Who Were Most Likely to Survive the Titanic Catastrophe?

Using logistic Regression

Ezequiel
Vincent
Luz

California State University Northridge
NSF 1842386

June 2019

Data Cleaning

Dropped Features

The first features to be dropped were:

- Passenger ID: Not relevant towards survival
- Ticket: Also not relevant, mostly random string/number
- Cabin: 77% of data in this column was missing

	# of Missing	Percent
Cabin	1014	0.774637
Age	263	0.200917
Embarked	2	0.001528
Fare	1	0.000764
Ticket	0	0.000000

Kept Features

- Survived: What we want to predict
- Pclass: Class of passenger
- Name: The title in names can be used
- Sex: Male or female
- Age: Age in years, can include values less than a year to zero
- SibSp: Number siblings / spouses
- Parch: Number of parents / children
- Fare: Fare paid to board
- Embarked: Location embarked from

Data Types I

- Survived:
  - Nominal datatype
  - 1 meaning survival
- Pclass:
  - Ordinal datatype
  - 1,2,3 that represents the status of the passenger, with 1 being high-class and 3 being low-class
- Name:
  - Nominal data type.
  - Information such as Titles can be pulled from this

Data Types II

- Age and Fare:
  - Continuous and quantitative data types
- SibSp and Parch:
  - Discrete quantitative data types
  - Can be used to create another feature to describe Family Size
- Sex and Embarked:
  - Nominal data types
  - Can be turned into and used as dummy variables

Missing Data

- Age, Embarked, and Fare
  - These 3 columns had some missing data but not a significant amount
  - The mode, most frequent value, of each column was used to fill in missing data for that column

	# of Missing	Percent
Cabin	1014	0.774637
Age	263	0.200917
Embarked	2	0.001528
Fare	1	0.000764
Ticket	0	0.000000

Unique Title Counts

Mr	757
Miss	260
Mrs	197
Master	61
Rev	8
Dr	8
Col	4
Major	2
Mlle	2
Ms	2
Sir	1
Capt	1
Jonkheer	1
Don	1
Mme	1
Lady	1
the Countess	1
Dona	1

- From the Name we could identify meaningful socioeconomic Title and do One Hot Encoding on them
- We see that there are not many repeats for other Titles than Mr, Miss, Mrs and Master
- Using them, would result in a over fitting
- We found that using the first 4 gives us better results

Feature Engineering II

Created a new feature named Family size

- Combined Sibsp (number of siblings) and Parch (parents) to form the new feature Family Size
- Family Size was shown to be highly correlated to Sibsp and Parch, so these two variables were later dropped
- This new combined feature, as well as having less features to work with, improves modeling


Pearson Correlation of Features
Age 1 -0.046 0.19 -0.11 0.026 -0.4 0.045 -0.19 -0.053 -0.18 0.15 -0.25 0.21 -0.35
Embarked -0.046 1 -0.24 0.045 -0.052 0.19 0.098 0.066 -0.17 0.068 0.094 -0.05 -0.058 0.017
Fare 0.19 -0.24 1 0.22 0.031 -0.56 -0.19 0.16 0.26 0.23 -0.19 0.091 0.14 0.012
Parch -0.11 0.045 0.22 1 0.0089 0.018 -0.21 0.37 0.082 0.79 -0.3 0.068 0.22 0.25
PassengerId 0.026 -0.052 0.031 0.0089 1 -0.038 0.013 -0.055 -0.005 -0.031 0.014 -0.05 0.034 0.0023
Pclass -0.4 0.19 -0.56 0.018 -0.038 1 0.12 0.061 -0.34 0.05 0.12 0.03 -0.18 0.095
Sex_code 0.045 0.098 -0.19 -0.21 0.013 0.12 1 -0.11 -0.54 -0.19 0.87 -0.67 -0.57 0.16
SibSp -0.19 0.066 0.16 0.37 -0.055 0.061 -0.11 1 -0.035 0.86 -0.24 0.08 0.065 0.33
Survived -0.053 -0.17 0.26 0.082 -0.005 -0.34 -0.54 -0.035 1 0.017 -0.55 0.33 0.34 0.085
FamilySize -0.18 0.068 0.23 0.79 -0.031 0.05 -0.19 0.86 0.017 1 -0.33 0.09 0.16 0.36
Mr 0.15 0.094 -0.19 -0.3 0.014 0.12 0.87 -0.24 -0.55 -0.33 1 -0.58 -0.49 -0.26
Miss -0.25 -0.05 0.091 0.068 -0.05 0.03 -0.67 0.08 0.33 0.09 -0.58 1 -0.21 -0.11
Mrs 0.21 -0.058 0.14 0.22 0.034 -0.18 -0.57 0.065 0.34 0.16 -0.49 -0.21 1 -0.093
Master -0.35 0.017 0.012 0.25 0.0023 0.095 0.16 0.33 0.085 0.36 -0.26 -0.11 -0.093 1
Age
Embarked
Fare
Parch
PassengerId
Pclass
Sex_code
SibSp
Survived
FamilySize
Mr
Miss
Mrs
Master
0.9
0.6
0.3
0.0
−0.3
−0.6

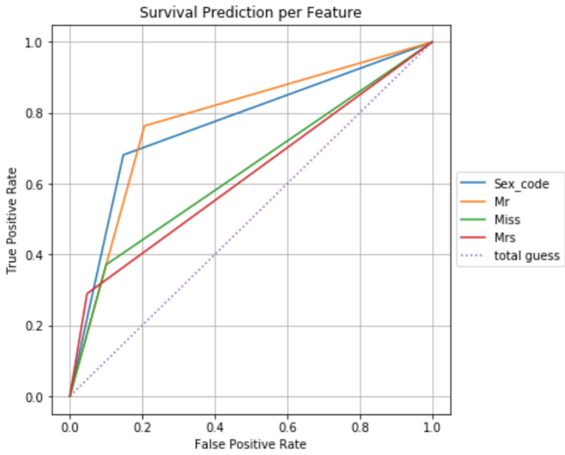
Model Testing

Cleaning is done.
We are left with a total of 10 Features:

- Mr, Mrs, Miss, and Master
- Sex_code
- FamilySize
- Fare
- Age
- Embarked
- Pclass

Feature Prediction Testing

Taking each feature individually, we test to look if they are a good predictor of Survival



Feature	Accuracy
Sex_code	0.786756
Title_Mr	0.782267
Title_Mrs	0.698092
Title_Miss	0.696970
Pclass	0.679012
Fare	0.666667
Embarked	0.636364
Title_Master	0.622896
Age	0.616162
FamilySize	0.616162

- Logistic regression requires the features to be scaled. We used the sklearn.preprocessing.StandardScaler
- Evaluate different combination of parameters by doing a Grid Search with Cross-Validation over the following processed datasets
  - Standardized Features (9 features)
  - PCA with 7 components
  - PCA with 8 components

```
Cumulative sum:  0.9601661693062497     with 7 components
Cumulative sum:  0.9933474854884274     with 8 components
```

Logistic Regression Parameters

- Logistic Regression has 5 different solvers, where each is able to perform different norm penalizations
- Below is our parameters grid

```
param_grid = [
    { 'solver':['liblinear'] },
    { 'solver':['newton-cg'], 'penalty':['l2', 'none'], 'multi_class':['ovr', 'multinomial', 'auto'] },
    { 'solver':['lbfgs'], 'penalty':['l2', 'none'], 'multi_class':['ovr', 'multinomial', 'auto'] },
    { 'solver':['saga'], 'l1_ratio':[0.5, 0.6, 0.75, 0.9], 'max_iter':[800], 'penalty':['elasticnet'] },
    { 'solver':['sag'], 'penalty':['l2'] }
]
```

Standardized Features Accuracy

```
Best parameters:
 {'l1_ratio': 0.75, 'max_iter': 800, 'penalty': 'elasticnet', 'solver': 'saga'}
 0.8305274971941639

0.8271604938271605 {'solver': 'liblinear'}
0.8271604938271605 {'multi_class': 'ovr', 'penalty': 'l2', 'solver': 'newton-cg'}
0.8282828282828283 {'multi_class': 'ovr', 'penalty': 'none', 'solver': 'newton-cg'}
0.8271604938271605 {'multi_class': 'multinomial', 'penalty': 'l2', 'solver': 'newton-cg'}
0.8282828282828283 {'multi_class': 'multinomial', 'penalty': 'none', 'solver': 'newton-cg'}
0.8271604938271605 {'multi_class': 'auto', 'penalty': 'l2', 'solver': 'newton-cg'}
0.8282828282828283 {'multi_class': 'auto', 'penalty': 'none', 'solver': 'newton-cg'}
0.8271604938271605 {'multi_class': 'ovr', 'penalty': 'l2', 'solver': 'lbfgs'}
0.8282828282828283 {'multi_class': 'ovr', 'penalty': 'none', 'solver': 'lbfgs'}
0.8271604938271605 {'multi_class': 'multinomial', 'penalty': 'l2', 'solver': 'lbfgs'}
0.8282828282828283 {'multi_class': 'multinomial', 'penalty': 'none', 'solver': 'lbfgs'}
0.8271604938271605 {'multi_class': 'auto', 'penalty': 'l2', 'solver': 'lbfgs'}
0.8282828282828283 {'multi_class': 'auto', 'penalty': 'none', 'solver': 'lbfgs'}
0.8282828282828283 {'l1_ratio': 0.5, 'max_iter': 800, 'penalty': 'elasticnet', 'solver': 'saga'}
0.8282828282828283 {'l1_ratio': 0.6, 'max_iter': 800, 'penalty': 'elasticnet', 'solver': 'saga'}
0.8305274971941639 {'l1_ratio': 0.75, 'max_iter': 800, 'penalty': 'elasticnet', 'solver': 'saga'}
0.8305274971941639 {'l1_ratio': 0.9, 'max_iter': 800, 'penalty': 'elasticnet', 'solver': 'saga'}
0.8271604938271605 {'penalty': 'l2', 'solver': 'sag'}
```

PCA Accuracy

- Using 7 components

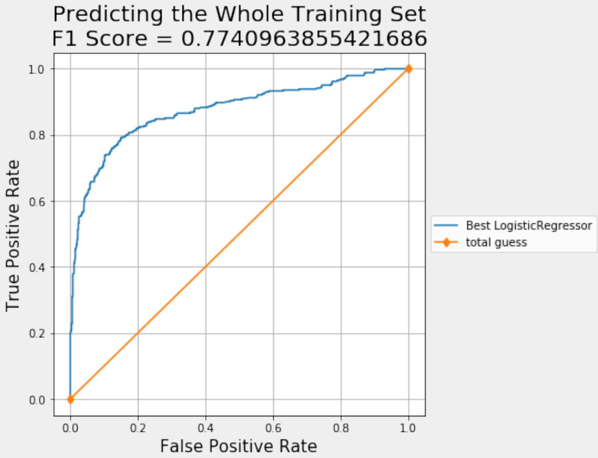
```
Best parameters:
 {'solver': 'liblinear'}
 0.8237934904601572
```

- Using 8 components

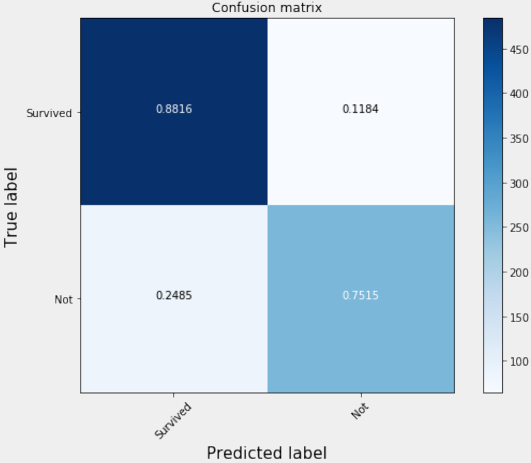
```
Best parameters:
 {'multi_class': 'ovr', 'penalty': 'none', 'solver': 'newton-cg'}
 0.8260381593714927
```

- Standarized Features results in better accuracy!

ROC Curve


Predicting the Whole Training Set
F1 Score = 0.7740963855421686
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Best LogisticRegressor
total guess

	Survived	Not
Survived	484	65
Not	85	257



Accuracy=0.8316 | Misclass=0.1684
Recall=0.8506 | Precision=0.8816

Kaggle Scores I

Kaggle Score: 0.58851

- Submitted a file with only passengers of Pclass 1 or 2 surviving

Kaggle Score: 0.76555

- Kaggle sample submission where all females on-board survived

Kaggle Scores II

Kaggle Score: 0.79425

- Used the "liblinear" logistic regression solver with all variables after cleaning

Kaggle Score: 0.79904

- Used the "saga" solver
- Elasticnet penalty and l1_ratio of 0.75
- "saga" looks to be the best solver to use

Kaggle Score: 0.80382

- "saga" solver used again, but this time the SibSp and Parch features were dropped
- Elasticnet penalty and l1_ratio of 0.75
- This resulted to be the best model

References

- https://www.kaggle.com/ldfreeman3/a-data-science-framework-to-achieve-99-accuracy
- https://www.kaggle.com/startupsci/titanic-data-science-solutions

Thanks!

NSF 1842386