Titanic Machine Learning from Disaster: Classification model investigation

Deirdre Boland 14 Feb 2024

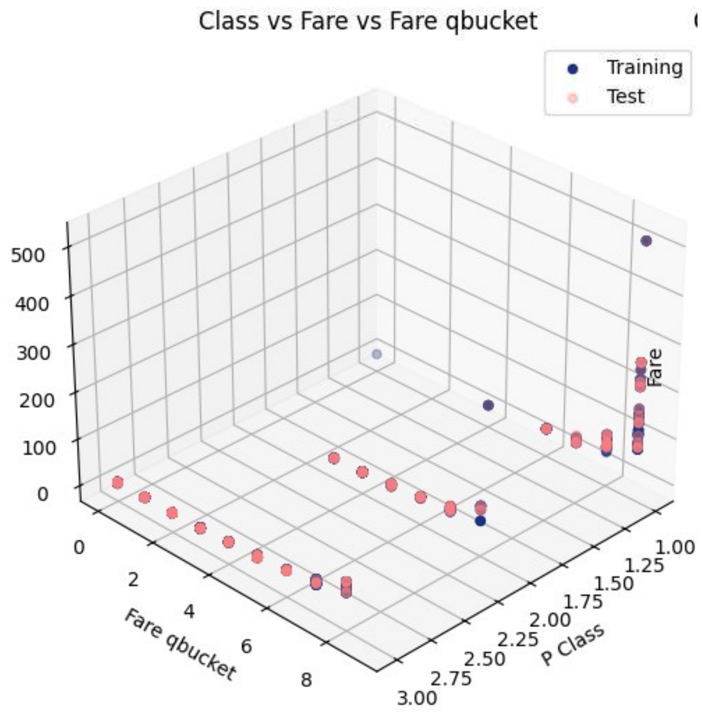
Objective and dataset

Objectives

- Fit Logistic Regression model to Titanic – Machine Learning from Disaster Kaggle competition data and measure accuracy.
- Compare performance of different classifier, not focussed on optimising for accuracy due to time limits

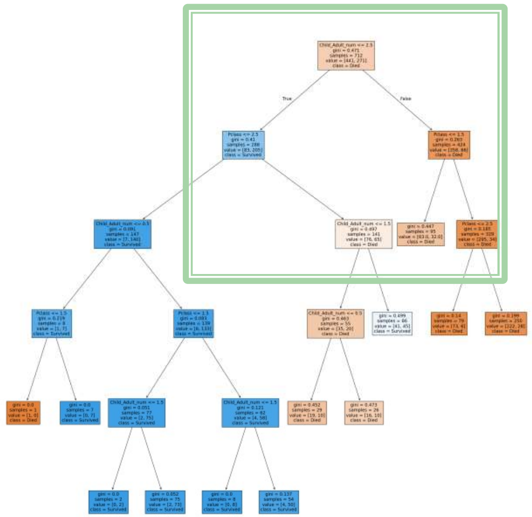
Data Preparation

- Cabin data had too many null values and ticket too many unique values to be useful - not used
  - Likely to vary between datasets which may be issue for model reuse, test transformation
- Sex was label encoded to binary (male = 1, female = 0)
- Embarked null values filled with "S" (most frequent) and label encoded to 0-2
- Age data investigation - Women and children first
  - Lot of null age data
  - generated new feature Child/Adult by decomposing Name and looking at Parent child (Parch) relationship and label encoded
    - child female – 0, child male - 1, adult female - 2, adult male -3
  - Age null filled with median and alternate using child/adult data to assign more appropriate age
    - Minimal difference due to small population of children
- Fare had some null data but also some zero values
  - Zero values - Traced back to 15 men with no apparent link, set to null
  - median Fare assigned to all nulls
  - Quartile bucket of Fare data (refer to image)



- Reference – Will Cukierski. Titanic - Machine Learning from Disaster. https://kaggle.com/competitions/titanic, 2012. Kaggle.

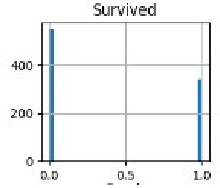
Decision Tree – Data Visualisation 1



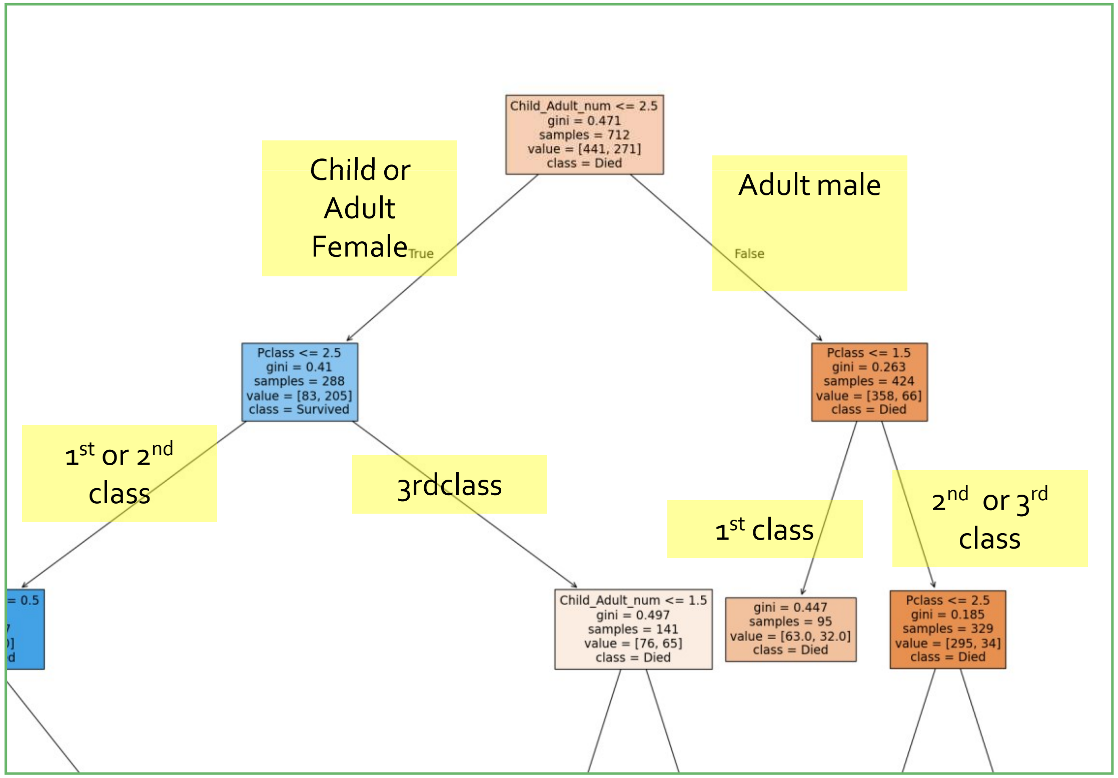
- Class and Child/Adult new feature only
  - not scaled, training data 80/20 test split
- Gini default model – accuracy 79%

Confusion Matrix

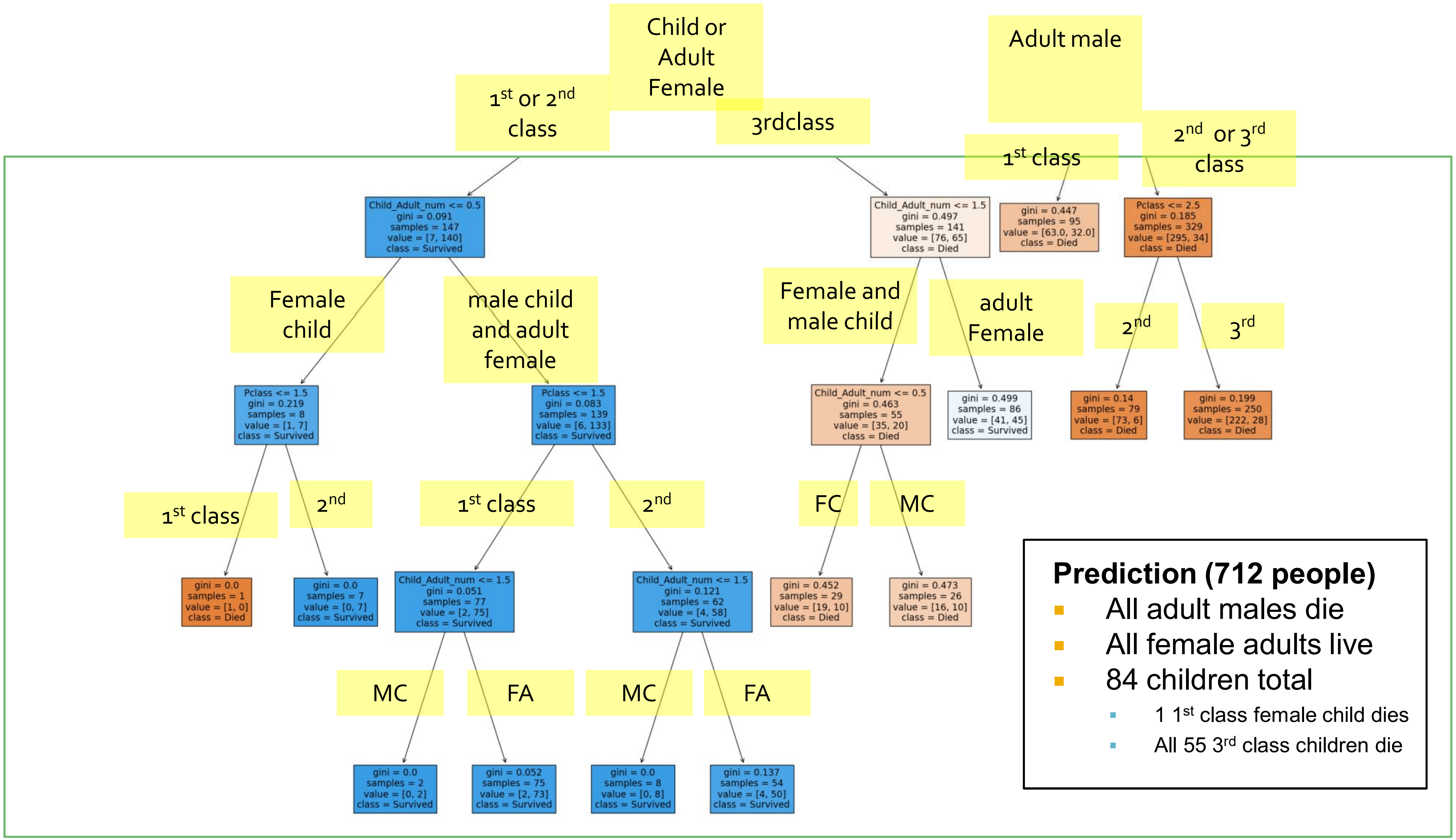
	Pred Die	Pred Live
Act (0) Died	([[0.88,	0.12],
Act (1) Lived	[0.347,	0.66]])



- 88% of the time will predict person died correctly
- 34.7 % of the time will predict person died when they survived
- 12% of time will predict someone lived when they died
- 66.5% of the time will predict person survived correctly



Decision Tree – Data Visualisation 2



Prediction (712 people)

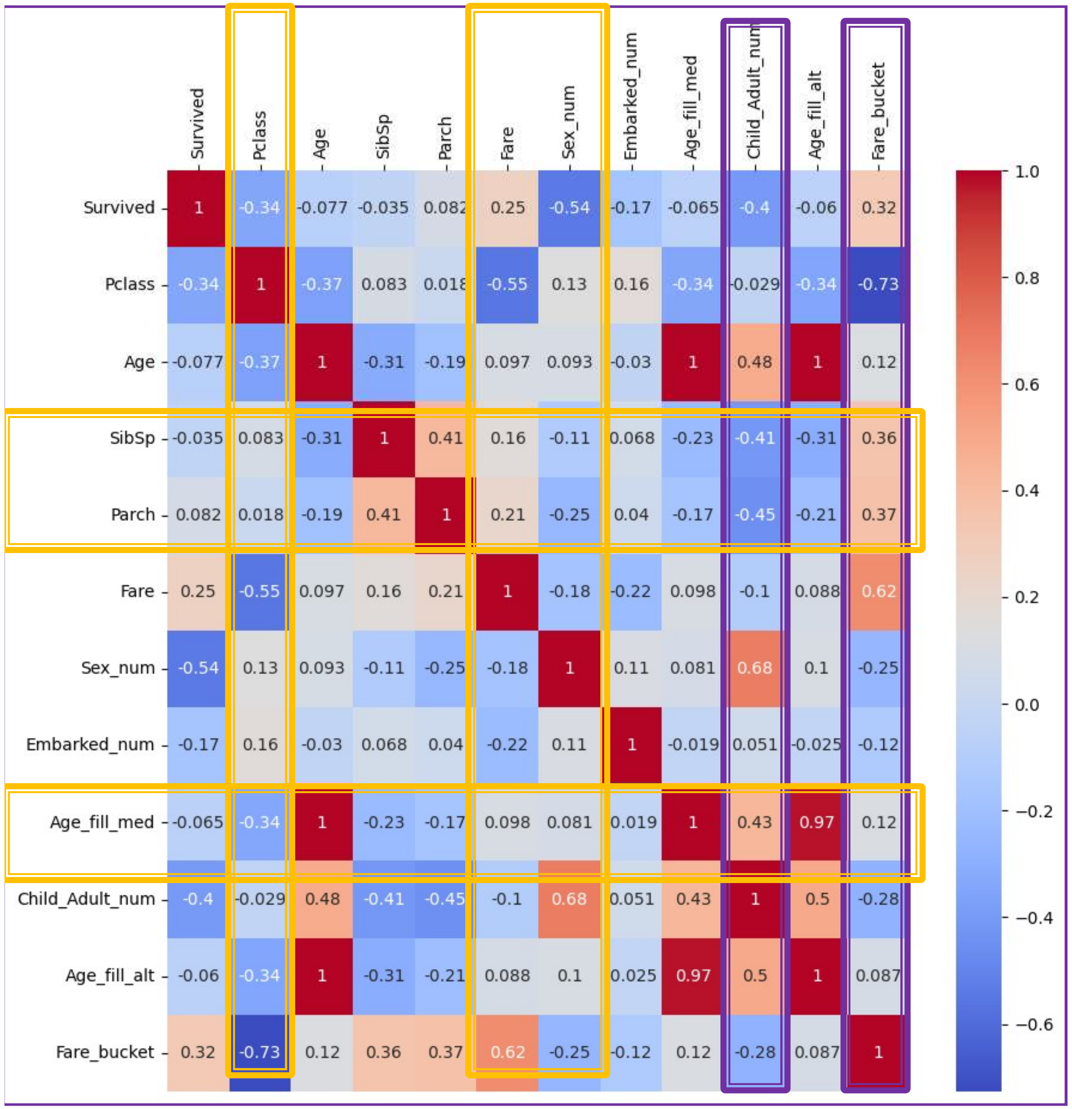
- All adult males die
- All female adults live
- 84 children total
  - 1 1st class female child dies
  - All 55 3rd class children die

Classifier Models compared

- All 80/20 test split of training data
- Comparable confusion matrix for most except Sigmoid and Bernoulli NB
  - K = 3 KNN model gives better prediction of survivors
- Due to imbalanced dataset models not as good at predicting those that survived
  - 60 % accuracy BeronulliNB couldn't predict anyone living

Model	**Logistic regression**	**KNN**		**SVM**			**Decision Tree**			**Naive Bayes**		
Features	Class, Child/Adult, Fare bucket								Class, Child/Adult	Class, Child/Adult, Fare bucket		
Scaling	Robust								none			
Other conditions		k = 3, error ~0.197	k = 9, error ~0.18	linear	sigmoid	rbf	gini/ entropy	gini with max depth	gini/ entropy /g with max depth	Gaussian	Multinomial	Bernoulli
Accuracy %	77	82	82	82	51	80	82	81	79	73	69	60
predict person died correctly (pink add to 100) %	88	87	94	91	61	91	94	90	88	83	77	100
predict person died when they survived (blue add to 100) %	39	27	35	31	64	37	35	32	34	42	45	100
predict someone lived when they died (pink add to 100) %	12	13	6	9	39	9	6	10	12	17	22	0
predict person survived correctly (blue add to 100) %	61	73	65	69	35	63	65	67	66	58	55	0
Comments					Proxy for neural networks				*gini presented	Bernouilli for binary data, Multinomial for discrete counts s (doesn't like negative values of scaled data), Gaussian assumes norm dist *tried text data but errored		

Feature Correlation



- Fare qbucket and Child/Adult encoded (new features purple), used instead of Fare and Sex encoded (original orange) :
  - Have a stronger correlation than original variable with sibling/spouse, parent/child relationships and age
- Haven't incorporated the port that passengers embarked from but only weak correlations to everything

Conclusion

- A simple unoptimised decision Tree predicts that all adult males die, all adult females live, all 3rd class children die
- Models 80+/-3% accuracy with comparable confusion matrix except:
  - SVM sigmoid
  - Naive Bayes, particularly Bernouli
- As per competition 100% accuracy possible if model all training data and fit test data (not used in modelling)
  - Further refinement of models and features