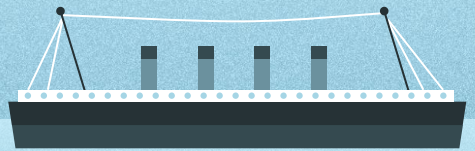
— Team Sigma —

TITANIC

Galvin Li, Dominic Matriccino, & Kevin Nguyen



1,451 miles

Titanic’s distance travelled

882 ft

Titanic’s length

1,517

Lives were lost

28°F

The ocean’s temperature

— Introduction —

Background: The Titanic infamously sank in 1912, killing hundreds of people due to a lack of lifeboats and safety procedures. Our dataset contains information about passengers on the Titanic. We want to see if there are any trends within certain predictors that can provide insight survivability and passenger fare.

- What were the optimal survival conditions?
- What type of passengers had the most/least expensive passenger fare?

Regression Problem: We want to determine which predictors significantly influence the cost of the fare and if they can accurately predict the passenger fare.

Classification Problem: We want to determine which predictors significantly influence passenger survivability and if they can accurately predict whether a passenger survived or not.

— Predictors —

Survival: 0 = died & 1 = survived

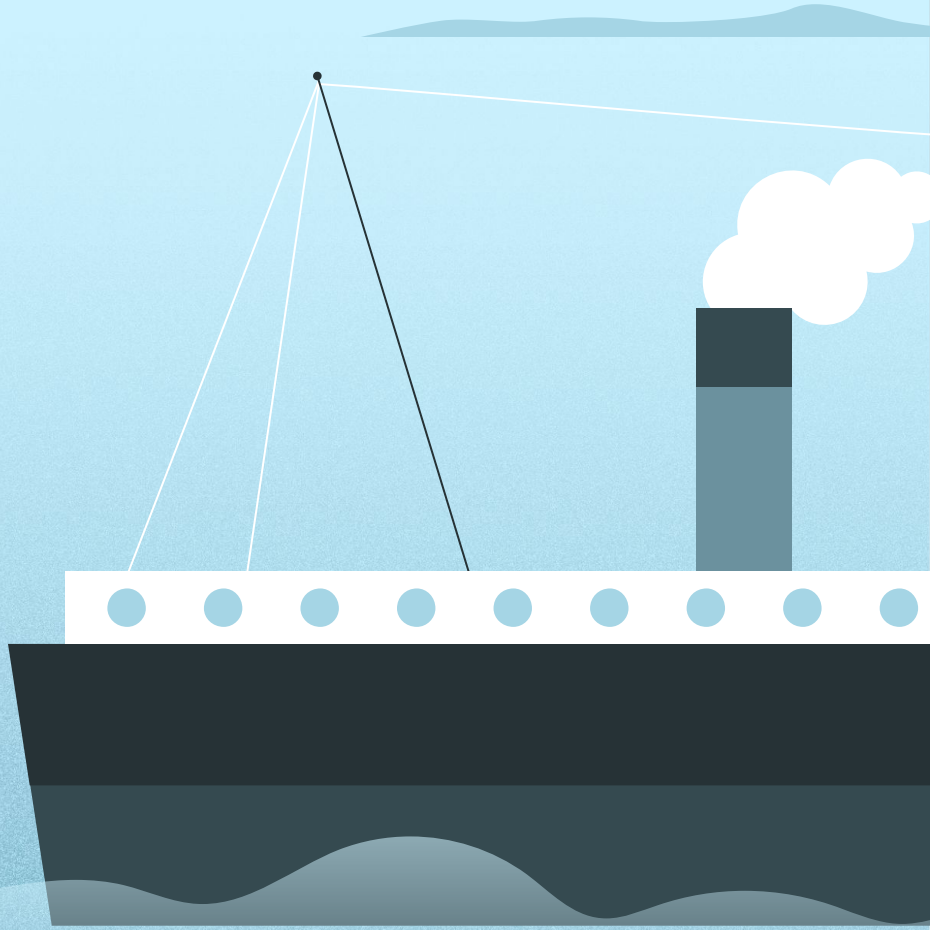
Age, Sex, and Fare

Pclass: The passenger's ticket class with levels 1, 2, & 3, denoting first class, second class, and third class, respectively

SibSp: The number of siblings and spouses the passenger travelled with (mistresses and fiancés excluded)

ParCh: The number of parents and children the passenger travelled with (nannies excluded)

Embarked: The port where the passenger boarded the Titanic, with levels C, Q, & S, denoting Cherbourg, Queenstown, and Southampton

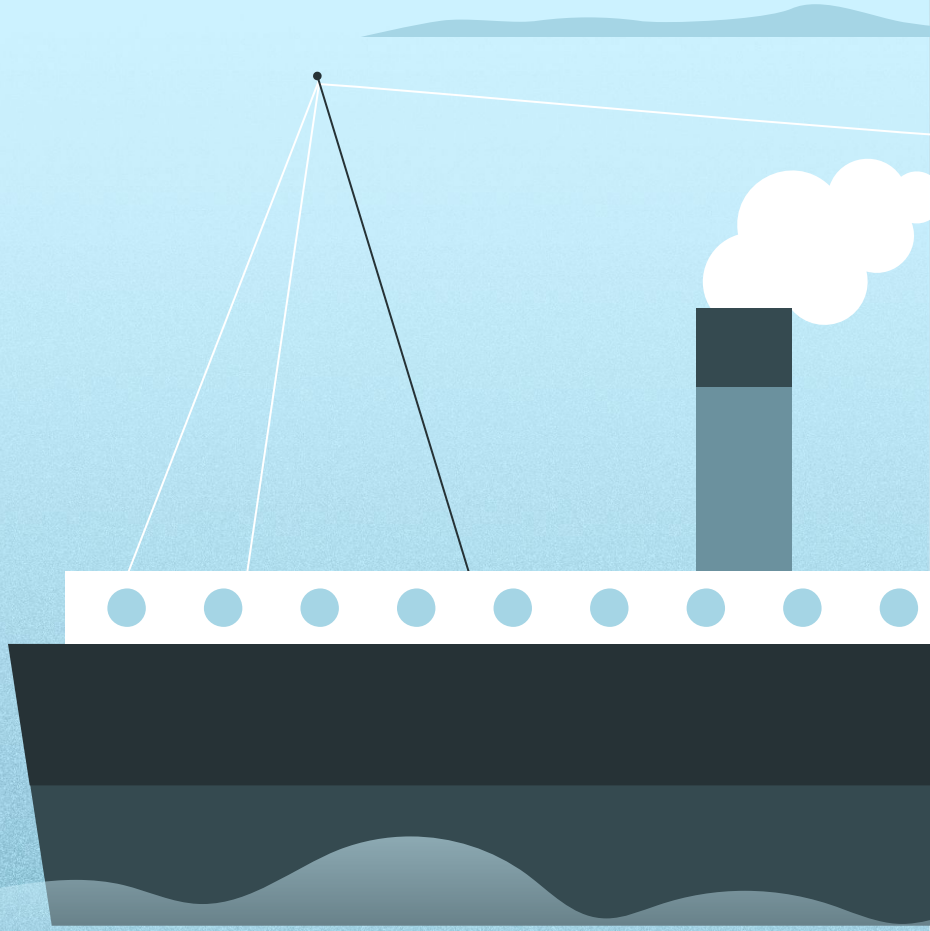


— Data Cleaning —

Removed predictors deemed unnecessary or had many missing values

Conducted research to find specific missing values

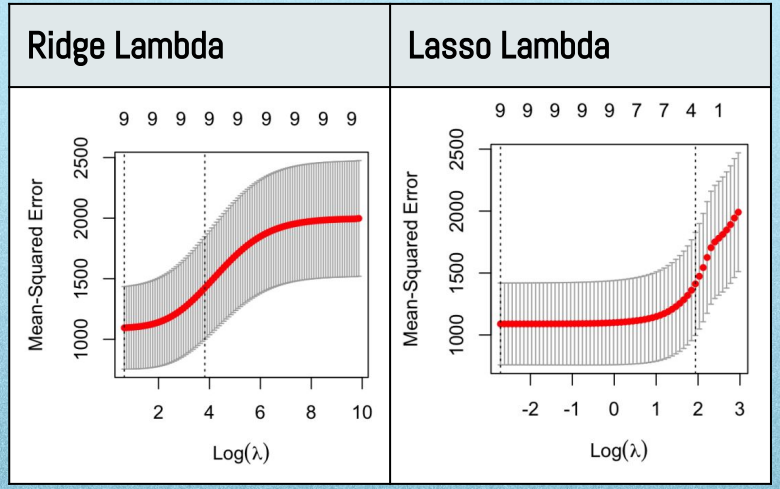
Performed 70/30 split on data into training set (623 observations) and test set (268 observations)



Regression Problem

— **Linear:** Choosing a Tuning Parameter —

- Ridge regression and lasso regression require a tuning parameter.
- 10-fold cross validation
- Ridge: 1.9292
- Lasso: 0.0662
- Lasso did not perform variable selection.



— Linear: Coefficients —

Ordinary Least Squares		Ridge		Lasso	
(Intercept)	84.17065181	(Intercept)	79.74395258	(Intercept)	83.63459779
Survived1	2.42799654	Survived1	3.73376898	Survived1	2.41821872
Pclass2	-55.67538242	Pclass2	-49.59379815	Pclass2	-55.34919744
Pclass3	-63.82541180	Pclass3	-57.83536302	Pclass3	-63.60080242
Sexmale	-4.85521247	Sexmale	-4.70008288	Sexmale	-4.77519873
Age	-0.09276794	Age	-0.08825074	Age	-0.08785925
SibSp	5.91449883	SibSp	5.66840616	SibSp	5.86926270
Parch	9.78408209	Parch	9.33682209	Parch	9.73838705
EmbarkedQ	-8.71219024	EmbarkedQ	-10.36478845	EmbarkedQ	-8.36146707
EmbarkedS	-10.56070301	EmbarkedS	-11.16773614	EmbarkedS	-10.34147756

— Linear Regression Methods —

- Naive (using training mean) performed the worst.
- OLS performed the best.

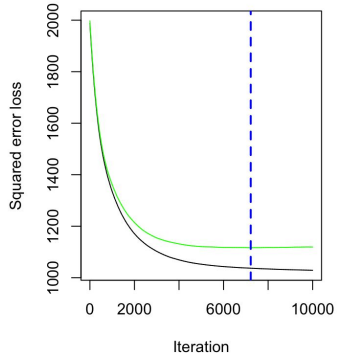
- Why does OLS beat ridge and lasso?
  - Bias-variance tradeoff

MSPE	**Naive Method**	**Ordinary Least Squares**	**Ridge**	**Lasso**
	3580.933	2249.013	2277.664	2251.337

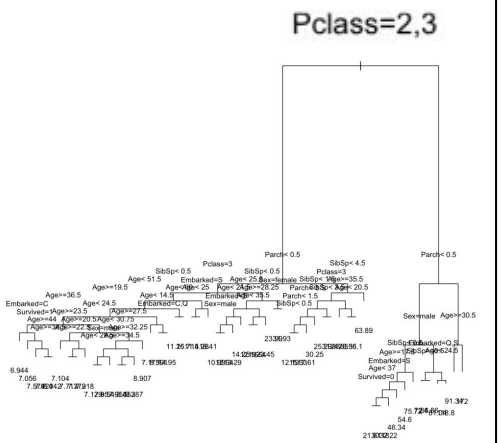
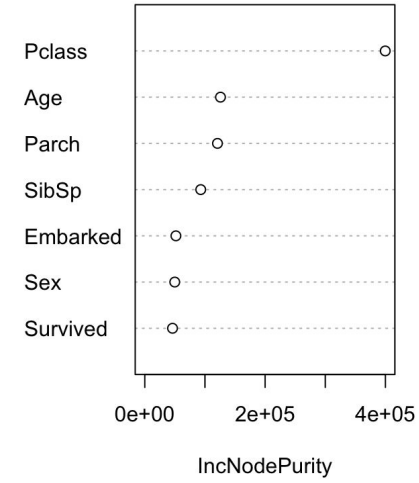
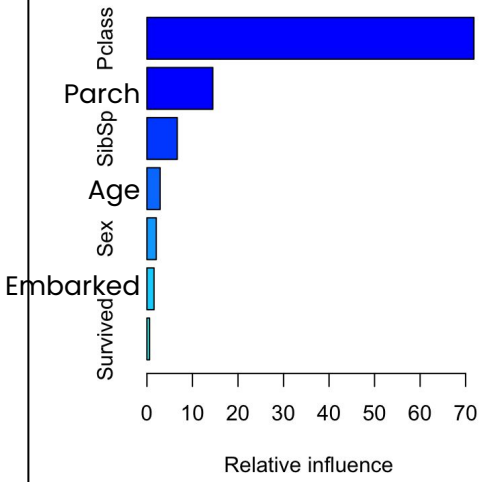
— **Nonlinear:** Choosing a Tuning Parameter —

- Decision tree's cost parameter and boosting's number of trees was found using 10-fold cross validation.
- Random forests mtry = 2.

Decision Tree	Boosting

`> tree_TuneR$bestTune`
`  cp`
`1  0`
	

— **Nonlinear:** Variable Importance —

Decision Tree	Random Forests	Boosting
		

— Nonlinear Regression Methods —

- All nonlinear methods performed better than linear methods.

- Boosting performed the worst - single split too simple?
- Random forests performed the best.
  - Allowed better splits than decision tree's greedy splitting.
  - Less overfitting compared to decision tree.

MSPE	**Decision Tree**	**Random Forests**	**Boosting**
	2201.030	2100.803	2299.528

Classification Problem

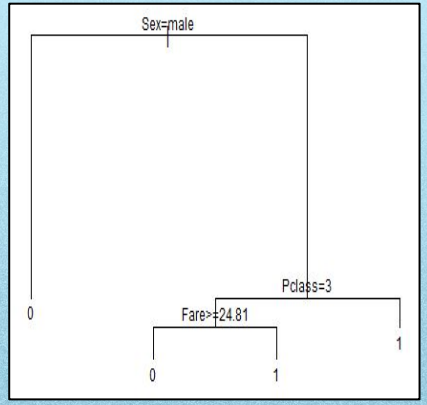
— Linear Classification Methods —

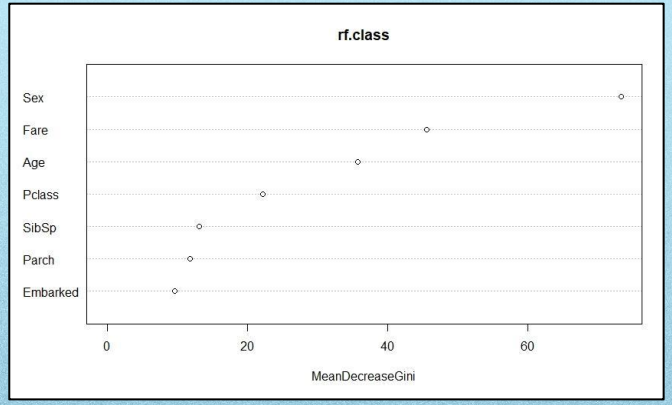
- QDA performs better likely because of its ability to pick up on nonlinearities in the feature space
- Logistic regression performed surprisingly poorly
- Our dataset was unbalanced so our FNR was consistently greater than FPR

Naive (Majority)	Logistic Regression	LDA	QDA
0.350746	0.2574627	0.2164179	0.2052239

— Nonlinear Classification Methods —

Misclassification Rate	Decision Tree	Random Forests	Boosting
	0.1902985	0.1828358	0.2164179

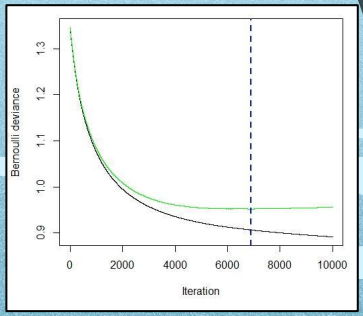




	rel.inf
Sex	57.163351
Pclass	15.902001
Fare	9.274288
Age	6.612239
Embarked	4.917296
SibSp	4.872435
Parch	1.258390

```
> tree_Tune$bestTune
   cp
23 0.022
```

```
No. of variables tried at each split: 2
```



— Conclusion and Additional Work —

- Additional information released about passengers who were on the Titanic is unlikely
- This data was collected before digitalization so the validity and reliability of the data is questionable
- The names of the passengers could be a useful source of information to do further research
- The increase in safety precautions regarding large cruise ships were likely a result of the Titanic
- Experimenting with different splits in the data or creating models with different combinations of predictors could improve our accuracy

Thank You For Listening!

Any questions?