Group 9: Correlation Analysis

Giuliano Piga, Erika Memije, Meyer Millman

June 21, 2019

Purpose

To analyze and display characteristics of variables to determine their effectiveness on data models.

Variables Considered

The variables that were chosen for both KNN and Decision Tree analysis were:

1. Pclass
2. Male
3. Female
4. "Embarked" Categories (C,Q,S)

```
[891 rows x 33 columns]

In [23]: del DF['Name']

In [24]: #DF

In [25]: from scipy.stats.stats import pearsonr

In [28]: count = 0
         features = ["Pclass", "Age", "SibSp","Parch", "Fare", "male", "female", "C","Q","S","I
         for x in features:
             print("Correlation and p: ", x, pearsonr(DF[x],DF["Survived"]))

Correlation and p:  Pclass (-0.33848103596101536, 2.53704738798042e-25)
Correlation and p:  Age (-0.06952761330099651, 0.037989220487832626)
Correlation and p:  SibSp (-0.03532249888573558, 0.29224392869817906)
Correlation and p:  Parch (0.08162940708348349, 0.0147992453747224)
Correlation and p:  Fare (0.25730652238496243, 6.120189341921873e-15)
Correlation and p:  male (-0.5433513806577553, 1.406066130879517e-69)
Correlation and p:  female (0.5433513806577552, 1.406066130879597e-69)
Correlation and p:  C (0.1682404312182332, 4.3971513298052554e-07)
Correlation and p:  Q (0.003650382683972173, 0.9133532352434973)
Correlation and p:  S (-0.15566027340439348, 3.0361110645208803e-06)
Correlation and p:  Mr (-0.5491991849030087, 2.4287826448462406e-71)
Correlation and p:  Mrs (0.3390402513843207, 2.0941266637294965e-25)
Correlation and p:  Miss (0.32709254908267793, 1.159990744524431e-23)


In [26]: Surv=np.array(DF['Survived'])
```

```
Out[29]: Mr              517
         Miss            182
         Mrs             125
         Master           40
         Dr                7
         Rev               6
         Mlle              2
         Major             2
         Col               2
         Don               1
         Mme               1
         Lady              1
         Capt              1
         the Countess      1
         Sir               1
         Ms                1
         Jonkheer          1
         Name: Name, dtype: int64

In [30]: data=np.array(DF[["male",'female','C','Q','S','Pclass','Survived']])

In [31]: data

Out[31]: array([[1, 0, 0, ..., 1, 3, 0],
                [0, 1, 1, ..., 0, 1, 1],
                [0, 1, 0, ..., 1, 3, 1],
                ...,
                [0, 1, 0, ..., 1, 3, 0],
                [1, 0, 1, ..., 0, 1, 1],
                [1, 0, 0, ..., 0, 3, 0]], dtype=int64)
```

Table of Values

Variable compared with "Survival"	p-value	Correlation Coefficient
Pclass	$2.573e^{-25}$	-0.3384
Female	$1.406e^{-69}$	0.5434
Fare	$6.120e^{-15}$	0.2573
Male	$1.406e^{-69}$	-0.5434
C	$4.397e^{-7}$	0.1682
Q	0.9133	0.0037

Why include "Embarked"?

```
In [46]: del td['Sex']

In [47]: #td

In [48]: names=[]
         for x in np.array(td['Name']):
             tokens=x.split(', ',maxsplit=2)
             names.append(tokens)

         namedata2=pd.DataFrame(names,columns=["Surname","title"])
         #namedata2

In [49]: names=[]
         for x in np.array(namedata2['title']):
             tokens=x.split('. ',maxsplit=1)
             names.append(tokens)

         titledata2=pd.DataFrame(names,columns=["title","name"])
         #titledata2

In [50]: td['Name']=titledata2['title']

In [51]: tdName=pd.get_dummies(td['Name'])
         tdEm=pd.get_dummies(td['Embarked'])

In [52]: frames2=[td,tdName,tdEm]

In [53]: TD=pd.concat(frames2,axis=1)

In [54]: del TD["Name"]
```

Model Improvement

1. Adding the "C,Q,S" features extracted from the Embarked variable improved fit for the Decision Tree.
2. Using the full "Embark" Variable could work better on other models.

Observations

1. The combinations of variables in this case in particular has a large effect on the model fit
2. No variable in the dataset had a high correlation with survival

No Free Lunch Theorem

1. There is no one model that works best for every problem. The assumptions of a great model for one problem may not hold for a different problem
2. Ultimately, using same predictors on different models will yield varying predictions. Based on the predictors you do use, some models may work better.