**Problem: Can we predict the probability of survival in a disaster?**

Using data from the sinking of the Titanic, we attempted to determine if we could use statistical models to predict the probability of survival. We determined that our problem was a grouping/classification problem, and hence focused on the following models:

- Logistic Regression
- K Nearest Neighbors
- Decision Trees
- Bagging a Decision Tree

From the data, we know that 38.4% of the people on the Titanic survived the sinking. For our purposes, this is the baseline we judge against. Though this was an academic exercise, the ideal is that the model we produce is deployable in future situations as a data product. Hence, since this model evaluates the probability of survival, success would be:

1) The model has significantly better than random accuracy (i.e. at least 75%)
2) The model has a great recall rate - in other words, its True Positive Rate is high, and we bias towards the True Positive Rate (i.e. perhaps greater than 70%).

Essentially, the model biases towards minimizing false negatives, since we assume that clients would prefer to allocate sufficient resources to respond to a disaster (save survivors) than come into a disaster with insufficient resources

Before going into discussing our models, some key assumptions and risks about the model are below.

*Assumptions:*

1) We are assuming that the variables in the data are predictors of survival;

2) We are assuming that the socio-economic distribution of the Titanic (pclass) is potentially meaningful and translatable to today (i.e. people making greater than $x are more likely to survive in a future disaster)

*Risks:*

1) We know from history that the Titanic did not have enough lifeboats or an orderly process for evacuation. Such process improvement may mean that using this model on other disasters is less meaningful, even if we are being strict and looking at sea-based disasters such as a sinking cruise ship.

2) The model may not be translatable to today's type of disasters.

3) Missing values and their imputation might actually skew results.

- Age is probably a predictor (bias towards saving the young, potentially). However, this data was missing for at least 100 of the passengers, and hence we dropped this data
- Levels on the Titanic could have been a significant predictor. However, only 200 rows had this data out of 700+ rows in the data, so we could not use this

4) Multicollinearity/Correlation of certain variables (Fare and PClass)

5) Missing relational data - the SibSp and ParentChild count variables neglect certain close familial relationships (close friend from a town; or mistress/fiancee).'

**Discussion of Results**

After cleaning the data, we ran several models on the data. We focused on Logistic Regression and Decision Trees for the following reasons:

1) Logistic Regression is parametric, and so we can use it to give a good probability estimate formula depending on the type of passenger
2) Decision Trees deploy if/then logic, essentially, and hence can also be rapidly deployable

One issue that may have made our Logistic Regression less deployable in the field is that since Age and Fare had a different scale by an order of 10 from the other variables, we used RobustScaler to scale them down. Hence, the coefficients for Logistic Regression are less intuition with scaled data than unscaled data.

Looking at our models, the best ones on accuracy were our optimized Decision Tree (pre-scaled) and K Nearest Neighbors. However, looking below, we can see how close in terms of accuracy most of our models were:

'Optimized KNN': 0.83240223463687146,
'Bagging Decision Tree': 0.83240223463687146,
‘Optimized Logistic on scaled data': 0.82681564245810057,
'Basic Logistic Regression': 0.8044692737430168,
'Optimized Decision Tree': 0.82122905027932958,
'Optimized Logistic': 0.81564245810055869,
'Optimized Bagging Decision Tree': 0.81005586592178769,
'Decision Tree': 0.75977653631284914,
'Logistic Regression without Features Tossed out by Decision Tree': 0.73743016759776536

However, we mentioned that we wanted a model that minimized false negatives. From that perspective, the Logistic Regressions outperformed, looking at selected confusion matrices on testing data (data we had held out from the overall data set based on a train/test split with stratification of our target variable).

*Confusion Matrix for Scaled Logistic Regression*

| | pred_not_survived | pred_survived |
|---|---|---|
| Did not survive | 96 | 10 |
| Survived | 21 | **52** |

*Confusion Matrix for Optimized KNN*

| | pred_not_survived | pred_survived |
|---|---|---|
| Did not survive | 98 | 8 |
| Survived | 22 | 51 |

*Bagging Decision Tree Confusion Matrix*

| | pred_not_survived | pred_survived |
|---|---|---|
| Did not survive | 99 | 7 |
| Survived | 23 | 50 |

In short, the accuracy shortfall for the Logistic Regression compared to Decision Trees and KNN were because the Logistic Regression biased in this case towards putting more people in the survived group, which is the type of bias we want in disaster planning. Hence, on that basis, it is a superior model.

To conclude, we determined the following:

1) We could evaluate the probability of survival based on Titanic data with reasonable accuracy compared to both the baseline in the data (38.4%) and random chance (50%)
2) Our Decision Trees are also reasonable out of the box solutions
3) However, our recall even on the Logistic Regression was 71% (52 survivors accurately predicted/73 total survivors in the test data). This is right at the cusp of what I mentioned an acceptable True Positive Rate.

**Followups:**

1. Can age be imputed for the passengers with missing values? This would expand the data.
2. Could Cabin data be imputed for the passengers where it is missing, perhaps off ticket data? If so, that could improve the model