NAVEEN SHARMA
naveenmca18@gmail.com
9986006479

Titanic : Machine Learning from Disaster

4th Aug 2019

OVERVIEW

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.

THE CLIENT

As such the project doesn’t have any specific client, but people or organization who are still studying about the Titanic and its passengers data they can leverage this analysis. EDA and model prediction can give greater insight of the features and their impact on the end class.

THE DATA

The data used in this project gathered from Kaggle. Kaggle has publicly available dataset that contains 891 entries in train.csv and 418 entries in test.csv. These files contains different columns e.g. Name,age, sex,fare, cabin. This data collected in .csv format.

The following files will be used in project . Data has been split into two groups.

1. **Train.csv :** This file contains metadata collected from kaggle and holds 891 entries with columns like Name,age, sex,fare, cabin etc. This file will be used for train our models.
2. **Test.csv:** This file contains metadata collected from kaggle and holds 418 entries with columns like Name,age, sex,fare, cabin etc. This file will be used to test our models

DATA COLLECTION

The Titanic dataset is freely available on kaggle website. (https://www.kaggle.com/c/titanic/data).

DATA WRANGLING

Overview

This section describes the various data cleaning and data wrangling methods applied on the Titanic dataset to explore the data and make it more precise to further analysis. These techniques will have the following sections.

Read From CSV Files

To analysis the data first I pulled the data from csv files. Train data from train.csv and Test data from test.csv.

Handling Null Values

Like most dataset this dataset also have some columns which were containing null values. We have to deal with these null values to make a good ML model. In Titanic dataset there are two columns which were having null values :

Age, Cabin, Embarked

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
PassengerId    891 non-null int64
Survived       891 non-null int64
Pclass         891 non-null int64
Name           891 non-null object
Sex            891 non-null object
Age            714 non-null float64
SibSp          891 non-null int64
Parch          891 non-null int64
Ticket         891 non-null object
Fare           891 non-null float64
Cabin          204 non-null object
Embarked       889 non-null object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
```

Handled these NaN value with different techniques like FFILL, BFILL, create Model.

Remove Not So Required Features

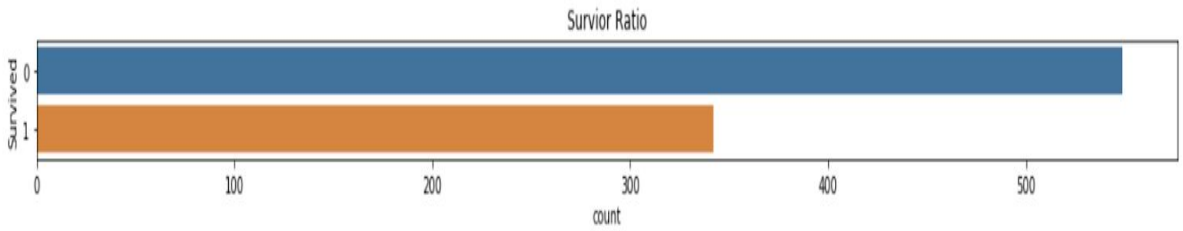
Some features such as PassengerId was not really required to analysis the data or build the model therefore that's dropped to reduce the dimension.

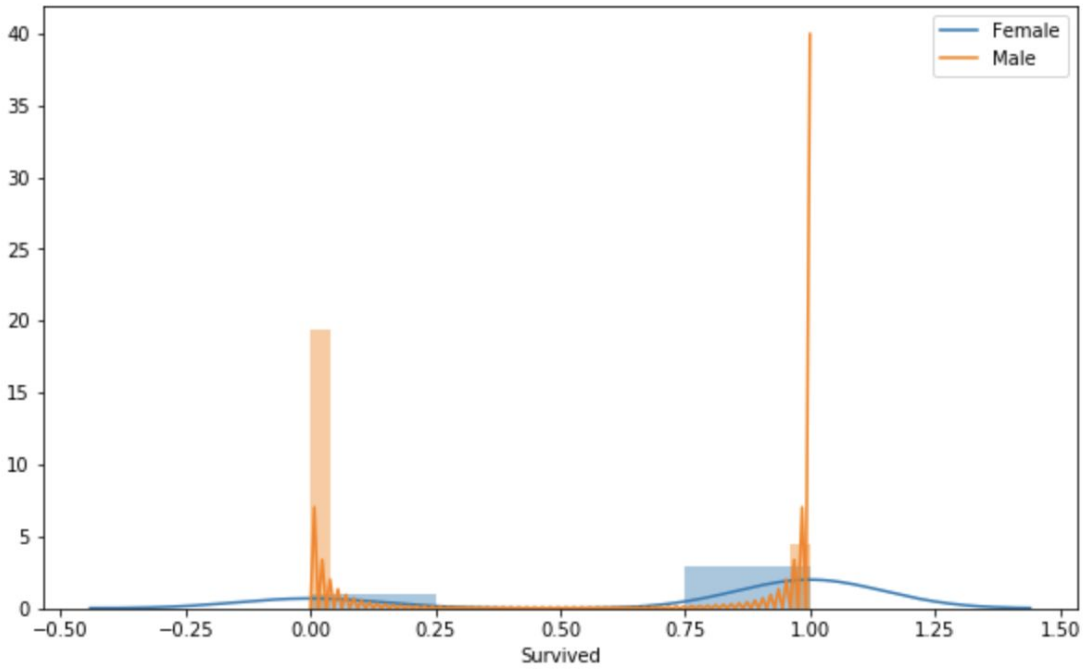
EXPLORATORY DATA VISUALIZATION

In this area, we will examine the various / different insight of data through visualization and descriptive statistics.

This will be quite informative.

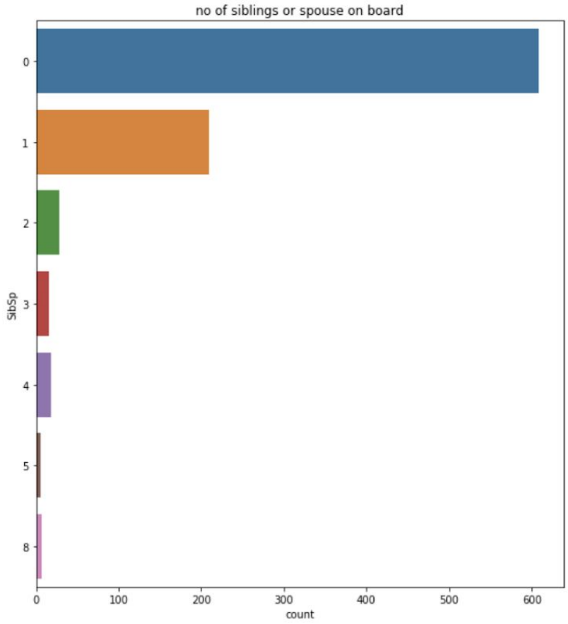
FeMale Vs Male Survival Ratio

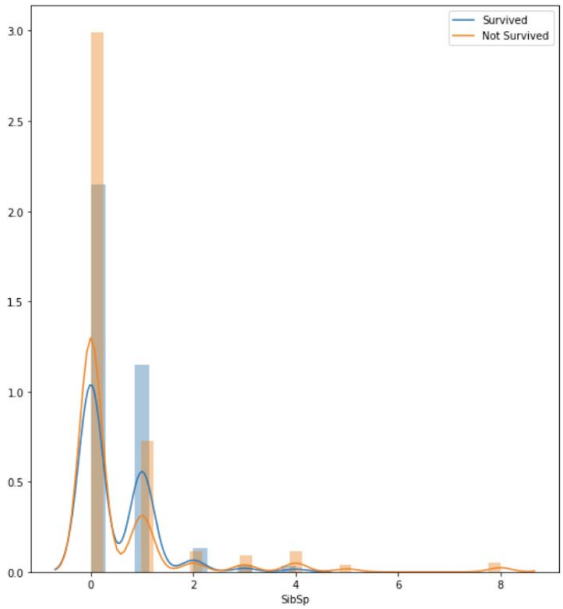




More women survived as compared to men because they were the one who were evacuated first.

Siblings and Parents on board

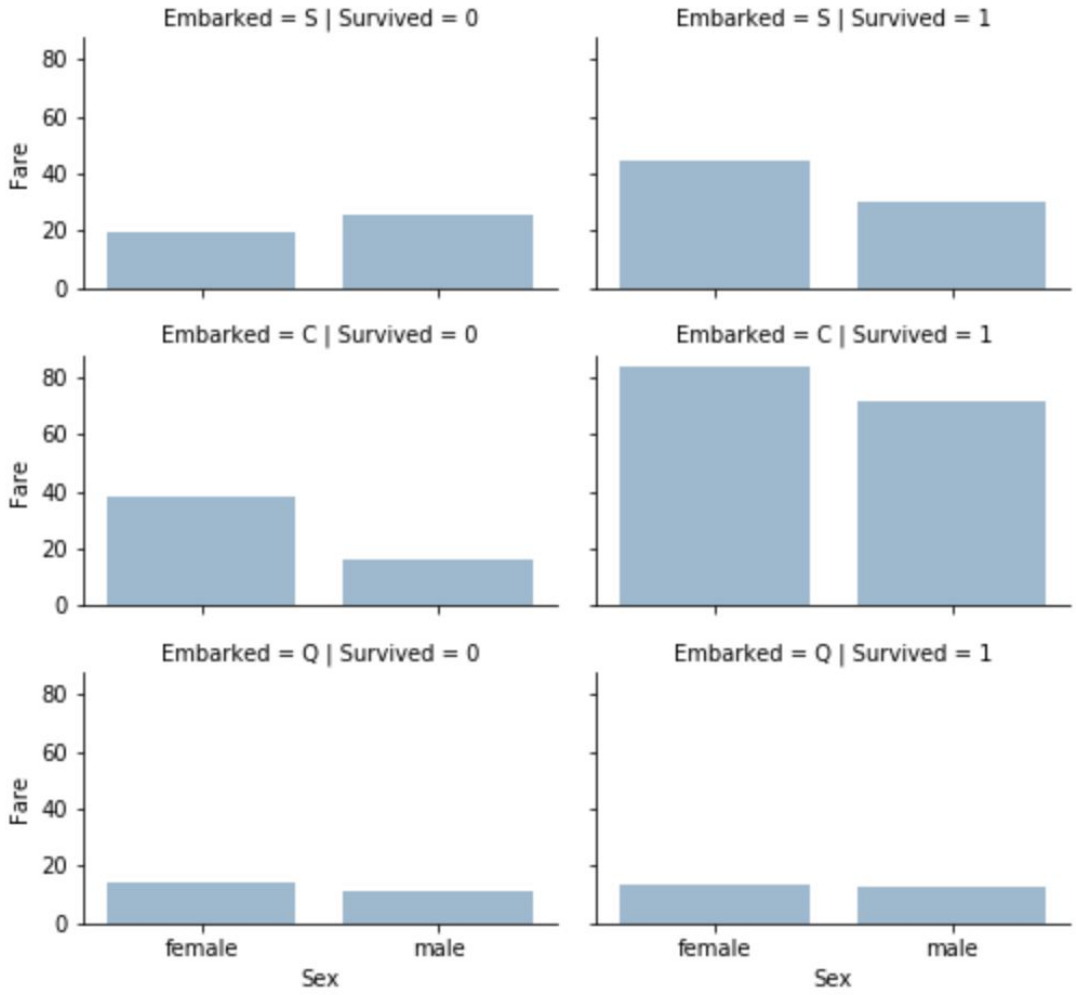




Passengers who were traveling solo survived most. I guess this could be obvious reason because when you are alone in any such situation, the only thing you first choose is to save yourself.

Passengers who were traveling with their family members were less likely to survive.

Survived vs Not Survived (P class / Sex / Age)



We can combine multiple features for identifying correlations using a single plot. This can be done with numerical and categorical features which have numeric values.

Pclass=3 had most passengers, however most did not survive. Confirms our classifying assumption #2.

Infant passengers in Pclass=2 and Pclass=3 mostly survived. Further qualifies our classifying assumption #2.

Most passengers in Pclass=1 survived. Confirms our classifying assumption #3.

Pclass varies in terms of Age distribution of passengers.

FEATURE ENCODING

As we are dealing with different kinds of feature we also encounter categorical features. Categorical features are really not helpful in building models because they aren't numeric. We have to deal with this situation and need to convert them into continuous variables.

There are different techniques are available in the market with their pros and cons. I've used ONE HOT ENCODING to convert the categorical data to continuous data.

In Titanic dataset following are the categorical features

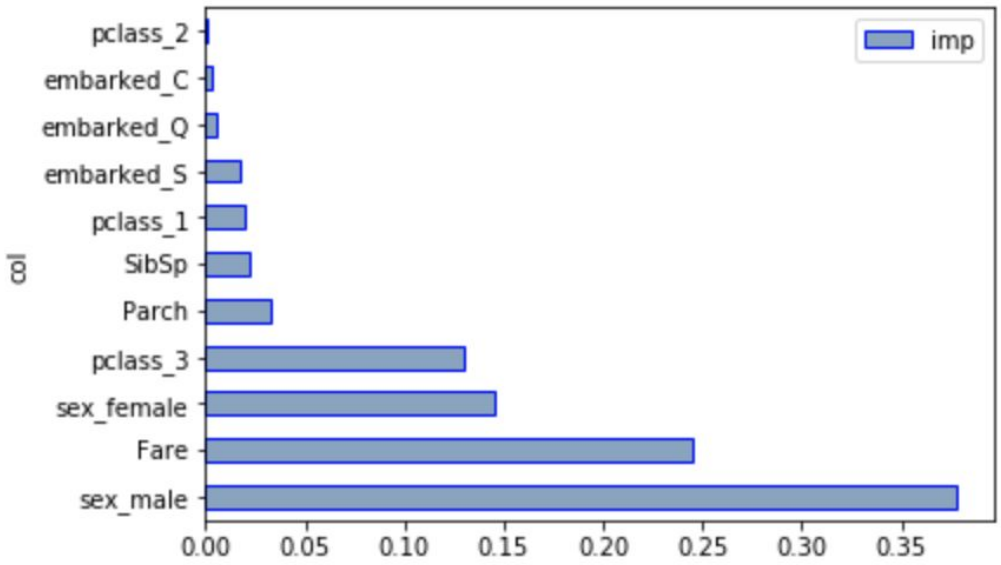
Sex, Embarked

Dataset insight after encoding

	Survived	SibSp	Parch	Fare	embarked_C	embarked_Q	embarked_S	sex_female	sex_male	pclass_1	pclass_2	pcla
0	0	1	0	7.2500	0	0	1	0	1	0	0	1
1	1	1	0	71.2833	1	0	0	1	0	1	0	0
2	1	0	0	7.9250	0	0	1	1	0	0	0	1
3	1	1	0	53.1000	0	0	1	1	0	1	0	0
4	0	0	0	8.0500	0	0	1	0	1	0	0	1

FEATURE IMPORTANCE

Feature importance is way to identify the most reavent feature for our best model creation. The model I choose is Gradient Boost Classifier with 80.65% accuracy.



MODELS

Logistic Regression

```
Accuracy:  79.98
Accuracy Cv :  79.42
```

Logistic Regression was used in the biological sciences in the early twentieth century. It was then used in many social science applications. Logistic Regression is used when the dependent variable(target) is categorical.

For example,

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

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

KNeighborsClassifier

```
Accuracy:  83.46
Accuracy Cv :  76.72
```

KNN can be used for both classification and regression predictive problems. However, it is more widely used in classification problems in the industry.

Random Forest

```
Accuracy:  91.23
Accuracy Cv :  79.64
```

Random Forest Classifier is ensemble algorithm. In next one or two posts we shall explore such algorithms. Ensembled algorithms are those which combines more than one algorithms of same or different kind for classifying objects

Gradient Boost

```
Accuracy:  86.61
Accuracy Cv 10 Folds:  80.65
```

Gradient boosting classifiers are a group of machine learning algorithms that combine many weak learning models together to create a strong predictive model. Decision trees are usually used when doing gradient boosting. Gradient boosting models are becoming popular because of their effectiveness at classifying complex datasets

Models Result

	Model	Score
3	GradientBoost	80.65
2	RandomForest	79.64
0	LogisticRegression	79.42
1	KNN	76.72

Clearly GradientBoot classifier algorithm is work better than others.

CONCLUSION

The report highlighted the process of data wrangling , inferential statistics , data visualization , feature engineering , predictive modeling performed on Titanic Dataset. All results and data insight are also highlighted. The highest accuracy model was Gradient boost classifier with 80.65% or accuracy.

Total 4 model was built

Logistic Regression

KNeighbour Classifier

Random Forest Classifier

Gradient Boost Classifier

80.65% is decent accuracy, still i see there are lots of room to tune this model to increase the accuracy.

Code : https://www.kaggle.com/navesharma9/titanic-survivor-analysis