Predict survival on the Titanic

In this Lab, we ask you to apply the tools of machine learning to predict which passengers survived the tragedy

Dataset

The dataset contains 891 observations of 12 variables:

- **PassengerId**: Unique ID for each passenger
- **Survived**: Survival (0 = No; 1 = Yes)
- **Pclass**: Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)
- **Name**: Name
- **Sex**: Sex
- **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)

```
# import os
# from google.colab import drive
# drive.mount('/content/drive', force_remount=False)
```

```
# imports
import warnings
warnings.filterwarnings('ignore')
# your code here
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

```
titanic = pd.read_csv('titanic.csv',index_col=0)# your code here
titanic.head()
```

	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket
PassengerId								
1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th	female	38.0	1	0	PC 17599

```
# print some info about the dataframe
# your code here
titanic.shape
```

```
(891, 11)
```

Looks like there are some Nan values, let's see how many for each column

```
titanic.isnull().sum()
```

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

Cabin contains a lot of Nan values, we'll drop this column
We'll replace the Nan values in **Age** with the age's median, and the ones in **Embarked** with **'S'**, which is the most frequent one in this column

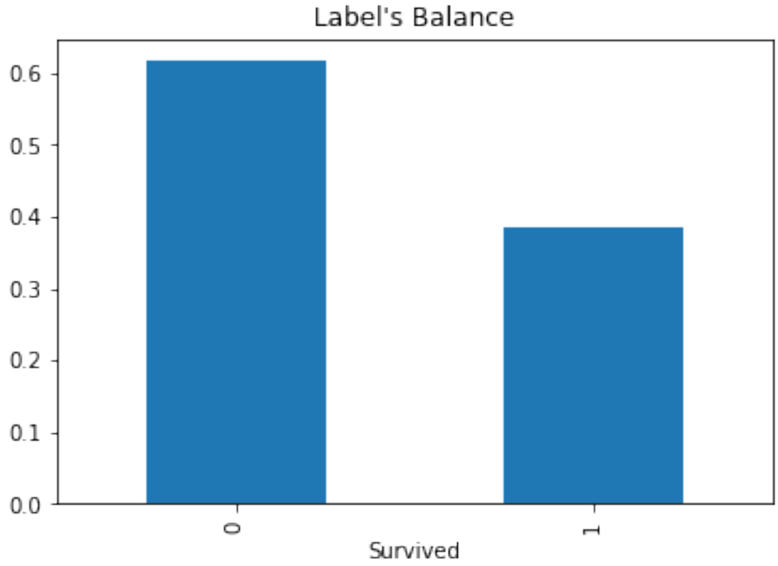
```
# your code here to drop Cabin
titanic.drop('Cabin',axis=1,inplace=True)
# check the fillna documentation: http://pandas.pydata.org/pandas-docs/stable/ge
titanic["Age"] = titanic["Age"].fillna(value=titanic["Age"].median())
titanic["Embarked"] = titanic["Embarked"].fillna(value="S")
titanic.isnull().sum()
```

```
Survived    0
Pclass      0
Name        0
Sex         0
Age         0
SibSp       0
Parch       0
Ticket      0
Fare        0
Embarked    0
dtype: int64
```

Visualization

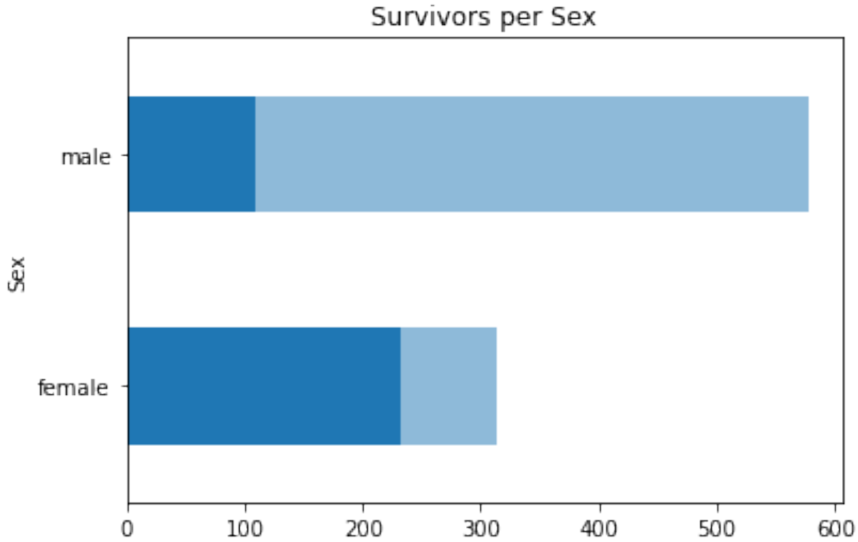
```
%matplotlib inline
import matplotlib.pyplot as plt
print ('survival rate =', titanic.Survived.mean())
(titanic.groupby('Survived').size()/titanic.shape[0]).plot(kind="bar",title="Lab
```

```
survival rate = 0.3838383838383838
<AxesSubplot:title={'center':"Label's Balance"}, xlabel='Survived'>
```



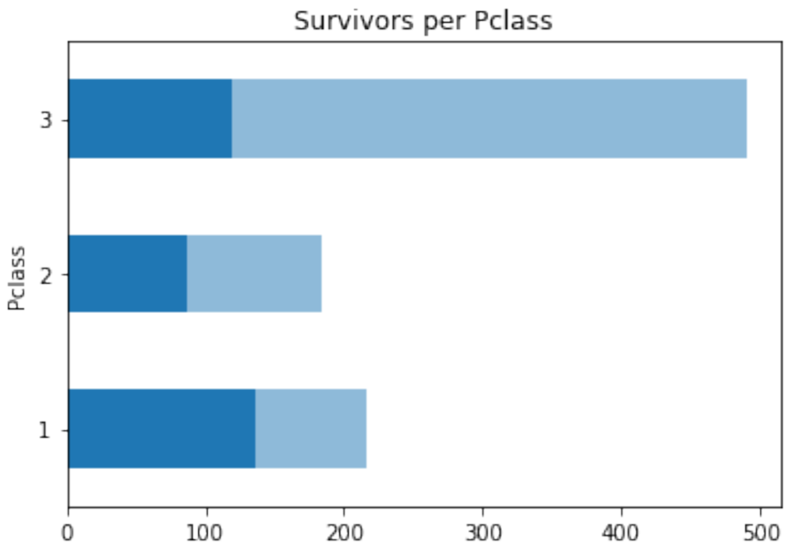
```
# make a function to plot survival against passenger attribute
def survival_rate(column,t):
    df=pd.DataFrame()
    df['total']=titanic.groupby(column).size()
    df['survived'] = titanic.groupby(column).sum()['Survived']
    df['percentage'] = round(df['survived']/df['total']*100,2)
    print(df)

    df['survived'].plot(kind=t)
    df['total'].plot(kind=t,alpha=0.5,title="Survivors per "+str(column))
    plt.show()


# Draw survival per Sex
survival_rate("Sex","barh")
```

```
        total  survived  percentage
Sex
female    314       233       74.20
male      577       109       18.89
```



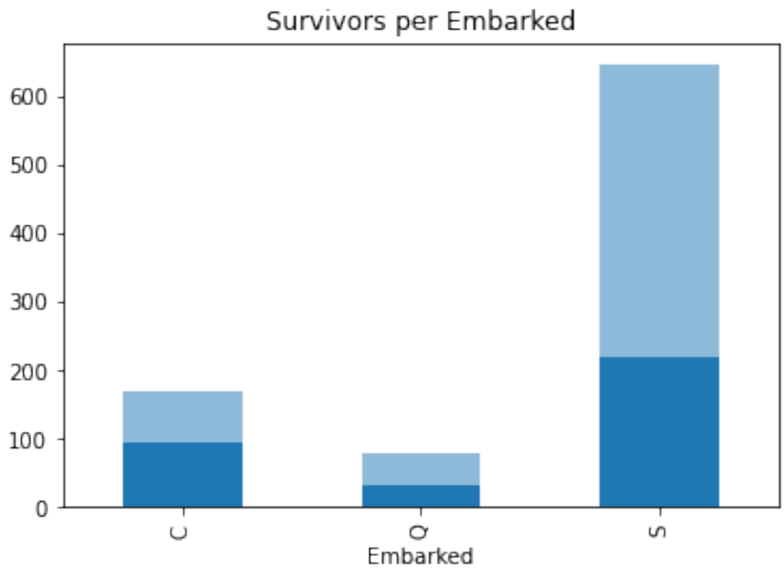
```
# Draw survival per Class
survival_rate("Pclass","barh")
```

```
        total  survived  percentage
Pclass
1         216       136       62.96
2         184        87       47.28
3         491       119       24.24
```



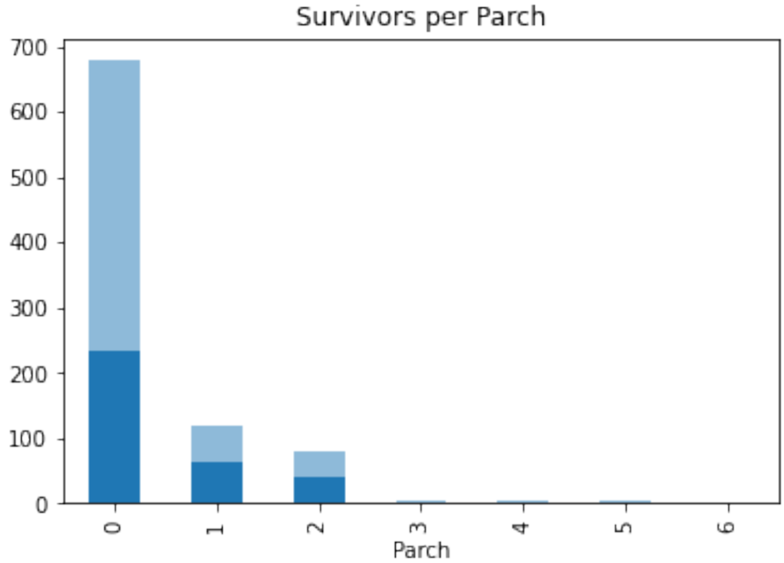

```
# Graph survived per port of embarkation
survival_rate("Embarked","bar")
```

```
          total  survived  percentage
Embarked
C           168        93       55.36
Q            77        30       38.96
S           646       219       33.90
```



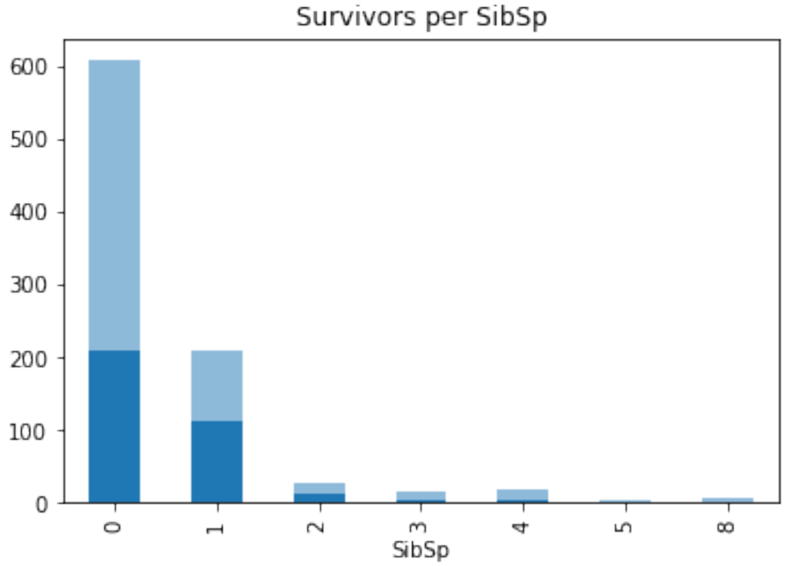
```
# Draw survived per Number of Parents/Children Aboard (Parch)
# your code here
survival_rate("Parch","bar")
```

Parch	total	survived	percentage
0	678	233	34.37
1	118	65	55.08
2	80	40	50.00
3	5	3	60.00
4	4	0	0.00
5	5	1	20.00
6	1	0	0.00



```
# Draw survived per Number of Siblings/Spouses Aboard (SibSp)
# your code here
survival_rate("SibSp","bar")
```

```
       total  survived  percentage
SibSp
0        608       210       34.54
1        209       112       53.59
2         28        13       46.43
3         16         4       25.00
4         18         3       16.67
5          5         0        0.00
8          7         0        0.00
```



Model training

Some of the columns don't have predictive power, so let's specify which ones are included for prediction

```
predictors = ["Pclass", "Sex", "Age", 'SibSp' ,'Parch', "Fare", "Embarked"]
```

We need now to convert text columns in **predictors** to numerical ones

```
for col in predictors: # Loop through all columns in predictors
    if titanic[col].dtype == 'object':  # check if column's type is object (text
        titanic[col] = pd.Categorical(titanic[col]).codes  # convert text to num

titanic.head()
```

	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	I
PassengerId									
1	0	3	Braund, Mr. Owen Harris	1	22.0	1	0	A/5 21171	7.
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th	0	38.0	1	0	PC 17599	71.

```
# Split the data into a training set and a testing set. Set: test_size=0.3, rand
# your code here
from sklearn.model_selection import train_test_split
y = titanic["Survived"]
x = titanic[predictors]
X_train, X_test, y_train, y_test = train_test_split(x,y,test_size=0.3, random_st

print ("train shape", X_train.shape, y_train.shape)
print ("test shape", X_test.shape, y_test.shape)
```

```
train shape (623, 7) (623,)
test shape (268, 7) (268,)
```

```
# import LogisticRegression from: http://scikit-learn.org/stable/modules/generat
# your code here
from sklearn.linear_model import LogisticRegression
clf = LogisticRegression(random_state=1)
clf.fit(X_train,y_train)
# your code here
train_score = clf.score(X_train,y_train)
test_score = clf.score(X_test,y_test)
print ('train accuracy =', train_score)
print ('test accuracy =', test_score)
```

```
train accuracy = 0.8073836276083467
test accuracy = 0.7723880597014925
```

Let's print the model's parameters

```
coeff = pd.DataFrame()
coeff['Feature'] = X_train.columns
coeff['Coefficient Estimate'] = pd.Series(clf.coef_[0])
coeff.loc[len(coeff)]=['Intercept',clf.intercept_[0]]
print (coeff)
```

```
     Feature  Coefficient Estimate
0     Pclass             -1.158692
1        Sex             -2.708761
2        Age             -0.040634
3      SibSp             -0.334012
4      Parch              0.071940
5       Fare             -0.000570
6   Embarked             -0.223307
7  Intercept              5.391544
```

We now need to predict class labels for the test set. We will also generate the class probabilities

```
# predict class labels for the test set
y_pred = clf.predict(X_test)
print (y_pred)
```

```
[1 0 1 1 1 0 0 1 1 1 0 1 0 0 1 0 0 0 0 1 0 0 1 0 1 0 1 1 0 1 1 0 0 1 0 1 0
 0 1 0 1 1 1 0 1 0 0 0 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 0 0 1 0 0 1 0 0 0 0 0
 1 0 1 0 0 1 0 0 0 0 1 0 0 0 1 1 0 0 0 1 0 1 0 1 0 0 1 0 0 1 1 0 0 0 0 0 0
 0 0 0 0 0 0 0 1 1 1 0 0 0 1 1 1 1 0 0 0 0 1 1 0 1 1 0 0 1 1 0 1 1 0 1 0 0
 1 0 1 0 0 1 0 0 0 0 1 0 0 0 1 0 0 1 1 0 0 0 1 1 1 0 1 0 0 0 1 0 1 1 0 0 1
 0 0 1 0 1 0 0 1 1 1 0 1 0 0 0 1 0 0 0 1 1 0 0 0 1 0 0 0 0 0 0 1 1 0 0 0
 0 0 0 0 1 0 1 0 0 0 0 0 0 1 1 1 1 0 0 0 0 0 0 0 1 1 0 1 0 0 0 1 1 1 0 1 0
 1 1 0 1 1 0 0 1 0]
```

```
# generate class probabilities : http://scikit-learn.org/stable/modules/generate
y_probs = clf.predict_proba(X_test)
print (y_probs)
```

```
 [0.96819807 0.03180193]
 [0.05269658 0.94730342]
 [0.931658   0.068342  ]
 [0.87604357 0.12395643]
 [0.77746925 0.22253075]
 [0.84945584 0.15054416]
 [0.93503218 0.06496782]
 [0.89643851 0.10356149]
 [0.04125003 0.95874997]
```

```
 [0.19533985 0.80466015]
 [0.77085873 0.22914127]
 [0.51933522 0.48066478]
 [0.87382014 0.12617986]
 [0.74132793 0.25867207]
 [0.72964262 0.27035738]
 [0.92394548 0.07605452]
 [0.70094191 0.29905809]
 [0.04415584 0.95584416]
 [0.51470922 0.48529078]
 [0.23425444 0.76574556]
 [0.70097333 0.29902667]
 [0.60811076 0.39188924]
 [0.83497752 0.16502248]
 [0.67771166 0.32228834]
 [0.81308544 0.18691456]
 [0.94205138 0.05794862]
 [0.39282351 0.60717649]
 [0.46814345 0.53185655]
 [0.33317311 0.66682689]
 [0.28618636 0.71381364]
 [0.9221556  0.0778444 ]
 [0.94286813 0.05713187]
 [0.83457972 0.16542028]
 [0.88251181 0.11748819]
 [0.92906434 0.07093566]
 [0.95096974 0.04903026]
 [0.92442744 0.07557256]
 [0.19105465 0.80894535]
 [0.22468389 0.77531611]
 [0.65007535 0.34992465]
 [0.09251543 0.90748457]
 [0.94833689 0.05166311]
 [0.9001962  0.0998038 ]
 [0.68228929 0.31771071]
 [0.40806752 0.59193248]
 [0.19561719 0.80438281]
 [0.27749947 0.72250053]
 [0.76966959 0.23033041]
 [0.27027965 0.72972035]
 [0.83450642 0.16549358]
 [0.44297783 0.55702217]
 [0.10338877 0.89661123]
 [0.92394548 0.07605452]
 [0.36586709 0.63413291]
 [0.14332527 0.85667473]
 [0.76730525 0.23269475]
 [0.80676685 0.19323315]
 [0.46946197 0.53053803]
 [0.76143753 0.23856247]]
```

As you can see, the classifier outputs two probabilities for each row. It's predicting a 1 (Survived) any time the probability in the second column is greater than 0.5. Let's visualize it all together.

```
pred = pd.DataFrame({
        "Survived_original": y_test,
        "Survived_predicted": y_pred,
        "Survived_proba": np.transpose(y_probs)[1]
        })
pred["Comparison"]= pred.Survived_original ==pred.Survived_predicted
pred.head()
```

	Survived_original	Survived_predicted	Survived_proba	Comparis
PassengerId				
863	1	1	0.860689	T
224	0	0	0.084543	T
85	1	1	0.873336	T
681	0	1	0.634082	Fa
536	1	1	0.922157	T

Confusion matrix

```
from sklearn import metrics
print (metrics.confusion_matrix(y_test, y_pred))
print (metrics.classification_report(y_test, y_pred))
```

```
[[129  24]
 [ 37  78]]
              precision    recall  f1-score   support

           0       0.78      0.84      0.81       153
           1       0.76      0.68      0.72       115

    accuracy                           0.77       268
   macro avg       0.77      0.76      0.76       268
weighted avg       0.77      0.77      0.77       268
```

As you can see, we can have the classification report for each class

K-Fold Cross Validation

```
# import cross_validation from: http://scikit-learn.org/stable/modules/generated
# your code here
from sklearn.model_selection import cross_val_score
clf = LogisticRegression(random_state=1)
scores = cross_val_score(clf, titanic[predictors], titanic["Survived"], scoring=
## see model
print(scores)
# Take the mean of the scores (because we have one for each fold)
print(scores.mean())
```

```
[0.7877095  0.78651685 0.78089888 0.76966292 0.82022472]
0.7890025735986442
```

When you are improving a model, you want to make sur that you are really doing it and not just being lucky. This is why it's good to work with cross validation instead of one train/test split.