Titanic: Machine Learning from Disaster – Kaggle Competition

Gareth Austen

Data Science

General Assembly

April 2015

Agenda

1. Defining the Problem
2. Overview of Kaggle Data and Data Manipulation
3. Model Selection and Methodology
4. Results of the Modeling process
5. Future Steps!

Overview of the Problem

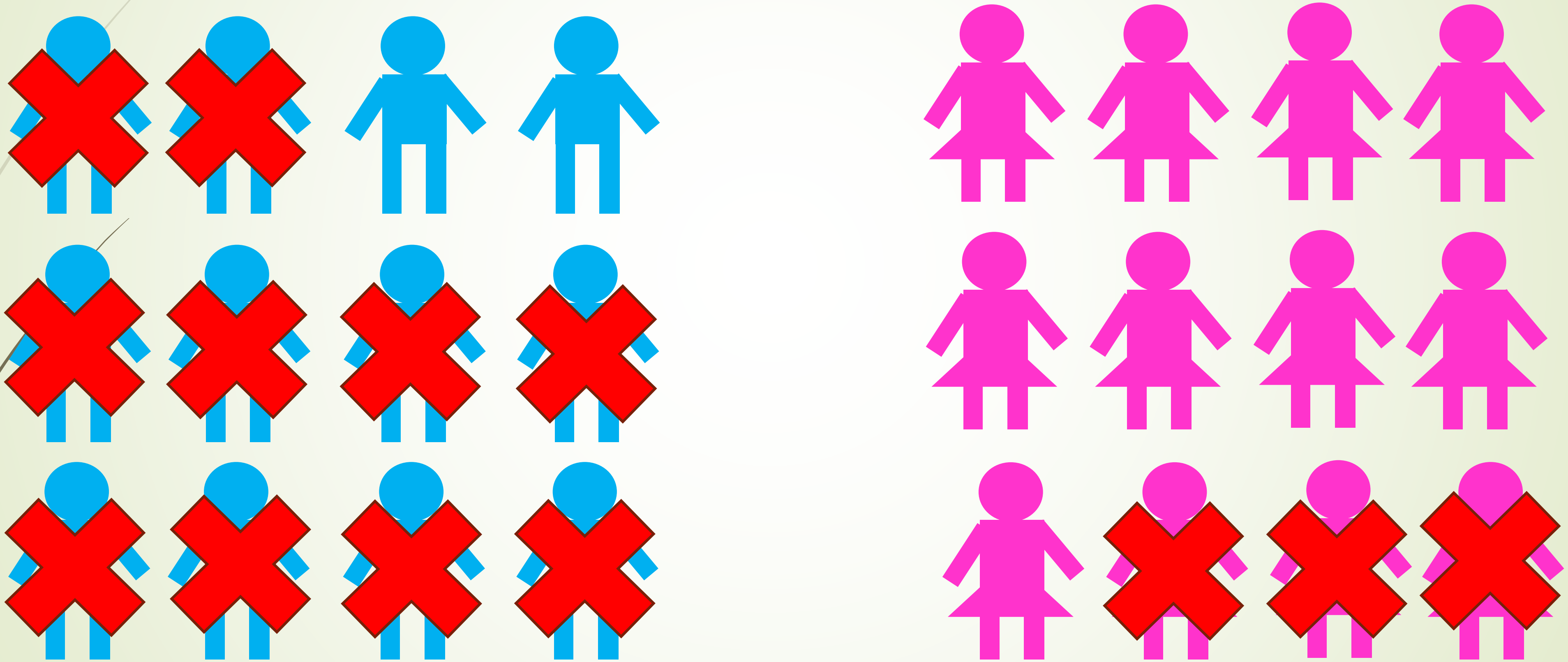
- One of the most infamous shipwrecks in history
- 1502 out of 2224 passengers were killed when the Titanic sunk
- Some groups are more likely to survive – eg women and children
- The Challenge: To predict passengers likelihood to survive

Overview of Kaggle Data and Data Manipulation

Overview of the Data

- Kaggle provides the data in the form of two csv's – one training and one test set
- Both datasets contain the following variables:
  - Survival
  - Pclass - Passenger Class
  - Name
  - Sex
  - Age
  - Sibsp
  - parch
  - Ticket
  - Fare
  - Cabin
  - Cabin Number
  - Port of Embarkation

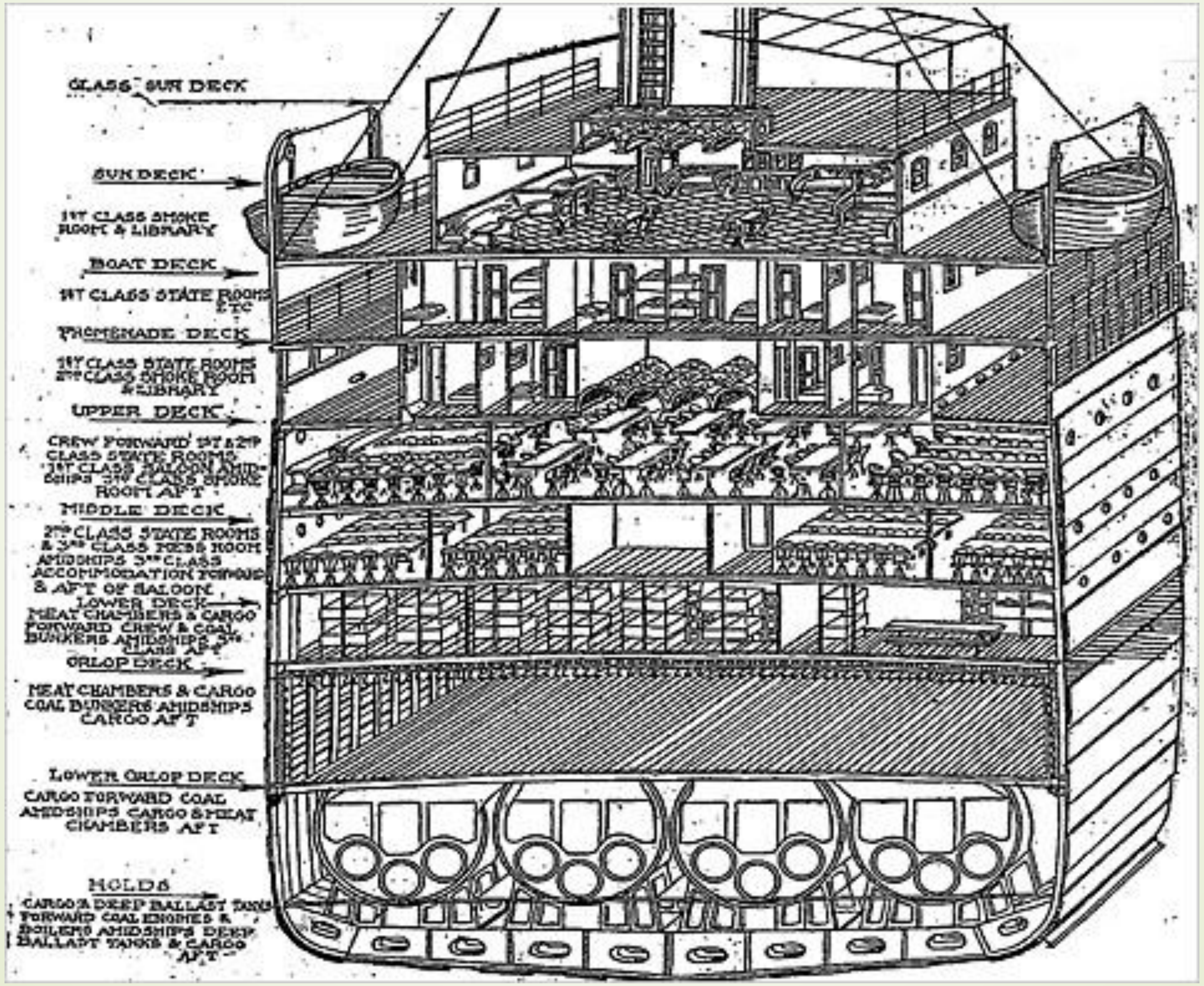
Male vs Female Survival Rate



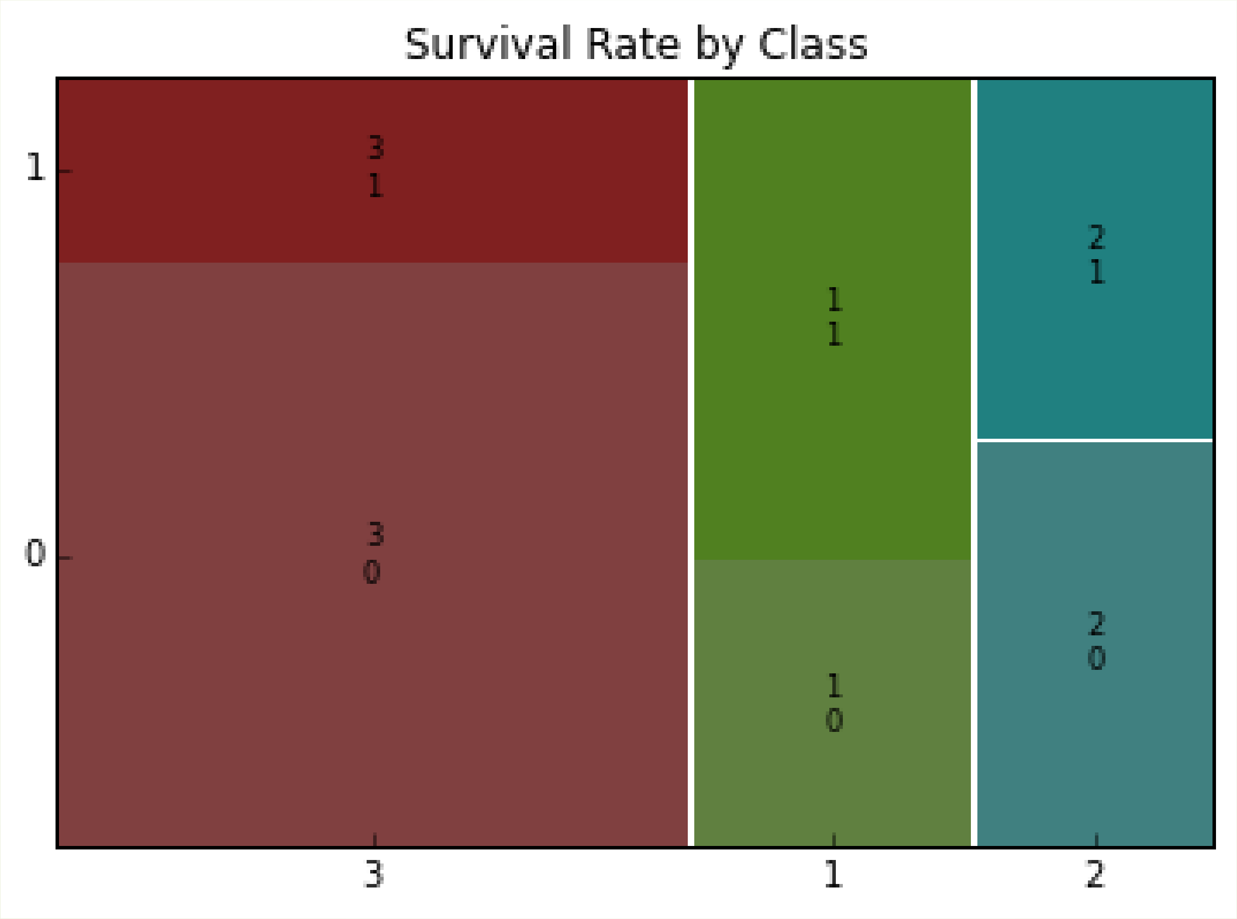
Male Survival Rate: 18%

Female Survival Rate: 75%

Cross – Section of the Titanic

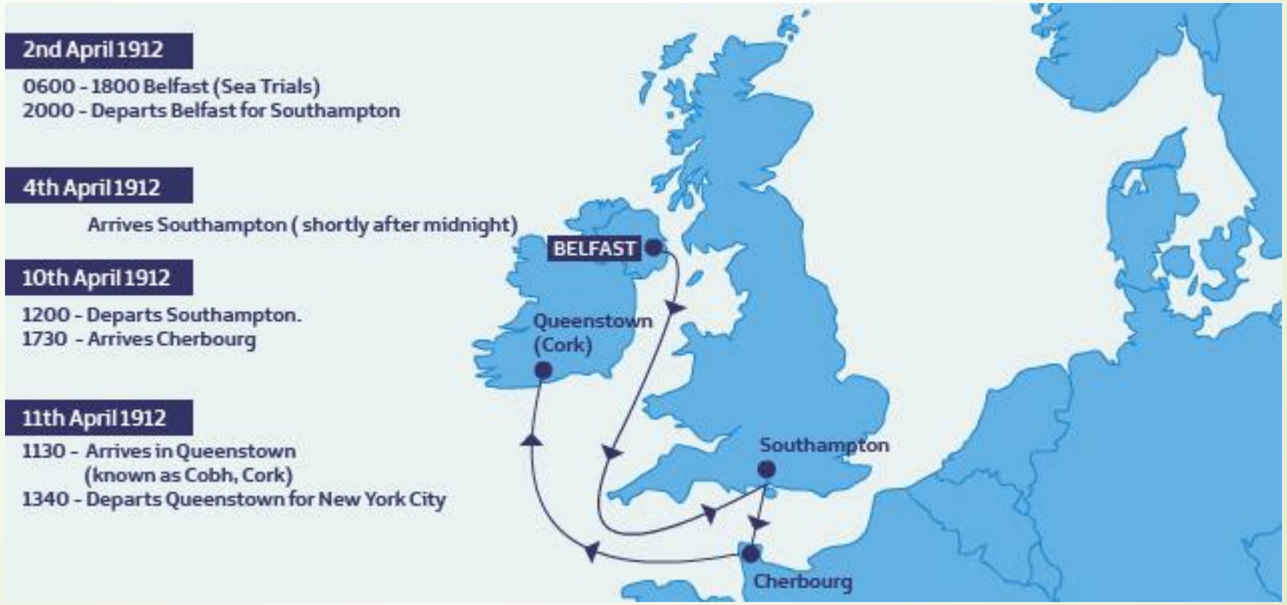

GLASS SUN DECK
SUN DECK
1ST CLASS SMOKE ROOM & LIBRARY
BOAT DECK
1ST CLASS STATE ROOMS ETC
PROMENADE DECK
1ST CLASS STATE ROOMS 2ND CLASS SMOKE ROOM & LIBRARY
UPPER DECK
CREW FORWARD 1ST & 2ND CLASS STATE ROOMS 1ST CLASS SALOON AMIDSHIPS 3RD CLASS SMOKE ROOM AFT
MIDDLE DECK
2ND CLASS STATE ROOMS & 3RD CLASS MESS ROOM AMIDSHIPS 3RD CLASS ACCOMMODATION FORWARD & AFT OF SALOON
LOWER DECK
MEAT CHAMBERS & CARGO FORWARD CREW & COAL BUNKERS AMIDSHIPS 3RD CLASS AFT
ORLOP DECK
MEAT CHAMBERS & CARGO COAL BUNKERS AMIDSHIPS CARGO AFT
LOWER ORLOP DECK
CARGO FORWARD COAL AMIDSHIPS CARGO & MEAT CHAMBERS AFT
HOLDS
CARGO & DEEP BALLAST TANKS FORWARD COAL ENGINES & BOILERS AMIDSHIPS DEEP BALLAST TANKS & CARGO AFT

Passenger Survival Rate by Class



Further Data Manipulation – Port of Embarkation

- The majority of passengers embarked at Southampton in England
- Two variables "Q_Port" and "C_Port" were created to indicate passengers who embarked at Cherbourg in Queenstown



2nd April 1912

0600 - 1800 Belfast (Sea Trials)
2000 - Departs Belfast for Southampton

4th April 1912

Arrives Southampton (shortly after midnight)

10th April 1912

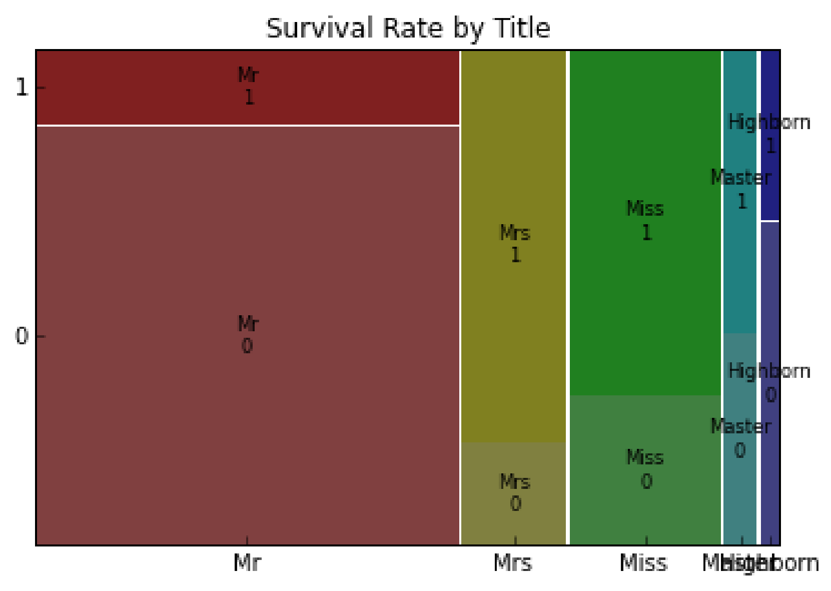
1200 - Departs Southampton.
1730 - Arrives Cherbourg

11th April 1912

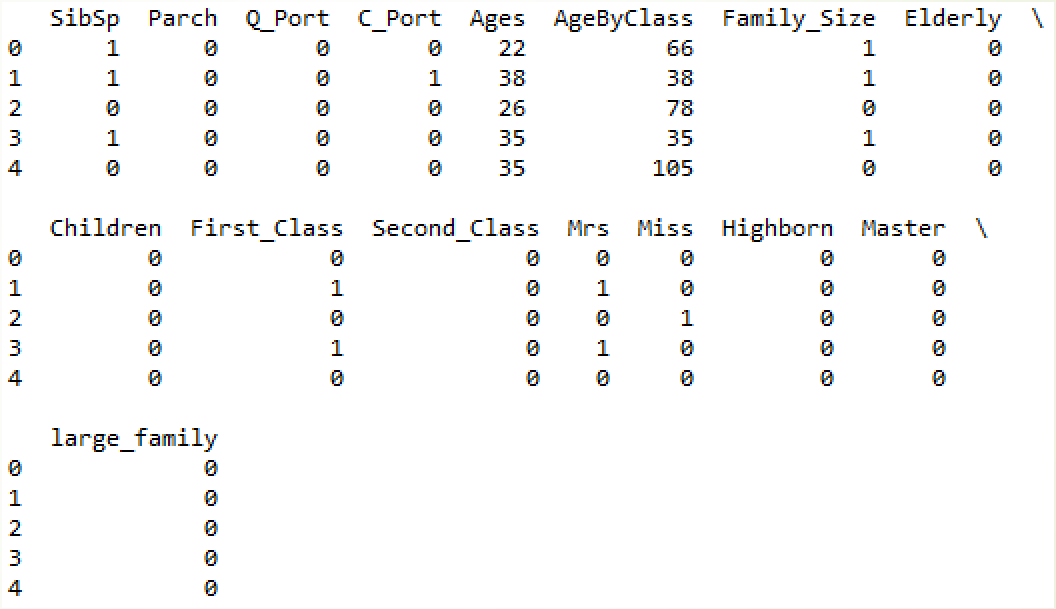
1130 - Arrives in Queenstown
(known as Cobh, Cork)
1340 - Departs Queenstown for New York City

Title Manipulation

- Split titles out from the names column in the data
- Why ?
  - To give more insight into the type of people on board
  - Some highborn people of many different titles
  - Using Title to try and improve average age is slightly more accurate
  - Allows the creation of some additional Categorical Variables



Final Modeling Dataset



```
   SibSp  Parch  Q_Port  C_Port  Ages  AgeByClass  Family_Size  Elderly  \
0      1      0       0       0    22          66            1        0
1      1      0       0       1    38          38            1        0
2      0      0       0       0    26          78            0        0
3      1      0       0       0    35          35            1        0
4      0      0       0       0    35         105            0        0

   Children  First_Class  Second_Class  Mrs  Miss  Highborn  Master  \
0         0            0             0    0     0         0       0
1         0            1             0    1     0         0       0
2         0            0             0    0     1         0       0
3         0            1             0    1     0         0       0
4         0            0             0    0     0         0       0

   large_family
0             0
1             0
2             0
3             0
4             0
```

Model Selection and Methodology

Model Selection

- What type of problem is it?
  - Binary Classification Problem
  - Dead (0) or Alive (1)

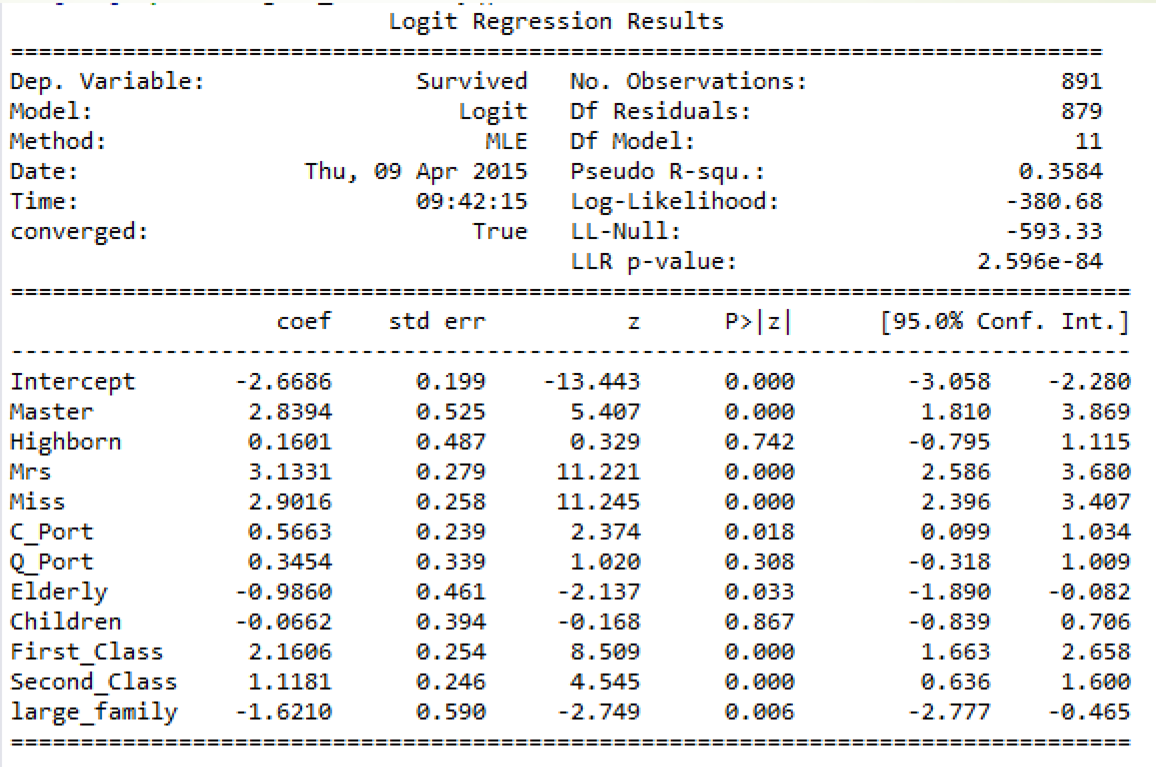
- Models Selected:
  - Logistic Regression
  - Random Forests
  - Boosting Trees

Methodology

- Logistic Regression showing poor results after initial testing
- Only optimize the parameters for Random Forest and Boosting Trees
- Used a combination of Recursive Feature Elimination and Grid Search to find the optimal parameters and features to be used in each model
- Random Forests:
  - Optimized for the best number of trees between a range of 10 and 750
  - 5 fold Cross-Validation
- Boosting Trees
  - Tuned the following parameters:
    - Learning Rate Range
    - Sub-Sampling Range
    - Number of Estimators Range

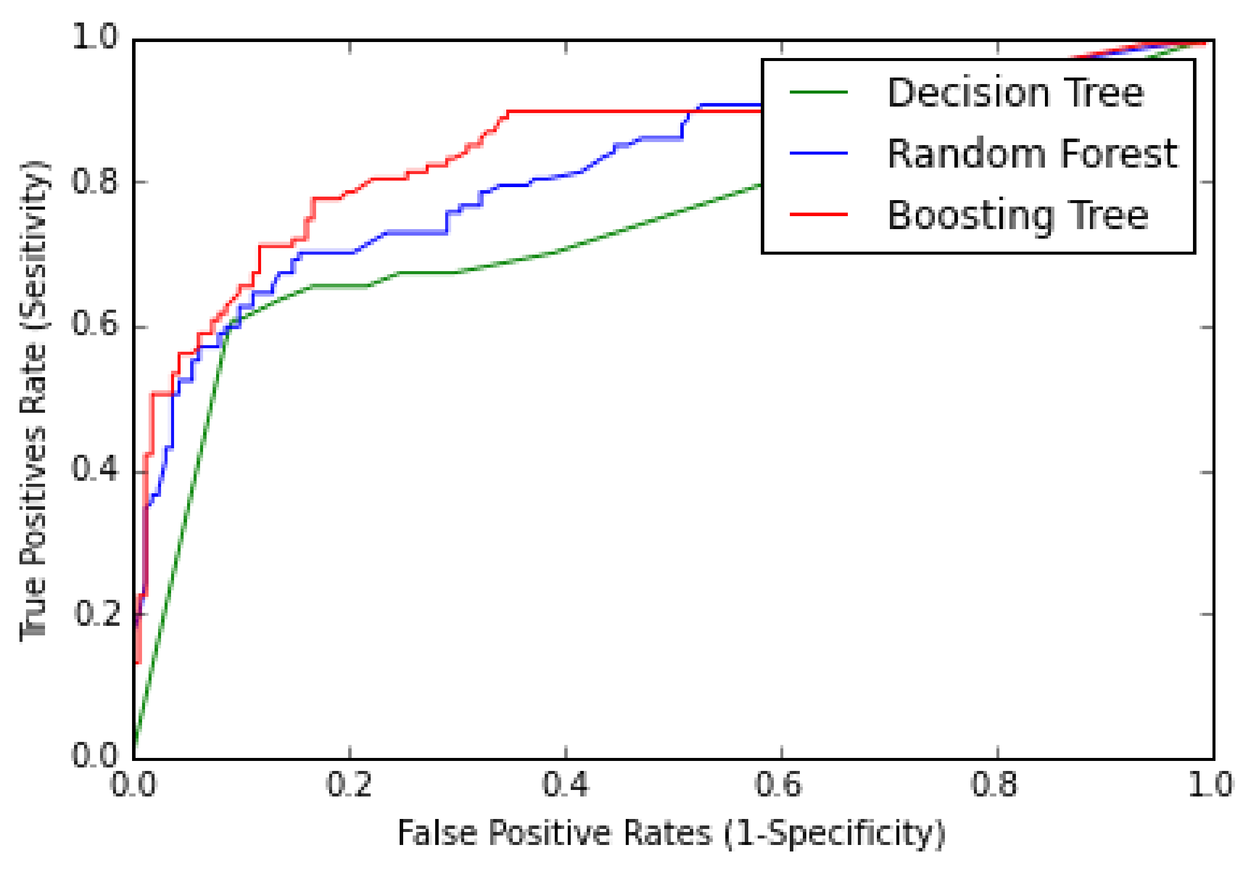
Modeling Results

Logistic Regression

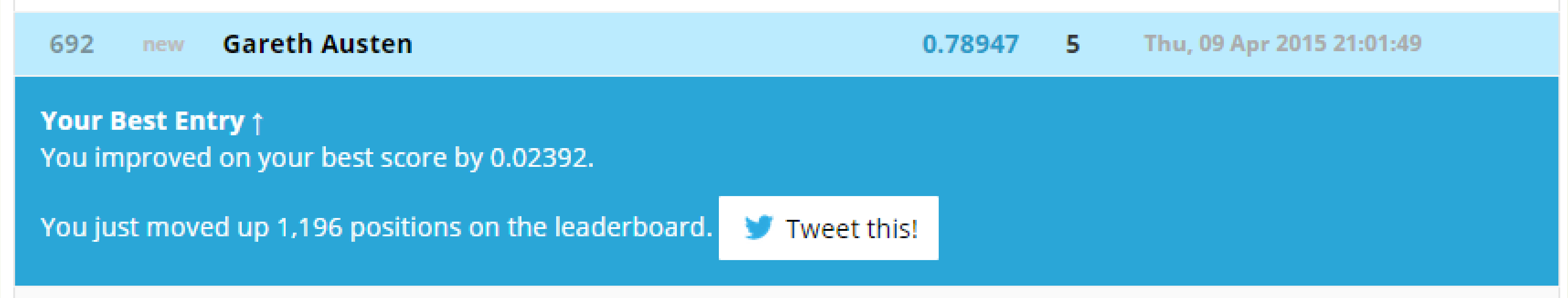


```
                           Logit Regression Results
==============================================================================
Dep. Variable:               Survived   No. Observations:                  891
Model:                          Logit   Df Residuals:                      879
Method:                           MLE   Df Model:                           11
Date:                Thu, 09 Apr 2015   Pseudo R-squ.:                  0.3584
Time:                        09:42:15   Log-Likelihood:                -380.68
converged:                       True   LL-Null:                       -593.33
                                        LLR p-value:                 2.596e-84
================================================================================
                   coef    std err          z      P>|z|      [95.0% Conf. Int.]
--------------------------------------------------------------------------------
Intercept       -2.6686      0.199    -13.443      0.000        -3.058    -2.280
Master           2.8394      0.525      5.407      0.000         1.810     3.869
Highborn         0.1601      0.487      0.329      0.742        -0.795     1.115
Mrs              3.1331      0.279     11.221      0.000         2.586     3.680
Miss             2.9016      0.258     11.245      0.000         2.396     3.407
C_Port           0.5663      0.239      2.374      0.018         0.099     1.034
Q_Port           0.3454      0.339      1.020      0.308        -0.318     1.009
Elderly         -0.9860      0.461     -2.137      0.033        -1.890    -0.082
Children        -0.0662      0.394     -0.168      0.867        -0.839     0.706
First_Class      2.1606      0.254      8.509      0.000         1.663     2.658
Second_Class     1.1181      0.246      4.545      0.000         0.636     1.600
large_family    -1.6210      0.590     -2.749      0.006        -2.777    -0.465
================================================================================
```

Random Forests vs Boosting Trees



Where I placed on Kaggle?



692 new Gareth Austen 0.78947 5 Thu, 09 Apr 2015 21:01:49

Your Best Entry ↑

You improved on your best score by 0.02392.

You just moved up 1,196 positions on the leaderboard.

Future Steps

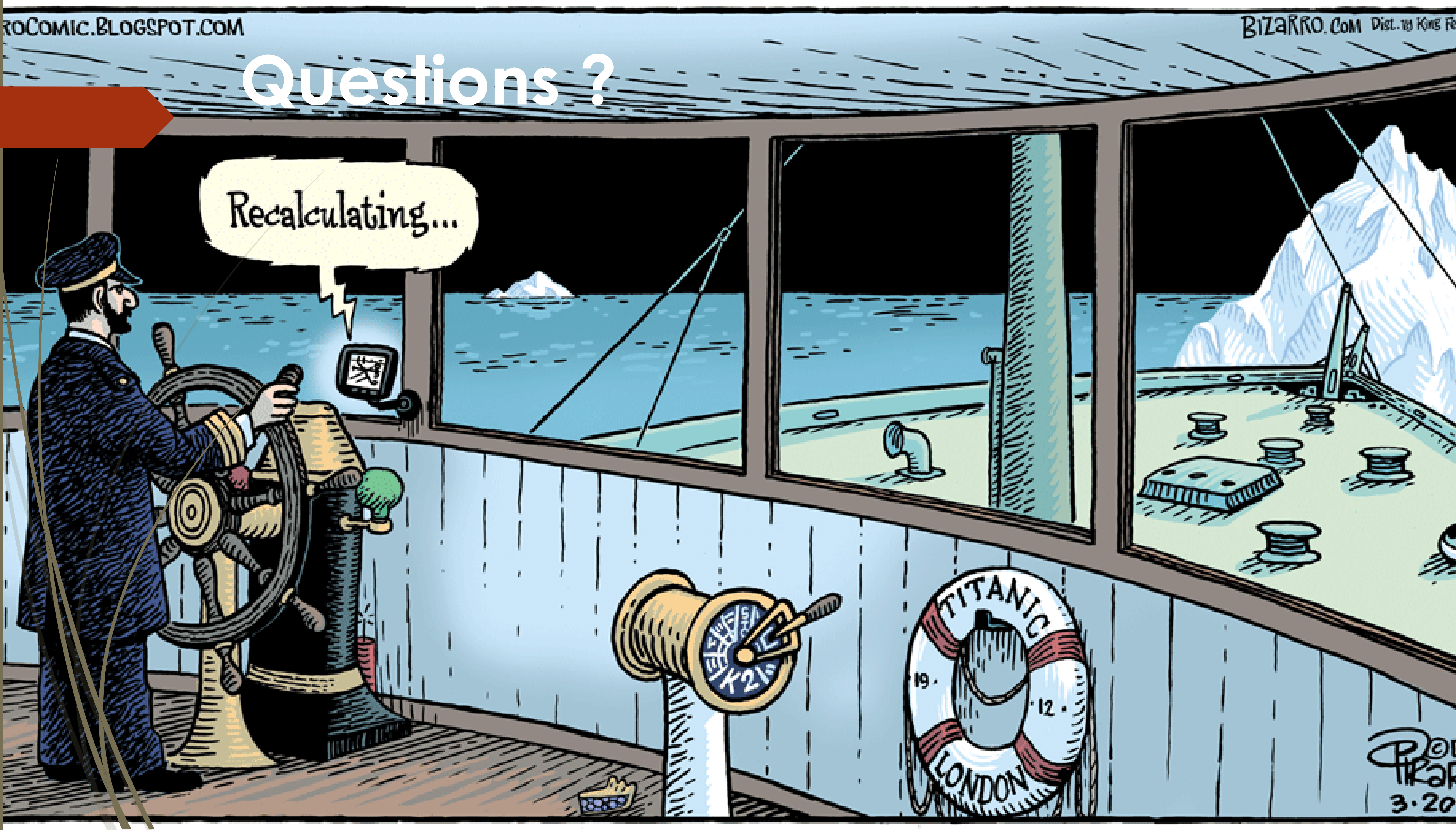
Future Steps and Key Takeaways

- Create further features using combinations of variables
  - Such as combining family size with surname
  - Look at using tickets and cabin numbers
  - Further explore the string variables to try and find other useful features

- Continue to optimize modeling parameters for Random Forests and Boosting Trees

- **Get a faster computer!!**

Questions ?


ROCOMIC.BLOGSPOT.COM
BIZARRO.COM Dist. by King Fe
Recalculating...
TITANIC
19
12
LONDON
Pirar
3·20