.204204	4.0	4.0
1	3	0	0	0.0	0.0	1.0	0.0	1.0	0.0	0.376877	2.0	1.0
1	1	1	0	0.0	0.0	1.0	0.0	1.0	0.0	0.297297	3.0	4.0
0	3	0	0	0.0	0.0	1.0	0.0	0.0	1.0	0.016517	3.0	1.0
...	...	...	...	...	...	...	...	...	...	...	...	...
0	2	0	0	0.0	0.0	1.0	0.0	0.0	1.0	0.609610	2.0	2.0
1	1	0	0	0.0	0.0	1.0	0.0	1.0	0.0	0.331832	1.0	3.0
0	3	1	2	0.0	0.0	1.0	0.0	1.0	0.0	0.439940	0.0	3.0
1	1	0	0	1.0	0.0	0.0	0.0	0.0	1.0	0.078078	2.0	3.0
0	3	0	0	0.0	1.0	0.0	0.0	0.0	1.0	0.238739	3.0	0.0

ows × 13 columns

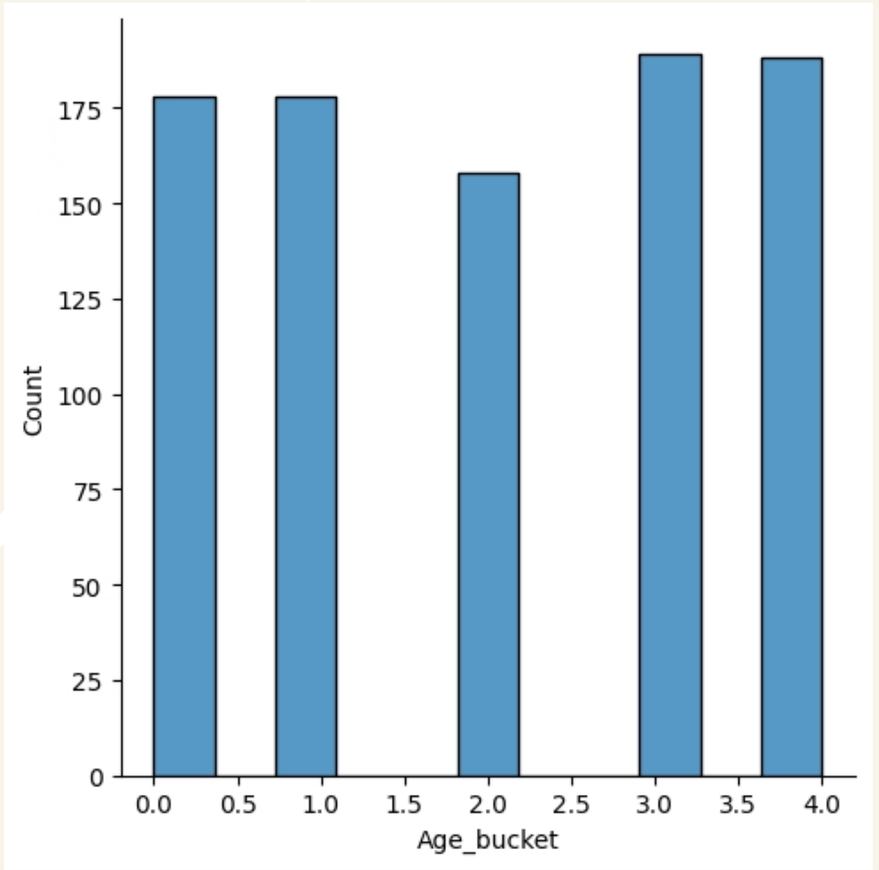
Before transformation

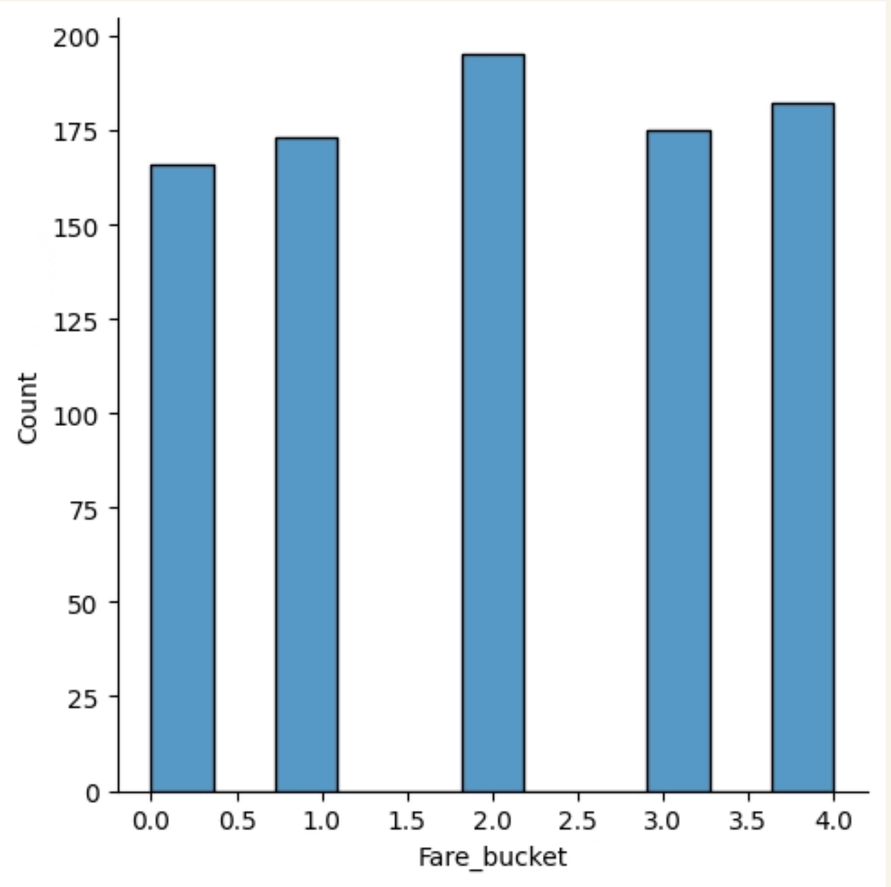




Skewed to the right, uneven distribution with a significant number of outliers.

After transformation

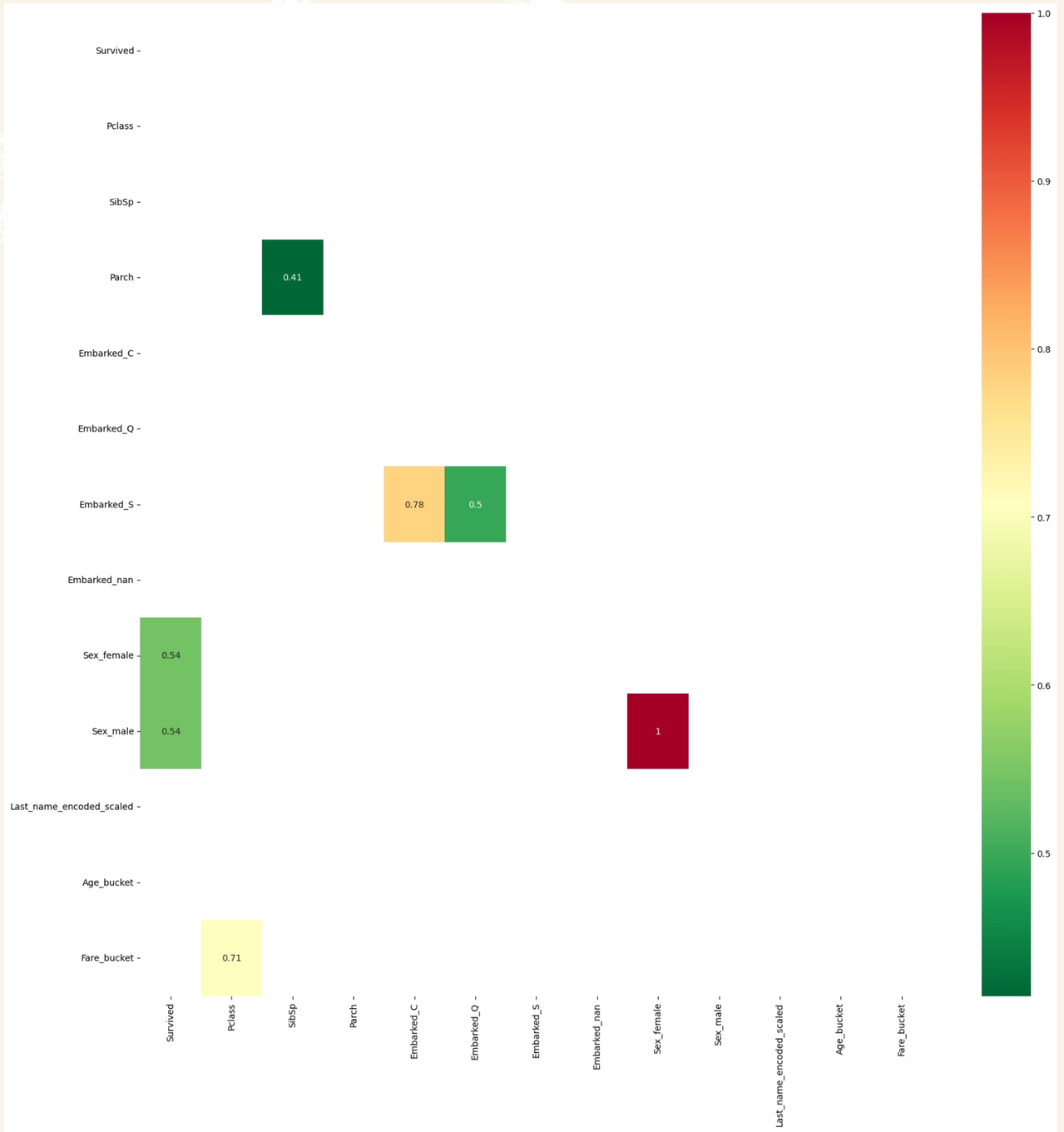




More uniform after bucket quantile processing, with outliers removed.

Correlation matrix

```
Survived                    1.000000
Sex_male                    0.543351
Sex_female                  0.543351
Pclass                      0.338481
Fare_bucket                 0.313809
Embarked_C                  0.168240
Embarked_S                  0.155660
Parch                       0.081629
Age_bucket                  0.066009
Embarked_nan                0.060095
Last_name_encoded_scaled    0.058100
SibSp                       0.035322
Embarked_Q                  0.003650
Name: Survived, dtype: float64
```



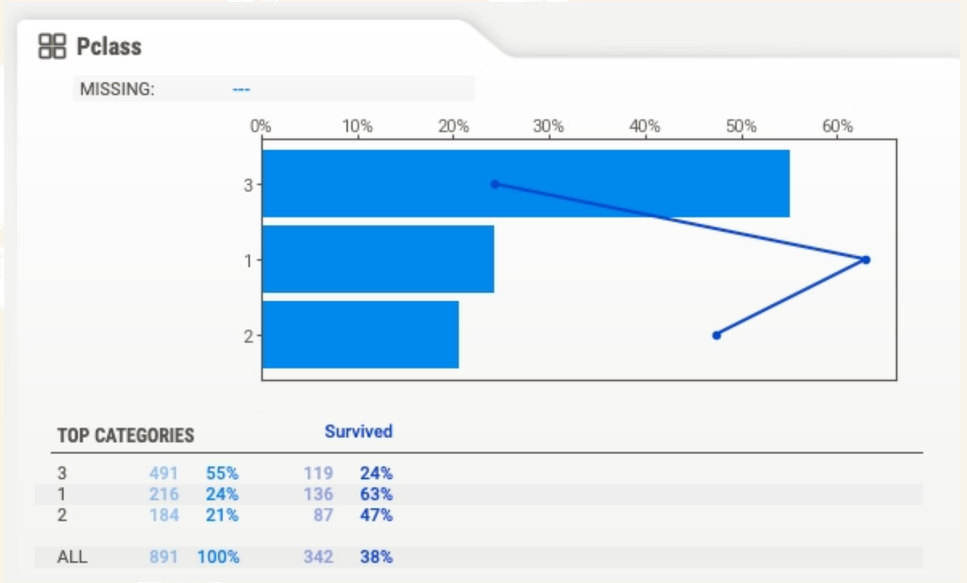
EDA

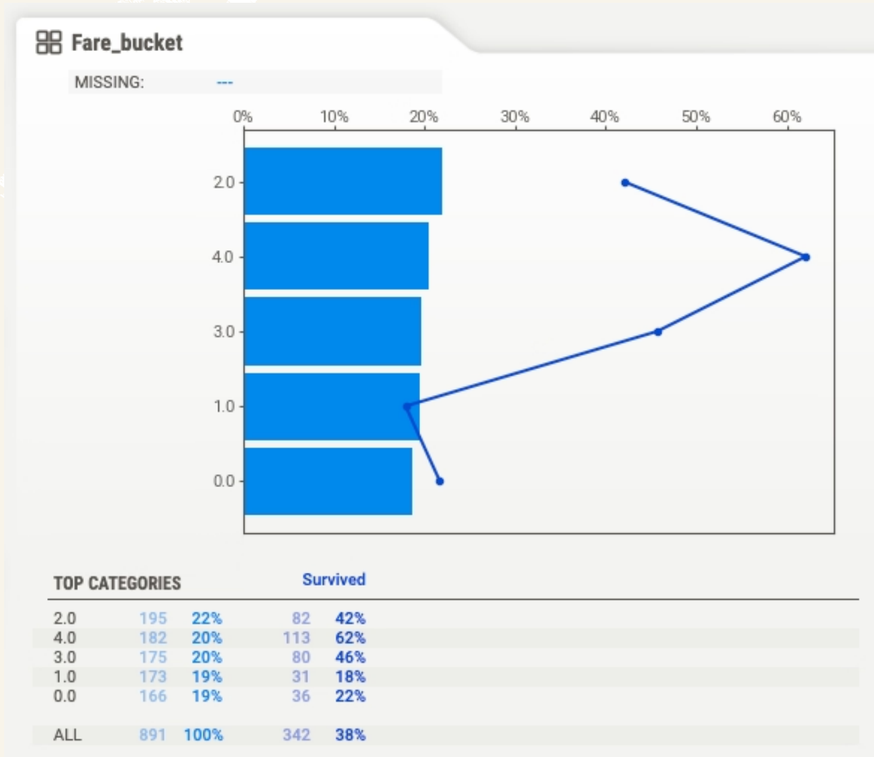
Pclass - P3 represents 55% of the data, with only 24% survival, whereas **P1 represents 24% of the data, with 63% survival.**

- **P1: Upper class**
- P2: Middle class
- P3: Lower class

Fare bucket - All categories are evenly distributed. **C4 has the highest survival rate of 62%**, while C1 has the lowest survival rate of 18%.

- C0: Between 0 and 7.8542
- C1: Between 7.8542 and 10.5
- C2: Between 10.5 and 21.6792
- C3: Between 21.6792 and 39.6875
- **C4: Between 39.6875 and 512.3292**



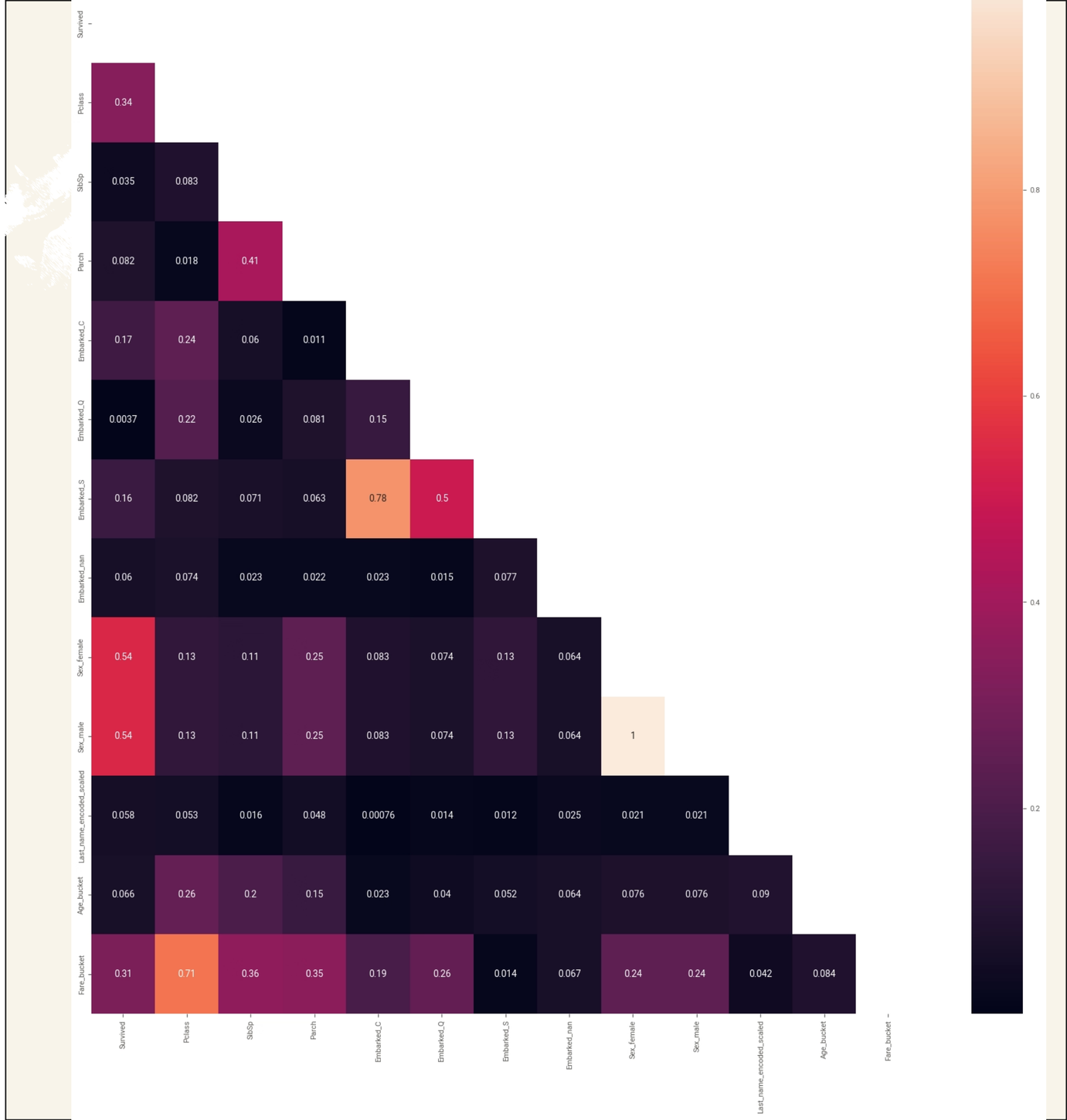


Correlation matrix

Pclass and fare correlation is 0.71

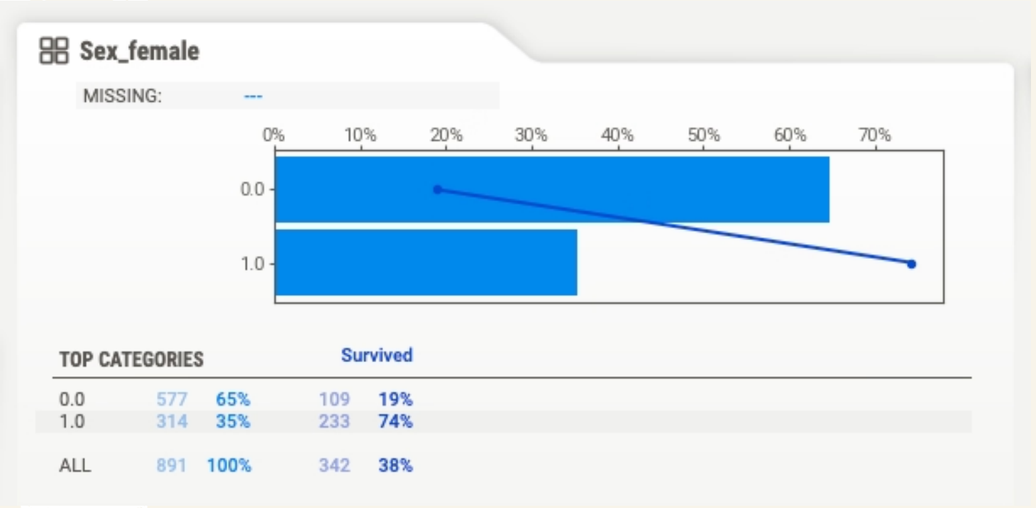
survival and Pclass is 0.34

Survival and Fare is 0.31



EDA

35% of the passengers were female, with a survival rate of 74%, while 65% were male, with only a 19% survival rate.



Evaluation

Mean Absolute Error (MAE) measures the absolute difference between predicted values and actual values. A low MAE indicates that the prediction is accurate.

Mean Squared Error (MSE) calculates the average squared discrepancy between predicted values and actual values. A low MSE suggests that the model's predictions exhibit minimal variance.

Root Mean Squared Error (RMSE) quantifies the average discrepancy between predicted and actual values. A low RMSE signifies high prediction accuracy.

R2 score represents the portion of variance in the target variable that can be anticipated from the features. Achieving a score of 1.00 for both training and testing datasets indicates that the model fits the data perfectly and can account for all the variability in the target variable.

KNN Classification score

```
Mean Absolute Error (MAE): 0.208955223880597
Root Mean Squared Error (RMSE): 0.45711620391383745
R2 Score: 0.12007504690431525
Mean Squared Error (MSE): 0.208955223880597
Accuracy 0.7910447761194029
```

Evaluation

Mean Absolute Error (MAE) measures the absolute difference between predicted values and actual values. A low MAE indicates that the prediction is accurate.

Mean Squared Error (MSE) calculates the average squared discrepancy between predicted values and actual values. A low MSE suggests that the model's predictions exhibit minimal variance.

Root Mean Squared Error (RMSE) quantifies the average discrepancy between predicted and actual values. A low RMSE signifies high prediction accuracy.

R2 score represents the portion of variance in the target variable that can be anticipated from the features. Achieving a score of 1.00 for both training and testing datasets indicates that the model fits the data perfectly and can account for all the variability in the target variable.

Decision tree(gini) score

```
Mean Absolute Error (MAE): 0.2537313432835821
Root Mean Squared Error (RMSE): 0.5037175233040658
R2 Score: -0.06848030018761708
Mean Squared Error (MSE): 0.2537313432835821
Accuracy 0.746268656716418
```

Decision tree(depth) score

```
Mean Absolute Error (MAE): 0.22388059701492538
Root Mean Squared Error (RMSE): 0.47316022340738384
R2 Score: 0.05722326454033788
Mean Squared Error (MSE): 0.22388059701492538
Accuracy 0.7761194029850746
```

Decision tree(entropy) score

```
Mean Absolute Error (MAE): 0.25
Root Mean Squared Error (RMSE): 0.5
R2 Score: -0.05276735459662274
Mean Squared Error (MSE): 0.25
Accuracy 0.75
```

Evaluation

Mean Absolute Error (MAE) measures the absolute difference between predicted values and actual values. A low MAE indicates that the prediction is accurate.

Mean Squared Error (MSE) calculates the average squared discrepancy between predicted values and actual values. A low MSE suggests that the model's predictions exhibit minimal variance.

Root Mean Squared Error (RMSE) quantifies the average discrepancy between predicted and actual values. A low RMSE signifies high prediction accuracy.

R2 score represents the portion of variance in the target variable that can be anticipated from the features. Achieving a score of 1.00 for both training and testing datasets indicates that the model fits the data perfectly and can account for all the variability in the target variable.

SVM linear

```
Mean Absolute Error (MAE): 0.2126865671641791
Root Mean Squared Error (RMSE): 0.4611795389695635
R2 Score: 0.10436210131332091
Mean Squared Error (MSE): 0.2126865671641791
Accuracy 0.7873134328358209
```

SVM sigmoid

```
Mean Absolute Error (MAE): 0.3805970149253731
Root Mean Squared Error (RMSE): 0.6169254532967279
R2 Score: -0.6027204502814256
Mean Squared Error (MSE): 0.3805970149253731
Accuracy 0.6194029850746269
```

SVM rbf

```
Mean Absolute Error (MAE): 0.2126865671641791
Root Mean Squared Error (RMSE): 0.4611795389695635
R2 Score: 0.10436210131332091
Mean Squared Error (MSE): 0.2126865671641791
Accuracy 0.7873134328358209
```

If I had more time, I would

- Further research ML algorithms to explore methods for enhancing the score.
- Experiment more with feature removal to refine the model.
- Consider using different classification models such as logistic regression and naive Bayes classification for comparison.

Thank you

Pitch

Want to make a presentation like this one?

Start with a fully customizable template, create a beautiful deck in minutes, then easily share it with anyone.

Create a presentation (It's free)