CMSE202 Section 1

Titanic Survivors' Prediction

Zhishan Li, Colin Williams, Titus Merriam

TITANIC

01

Our Question

Scientific question that we try to answer

02

Data & Model

The model that is applied to the chosen topic

03

Computational Techniques

Methods(Python libraries/packages) that are used

04

Answers

Results that we arrive at

05

Difficulties or Complications

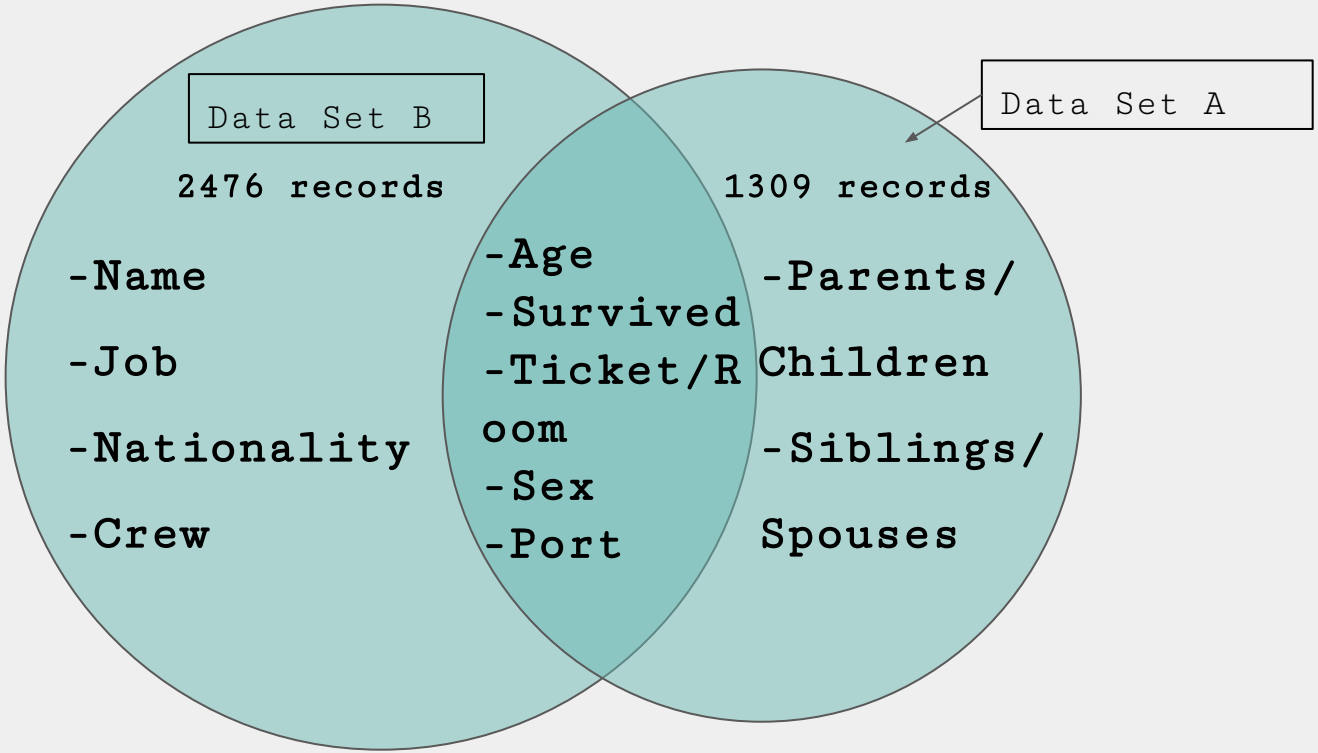
Difficulties that we faced and how we overcome

Questions to Answer

- **Can the survivors of the titanic disaster be accurately predicted based on their features?**
  i.e. Gender, class,nationality, etc.
- **Which features are more important to the outcome of a passenger?**

Data Background

- Two data sets are used



Models

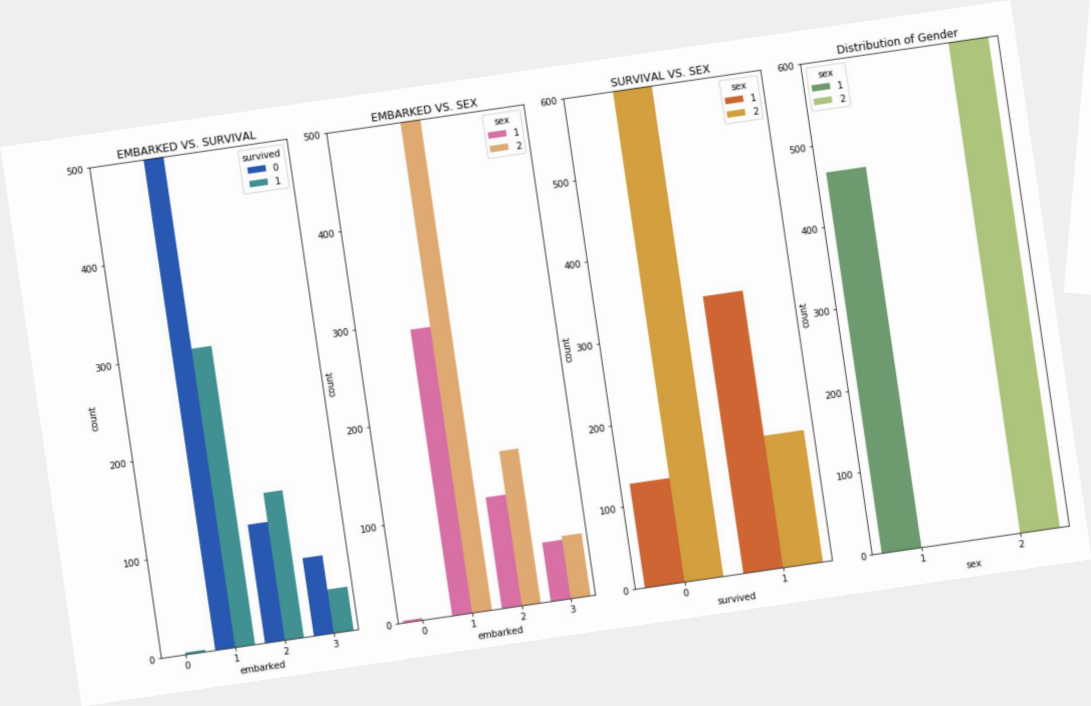
-Categorical Data→ Numerical Data

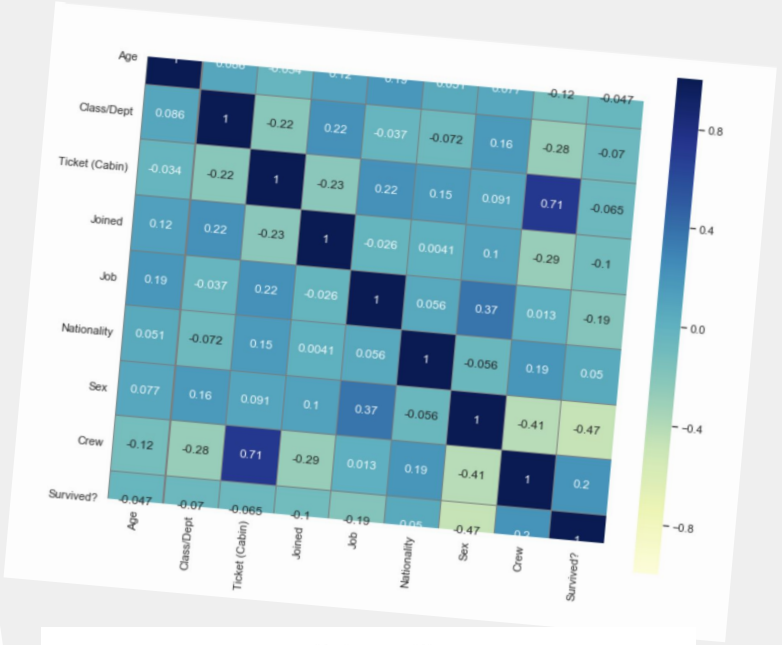
    -NA/Missing Value→ 0

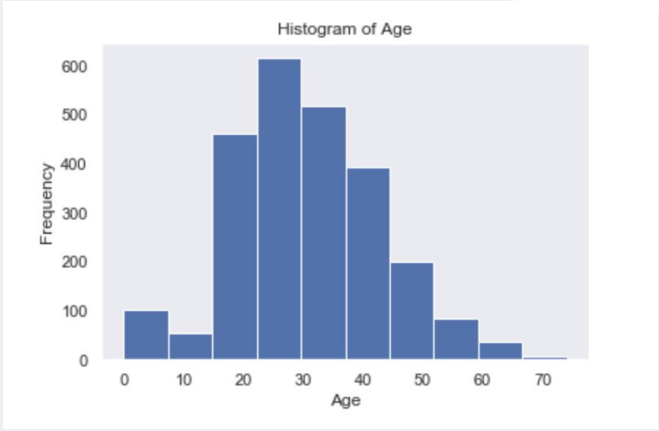
    -Embarking ports, cabin markers, and home. dest→ unique integer values

    -Sex: Female=1, Male=2

Initial Analytics







Computational Techniques

1. Random Forest Classifier

- Consist of a large number of individual decision trees→ an ensemble
  - Each individual tree in the random forest spits out a class prediction
  - The class with the most votes becomes the model’s prediction

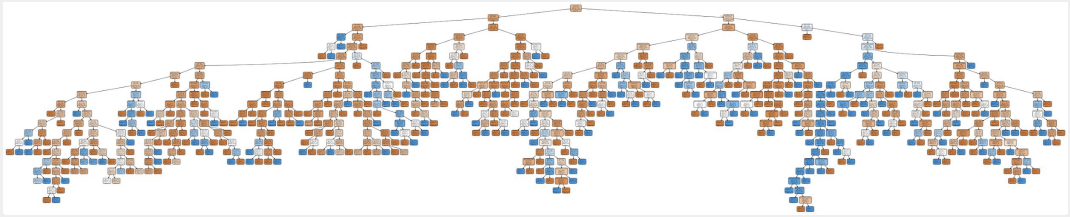
Computational Techniques

1. Random Forest Classifier

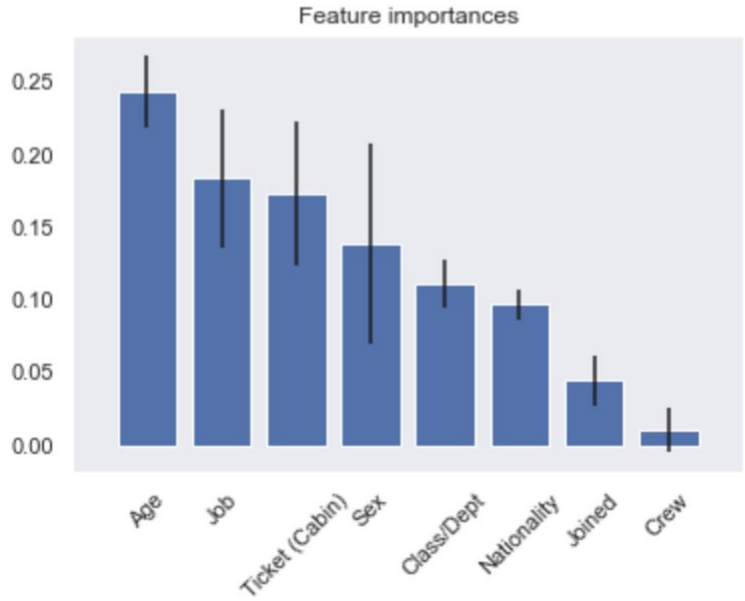
a. First Random Forest

i. Accuracy: 0.8368336025848142

ii. Using 8 features with ranking(see the bar graph on the right)



```
Feature ranking:
1. feature Age (0.243074)
2. feature Job (0.183479)
3. feature Ticket (Cabin) (0.172788)
4. feature Sex (0.138296)
5. feature Class/Dept (0.110925)
6. feature Nationality (0.096564)
7. feature Joined (0.044351)
8. feature Crew (0.010523)
```



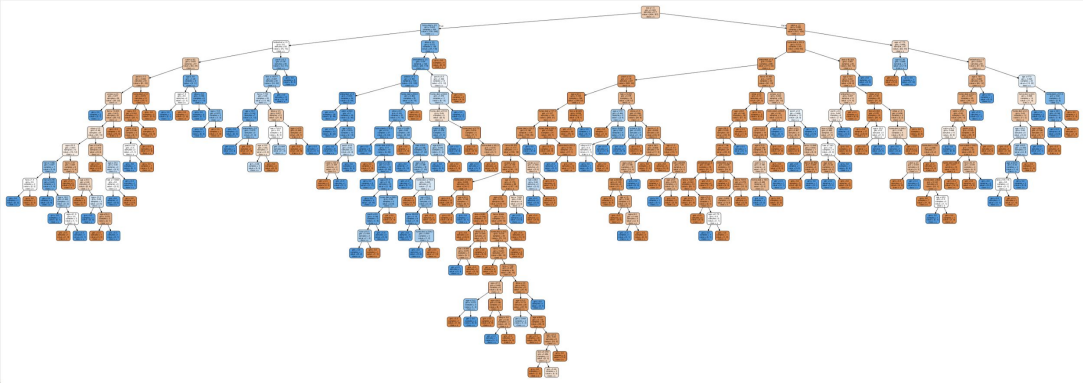
Computational Techniques

1. Random Forest Classifier

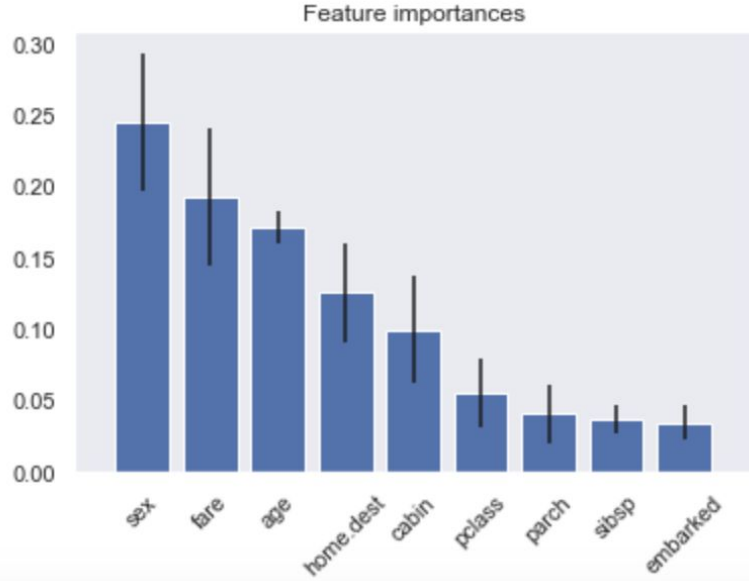
b. Second Random Forest

i. Accuracy: 0.7896341463414634

ii. Using 9 features with ranking(see the bar graph on the right)



```
Feature ranking:
1. feature sex (0.245039)
2. feature fare (0.192491)
3. feature age (0.170954)
4. feature home.dest (0.125303)
5. feature cabin (0.099419)
6. feature pclass (0.055464)
7. feature parch (0.040739)
8. feature sibsp (0.036215)
9. feature embarked (0.034376)
```



Computational Techniques

1. Random Forest Classifier

c. After two Random Forest, Now...

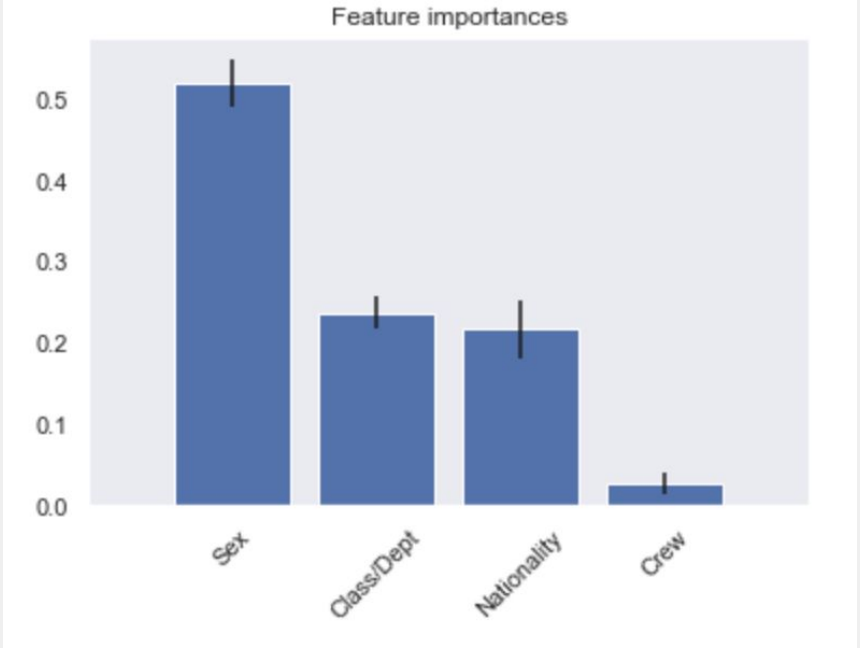
** Try RF again by removing 1~5 features & keep the best score

** Using combination of 3 to 7 features of the 8 total features (full data set)

i. Accuracy: 0.8539579967689823

ii. Achieved with just 4 features (see on the right)

```
Feature ranking:
1. feature Sex (0.518537)
2. feature Class/Dept (0.237046)
3. feature Nationality (0.216393)
4. feature Crew (0.028023)
```



Computational Techniques

2. Support Vector Machines

- Supervised Machine Learning Algorithm - sklearn
- Segregate 2 Categories: Survived or Died
  - by a hyperplane/line

Computational Techniques

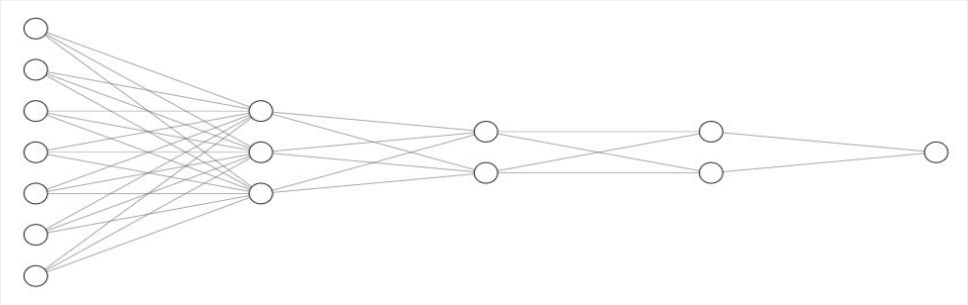
3. Stochastic Gradient Descent

- Implements regularized linear models with *stochastic gradient descent (SGD)* learning - used sklearn
  - The gradient of the loss is estimated each sample at a time
  - The model is updated along the way with a decreasing learning rate

Computational Techniques

4. Neural Network

- Use Keras to build a neural network
  - 7 feature input, 3 hidden layers, 1 node output.
  - Included Rectified Linear Unit, Linear, and Sigmoid activations.



Result Comparison

	Random Forest	Support Vector Machine	SGD Classifier	Neural Network
Accuracy (0.25 Training Size)	0.8540	0.7625	0.7512	0.7415

Answer to the Question

- The best accuracy we reach is the Random Forest classifier which gave 85.4% accuracy.
- We can predict the survivors of Titanic disaster based on the top 4 features(from the highest to lowest):
  - Sex, Class, Nationality, Crew

Difficulties & Complications

- Full data set accessibility and missing values
  - Purchase the full data set
  - Fill in missing values with a default value or remove columns entirely
- Using categorical data
  - Assign a number or omit from calculations

Thank You

Questions?