final_project

December 6, 2019

1. Introduction

0.0.1 1.1. Data Dictionary

- `PassengerId` : the unique id of the row and it doesn't have any effect on `Survived`.
- `Survived` : binary (**0** or **1**);
  - **1 = Survived**
  - **0 = Not Survived**
- `Pclass` (Passenger Class) : the socio-economic status of the passenger. It is a categorical ordinal feature which has **3** unique values (**1**, **2** or **3**);
  - **1 = Upper Class**
  - **2 = Middle Class**
  - **3 = Lower Class**
- `Name`, `Sex` and `Age` features are self-explanatory.
- `SibSp` : the total number of the passengers' siblings and spouse.
- `Parch` : the total number of the passengers' parents and children.
- `Ticket` : the ticket number of the passenger.
- `Fare` : the passenger fare.
- `Cabin` : the cabin number of the passenger.
- `Embarked` is port of embarkation. It is a categorical feature and it has **3** unique values (**C**, **Q** or **S**);
  - **C = Cherbourg**
  - **Q = Queenstown**
  - **S = Southampton**

0.0.2 1.2. Libarary

0.0.3 1.3. Loading the Dataset

```
There are 418 samples in test_data
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1309 entries, 0 to 1308
Data columns (total 12 columns):
Age            1046 non-null float64
Cabin          295 non-null object
Embarked       1307 non-null object
Fare           1308 non-null float64
Name           1309 non-null object
Parch          1309 non-null int64
```

```
PassengerId    1309 non-null int64
Pclass         1309 non-null int64
Sex            1309 non-null object
SibSp          1309 non-null int64
Survived       891 non-null float64
Ticket         1309 non-null object
dtypes: float64(3), int64(4), object(5)
memory usage: 122.8+ KB
```

```
/Users/kenxu/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:2:
FutureWarning: Sorting because non-concatenation axis is not aligned. A future
version
of pandas will change to not sort by default.

To accept the future behavior, pass 'sort=False'.

To retain the current behavior and silence the warning, pass 'sort=True'.
```

```
    Age Cabin Embarked     Fare  \
0  22.0   NaN        S   7.2500
1  38.0   C85        C  71.2833
2  26.0   NaN        S   7.9250
3  35.0  C123        S  53.1000
4  35.0   NaN        S   8.0500

                                                 Name  Parch  PassengerId  \
0                             Braund, Mr. Owen Harris      0            1
1  Cumings, Mrs. John Bradley (Florence Briggs Th...      0            2
2                              Heikkinen, Miss. Laina      0            3
3        Futrelle, Mrs. Jacques Heath (Lily May Peel)      0            4
4                            Allen, Mr. William Henry      0            5

   Pclass     Sex  SibSp  Survived            Ticket
0       3    male      1       0.0         A/5 21171
1       1  female      1       1.0          PC 17599
2       3  female      0       1.0  STON/O2. 3101282
3       1  female      1       1.0            113803
4       3    male      0       0.0            373450
```

0.1 2. Missing Values

[6]:
```
Age            263
Cabin         1014
Embarked         2
Fare             1
```

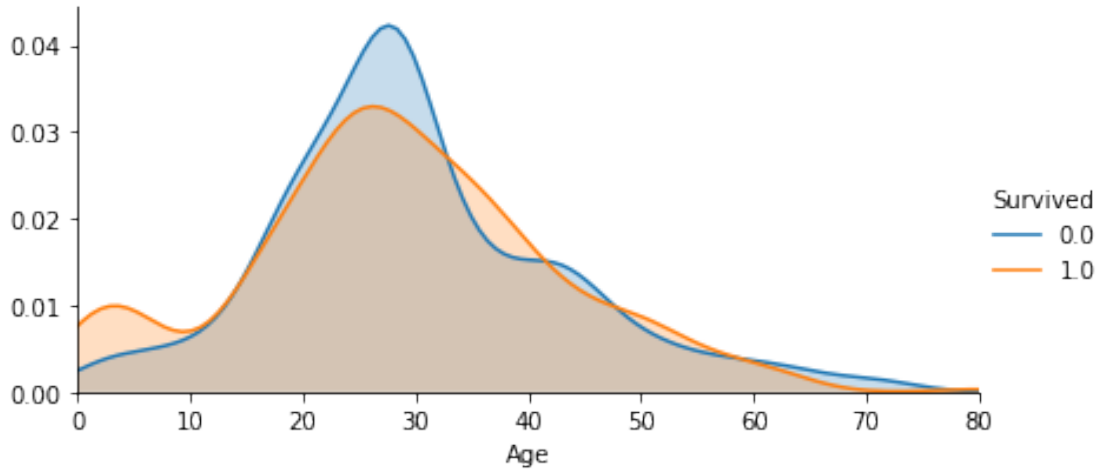
```
Name              0
Parch             0
PassengerId       0
Pclass            0
Sex               0
SibSp             0
Survived        418
Ticket            0
dtype: int64
```

0.1.1 2.1. Age

The age feature has 263 null values. So we use random forest regression model to simulate the value.Fill missing values using Random Forest.The features we use here are sex, pclass, Parch, SibSp.

```
/Users/kenxu/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:5:
FutureWarning: Method .as_matrix will be removed in a future version. Use
.values instead.
  """
/Users/kenxu/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:6:
FutureWarning: Method .as_matrix will be removed in a future version. Use
.values instead.
```

[8]: `<seaborn.axisgrid.FacetGrid at 0x10a49bb70>`



0

0.1.2 2.2. Fare

[10]:
```
      Age Cabin Embarked  Fare                Name  Parch  PassengerId  \
1043  60.5   NaN        S   NaN  Storey, Mr. Thomas      0         1044

      Pclass   Sex  SibSp  Survived Ticket
1043       3  male      0       NaN   3701
```

0.1.3 2.3. Embarked

[12]:
```
      Age Cabin Embarked  Fare                                        Name  \
61   38.0   B28      NaN  80.0                        Icard, Miss. Amelie
829  62.0   B28      NaN  80.0  Stone, Mrs. George Nelson (Martha Evelyn)

     Parch  PassengerId  Pclass     Sex  SibSp  Survived  Ticket
61       0           62       1  female      0       1.0  113572
829      0          830       1  female      0       1.0  113572
```

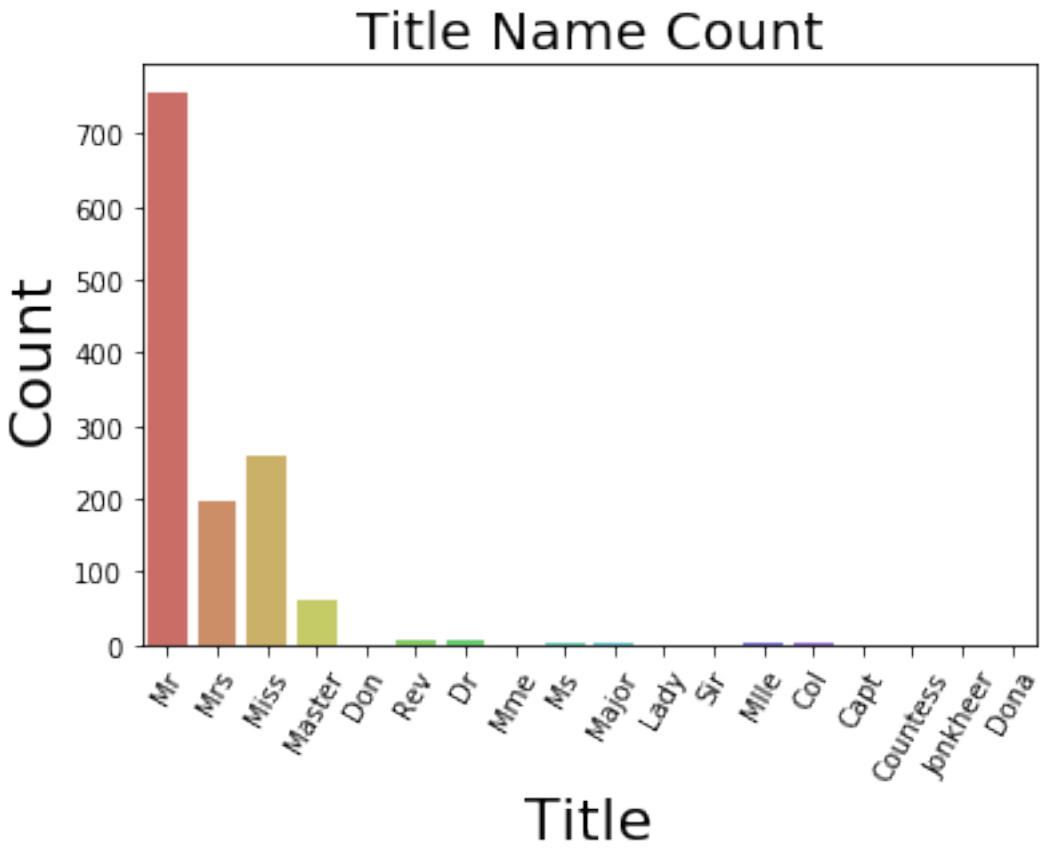
When I googled **Stone, Mrs. George Nelson (Martha Evelyn)**, I learned that **Mrs Stone** boarded the Titanic in **Southampton** on 10 April 1912 and was travelling in first class with her maid **Amelie Icard** in this page Martha Evelyn Stone: Titanic Survivor.

3. Feature Engineering

0.1.4 3.1. Title Extraction

Here I refer to a very interesting kernal for title extraction:Titanic [EDA] + Model Pipeline + Keras NN.

[14]:
```
0                              Braund, Mr. Owen Harris
1    Cumings, Mrs. John Bradley (Florence Briggs Th…
2                               Heikkinen, Miss. Laina
3         Futrelle, Mrs. Jacques Heath (Lily May Peel)
4                             Allen, Mr. William Henry
Name: Name, dtype: object
```

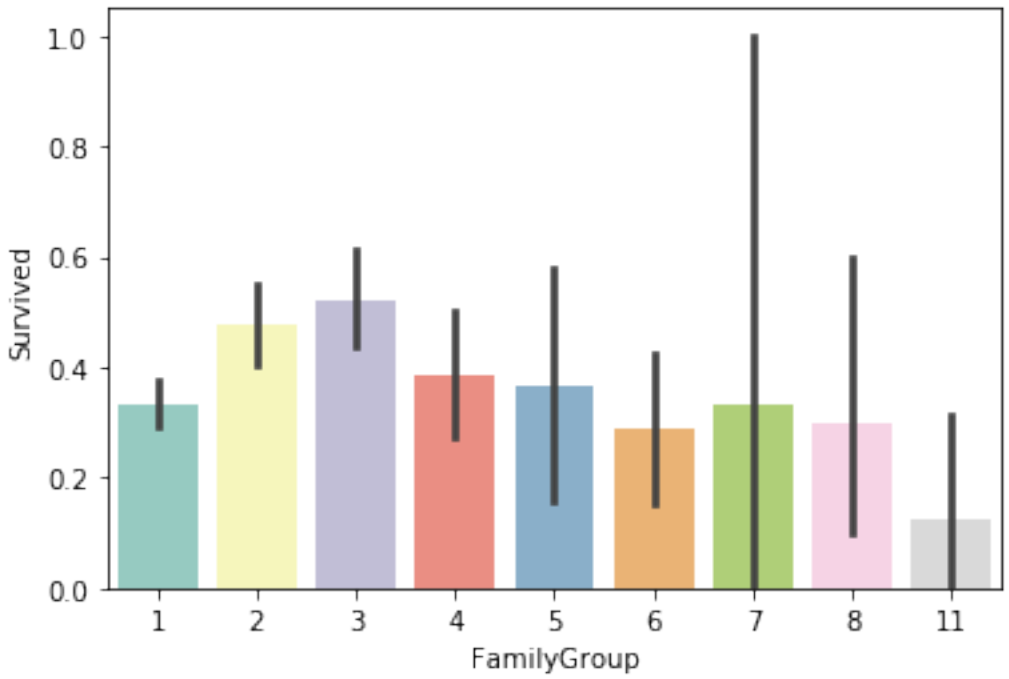



```
␣
↪---------------------------------------------------------------------------

       NameError                                 Traceback (most recent call␣
↪last)

       <ipython-input-1-1093a9851e49> in <module>
        22
        23 # we map each title to correct category
   ---> 24 all_data['Title'] = all_data.Title.map(Title_Dictionary)
        25 all_data['Title'].value_counts()


       NameError: name 'all_data' is not defined
```

0.1.5 3.2. Surname

We could make the exception list of male and female. Then change the features of test surname in the list into the dead features or survived features according to the list.

```
{'Goodwin', 'Turpin', 'Danbom', 'Barbara', 'Bourke', 'Jussila', 'Lefebre',
'Boulos', 'Palsson', 'Rosblom', 'Van Impe', 'Strom', 'Oreskovic', 'Attalah',
'Canavan', 'Sage', 'Ford', 'Cacic', 'Johnston', 'Robins', 'Arnold-Franchi',
'Lahtinen', 'Panula', 'Skoog', 'Olsson', 'Zabour', 'Lobb', 'Rice', 'Caram',
'Elias', 'Vander Planke', 'Ilmakangas'}
{'Bishop', 'Cardeza', 'Bradley', 'Chambers', 'Kimball', 'Beane', 'Beckwith',
'Frauenthal', 'Jussila', 'McCoy', 'Harder', 'Frolicher-Stehli', 'Jonsson',
'Moubarek', 'Goldenberg', 'Nakid', 'Dick', 'Taylor', 'Duff Gordon', 'Daly',
'Greenfield'}
```



[19]:

	Age	Cabin	Embarked	Fare	Name	Parch	PassengerId
0	22.0	NaN	S	7.2500	Braund, Mr. Owen Harris	0	1
1	38.0	C85	C	71.2833	Cumings, Mrs. John Bradley (Florence Briggs Th…	0	2
2	26.0	NaN	S	7.9250	Heikkinen, Miss. Laina	0	3
3	35.0	C123	S	53.1000			
4	35.0	NaN	S	8.0500			

```
3       Futrelle, Mrs. Jacques Heath (Lily May Peel)      0           4
4                            Allen, Mr. William Henry      0           5

   Pclass     Sex  SibSp  Survived            Ticket Title    Surname  \
0       3    male      1       0.0         A/5 21171    Mr     Braund
1       1  female      1       1.0          PC 17599   Mrs    Cumings
2       3  female      0       1.0  STON/O2. 3101282  Miss  Heikkinen
3       1  female      1       1.0            113803   Mrs   Futrelle
4       3    male      0       0.0            373450    Mr      Allen

   FamilyGroup
0            2
1            2
2            1
3            2
4            2
```

[20]:
```
Andersson    11
Sage         11
Goodwin       8
Asplund       8
Davies        7
             ..
Baccos        1
Jansson       1
Pernot        1
Beesley       1
Anderson      1
Name: Surname, Length: 875, dtype: int64

/Users/kenxu/anaconda3/lib/python3.6/site-packages/ipykernel_launcher.py:1:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  """Entry point for launching an IPython kernel.
```

[61]:

	Surname	Sex	Age	Title	Survived	SurnameSurvival
13	Andersson	male	39.000000	Mr	0.0	0.500000
17	Williams	male	32.809148	Mr	1.0	0.250000
62	Harris	male	45.000000	Mr	0.0	0.333333
146	Andersson	male	27.000000	Mr	1.0	0.500000
155	Williams	male	51.000000	Mr	0.0	0.250000
219	Harris	male	30.000000	Mr	0.0	0.333333
224	Hoyt	male	38.000000	Mr	1.0	0.500000

249	Carter	male	54.000000	Officer	0.0	0.500000
304	Williams	male	28.421211	Mr	0.0	0.250000
390	Carter	male	36.000000	Mr	1.0	0.500000
428	Flynn	male	28.421211	Mr	0.0	0.333333
550	Thayer	male	17.000000	Mr	1.0	0.500000
570	Harris	male	62.000000	Mr	1.0	0.333333
572	Flynn	male	36.000000	Mr	1.0	0.333333
645	Harper	male	48.000000	Mr	1.0	0.500000
692	Lam	male	28.421211	Mr	1.0	0.500000
698	Thayer	male	49.000000	Mr	0.0	0.500000
735	Williams	male	28.500000	Mr	0.0	0.250000
793	Hoyt	male	42.919311	Mr	0.0	0.500000
825	Flynn	male	28.421211	Mr	0.0	0.333333
826	Lam	male	28.421211	Mr	0.0	0.500000
848	Harper	male	28.000000	Officer	0.0	0.500000

0.1.6 3.3. Family size

The kernal Titanic [EDA] + Model Pipeline + Keras NN also provide a good idea about the familysize.

We can see that the families with size 2 to 4 have relatively higher survival rate, so we can label the family size with 3 different type.

0.1.7 3.4. Cabin and Deck

There are many null values in 'Cabin' features. For the better predictions, it is believed to delete the feature of 'Cabin'. But here we need this feature because when the ship sinks, certain parts of the ship have different probability drown in water. So we deal to make a new feature to substitute the feature.

Here I refer to another kernal's engineering on `cabin`(Titanic: Tutorial, Encoding, Feature Eng, 81.8%) and simplified the code.

0.1.8 3.5. Ticket Group

4. Modeling

4.1. Random Forest

0.1.9 Model interpretation

4.2. XGBoost

```
/Users/kenxu/anaconda3/bin:/Users/kenxu/anaconda3/condabin:/Library/Frameworks/P
ython.framework/Versions/3.7/bin:/anaconda3/bin:/usr/bin:/bin:/usr/sbin:/sbin:/u
sr/local/bin:/Library/Java/JavaVirtualMachines/jdk1.8.0_231.jdk/Contents/Home:/L
ibrary/Java/JavaVirtualMachines/jdk1.8.0_231.jdk/Contents/Home/bin#:/Users/kenxu
/zeppelin-0.8.2-bin-all/bin
```