Predicting Survival on the Titanic

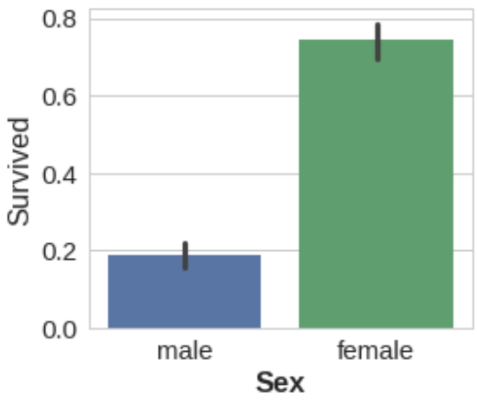
Systematically Building a Machine-Learning Pipeline.

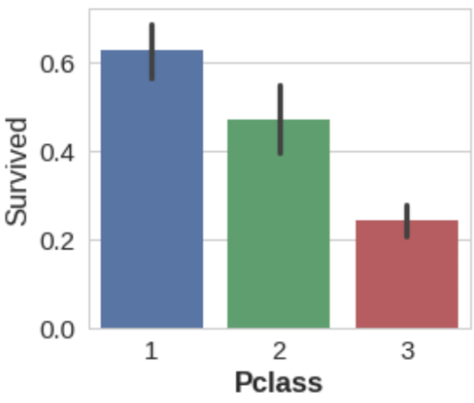
Data Dictionary

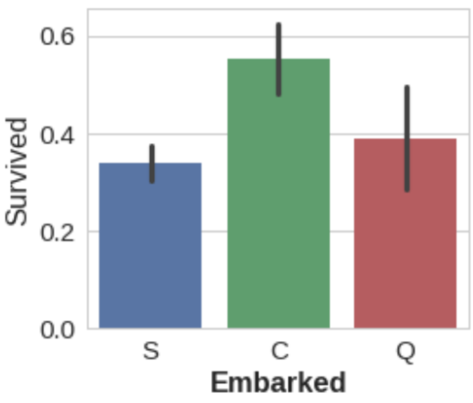
> A few notes below about the meaning of the features in the raw dataset:
>
> - **Survival**: 0 = False (Deceased), 1 = True (Survived).
> - **Pclass**: Passenger ticket class; 1 = 1st (upper class), 2 = 2nd (middle class), 3 = 3rd (lower class).
> - **SibSp**: Passenger's total number of siblings (including step-siblings) and spouses (legal) aboard the Titanic.
> - **Parch**: Passenger's total number of parents or children (including stepchildren) aboard the Titanic.
> - **Embarked**: Port of Embarkation, where C = Cherbourg, Q = Queenstown, S = Southampton.
> - **Age**: Ages under 1 are given as fractions; if the age is estimated, it is in the form of xx.5.

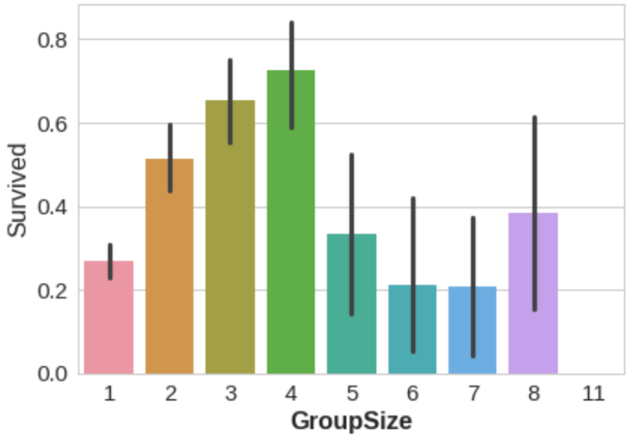
Univariate Feature Plots

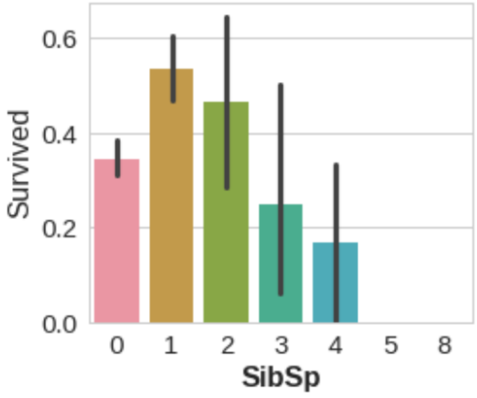
(Examples)

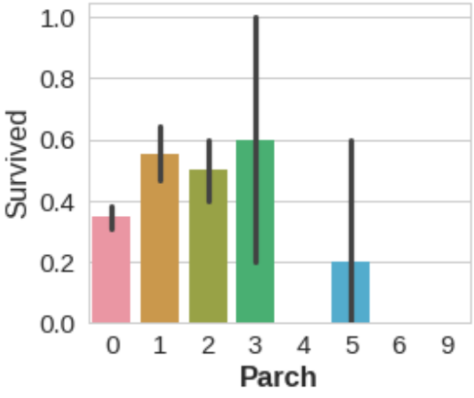


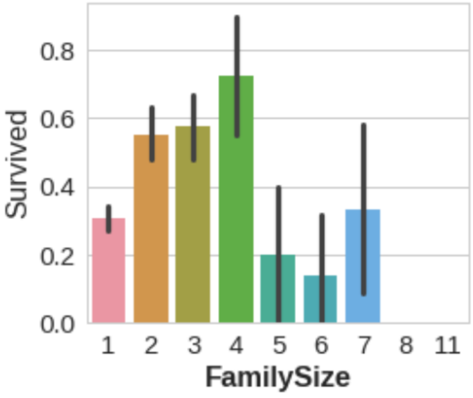


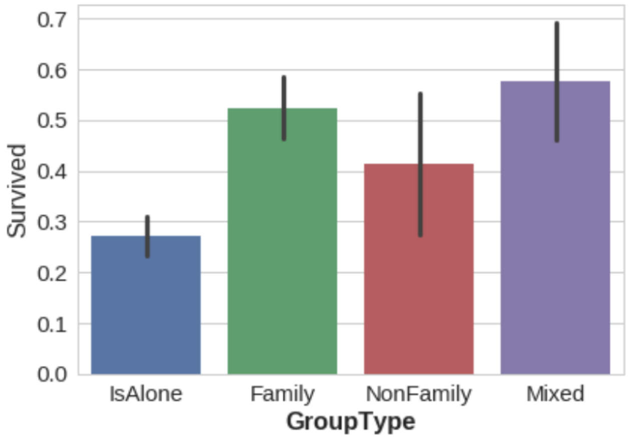


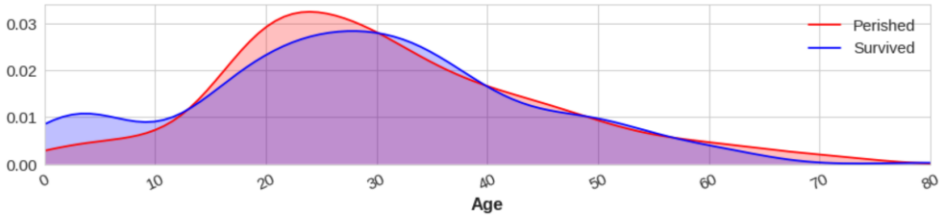


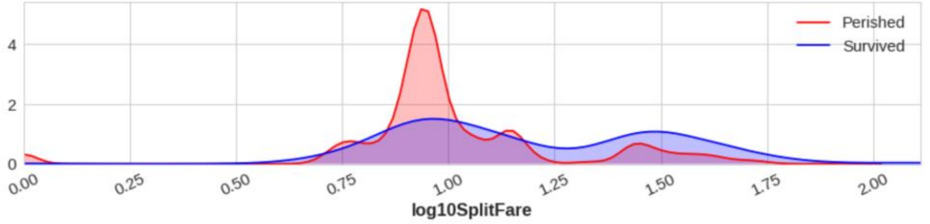


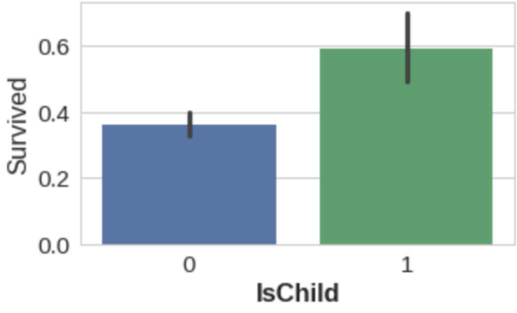








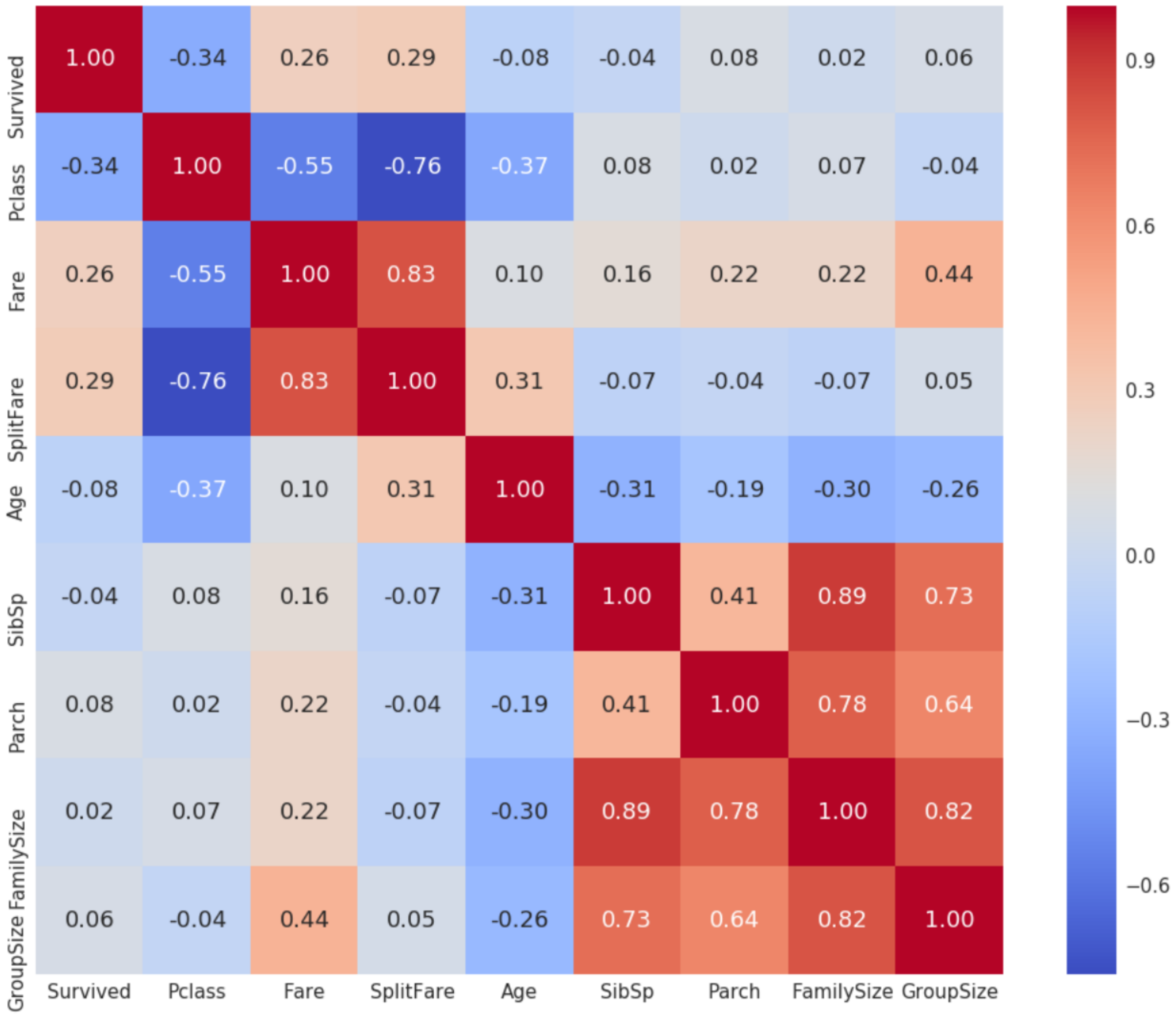




Multivariate Feature Exploration

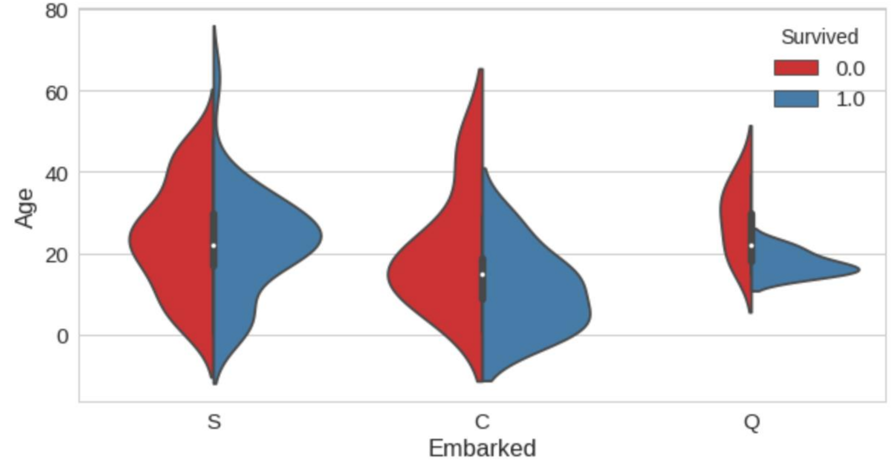
(Examples)

Feature Correlation Matrix

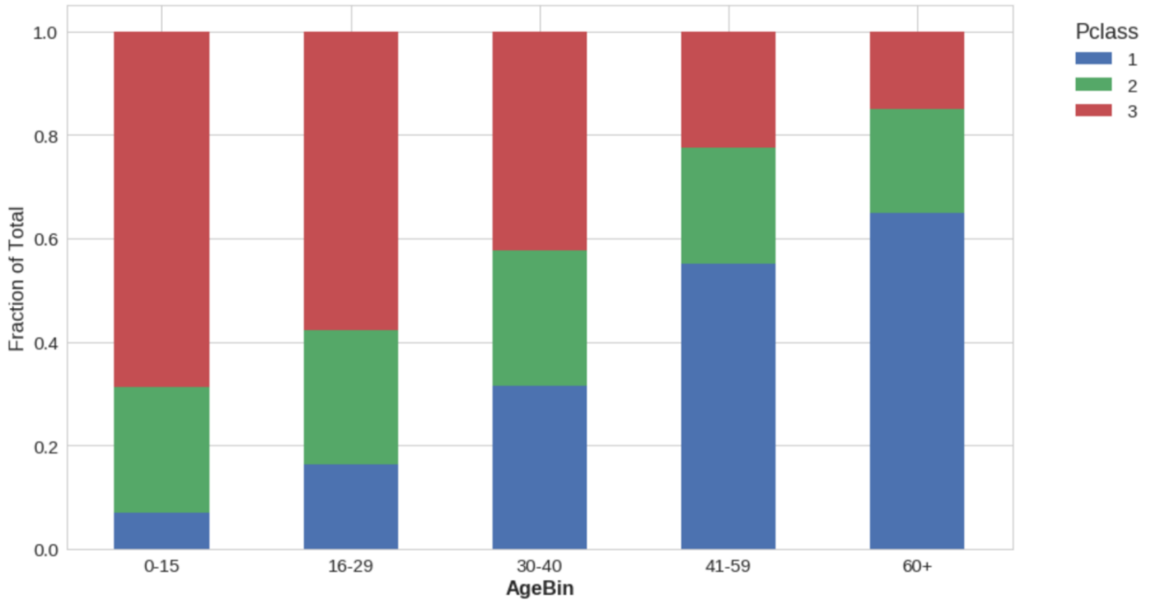


	Survived	Pclass	Fare	SplitFare	Age	SibSp	Parch	FamilySize	GroupSize
Survived	1.00	-0.34	0.26	0.29	-0.08	-0.04	0.08	0.02	0.06
Pclass	-0.34	1.00	-0.55	-0.76	-0.37	0.08	0.02	0.07	-0.04
Fare	0.26	-0.55	1.00	0.83	0.10	0.16	0.22	0.22	0.44
SplitFare	0.29	-0.76	0.83	1.00	0.31	-0.07	-0.04	-0.07	0.05
Age	-0.08	-0.37	0.10	0.31	1.00	-0.31	-0.19	-0.30	-0.26
SibSp	-0.04	0.08	0.16	-0.07	-0.31	1.00	0.41	0.89	0.73
Parch	0.08	0.02	0.22	-0.04	-0.19	0.41	1.00	0.78	0.64
FamilySize	0.02	0.07	0.22	-0.07	-0.30	0.89	0.78	1.00	0.82
GroupSize	0.06	-0.04	0.44	0.05	-0.26	0.73	0.64	0.82	1.00

Age distributions for different embarkment ports.



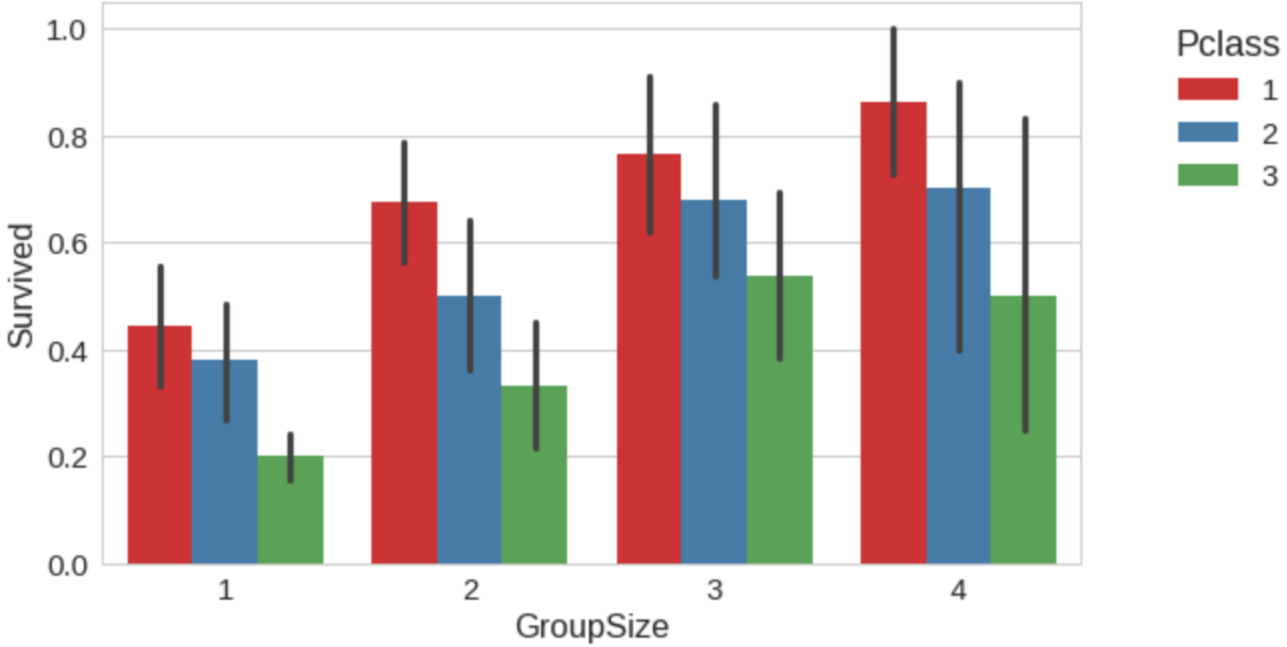
Class distributions among Age bins.



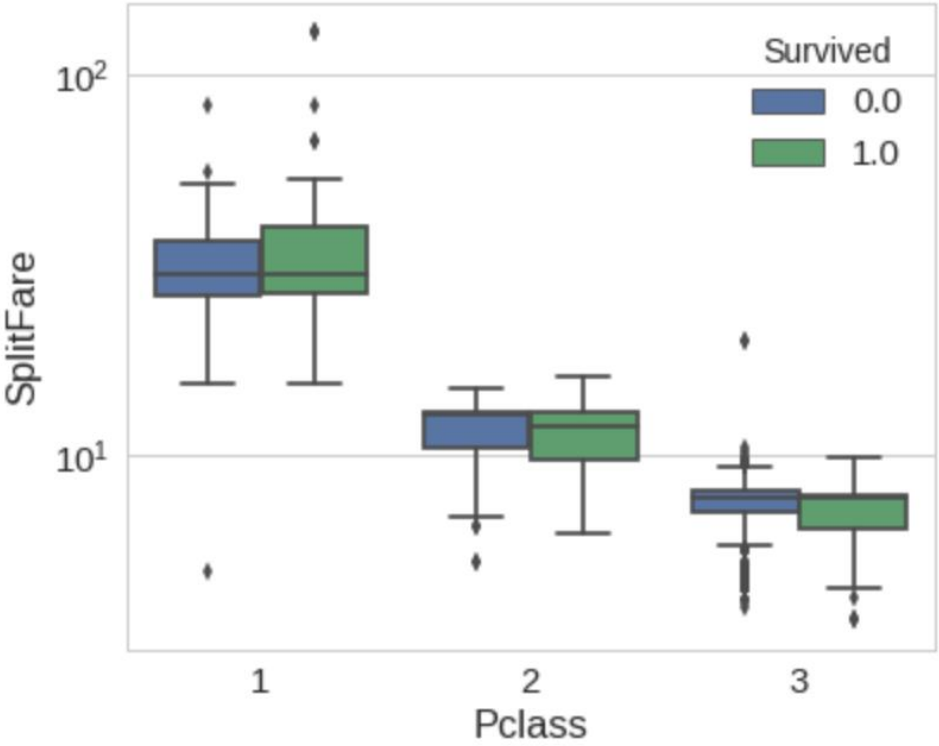
Multivariate Feature Exploration

(More Examples)

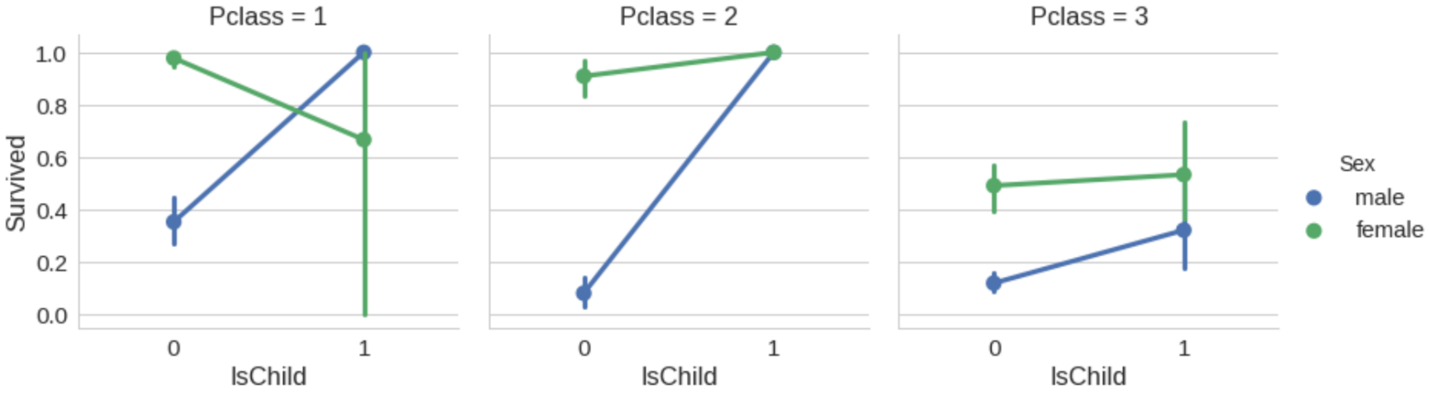
Survival, GroupSize, and Pclass.



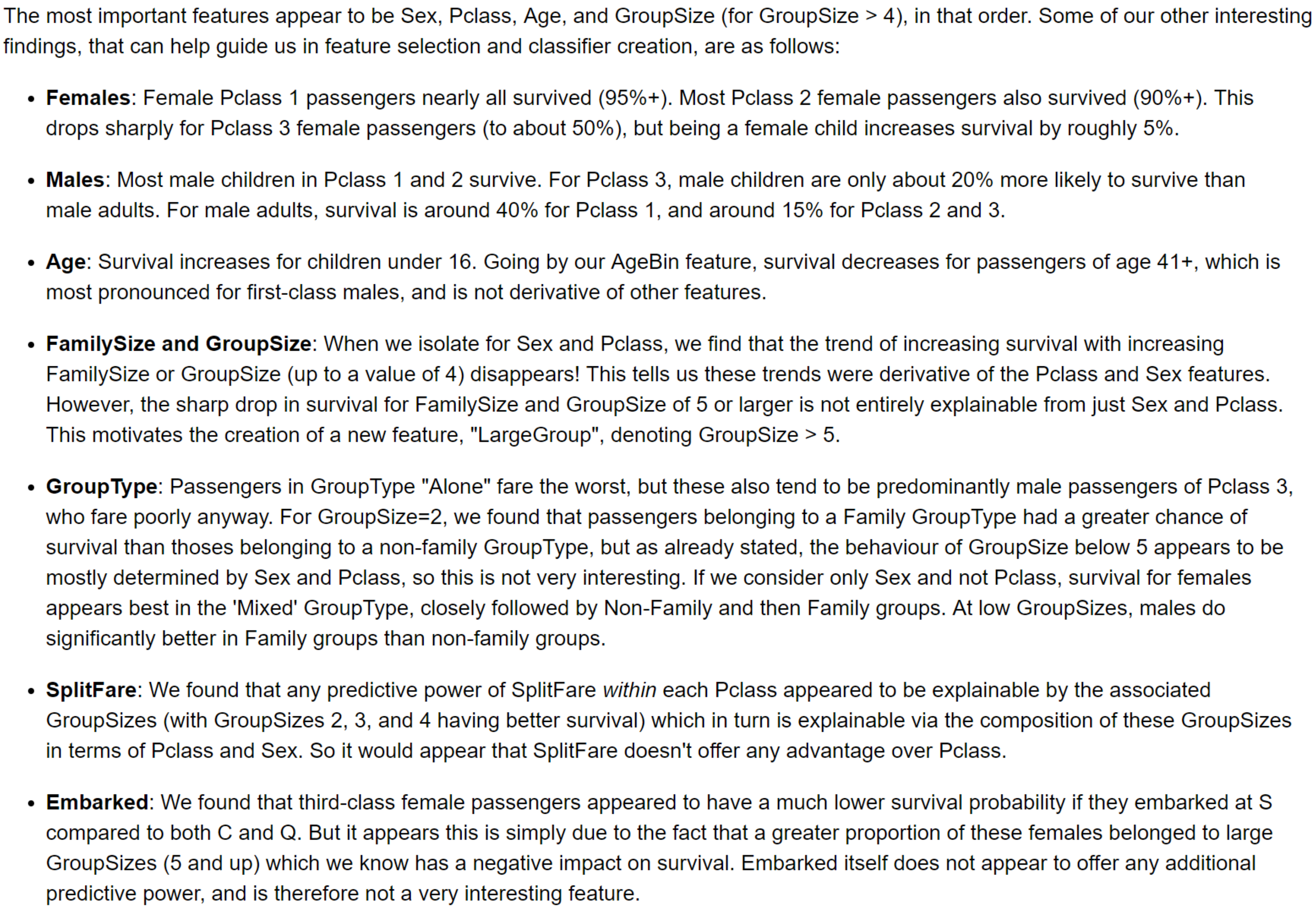
SplitFare vs Pclass showing quartiles.



Impact of being a child on survival, for different classes and sexes.



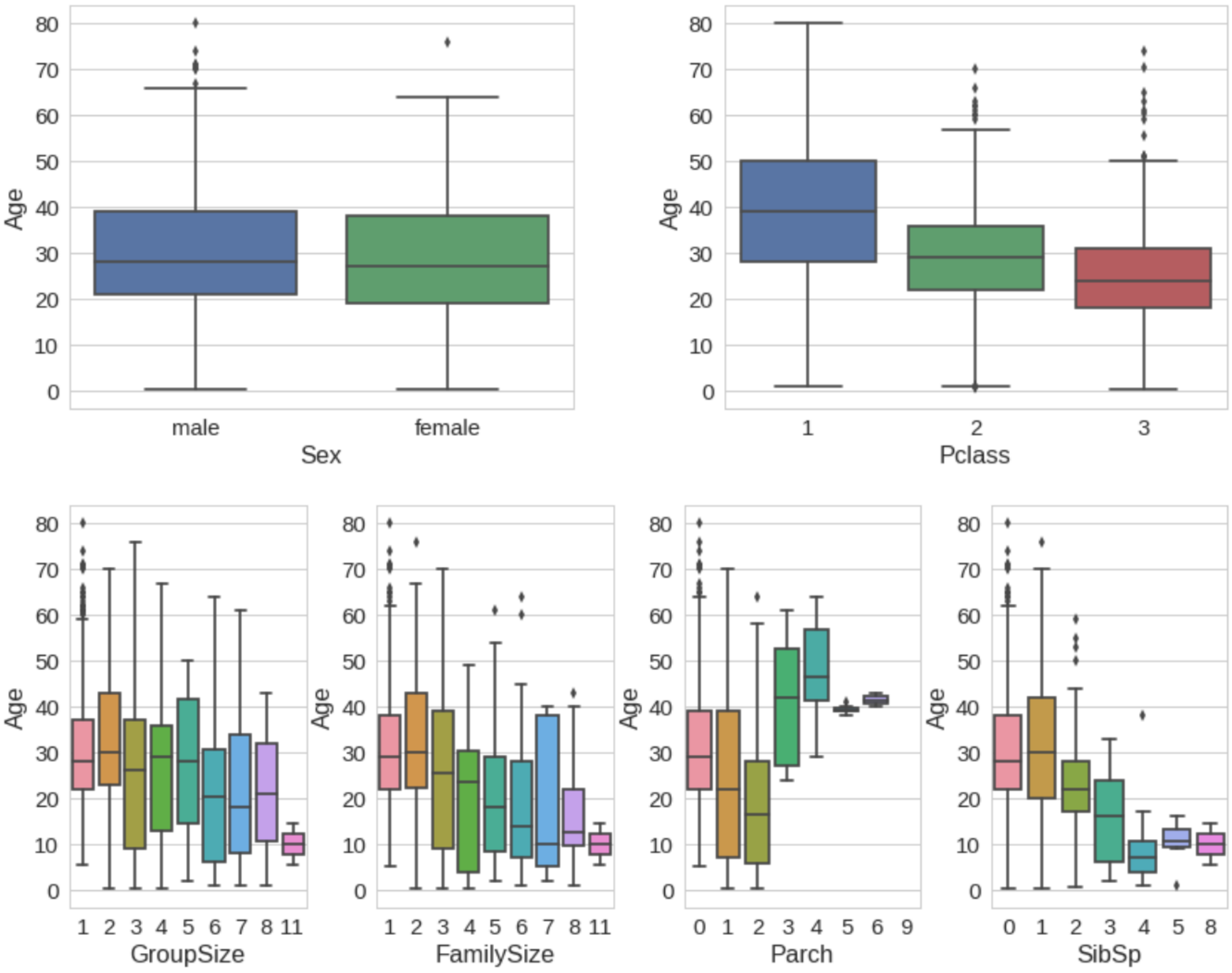
Key Findings of Exploratory Data Analysis

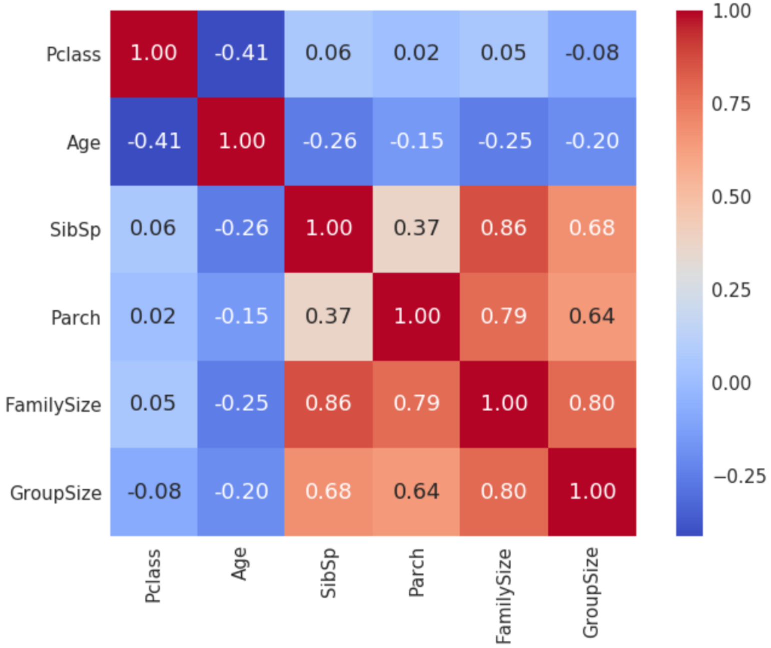


The most important features appear to be Sex, Pclass, Age, and GroupSize (for GroupSize > 4), in that order. Some of our other interesting findings, that can help guide us in feature selection and classifier creation, are as follows:

- **Females**: Female Pclass 1 passengers nearly all survived (95%+). Most Pclass 2 female passengers also survived (90%+). This drops sharply for Pclass 3 female passengers (to about 50%), but being a female child increases survival by roughly 5%.
- **Males**: Most male children in Pclass 1 and 2 survive. For Pclass 3, male children are only about 20% more likely to survive than male adults. For male adults, survival is around 40% for Pclass 1, and around 15% for Pclass 2 and 3.
- **Age**: Survival increases for children under 16. Going by our AgeBin feature, survival decreases for passengers of age 41+, which is most pronounced for first-class males, and is not derivative of other features.
- **FamilySize and GroupSize**: When we isolate for Sex and Pclass, we find that the trend of increasing survival with increasing FamilySize or GroupSize (up to a value of 4) disappears! This tells us these trends were derivative of the Pclass and Sex features. However, the sharp drop in survival for FamilySize and GroupSize of 5 or larger is not entirely explainable from just Sex and Pclass. This motivates the creation of a new feature, "LargeGroup", denoting GroupSize > 5.
- **GroupType**: Passengers in GroupType "Alone" fare the worst, but these also tend to be predominantly male passengers of Pclass 3, who fare poorly anyway. For GroupSize=2, we found that passengers belonging to a Family GroupType had a greater chance of survival than thoses belonging to a non-family GroupType, but as already stated, the behaviour of GroupSize below 5 appears to be mostly determined by Sex and Pclass, so this is not very interesting. If we consider only Sex and not Pclass, survival for females appears best in the 'Mixed' GroupType, closely followed by Non-Family and then Family groups. At low GroupSizes, males do significantly better in Family groups than non-family groups.
- **SplitFare**: We found that any predictive power of SplitFare *within* each Pclass appeared to be explainable by the associated GroupSizes (with GroupSizes 2, 3, and 4 having better survival) which in turn is explainable via the composition of these GroupSizes in terms of Pclass and Sex. So it would appear that SplitFare doesn't offer any advantage over Pclass.
- **Embarked**: We found that third-class female passengers appeared to have a much lower survival probability if they embarked at S compared to both C and Q. But it appears this is simply due to the fact that a greater proportion of these females belonged to large GroupSizes (5 and up) which we know has a negative impact on survival. Embarked itself does not appear to offer any additional predictive power, and is therefore not a very interesting feature.

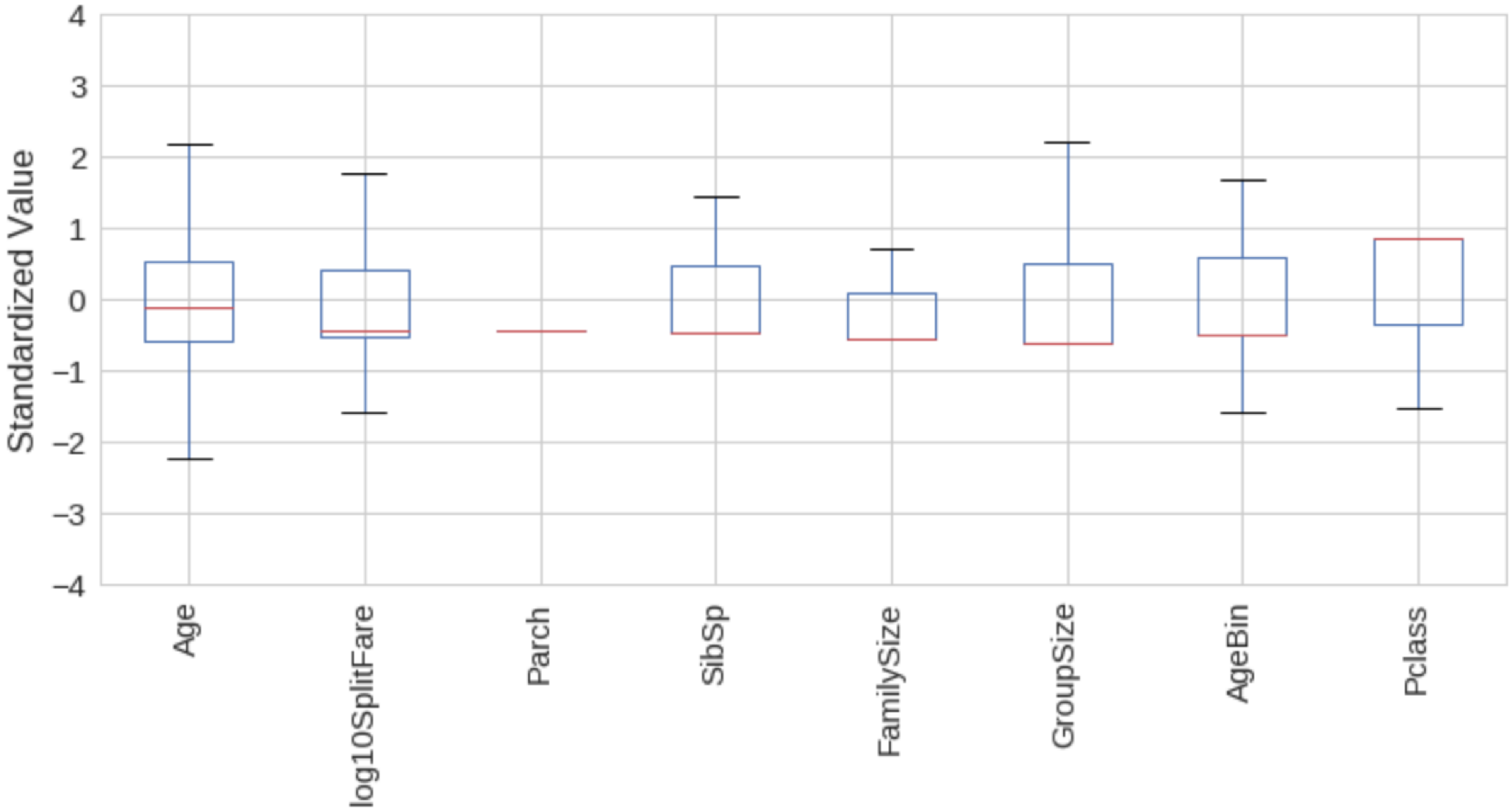
Using Feature Correlations to Impute Missing Age Entries





	Pclass	Age	SibSp	Parch	FamilySize	GroupSize
Pclass	1.00	-0.41	0.06	0.02	0.05	-0.08
Age	-0.41	1.00	-0.26	-0.15	-0.25	-0.20
SibSp	0.06	-0.26	1.00	0.37	0.86	0.68
Parch	0.02	-0.15	0.37	1.00	0.79	0.64
FamilySize	0.05	-0.25	0.86	0.79	1.00	0.80
GroupSize	-0.08	-0.20	0.68	0.64	0.80	1.00

Checking Ordinal & Continuous-Variable Feature Distributions After Standardization:



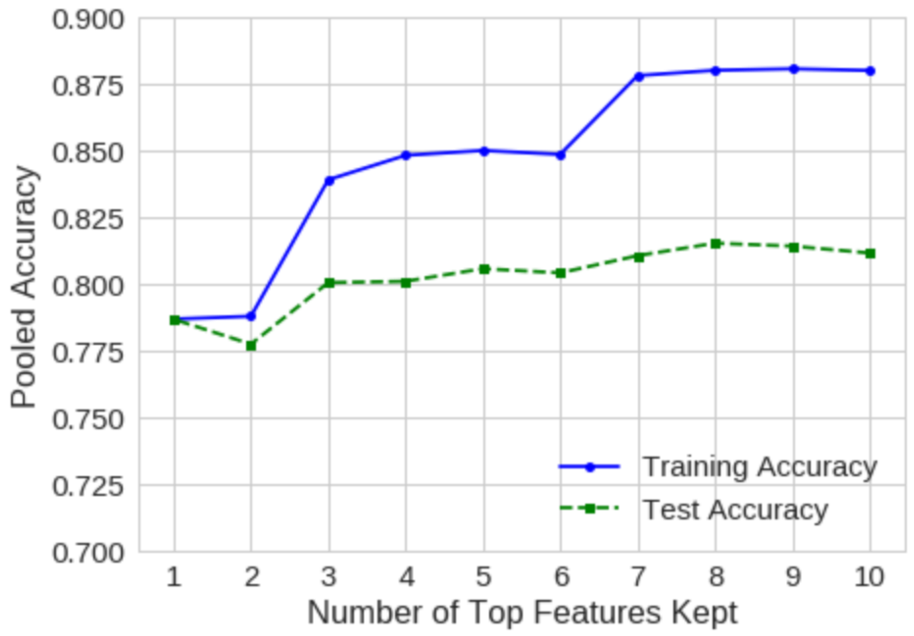
Note: the reason for the odd appearance of the Parch boxplot is merely that the boundaries of the first, second, and third quartiles all have the same value, since we have Parch=0 for more than 75% of all instances (prior to standardization).

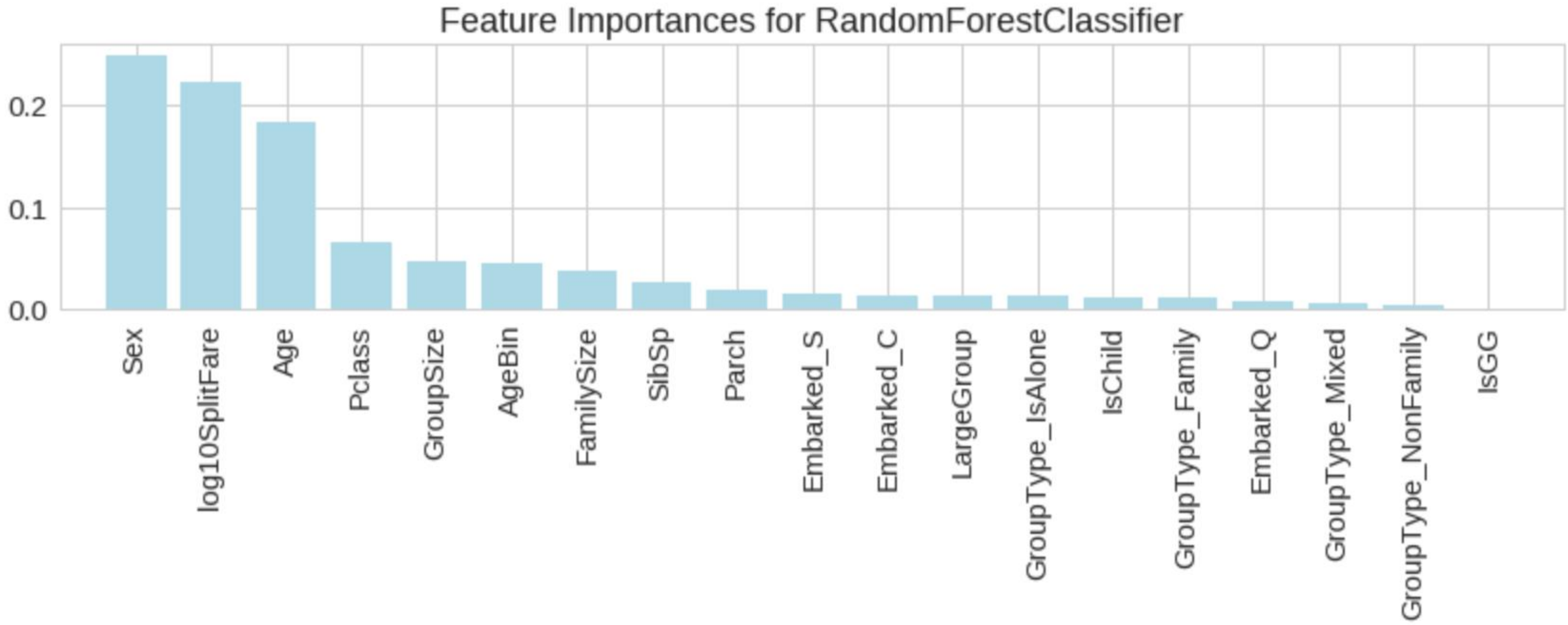
Exploring Feature Importance

Via ***recursive feature elimination (RFE)*** *on full feature set:*

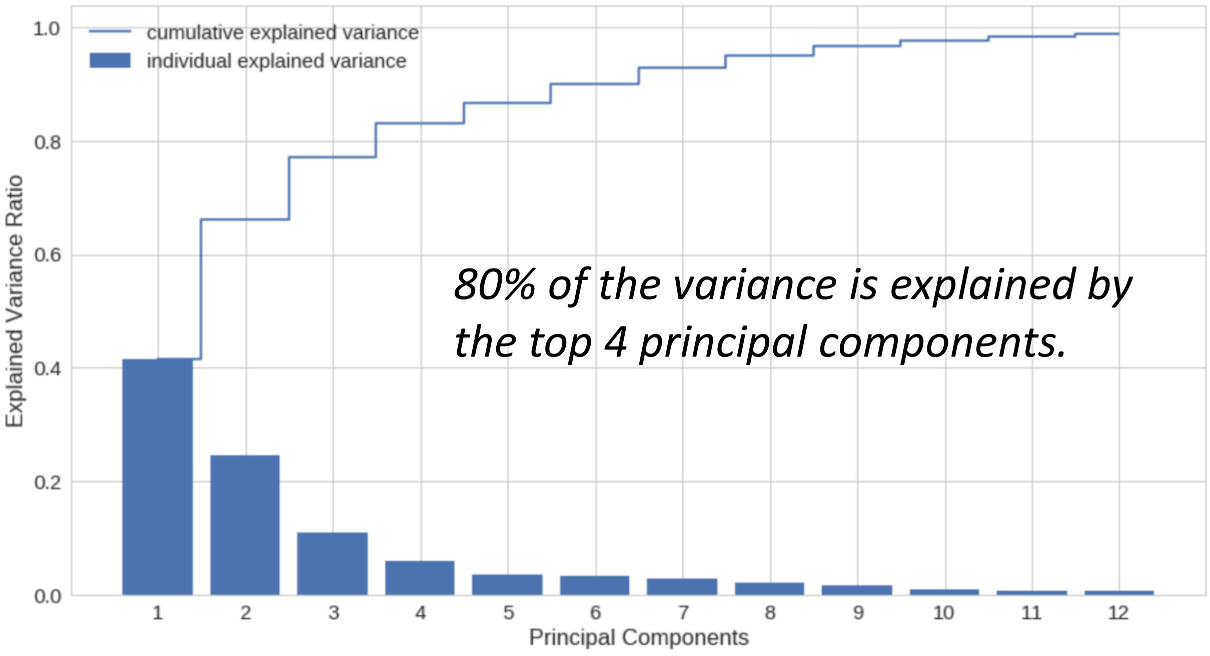
Feature Name	Overall Ranking
Sex	1
Pclass	2
log10SplitFare	3
SibSp	4
FamilySize	5
GroupSize	6
Age	7
LargeGroup	8
GroupType_Family	9
GroupType_NonFamily	10

Overfitting (poor generalization) yielded by this approach!

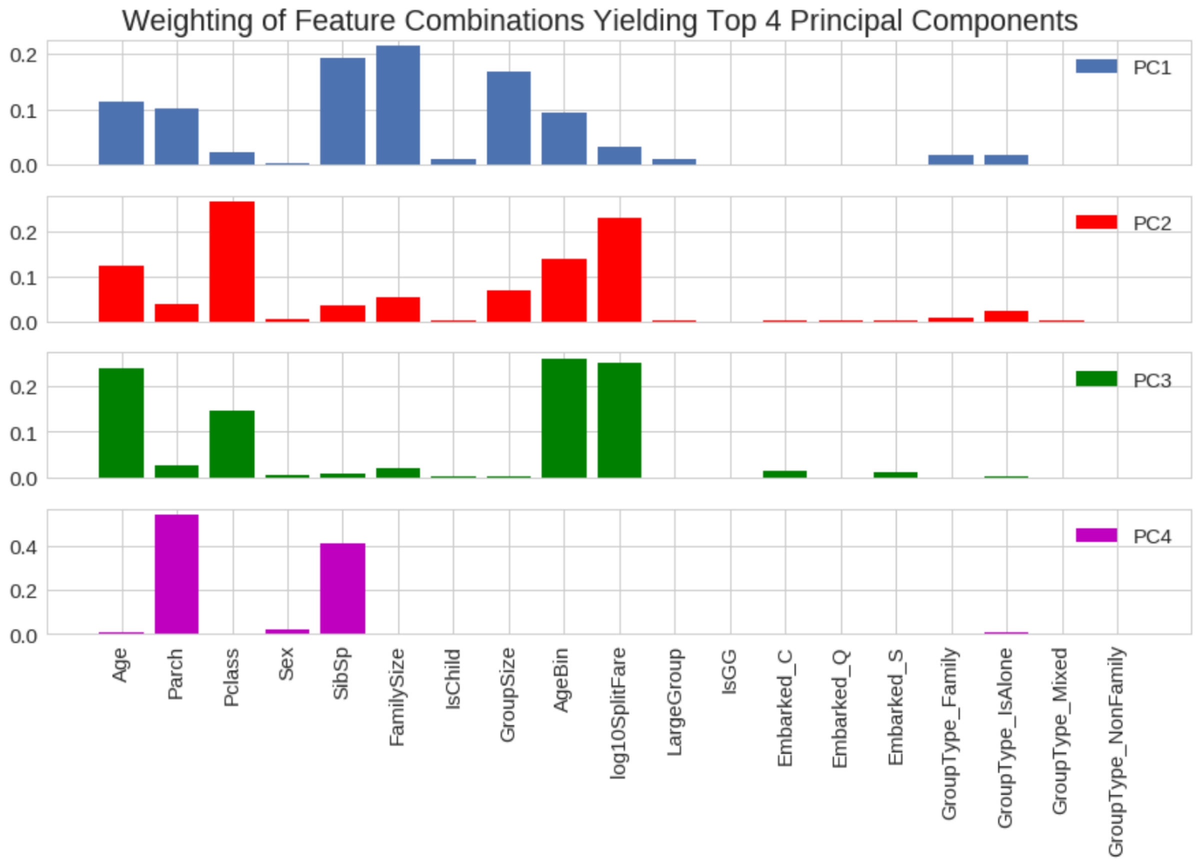


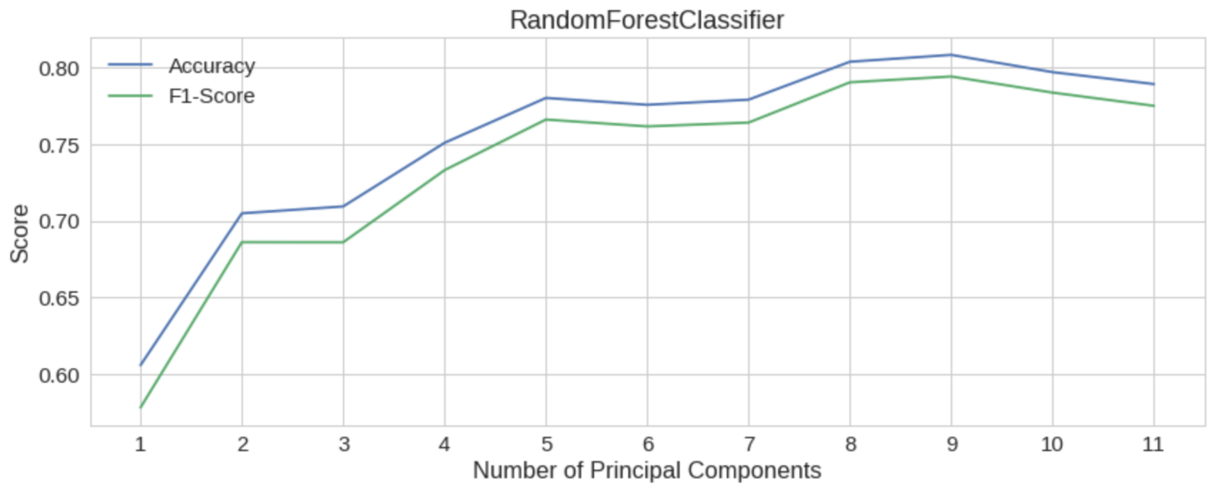


Principal Component Analysis (PCA)



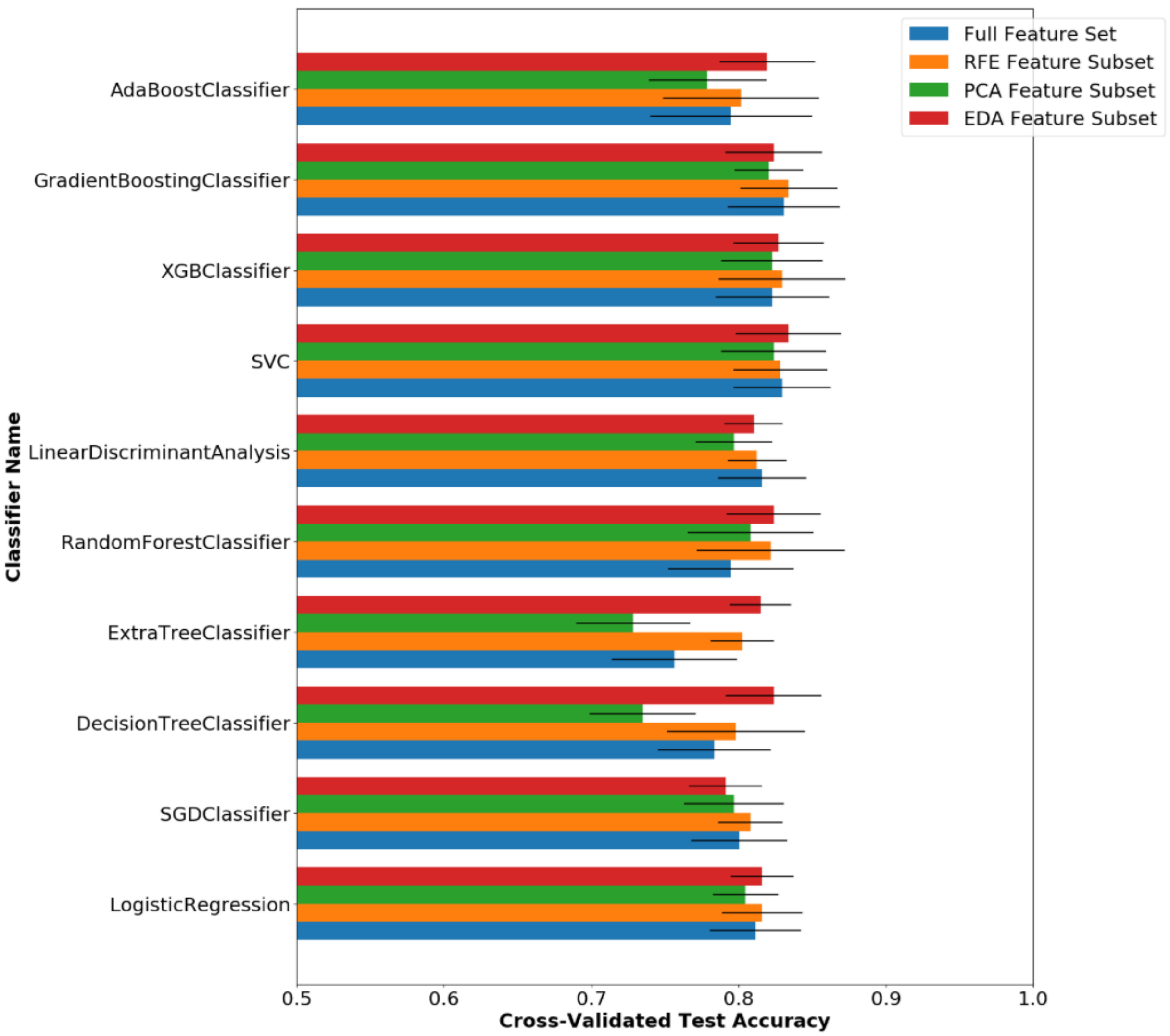
80% of the variance is explained by the top 4 principal components.

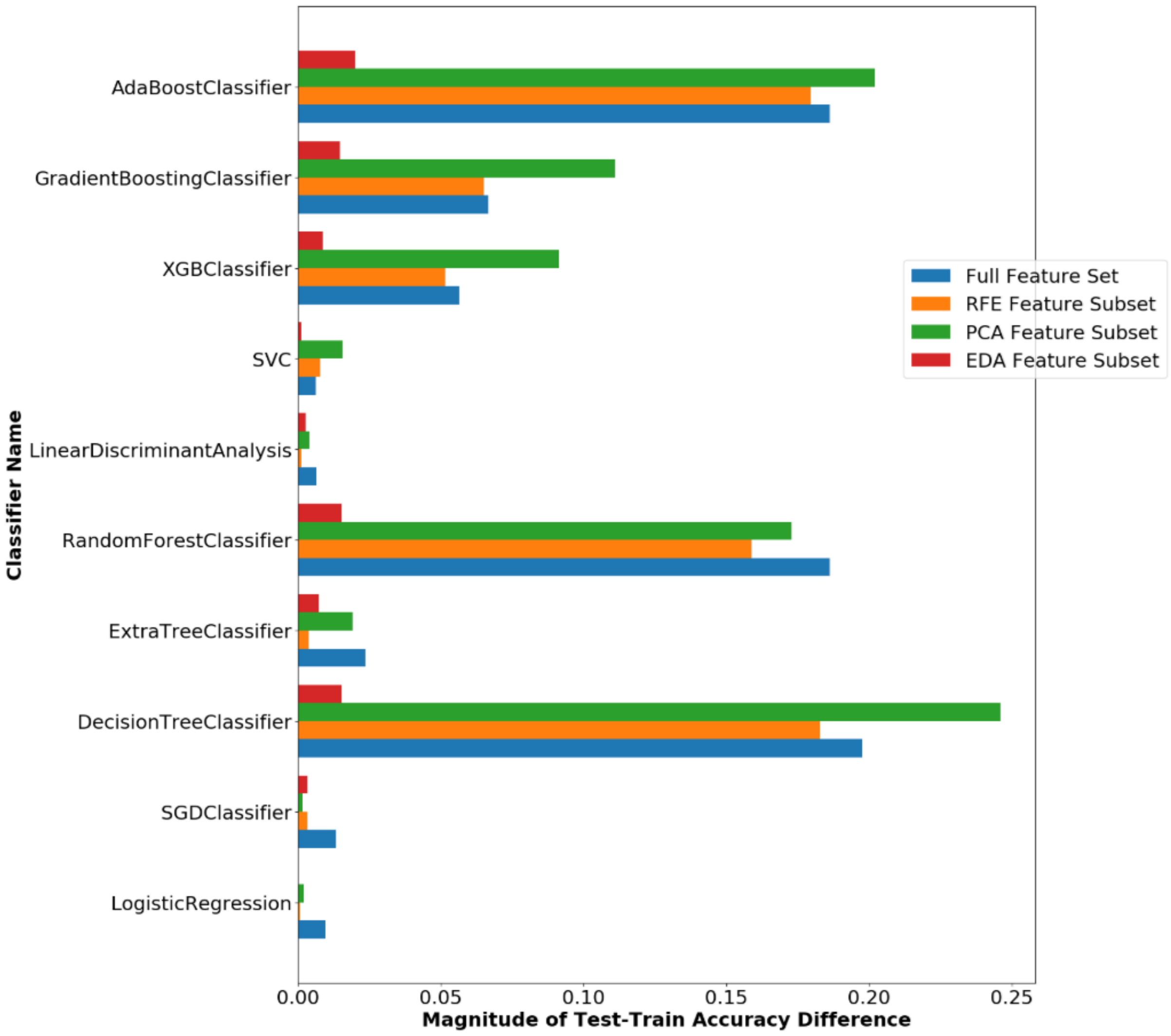




Testing classifier performance vs. number of kept components (example).

Comparison of Classifier Performance using Different Feature Sets

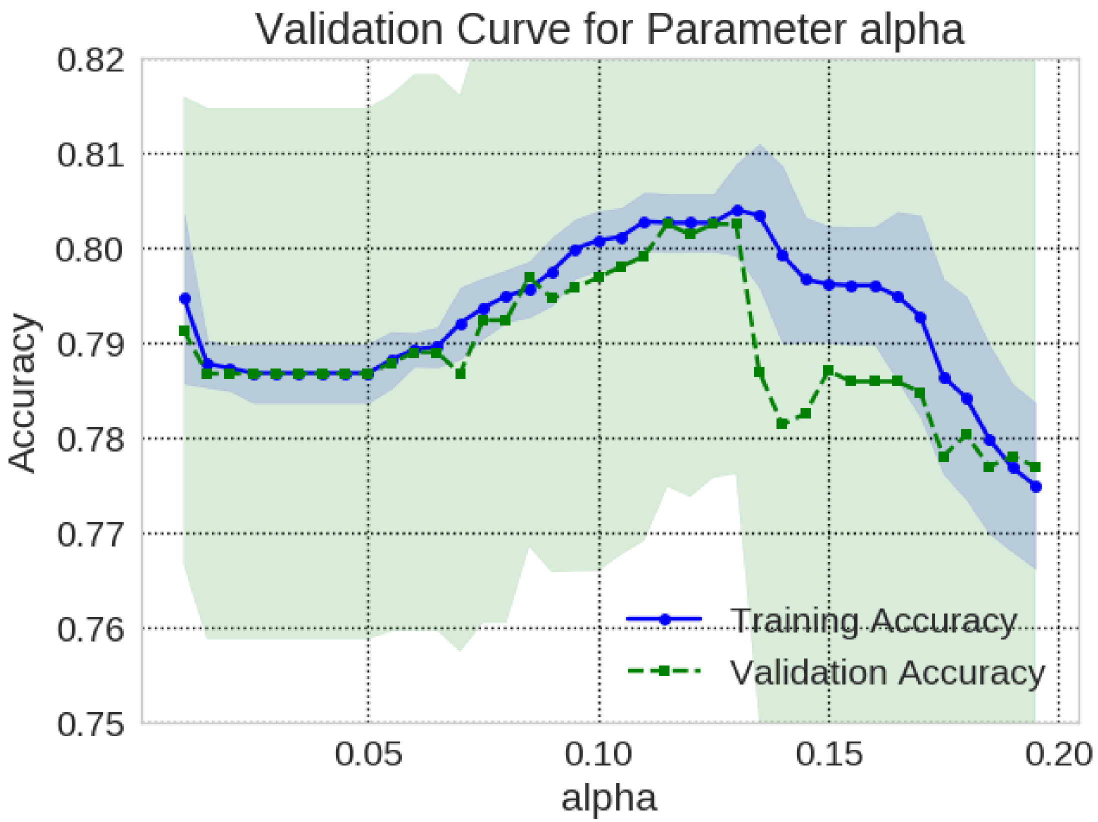




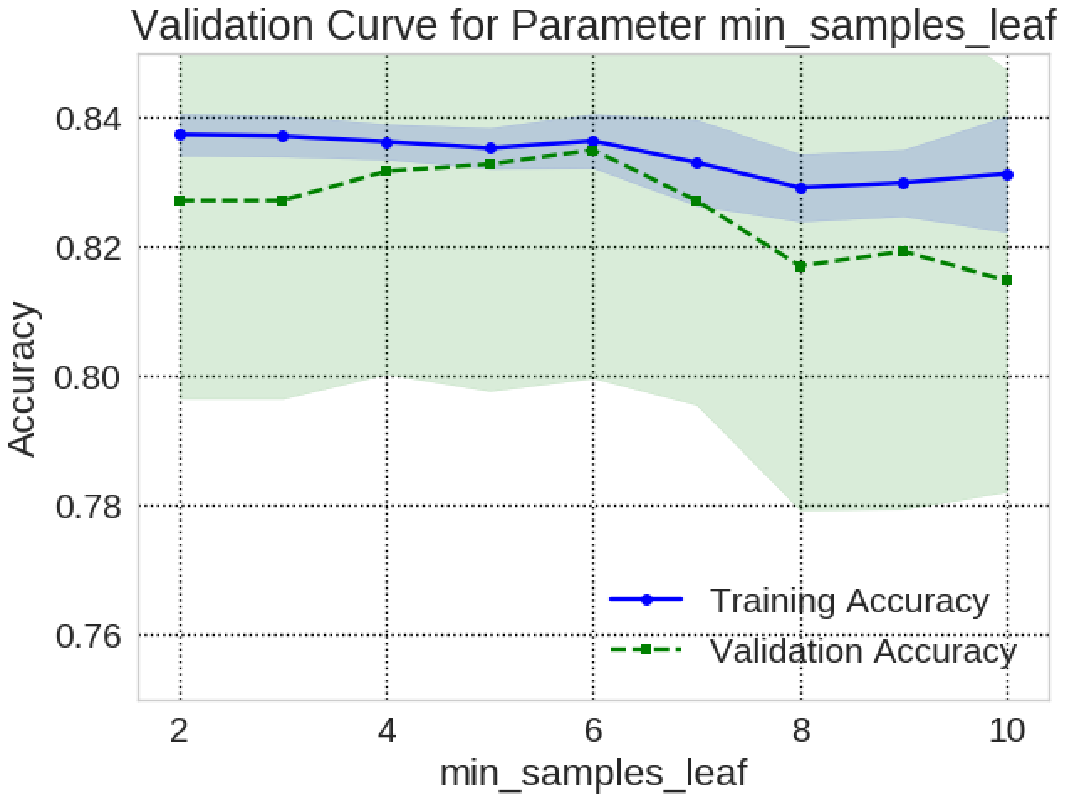
(Smaller difference means better generalization)

Example Validation Curves from Hyperparameter Tuning:

SGD classifier, regularization strength (alpha):

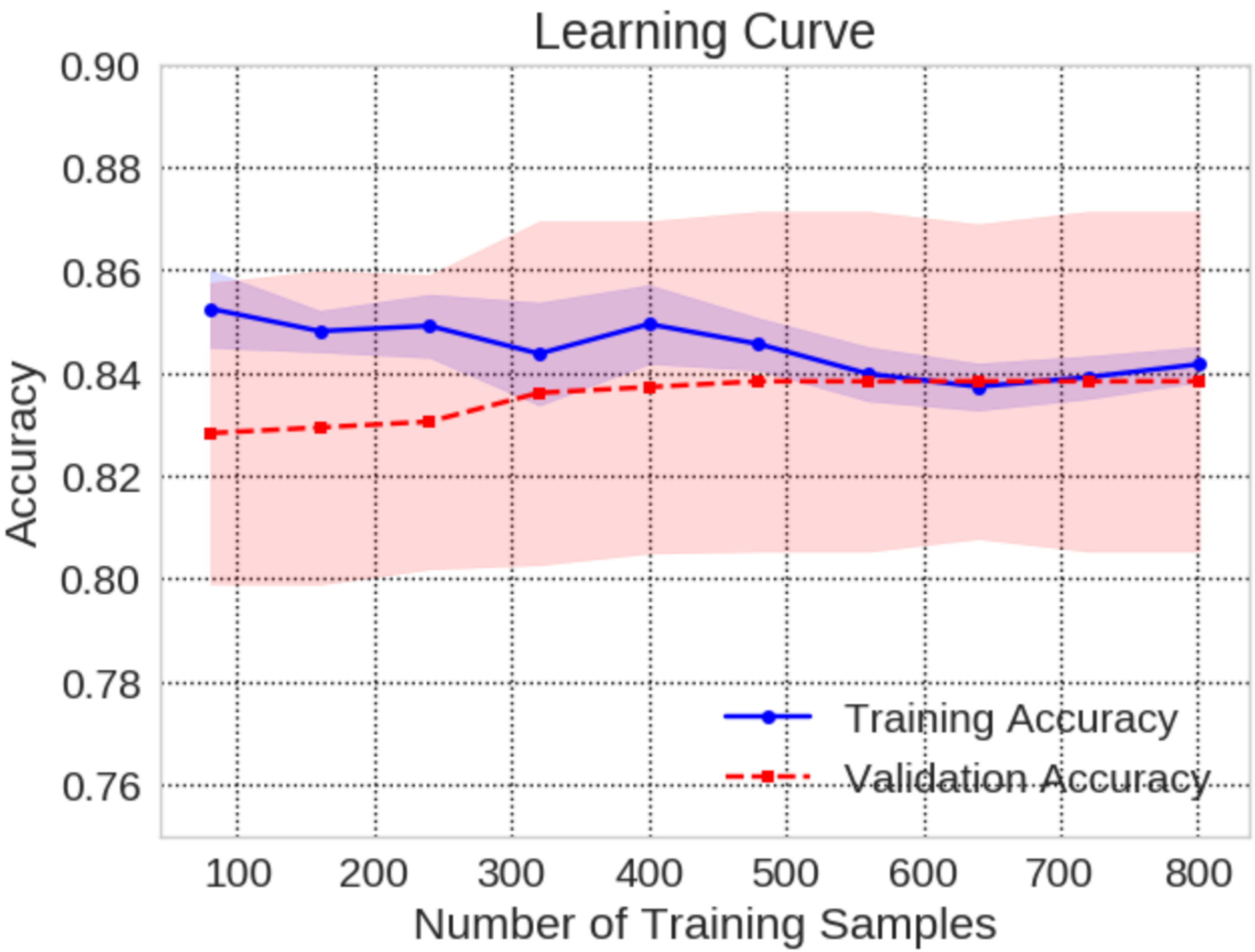


Decision tree classifier, minimum number of samples needed to support a leaf:



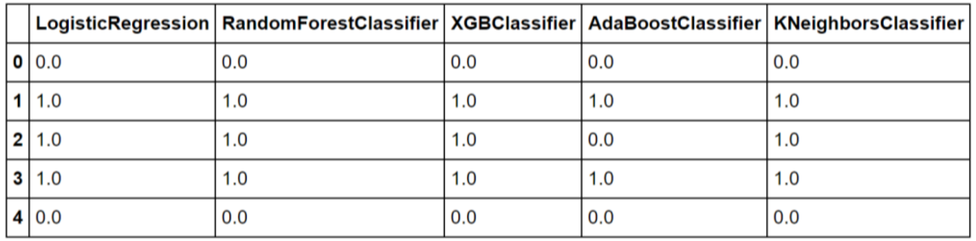
Performance of Stacked Classifier (XGB as Layer 2)

```
Accuracy: 0.8384 (+/- 0.0330) [Stacking with XGBClassifier]
```



The small difference between train and validation accuracies indicates good generalization.

Example out-of-fold predictions from base classifier set, used as input features for 2nd layer (XGB):



	LogisticRegression	RandomForestClassifier	XGBClassifier	AdaBoostClassifier	KNeighborsClassifier
0	0.0	0.0	0.0	0.0	0.0
1	1.0	1.0	1.0	1.0	1.0
2	1.0	1.0	1.0	0.0	1.0
3	1.0	1.0	1.0	1.0	1.0
4	0.0	0.0	0.0	0.0	0.0