Matt Thomson
April 27th, 2017
Titanic Final Project

Introduction

This project focuses on predicting the survival of the RMS Titanic's 2224 passengers. Unfortunately 1502 (67.54%) of the passengers did not survive. This project explores the data of the passengers and builds a predictive model whose goal is to most accurately predict the outcome of survival based on the known data.

Description of the Data

Feature	Data Type	Description
PassengerId	Unique	The ID of the passenger
Survived	Binary	Survival (0 = No, 1 = Yes)
Pclass	Ordinal	Class of the Ticket (1 = 1st class (upper), 2 = 2nd class (middle), 3 = 3rd class (lower))
Name	Unique	Name of the passenger
Sex	Binary	Sex of the passenger (female, male)
Age	Continuous	Age of the passenger in years
SibSp	Discrete	Number of siblings and spouses on the Titanic
Parch	Discrete	Number of parents and children on the Titanic
Ticket	Discrete	Ticket Number
Fare	Continuous	Cost of the ticket
Cabin	Discrete	Cabin Number
Embarked	Nominal	Port the passenger embarked from.

Feature	Completeness	Uniqueness
PassengerId	100%	100%
Survived	100%	0%
Pclass	100%	0%
Name	100%	100%
Sex	100%	0%
Age	80.13%	N/A
SibSp	100%	0.01%
Parch	100%	0.01%
Ticket	100%	76.43%
Fare	100%	N/A
Cabin	22.90%	22.90%
Embarked	99.78%	0%

Looking at the completeness and uniqueness of the dataset will give us foresight into what datasets will need to be rebuilt or removed.

Preprocessing of Data

This process was broken into four major steps, Feature Reconstruction, Feature Transformation, Feature Generation, and Feature Reduction, where some or all of them were applied to each feature or set of features.

Feature Reconstruction

The goal of feature reconstruction is to most accurately complete missing values in the dataset. The Age, Fare, and Embarked features had to be reconstructed in this manner.

It was decided that the Age feature is to be reconstructed based on the mean age of the person's Sex and Class features. For example, if the person is a female in second class without

an age, the age for this person would be reconstructed as the mean age of all females in second class.

The Fare feature was reconstructed in the same way as the Age feature.

The Embarked feature was reconstructed as the mode of the Embarked feature.

Feature Transformation

The goal of Feature Transformation is to map the current values in the feature to a set of values that are able to be handled by the classifiers. This is often a set of small integers.

Often times the data needs to be converted from continuous data to a discrete set. For this dataset, almost every feature underwent a feature transformation.

Feature Generation

The goal of Feature Generation is to take one or more features and combine them in such a way that they create one impactful feature. This project creates three new features, Sex-Class, Family, and Title.

Sex-Class is all combinations of the Sex and Pclass features. This feature more distinctly separates people.

Family Size is the number of spouses and siblings (SibSp) plus the number of parents and children (Parch) plus the person to get an estimated family size.

Title is the title that the person goes by. For example, Mr., Mrs, Miss, etc. This feature was inspired by Megan L. Risdal's kernel report.[1] Although the implementation is a bit different, the idea behind it is largely the same.

Feature Reduction

The goal of Feature Reduction is to reduce the dimensionality of the dataset before running it through various classifiers. This improves performance times of the classifiers as well as hopes to avoid problems related to the curse of dimensionality.

[1] https://www.kaggle.com/mrisdal/titanic/exploring-survival-on-the-titanic

Many features were ultimately removed from the dataset, but most persisted through Feature Generation or Feature Transformation. In the end only the Embarked feature remained unmodified.

Details of the Feature Manipulation

Sex

The Sex feature is one of the most influential features in this dataset. Care was taken to not use this influence to skew any other features. In the end Sex was combined with Pclass to generate a new, nominal, feature called SexClass. The original feature, Sex, was removed from the dataset during the Feature Reduction stage.

Pclass

The class feature was used to create the SexClass feature as stated above. It was also removed from the dataset during the Feature Reduction stage.

SexClass

This feature was used to help reconstruct the Age and Fare features. The mean values of the respective features were used in place of erroneous or empty pieces of data.

Age

This feature was transformed from a continuous feature to two different discrete features, AgeRange and AgeFreq. The distribution in the AgeRange feature was based on equal years per range, whereas the AgeFreq feature was based on equal quantities of people per range.

It was analyzed that the AgeFreq distribution gave a more informative separation of age than AgeRange did. Therefore AgeFreq feature replaced the Age feature in the dataset.

Fare

This feature underwent the same distributions and analysis as the Age feature. It was also concluded that the FareFreq distribution was more telling than the FareRange feature. Therefore the FareFreq feature replaced the Fare feature in the dataset.

SibSp and Parch

Neither of these discrete features needed any transformations, however they were used to create a new feature of Family Size. They were both removed from the dataset during the Feature Reduction stage.

Embarked

This feature had to have some fields reconstructed. This was done so by taking the mode of the feature and applying it to erroneous or empty fields.

The feature was also transformed to a discrete integer set for processing by classifiers.

Name

The Name feature was transformed to only represent the title of the person. This created a fairly diverse set of titles. It was decided to condense most of the infrequent titles into an other category. The Name feature was removed from the dataset during the Feature Reduction stage.

Experimentation

Many experiments were done with various combinations of features to attempt to figure out the best use of the features and classifiers.

Each feature combination was performed on 17 different classifiers.

Features of Experiment 1: Embarked, SexClass, AgeFreq, FareFreq, FamilySize, Title
Features of Experiment 2: Exp 1 + Sex
Features of Experiment 3: Exp 1 + Pclass
Features of Experiment 4: Exp 1 - Title
Features of Experiment 5: Exp 1 - SexClass
Features of Experiment 6: Exp 1 - FamilySize

Classifier	Exp 1	Exp 2	Exp 3	Exp 4	Exp 5	Exp 6
Decision Tree	90.57	90.57	90.57	89.79	89.34	88.44
Random Forest	90.24	89.90	90.12	89.56	89.23	87.99
KNN-3	87.32	87.32	87.43	86.42	86.08	83.95
KNN-4	85.86	85.86	85.86	85.97	85.07	85.19
KNN-5	85.82	85.52	85.63	84.85	84.06	85.86
SVM	84.62	84.51	84.51	84.62	83.73	84.62
KNN-7	84.62	84.62	84.74	83.73	83.39	84.62
NN (15, 2)	83.50	84.96	85.07	84.62	83.05	83.39
NN (12, 2)	83.39	85.41	83.95	82.38	82.83	81.93
NN (12, 2, 2)	83.28	83.28	84.40	82.27	81.26	82.27
NN (8, 2)	81.59	82.27	61.62	83.28	61.62	61.62
NN (5, 5, 5)	81.26	81.82	81.26	75.31	82.49	78.79
Linear SVC	80.58	82.60	82.60	80.02	81.37	78.68
Naive Bayes	79.69	79.24	77.33	77.78	80.81	78.68
Stochastic Gradient Descent	79.80	80.02	80.70	74.41	78.00	78.68
Perceptron	78.00	70.71	79.80	79.91	75.08	78.79
NN (8, 5)	61.62	82.94	81.48	82.04	82.49	78.79

Results and Analysis

As it turns out, all of the features provide a positive influence on the classifiers in general. This means that the preprocessing was a large success. Further validation on the feature set that was used for experiment one was performed.

The training set provided had 300 random records used as testing data and the rest used as training data to perform cross validation. And upon running 10 cross validations of this nature and taking the mean of the scores for each classifier, all classifiers except Stochastic Gradient Descent and NN(12, 2) had similar mean scores.

Classifier	Results	Cross Validation
Decision Tree	90.57	91.86
Random Forest	90.24	91.41
KNN-3	87.32	86.532
KNN-4	85.86	85.686
KNN-5	85.82	85.50
SVM	84.62	85.08
KNN-7	84.62	84.76
NN (15, 2)	83.50	84.28
NN (12, 2, 2)	83.28	84.08
NN (8, 2)	81.59	81.85
NN (5, 5, 5)	81.26	81.81
Linear SVC	80.58	80.73
Naive Bayes	79.69	79.56
Stochastic Gradient Descent	79.80	75.33
Perceptron	78.00	74.63
NN (12, 2)	83.39	73.74
NN (8, 5)	61.62	61.39

Conclusion

I believe that the Decision Tree and Random Forest classifiers did consistently well because of the very high survival rate given Female and first or second class. Since this was 161 of 342 (47.08%) of the survivors in the training data, this made the first decision very easy for these two classifiers. There were other strong survival rates within various branches of the decision tree.

It is important to note that this is subject to overfitting. And for this reason the Random Forest classifier was chosen for the final model. The relatively small dimensionality also lends itself to decision tree classifiers. These reasons are why I believe the decision tree based classifiers always performed very well on this dataset.

Cabin was one feature that I wish there were more datapoints for. I have to imagine that the location of one's cabin influenced survival. My intuition is that the cabins on the break line of the Titanic would have a lower overall survival rate. The cabins that were on the first half that sunk may also have a lower survival rate.

Source Code

All source code can be found at https://github.com/redice44/titanic.