Experiment No. 2
Analyze the Titanic Survival Dataset and Apply appropriate Regression Technique
Date of Performance:29/07/24
Date of Submission:

Aim: Analyze the Titanic Survival Dataset and Apply appropriate Regression Technique.

Objective: Able to perform various feature engineering tasks, apply logistic regression on the given dataset and maximize the accuracy.

Theory:

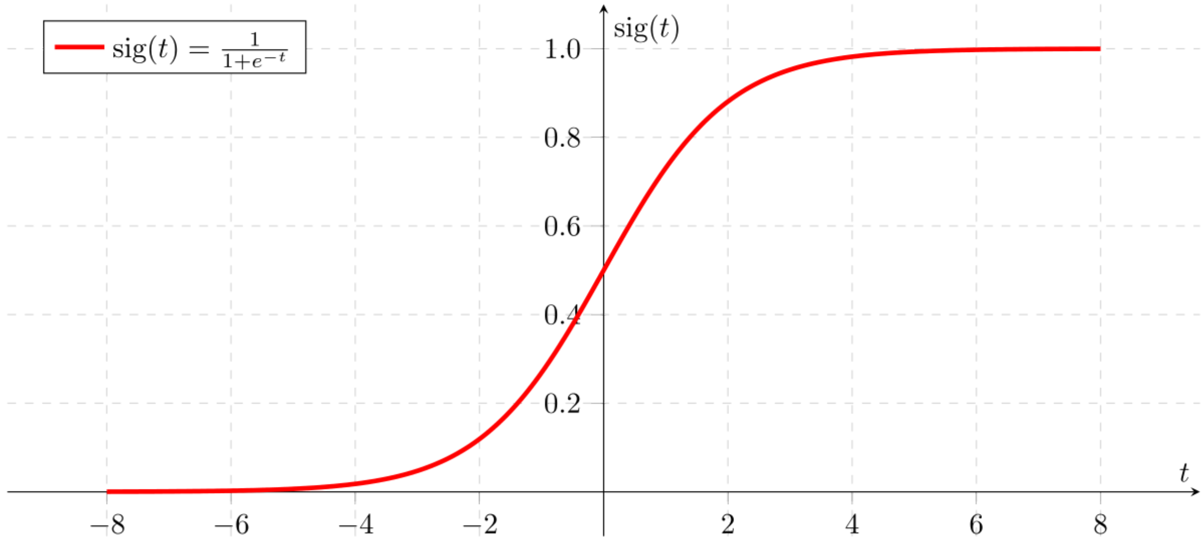
Logistic Regression was used in the biological sciences in early twentieth century. It was then used in many social science applications. Logistic Regression is used when the dependent variable(target) is categorical and is binary in nature. In order to perform binary classification the logistic regression techniques makes use of Sigmoid fuction.

For example,

To predict whether an email is spam (1) or (0)

Whether the tumor is malignant (1) or not (0)

Consider a scenario where we need to classify whether an email is spam or not. If we use linear regression for this problem, there is a need for setting up a threshold based on which classification can be done. Say if the actual class is malignant, predicted continuous value 0.4 and the threshold value is 0.5, the data point will be classified as not malignant which can lead to serious consequence in real time.



From this example, it can be inferred that linear regression is not suitable for classification problem. Linear regression is unbounded, and this brings logistic regression into picture. Their value strictly ranges from 0 to 1.

Dataset:

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).

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd

sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# 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

Variable Notes

pclass: A proxy for socio-economic status (SES)

1st = Upper, 2nd = Middle, 3rd = Lower

age: Age is fractional if less than 1. If the age is estimated, is it in the form of xx.5

sibsp: The dataset defines family relations in this way...,

Sibling = brother, sister, stepbrother, stepsister

Spouse = husband, wife (mistresses and fiancés were ignored)

parch: The dataset defines family relations in this way...

Parent = mother, father

Child = daughter, son, stepdaughter, stepson

Some children travelled only with a nanny, therefore parch=0 for them.

Code:

```
import pandas as pd
from sklearn.preprocessing import LabelEncoder
from sklearn.linear_model import
LogisticRegression from sklearn.model_selection
import train_test_split from sklearn.metrics
import accuracy_score

data = pd.read_csv("/content/titanic.csv")
print(data)
```

```
     PassengerId  Survived  Pclass  \
0    1         0       3
1    2         1       1
2    3         1       3
3    4         1       1
4    5             0          3..
...        ...      ...
886          887         0       2
887          888         1       1
888          889         0       3
889          890         1       1
890          891         0       3
```

```
                                                Name     Sex   Age
SibSp
\
                             Braund, Mr. Owen Harris    male  22.0

    Cumings, Mrs. John Bradley (Florence Briggs Th...  female  38.0

                              Heikkinen, Miss. Laina  female  26.0

        Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0

                            Allen, Mr. William Henry    male  35.0
```

```
..                                                          ...      ...    ...
..
386                              Montvila, Rev. Juozas    male  27.0

387                       Graham, Miss. Margaret Edith  female  19.0

388           Johnston, Miss. Catherine Helen "Carrie"  female   NaN

389                              Behr, Mr. Karl Howell    male  26.0

390                                Dooley, Mr. Patrick    male  32.0

Parch            Ticket     Fare Cabin Embarked
0        0         A/5 21171   7.2500   NaN        S
1        0          PC 17599  71.2833   C85        C
```

```
2     0  STON/O2. 3101282   7.9250   NaN        S
3     0            113803  53.1000  C123        S
4     0            373450   8.0500   NaN        S
..      ...               ...      ...   ...      ...
886      0            211536  13.0000   NaN        S
887      0            112053  30.0000   B42        S
888      2        W./C. 6607  23.4500   NaN        S
889      0            111369  30.0000  C148        C
890      0            370376   7.7500   NaN        Q
[891 rows x 12 columns]
```

```
le=LabelEncoder()
le.fit(data["Sex"])
data["Sex"]=le.transform(data["Sex"])
print(data["Sex"])
0     1
1     0
2     0
3     0
4     1
..
886    1
887    0
888    0
889    1
890    1
Name: Sex, Length: 891, dtype: int64
```

```
data["Age"].fillna(data["Age"].mean(), inplace=True) x =
data[["Pclass", "Sex", "Age", "SibSp", "Parch", "Fare"]]
y = data["Survived"]

model = LogisticRegression()
x_train, x_test, y_train, y_test
=
train_test_split(x,y,random_state=10,test_size=0
.1) model.fit(x_train,y_train) y_pred =
model.predict(x_test)
print(accuracy_score(y_test,y_pred))

0.8
```

Conclusion:

a machine learning process to predict survival rates on the Titanic using a logistic regression model. The data was preprocessed by filling in missing values and encoding categorical features like gender. The model was then trained and tested, yielding an accuracy of 80%. This indicates the model's capability to predict survival based on factors like passenger class, age, gender, and fare..