Machine Learning Project

Xin Li, xxl162230
Xiaoyu Zhang, xxz173130

1. Introduction

The sinking of the RMS Titanic is one of the most infamous shipwrecks in history. One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, children, and the upper-class.
The problem is to complete the analysis of what sorts of people were likely to survive. In particular, apply the tools of machine learning to predict which passengers survived the tragedy.
Basic approaches include missing data imputation, feature engineering, logistic regression and support vector machine.

2. Problem Definition

Task Definition

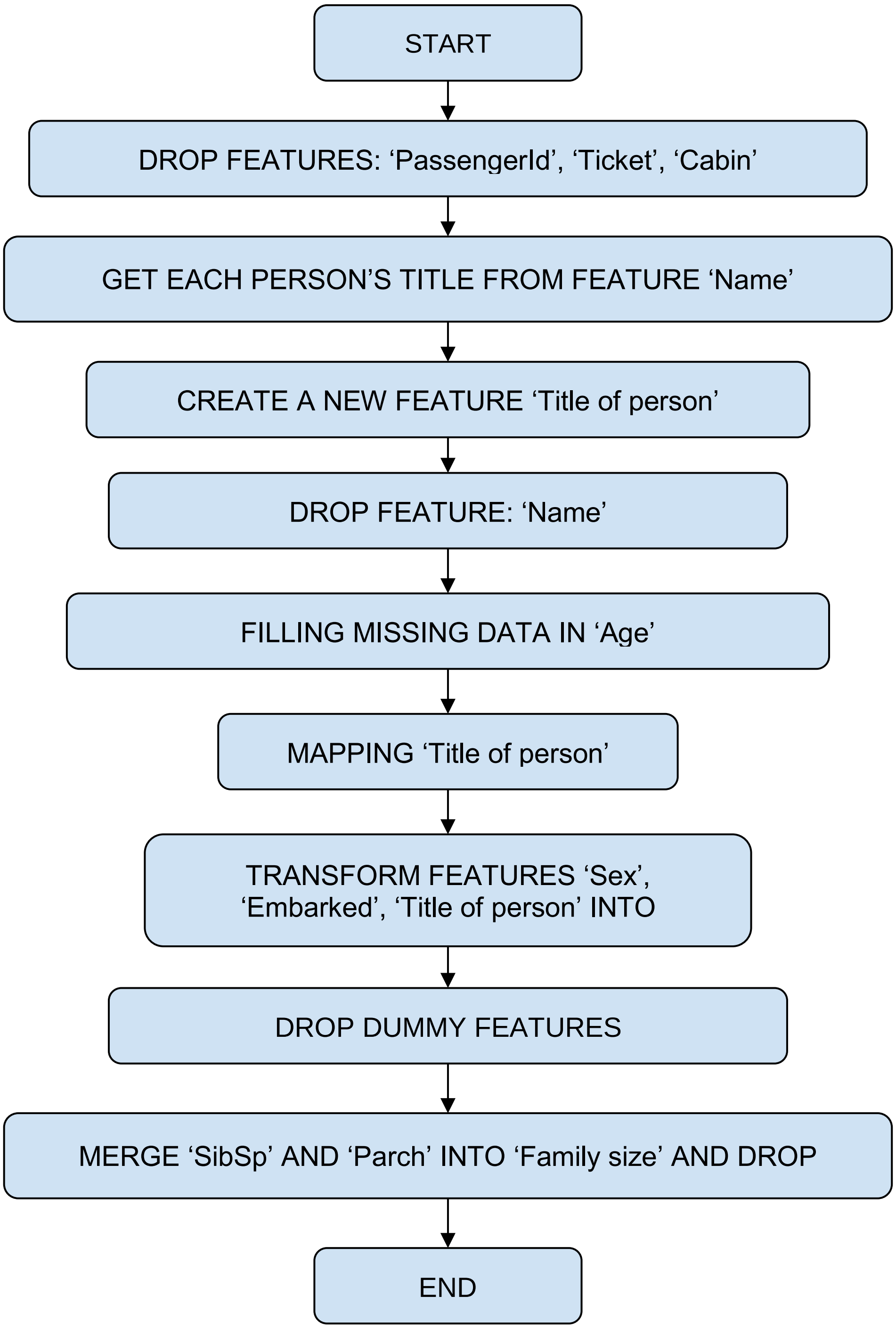
The data which problem provides has been split into training and test set. For the test set, problem does not provide the ground truth for each passenger. It is our job to predict these outcomes. In data set, we have some features as below.

Variable	**Definition**
survival	Survival
pclass	Ticket class
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

3. Experimental Evaluation

3.1 Methodology

1.The flowchart of data processing is shown below.



flowchart of data processing

2.In the first step, we use pandas.read_csv to read training data set, see what features are included and what the type they are.

```
E:\Anaconda3\envs\tensorflow\python.exe E:/python3/ml_proj/training_part.py
Content in raw data:

   PassengerId  Survived  Pclass  \
0            1         0       3
1            2         1       1
2            3         1       3
3            4         1       1
4            5         0       3

                                                Name     Sex   Age  SibSp  \
0                            Braund, Mr. Owen Harris    male  22.0      1
1  Cumings, Mrs. John Bradley (Florence Briggs Th...  female  38.0      1
2                             Heikkinen, Miss. Laina  female  26.0      0
3       Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0      1
4                           Allen, Mr. William Henry    male  35.0      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
```

Overview of data set

From the result, we know:
'Sex' and 'Embarked' features are categorical. Categorical feature should be encoded.

3.Then we use describe() function to generate the descriptive statistic to get the information about features of data set.

```
Data describe:

       PassengerId    Survived      Pclass         Age       SibSp  \
count   891.000000  891.000000  891.000000  714.000000  891.000000
mean    446.000000    0.383838    2.308642   29.699118    0.523008
std     257.353842    0.486592    0.836071   14.526497    1.102743
min       1.000000    0.000000    1.000000    0.420000    0.000000
25%     223.500000    0.000000    2.000000   20.125000    0.000000
50%     446.000000    0.000000    3.000000   28.000000    0.000000
75%     668.500000    1.000000    3.000000   38.000000    1.000000
max     891.000000    1.000000    3.000000   80.000000    8.000000

            Parch        Fare
count  891.000000  891.000000
mean     0.381594   32.204208
std      0.806057   49.693429
min      0.000000    0.000000
25%      0.000000    7.910400
50%      0.000000   14.454200
75%      0.000000   31.000000
max      6.000000  512.329200
```

Descriptive statistic

4.To prepare the data, we need to deal with missing data. We use draw_missing_data_table function to see the missing data. Sum up all missing date and show the number of missing data of each feature. The results are shown as below.

```
Missing data in training file:
Cabin           687
Age             177
Embarked          2
Fare              0
Ticket            0
Parch             0
SibSp             0
Sex               0
Name              0
Pclass            0
Survived          0
PassengerId       0
dtype: int64
```

```
Missing data in testing file:
Cabin           326
Age              86
Embarked          0
Fare              0
Ticket            0
Parch             0
SibSp             0
Sex               0
Name              0
Pclass            0
PassengerId       0
dtype: int64
```

Missing data

From the result, we get:
(1)'Cabin' has too many missing values. We decide to delete this feature.
(2)'Age' can be imputed. For now, we'll use data_proc(object) function in data_processing step. Object varys 0,1,2,3,4,5 and each of them represent a different method to filling gaps in a feature. Later, we will revise this function.

5.Drop features
From above result, we drop ‘Cabin’ feature.

6.Based on our experience, some feature, like ‘Ticket’ and ‘PassengerId’, have no correlation with the outcomes. So we decide to delete these two features.

7.To complete the imputation of ‘Age’ missing data, our method is data_proc(object) function. Different object represents different method as shown below.
  '0' -- drop vacant value
  '1' -- using mean value
  '2' -- using median value
  '3' -- using a previous value
  '4' -- using a next value
  '5' -- using the average age of corresponding title

When ‘object’ equals 5, method is to estimate the missing values based on known relationships. In this case, we can do this by using the information in the variable ‘Name’. Looking to 'Name' values, we can see person's name and title. Person's title is a relevant information to estimate ages. To give an example, we know that a person with the title 'Master' is someone under 13 years old, since 'a boy can be addressed as master only until age 12'. Therefore, employing the information in 'Name' we can improve our imputation method.
Method step are:
Extract titles from ‘Name’. Decrease title categories into 5(Mrs., Ms., Mr., Mrs., Other)
For each title, get people’s average age and use it to fill missing values.

8.Then transform feature ‘Sex’, ‘Embarked’, ‘Title’ into categorical. Change 'male' and 'female' in Sex into 1 and 0, 1 means 'male', 0 means 'female'. Delete ‘Embark_C’ and keep ‘Embark_Q’ and ‘Embark_S’.

9.Drop dummy features. We merge ‘SibSp’ and ‘Parch’ into new feature ‘Family size’.Then delete ‘SibSp’, ‘Parch’ and ‘Name’ features.

So far, we finish the data processing step.

For training step, we plan to use two methods: Logistic Regression and SVM. Use two different machine learning algorithm and use different data processing methods (object varies 0,1,2,3,4,5) to train the model. We use cross-validation to get the accuracy of each case.

3.2 Results

Here are the accuracy of Logistic Regression and SVM when using different data processing method.

Object / Method	0	1	2	3	4	5
Logistic Regression	0.8111	0.8212	0.8212	0.8212	0.8212	0.8156
SVM	0.8321	0.8268	0.8268	0.8268	0.8044	0.7932

4. Conclusion

After analysis of problem and compare the accuracy, using SVM method and data_proc(object=0) which means drop vacant value is the best way to predict outcomes of test data set.

5.References

Problem link:
https://www.kaggle.com/c/titanic