Supervised Machine Learning: Classification

Course Project

Evgeny Zorin

6. October 2021

Titanic - Machine Learning from Disaster

Abstract

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.
This project will try to answer the question: “what sorts of people were more likely to survive?” using passenger data (e.g. name, age, gender, socio-economic class, etc).

Main Objective

The main objective of the analysis is to interpret the features in the dataset, which can be used to build the most accurate predictive model based on the classification algorithms.

Data Summary

The data was taken from Kaggle and it consists of two separated datasets:

1. Training dataset contains the 891 observations of Titanic passengers and 12 features (including target feature) as described in Table 1.
2. Testing dataset contains the 418 observations of Titanic passengers and the same 11 features (without target feature) as in Train dataset.

There are 177 null values in the training dataset and 86 null values in the testing dataset in **Age** column, 687 null values in the training dataset and 327 null values in the testing dataset in **Cabin** column, 2 null values in the training dataset in **Embarked** column and 1 null value in the testing dataset in **Fare** column. There are no duplicated values in these two datasets.

Table 1: Columns and Descriptions

Column	Description
PassengerId	Passenger ID
Survived	Survival: 0 = No, 1 = Yes
Pclass	Ticket class: 1 = 1st, 2 = 2nd, 3 = 3rd
Name	Name of the passenger
Sex	Sex: Female, Male
Age	Age of the passenger in years
SibSp	Number of siblings / spouses aboard the Titanic
Parch	Number of parents / children aboard the Titanic
Ticket	Ticket number
Fare	Passenger fare
Cabin	Cabin number
Embarked	Port of Embarkation: C = Cherbourg, Q = Queenstown, S = Southampton

Summary Statistics

Summary Statistics of Numerical Features

Testing dataset is represented by 891 passengers, which is 40% of the total number of Titanic passengers (2,224). Target feature **Survived** is a categorical feature with two values: 0 = didn't survived, 1 = survived. In the testing dataset only 38.4% of passengers survived. Age feature is also a categorical feature and is represented by a distribution from 0 to 80 years old. The average age of the passengers is 30 years old. SibSp and Parch features show that more than 75% of the passengers traveled alone. Fare feature shows that about 75% of the tickets were under $31, and the maximum price for a single ticket was $512. It can be assumed that tickets between these values were purchased by the passengers traveling in first class. Summary statistics of the numerical features is shown in Table 2.

Table 2: Summary Statistics of Numerical Features

	count	mean	std	min	25%	50%	75%	max
PassengerId	891.0	446.000000	257.353842	1.00	223.5000	446.0000	668.5	891.0000
Survived	891.0	0.383838	0.486592	0.00	0.0000	0.0000	1.0	1.0000
Pclass	891.0	2.308642	0.836071	1.00	2.0000	3.0000	3.0	3.0000
Age	714.0	29.699118	14.526497	0.42	20.1250	28.0000	38.0	80.0000
SibSp	891.0	0.523008	1.102743	0.00	0.0000	0.0000	1.0	8.0000
Parch	891.0	0.381594	0.806057	0.00	0.0000	0.0000	0.0	6.0000
Fare	891.0	32.204208	49.693429	0.00	7.9104	14.4542	31.0	512.3292

Summary Statistics of Categorical Features

Name feature represents all the passengers on board. Sex feature is a categorical feature with two values: male (65.2% passengers) and female (34.8%). Ticket feature has 210 duplicated values (23.6%). Cabin feature also has 83.5% of duplicated values as well as about 77% of missing values. Embarked feature is a categorical feature with 3 values and the most represented value is 'S' with 644 observations. Summary statistics of the categorical features is shown in Table 3.

Table 3: Summary Statistics of Categorical Features

	count	unique	top	freq
Name	891	891	Mitchell, Mr. Henry Michael	1
Sex	891	2	male	577
Ticket	891	681	CA. 2343	7
Cabin	204	147	C23 C25 C27	4
Embarked	889	3	S	644

Data Wrangling

Data wrangling workflow consists of two main processes:

- Missing Values Handling
- Data Cleaning

Missing Values Handling

There are a lot of missing values in Age feature in the training and testing datasets. It would be wrong to fill them as an average or a most common value. The best way to solve this problem is to categorize Age feature, to get the average value for each category, and then to fill the missing values with these average values. To perform this operation will be used Name feature which includes the title (e.g. Mr, Mrs, Miss, Dr, etc.).
There are only two missing values in Embarked feature in the training dataset. It was decided to fill them with the most frequent value which is "S".
There is only one missing value in Fare feature in the testing dataset. It was decided to fill it with the most common value.

Data Cleaning

Features such as **PassengerId, Name, Ticket** and **Cabin** were dropped from the dataset, since they don't provide any useful information for the prediction or interpretation.

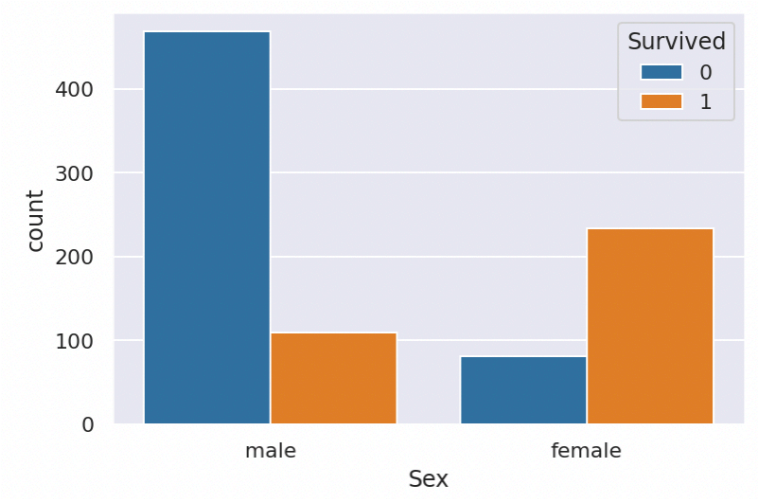
Exploratory Data Analysis

Understanding the distribution of independent features should help to build a more accurate prediction model.

Sex

Despite the fact that females represent only a third of the total number of passengers, their survival rate (74.2%) is almost 4 times higher than that of males (18.9%). Based on this, it can be assumed that Sex feature will have the greatest impact on prediction. Distribution of Sex feature is shown in Figure 1.

Figure 1: Sex Distribution



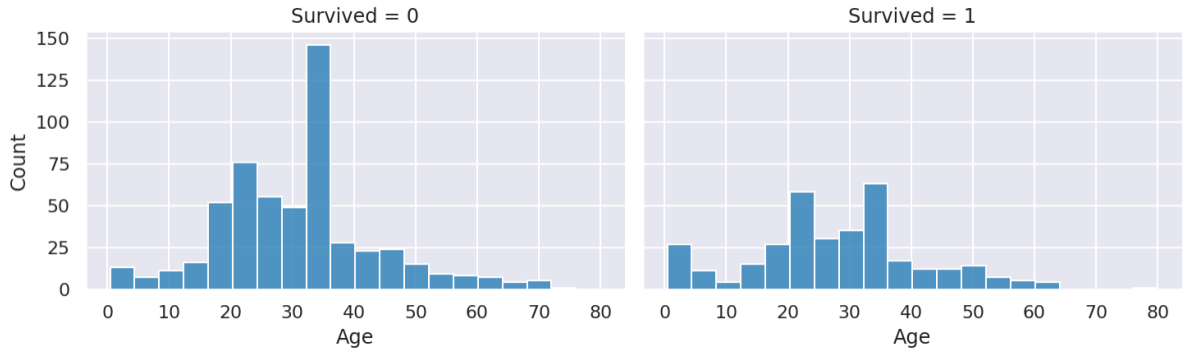
Survived	0	1	All
Sex			
female	81	233	314
male	468	109	577
All	549	342	891

Survived	
Sex	
female	0.742038
male	0.188908

Age

To analyze Age feature, which has continuous values, it was first splitted into 5 categories of 16 years each. Distribution of this feature (Figure 2) shows that passengers under 16 years old have the highest (54.8%), and passengers 64 years old and above, have the lowest survival rate (9.1%). Passengers between 16 and 64 years old have survival rate between 32% and 43.5%. It can be assumed that Age feature will also have high impact on prediction.

Figure 2: Age Distribution



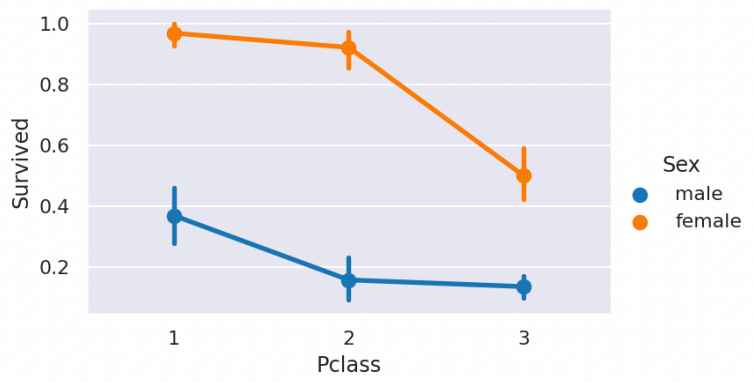
Survived	0	1	All
Age			
(0.34, 16.336]	47	57	104
(16.336, 32.252]	232	150	382
(32.252, 48.168]	221	104	325
(48.168, 64.084]	39	30	69
(64.084, 80.0]	10	1	11
All	549	342	891

Survived	
Age	
(0.34, 16.336]	0.548077
(16.336, 32.252]	0.392670
(32.252, 48.168]	0.320000
(48.168, 64.084]	0.434783
(64.084, 80.0]	0.090909

Pclass

Distribution of Pclass feature (Figure 3) clearly shows that the higher the socio-economic class of passengers, the higher their survival rate. It can be assumed that Pclass feature will have very high impact on prediction.

Figure 3: Pclass Distribution



Survived	0	1	All
Pclass			
1	80	136	216
2	97	87	184
3	372	119	491
All	549	342	891

	Survived
Pclass	
1	0.629630
2	0.472826
3	0.242363

Embarked

Distribution of Embarked feature (Figure 4) shows that passengers who have boarded at the port of Cherbourg have the highest survival rate. This feature alone does not provide any objective information, so later it will be required to cross-analyze this feature with Pclass feature. It can be assumed that such a high survival rate is probably due to the large number of first-class passengers boarding in this port.

Figure 4: Embarked Distribution

Survived	0	1	All
Embarked			
C	75	93	168
Q	47	30	77
S	427	219	646
All	549	342	891

	Survived
Embarked	
C	0.553571
Q	0.389610
S	0.339009

SibSp

Distribution of SibSp feature (Figure 5) shows that passengers who traveled without siblings or spouses had a lower survival rate than passengers who traveled with family members. It can be assumed that most of these passengers were from the third class.

Figure 5: SibSp Distribution

Survived	0	1	All
SibSp			
0	398	210	608
1	97	112	209
2	15	13	28
3	12	4	16
4	15	3	18
5	5	0	5
8	7	0	7
All	549	342	891

	Survived
SibSp	
0	0.345395
1	0.535885
2	0.464286
3	0.250000
4	0.166667
5	0.000000
8	0.000000

Parch

Distribution of Parch feature (Figure 6) shows similar results to SibSp feature, where passengers who traveled with family members have a higher survival rate. It is likely to create a new feature that represents the family size of the passenger on board by combining these two features.

Figure 6: Parch Distribution

Survived	0	1	All
Parch			
0	445	233	678
1	53	65	118
2	40	40	80
3	2	3	5
4	4	0	4
5	4	1	5
6	1	0	1
All	549	342	891

	Survived
Parch	
0	0.343658
1	0.550847
2	0.500000
3	0.600000
4	0.000000
5	0.200000
6	0.000000

Fare

To analyze Fare feature, which has continuous values, it was decided to divide it into four quantiles. The results (Figure 7) clearly show that the higher the fare, the higher the survival rate. This supports the assumption that passengers from the high socio-economic class have the highest survival rate.

Figure 7: Fare Distribution

Survived	0	1	All
Fare			
(-0.001, 7.91]	179	44	223
(7.91, 14.454]	156	68	224
(14.454, 31.0]	121	101	222
(31.0, 512.329]	93	129	222
All	549	342	891

	Survived
Fare	
(-0.001, 7.91]	0.197309
(7.91, 14.454]	0.303571
(14.454, 31.0]	0.454955
(31.0, 512.329]	0.581081

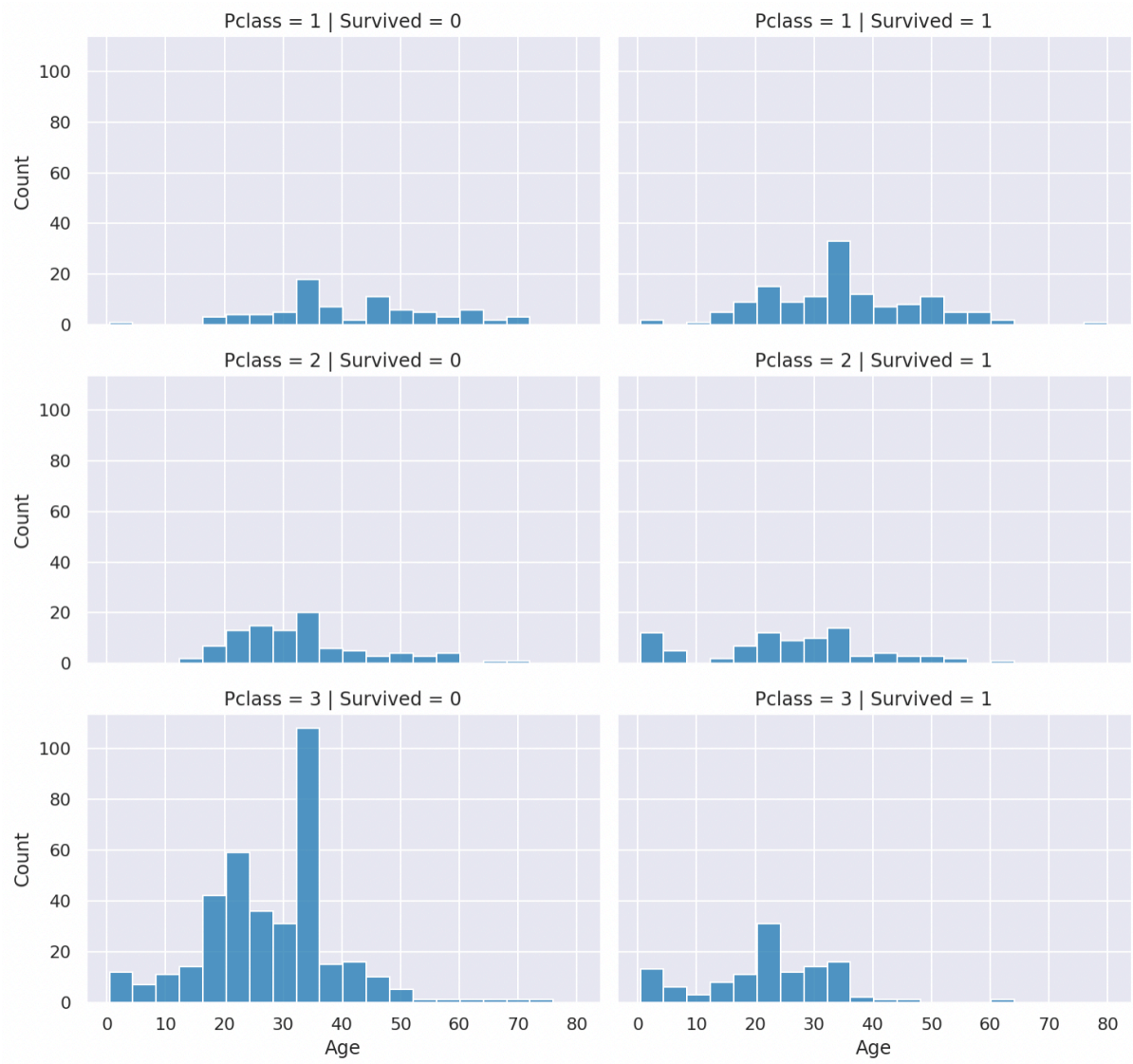
Exploratory Data Cross-Analysis

Additionally it was decided to analyze the combination of multiple features to find some patterns and correlation between them.

Age + Pclass

Distribution of Age and Pclass features (Figure 8) shows that the majority of passengers who did not survive were between the ages of 30 and 40 in all three socio-economic classes. In contrast, the majority of passengers who survived from the first class were also between the ages of 30 and 40.

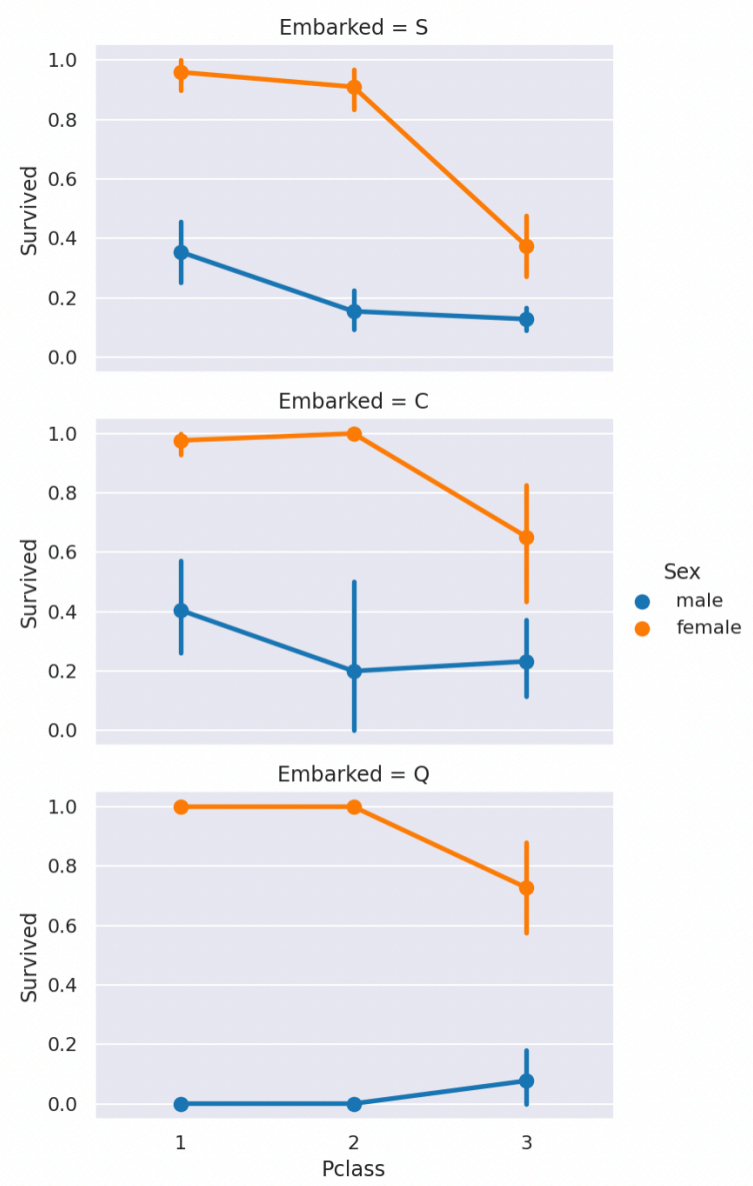
Figure 8: Age and Pclass Distribution



Pclass + Embarked

Distribution of Pclass and Embarked features (Figure 9) shows that, as previously assumed, passengers who boarded at the port of Cherbourg have the highest survival rate. Moreover, this is true for passengers in all three classes. Females in first and second classes have a survival rate between 90% and 100%, regardless of port of embarkation. Surprisingly, all the males in first and second classes who boarded at the port of Queenstown did not survive. It can be assumed that port of embarkation is a minor feature and its impact on the prediction will be relatively weak, while the main features are Sex and Pclass.

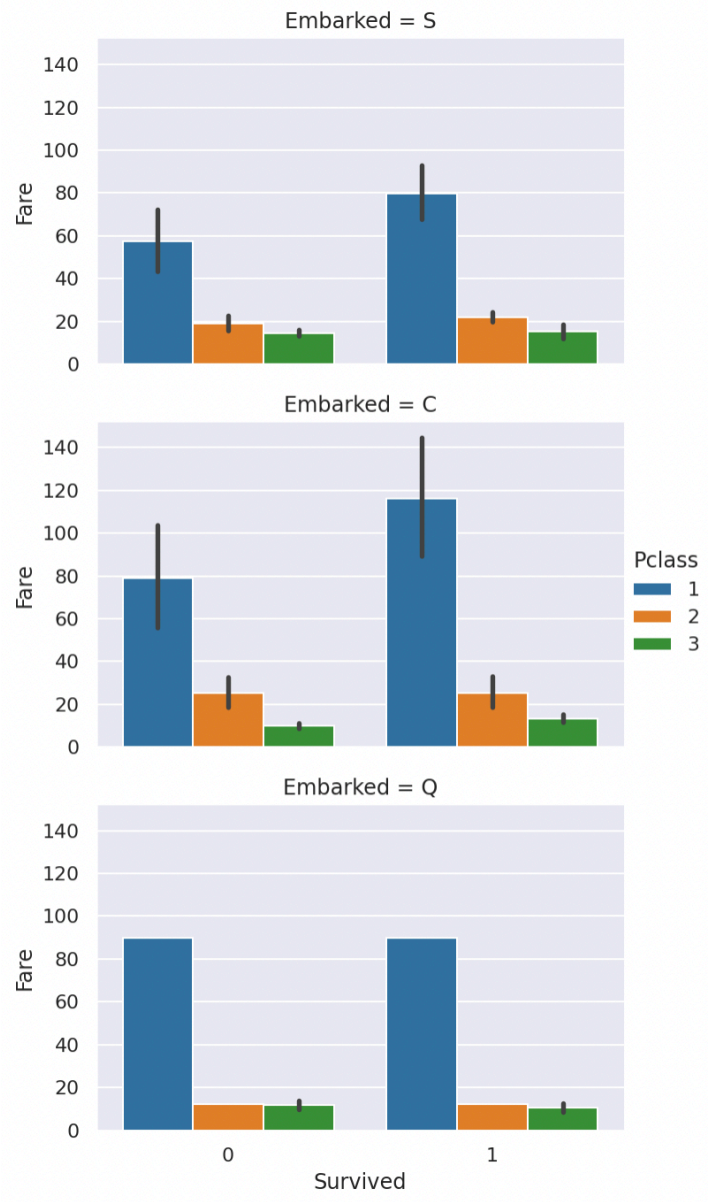
Figure 9: Pclass and Embarked Distribution



Embarked + Fare

Distribution of Embarked and Fare features (Figure 10) shows that port of embarkation correlates with survival rate. But this is true only for the first class passengers. At the ports of Cherbourg and Southampton, it was possible to purchase more expensive first class fares. Accordingly, whoever paid more survived. For the other classes, there is no relationship between port of embarkation and fares.

Figure 10: Embarked and Fare Distribution



Feature Engineering

As mentioned in the analysis above, SibSp and Parch features show a similar distribution. Therefore, it was decided to combine them into a new feature called Fsize, which represents family size. It can be assumed that a passenger with up to 3 family members on board has a higher survival rate than a passenger traveling alone. In the opposite, a passenger with 4 and more family members on board has the lowest survival rates.

New Feature Creating: Fsize

Distribution of Fsize feature (Figure 11) confirms that passengers traveling alone have a lower survival rate (30%) compared to passengers with three family members on board (72%). Starting with the fourth family member, survival rate drops to 20%. Therefore, it was decided to split this feature into 5 categories (from 0 to 4 family members on board).

Figure 11: Fsize Distribution

Survived	0	1	All
Fsize			
0	374	163	537
1	72	89	161
2	43	59	102
3	8	21	29
4	12	3	15
5	19	3	22
6	8	4	12
7	6	0	6
10	7	0	7
All	549	342	891

	Survived
Fsize	
0	0.303538
1	0.552795
2	0.578431
3	0.724138
4	0.200000
5	0.136364
6	0.333333
7	0.000000
10	0.000000

Ordinal Encoding

To build a predictive model, it is required to encode all categorical and continuous features into numerical features. Encoding categorical features converts each value in a feature into a numeric value, starting from zero. Encoding continuous features, these features are divided into categories, and then each category is converted into a numeric value.

Sex, Embarked, Title

Ordinal encoding was performed on these categorical features as follows:

Sex: male = 0, female = 1
Embarked: S = 0, C = 1, Q = 2
Title: Master = 0, Miss = 1, Mr = 2, Mrs = 3, Other = 4

Fsize

Ordinal encoding was performed on this numerical features as follows:

0 family members = 0
1 family members = 1
2 family members = 2
3 family members = 3
4 and more family members = 4

Age

Ordinal encoding was performed on this continuous feature as follows:

0 - 16 years old = 0
16 - 32 years old = 1
32 - 48 years old = 2
48 - 64 years old = 3
64 - 80 years old = 4

Fare

Ordinal encoding was performed on this continuous feature as follows:

$0 - $7.91 = 0
$7.91 - $14.454 = 1
$14.454 - $31 = 2
$31 - $512 = 3

Samples of training and testing datasets after encoding of all independent features are shown in Table 4 and Table 5.

Table 4: Training Dataset Sample

	Survived	Pclass	Sex	Age	Fare	Embarked	Title	Fsize
0	0	3	0	1	0	0	2	1
1	1	1	1	2	3	1	3	1
2	1	3	1	1	1	0	1	0
3	1	1	1	2	3	0	3	1
4	0	3	0	2	1	0	2	0
5	0	3	0	2	1	2	2	0
6	0	1	0	3	3	0	2	0
7	0	3	0	0	2	0	0	4
8	1	3	1	1	1	0	3	2
9	1	2	1	0	2	1	3	1

Table 5: Testing Dataset Sample

	Pclass	Sex	Age	Fare	Embarked	Title	Fsize
0	3	0	2	0	2	2	0
1	3	1	2	0	0	3	1
2	2	0	3	1	2	2	0
3	3	0	1	1	0	2	0
4	3	1	1	1	0	3	2
5	3	0	0	1	0	2	0
6	3	1	1	0	2	1	0
7	2	0	1	2	0	2	2
8	3	1	1	0	1	3	0
9	3	0	1	2	0	2	2

Modeling and Prediction

To build a predictive model, the data will be trained on the following classification algorithms:

- Logistic Regression
- Support Vector Machine
- K-Nearest Neighbors
- Decision Tree
- Random Forest

At the end, feature coefficients will be examined and interpreted to understand their impact on the prediction.

Logistic Regression

Logistic Regression is a useful model to run early in the workflow. Logistic regression measures the relationship between the categorical dependent feature and one or more independent features by estimating probabilities using a logistic function, which is the cumulative logistic distribution.
The accuracy score obtained by the model based on the training dataset is 79.69.

Support Vector Machine

Support Vector Machine (SVM) is supervised learning model with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training samples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new test samples to one category or the other, making it a non-probabilistic binary linear classifier.
The accuracy score obtained by the model based on the training dataset is 83.17, which is higher than Logistic Regression model.

K-Nearest Neighbors

K-Nearest Neighbors algorithm (K-NN) is a non-parametric method used for classification and regression. A sample is classified by a majority vote of its neighbors, with the sample being assigned to the class most common among

its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.
The accuracy score obtained by the model based on the training dataset is 85.07, which is higher than previous two models.

Decision Tree

Decision Tree model uses a decision tree as a predictive model which maps features (tree branches) to conclusions about the target value (tree leaves). Tree models where the target variable can take a finite set of values are called classification trees; in these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees.
The accuracy score obtained by the model based on the training dataset is 88.78, which is the highest among models evaluated so far.

Random Forest

Random Forest is one of the most popular models. Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees (n_estimators=100) at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.
The accuracy score obtained by the model based on the training dataset is also 88.78.

Model Evaluation and Prediction

Evaluating all five models based on their accuracy scores (Table 6) it can be seen that both the Decision Tree and Random Forest models have the highest accuracy scores. It was decided to use the Random Forest model for prediction because it corrects the Decision Tree habit of overfitting the training set.

Table 6: Model Evaluation

	Model	Score
0	Decision Tree	0.887767
1	Random Forest	0.887767
2	K-Nearest Neighbors	0.850730
3	Support Vector Machine	0.831650
4	Logistic Regression	0.796857

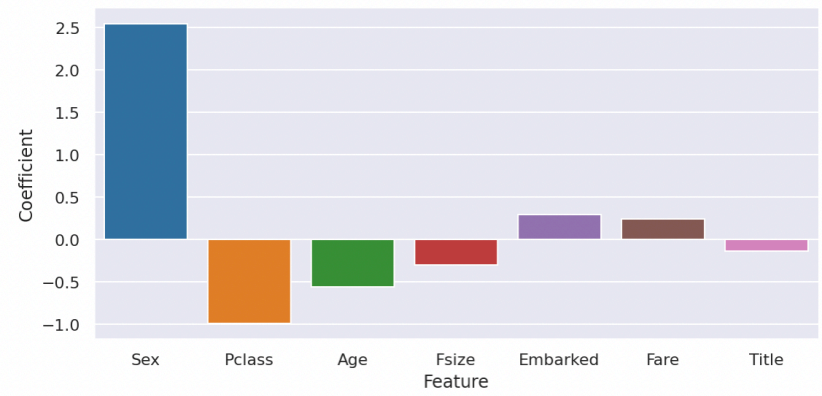
Key Findings and Insights

Logistic Regression will be used to validate the assumptions and decisions made in the feature engineering by calculating the coefficient of the features in the decision function.

Positive coefficients increase the log-odds of the response (and thus increase the probability), and negative coefficients decrease the log-odds of the response (and thus decrease the probability).

Distribution of the coefficients is shown in Figure 12.

Figure 12: Coefficients Distribution



	Feature	Coefficient
0	Sex	2.540811
1	Pclass	-0.994058
2	Age	-0.559350
3	Fsize	-0.305956
4	Embarked	0.288323
5	Fare	0.236309
6	Title	-0.142510

Features Interpretation

The Sex feature has the greatest impact on the survival rate. Not surprising when considering that of 34.8% of all females, more than 74% survived, while of 65.2% of all males, less than 19% survived.

The Pclass feature has the greatest negative impact on the survival rate. The lower the ticket class, the lower the chance to survive.
The Age feature has a relatively low coefficient. This is probably due to the fact that it was splitted into only 5 categories of 16 years each, and increasing the number of categories could have increased the coefficient.
Surprisingly, the Fsize feature has a negative impact on the survival rate. From the analysis of this feature, it was seen that the larger the family size, the higher the survival rate. The coefficient of the feature indicates the opposite. Probably another feature should have been created, consisting of two categories - whether the passenger travels alone or not.
The Embarked and Fare features confirm the assumption that wealthier passengers have a higher survival rate, although these features do not have a strong enough impact on the survival rate.
The Title feature also confirms the assumption that children and females have a higher survival rate. This feature has the weakest impact on the survival rate.

Next Steps

The next steps to achieve better model prediction would be to create other artificial features such as **Pclass*Age** and **Pclass*Embarked** to better understand the impact of these two features together on the target feature, and to split Fsize feature into only two categories - whether the passenger travels alone or not. Using GridSearchCV to find the optimal parameters and Ensembling such as Voting Classifier, Bagging and Boosting would help to achieve better results either.

Appendix

GitHub URL:
https://github.com/evgenyzorin/IBM-Machine-Learning/tree/main/Classification