Spaceship Titanic

Predict which passengers are transported to an alternate dimension

OUR TEAM

Anmol Agrawal
EID:aa96674

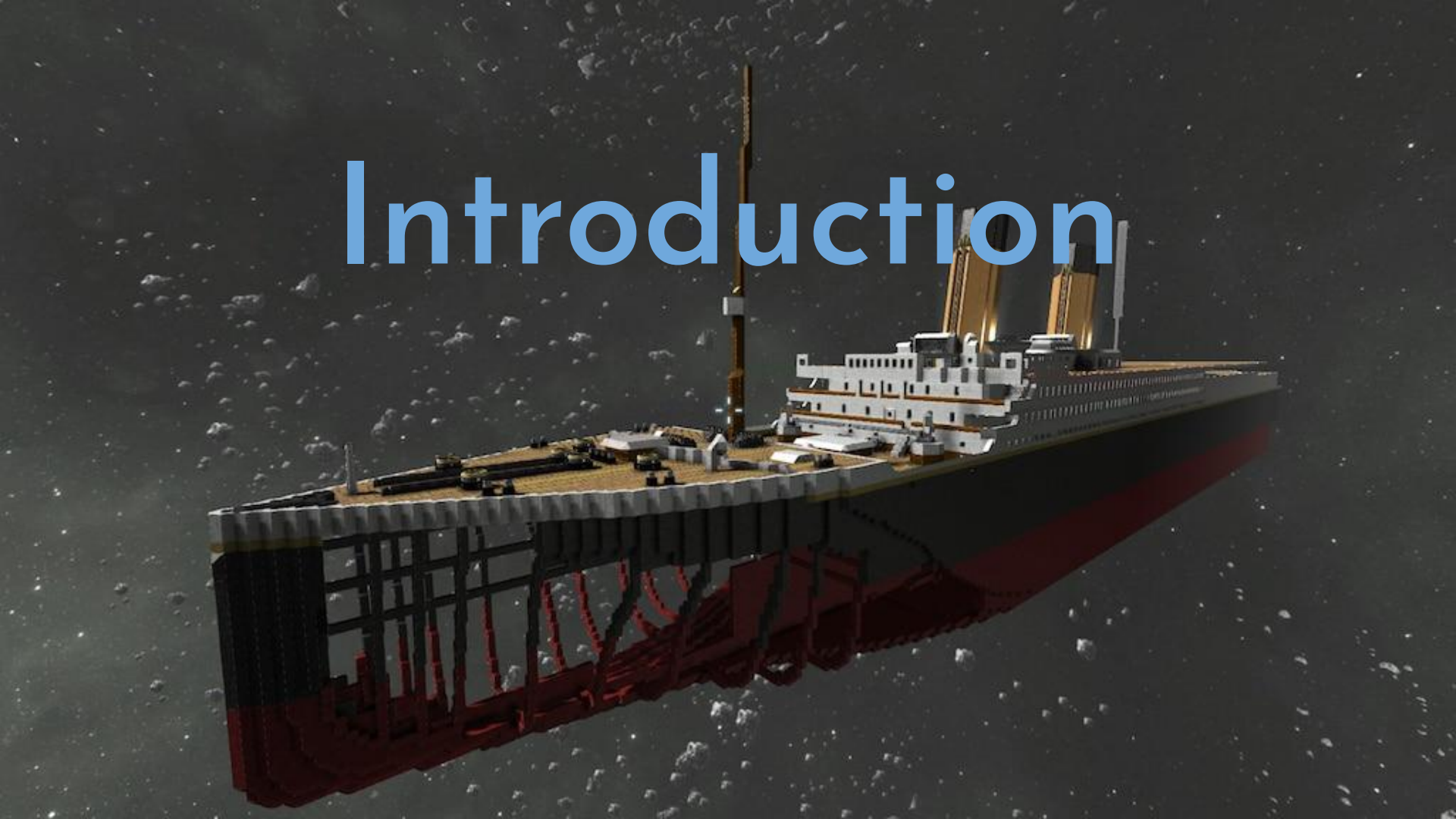
Anvesh Karangula
EID:ak49956

Evan Hadd
EID: eh28834

Lisa Ko
EID:hk25297

Bi dy)
Yang
EID:by4325

Introduction



OUR PREVIEW

Exploratory Data Analysis

01

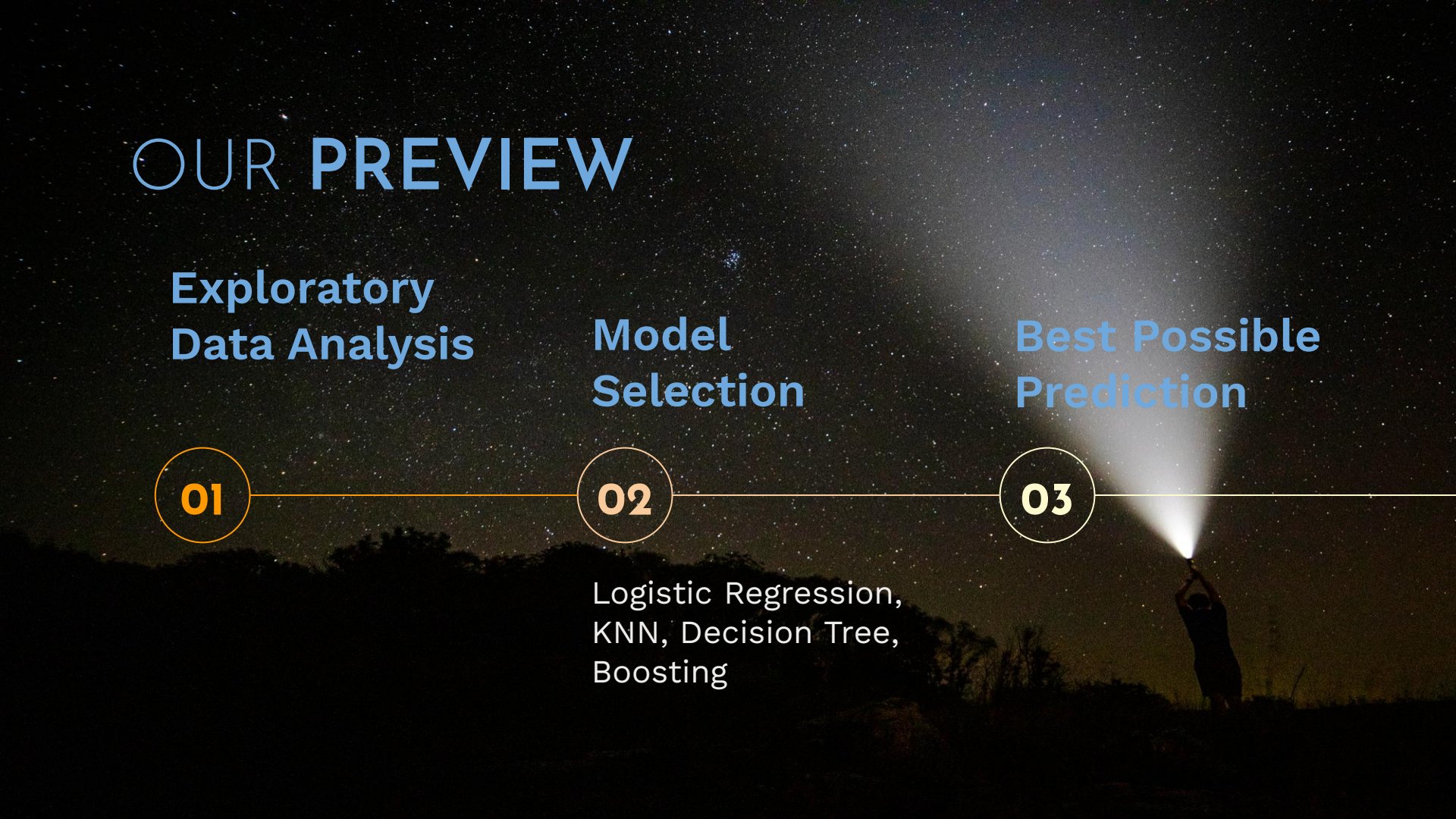
Model Selection

02

Logistic Regression, KNN, Decision Tree, Boosting

Best Possible Prediction

03




Exploratory
Data Analysis

- `PassengerId` - Unique ID for each passenger
- `HomePlanet` - The planet the passenger departed from
- `CryoSleep` - Whether the passenger elected to be put into suspended animation for the duration of the voyage. Passengers in cryosleep are confined to their cabins
- `Cabin` - The cabin number where the passenger is staying. Takes the form `deck/num/side`, where `side` can be either P for *Port* or S for *Starboard*
- `Destination` - The planet the passenger will be debarking to
- `Age` - The age of the passenger
- `VIP` - Whether the passenger has paid for special VIP service during the voyage
- `RoomService`, `FoodCourt`, `ShoppingMall`, `Spa`, `VRDeck` - Amount the passenger has billed at each of the *Spaceship Titanic*'s many luxury amenities
- `Name` - The first and last names of the passenger

Filling Missing Values of Expenditures

"Are you asleep?"

'Cyrosleep'

Yes

I don't know

No

Fill with 0

Delete whole row

Fill with Mean

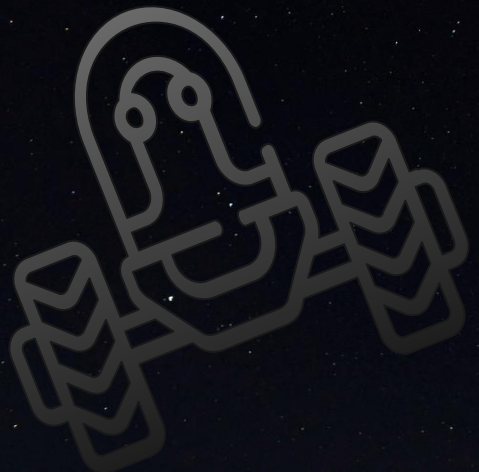
DATA EXTRACTION

Split 'Cabin' Column

- Deck
- Deck number
- Sign

Convert Categorical data

- To Factor data type
- Recode as 0 or 1



01

Logistic Regression

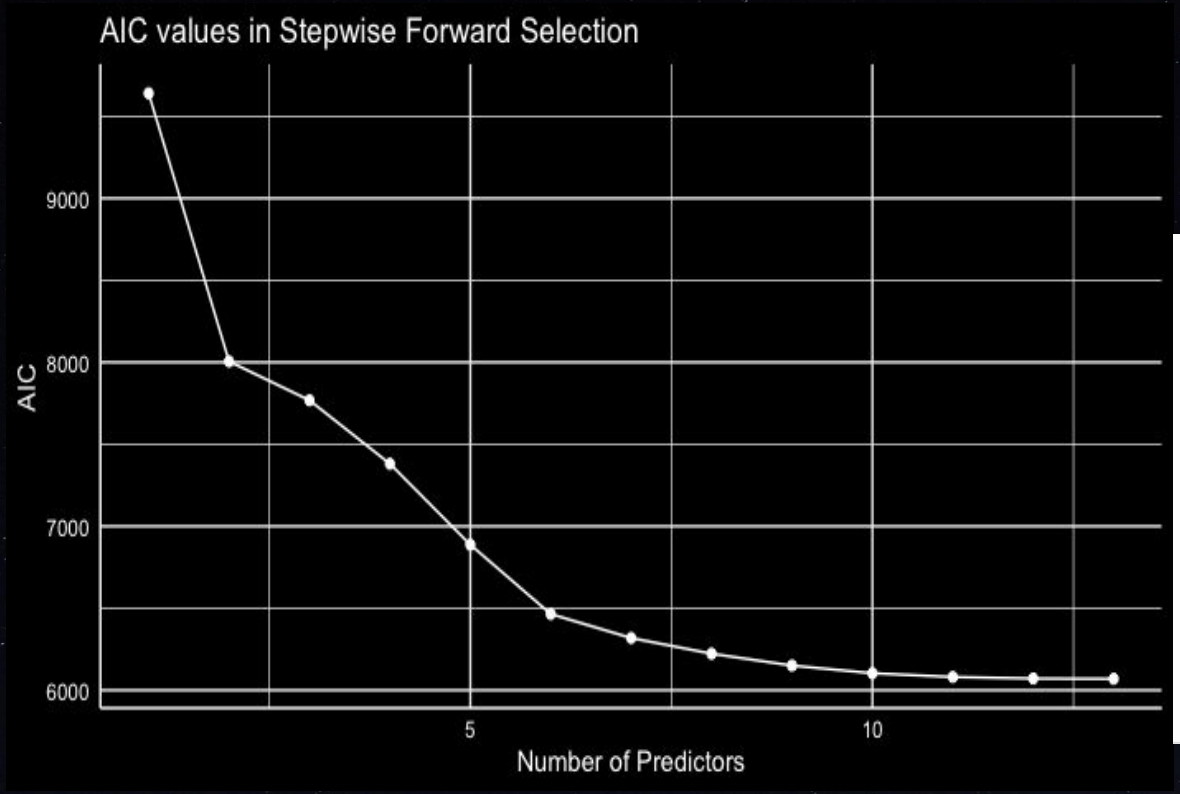
Logistic Regression

Model created from Step function:

Transported = CryoSleep + Spa + HomePlanet + VRDeck + RoomService + FoodCourt + Deck + Side + ShoppingMall + Destination + Age

AIC: 7363.3

Logistic Regression




```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 612 125
         1 209 576

               Accuracy : 0.7806
                 95% CI : (0.7589, 0.8011)
    No Information Rate : 0.5394
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5622
```

02

KNN

What we did

- Converted categorical variables to dummy variables
- Normalized numerical data (Not to skew the results)
- Response variable as a factor
- Used recursive feature elimination to pick the important predictors (CryoSleep, Spa, RoomService, VRDeck, FoodCourt)
- Tried various k-fold CV's to find the optimal error rate

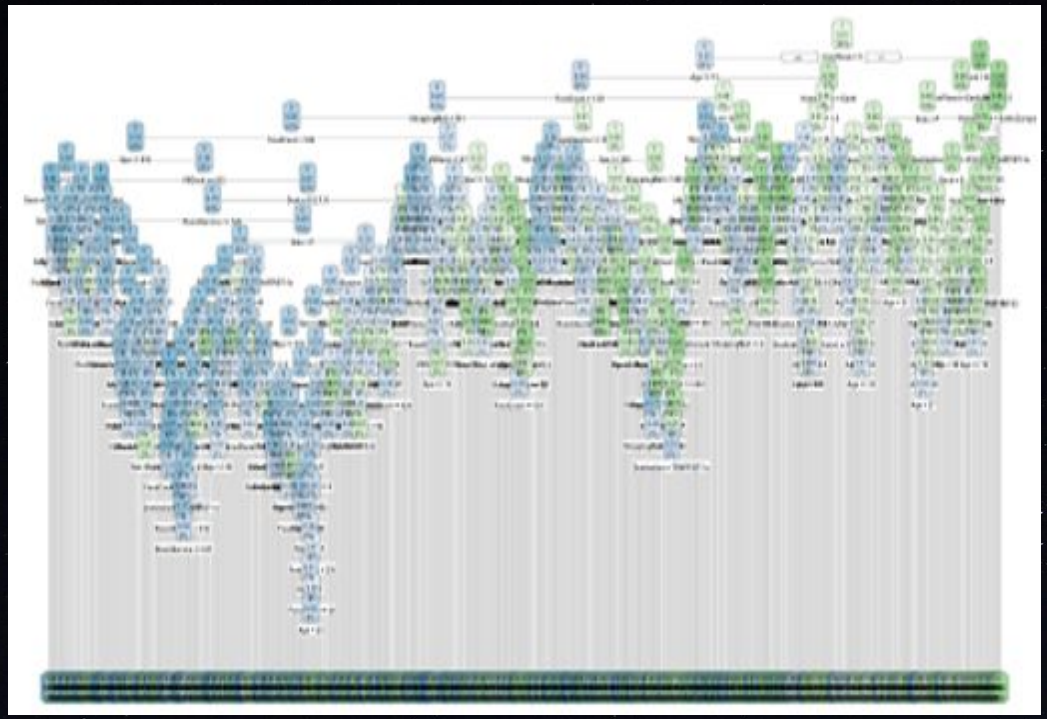
KNN

K value with minimum error rate	25
Minimum error rate	0.2112
Accuracy	78.88 %
K-fold CV selected	8

03

Decision Tree

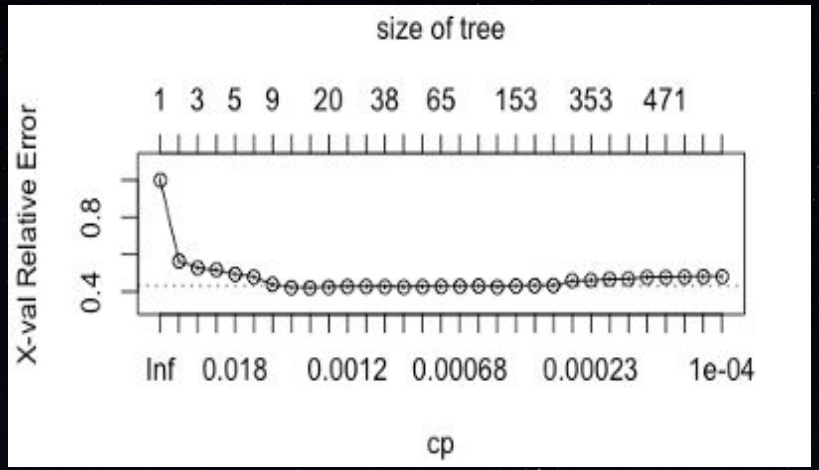
Decision Tree/Pruning



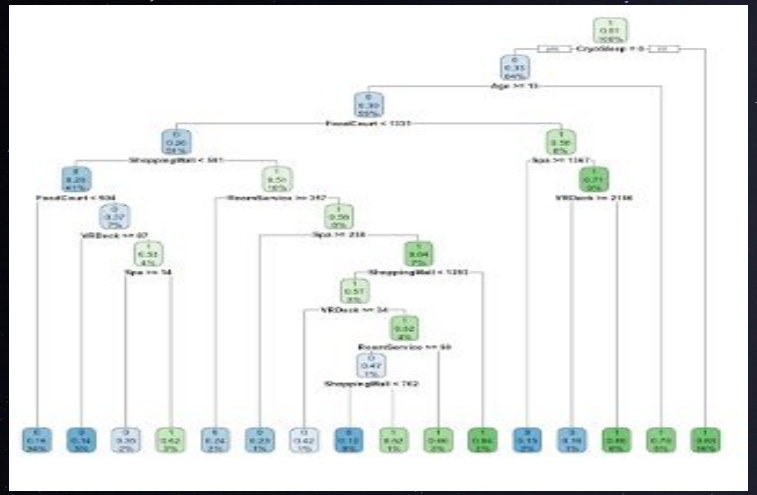
Accuracy Level of Big Tree on Training set: **0.9128447**
POTENTIAL OVERFITTING!!!!

Accuracy Level on Test set: **0.7486726**
BIG TREE WAS A DEFINITE OVERFITTING

Decision Tree/Pruning

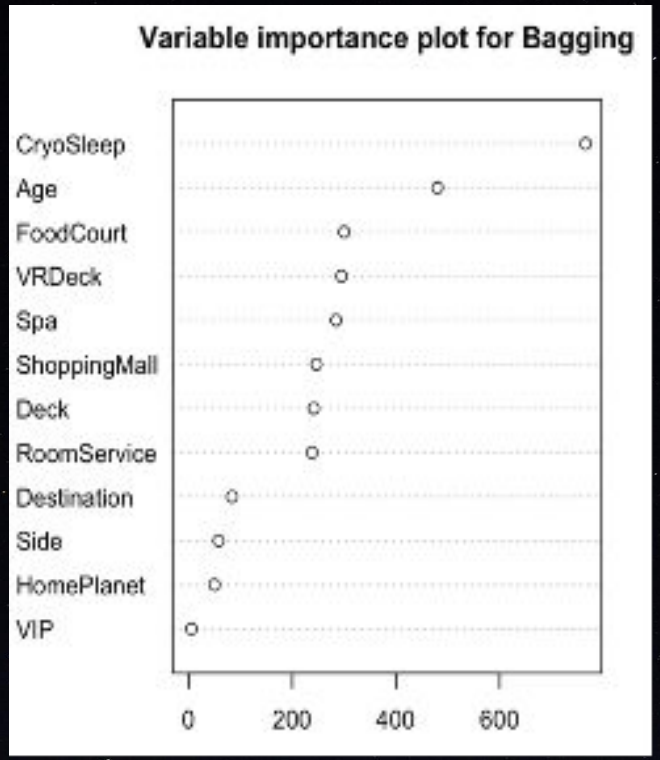


Best alpha: **0.0015**
Size of Pruned Tree: **16**



Accuracy of Pruned Tree on test set: **0.7699115**

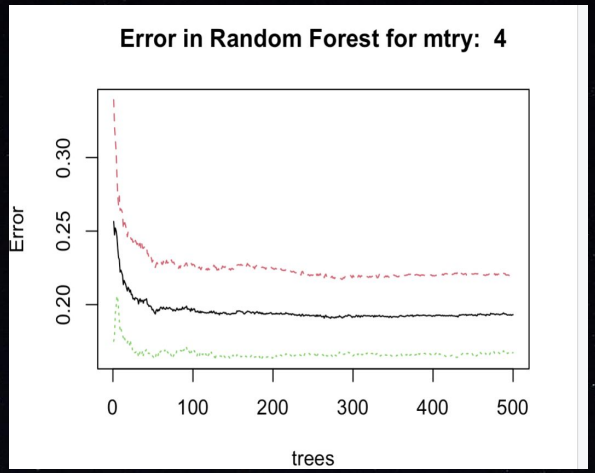
Bagging

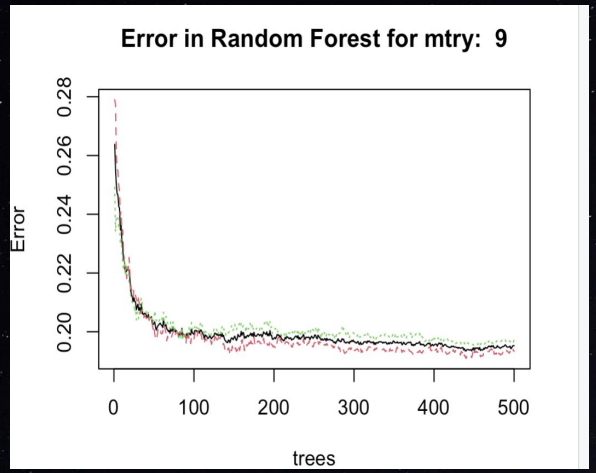


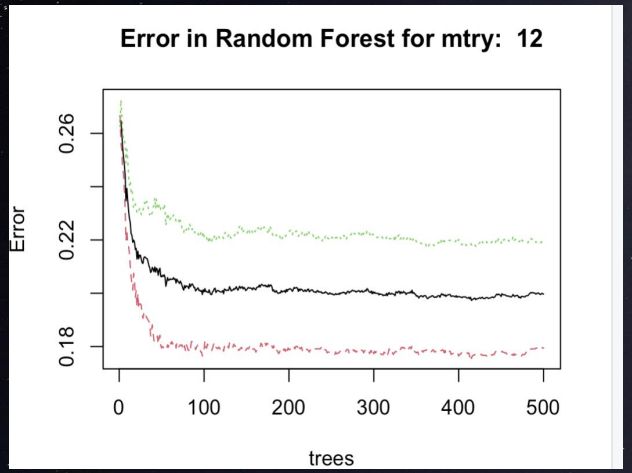
> cat('Accuracy of bagging on test set: ',accuracy_v_bag,'\n')

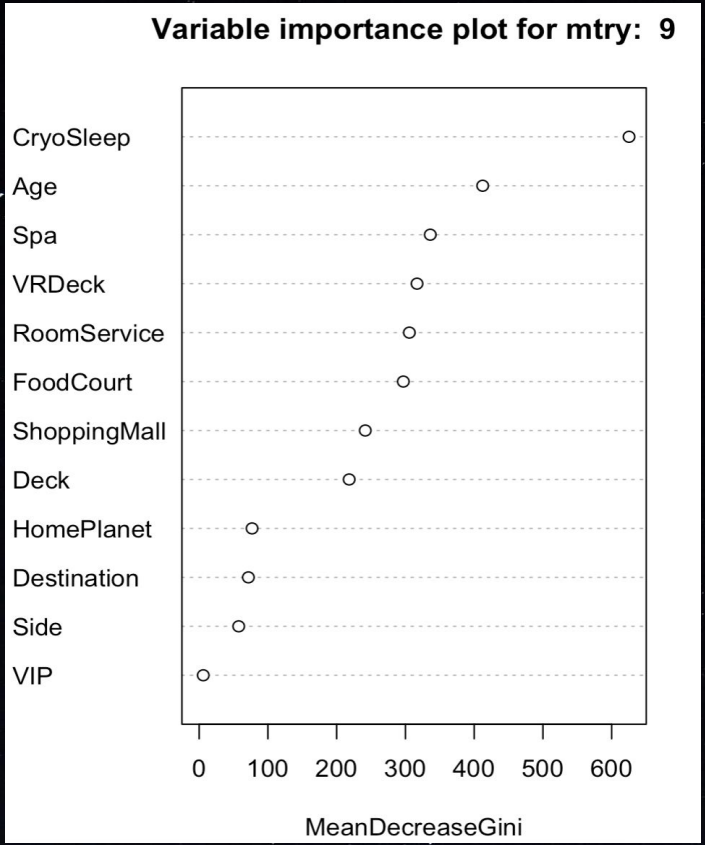
> Accuracy of bagging on test set: **0.7758112**

Random Forest









> cat('Best mtry:', best_mtry)

Best mtry: **9**> cat('Best Accuracy:', best_accuracy)

Best Accuracy: **0.7899705**

04

Xtreme Gradient Boosting

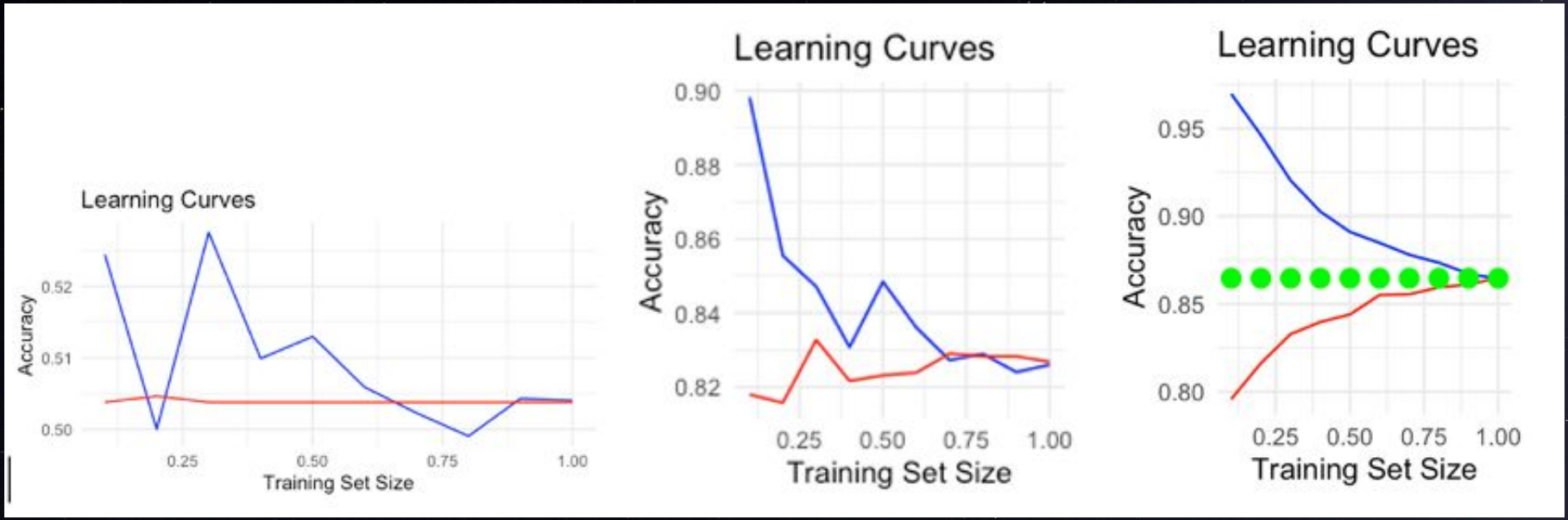
xgBoosting

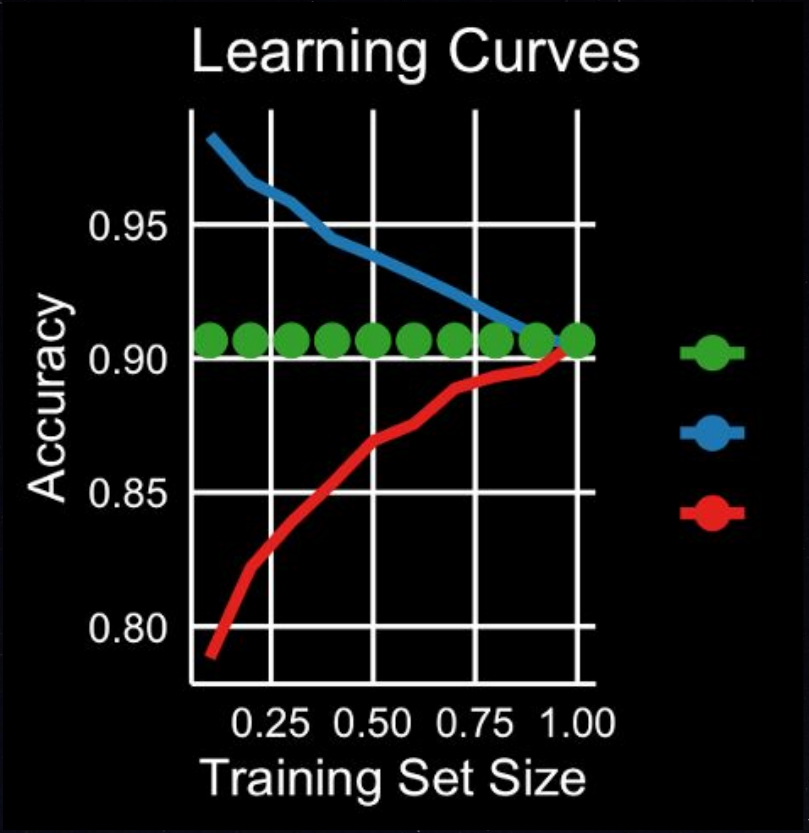
- **Hyperparameter Tuning and Cross-Validation**
- **Model training and evaluation**
- **Best model selection and testing**
- **Visualization**

eta	Learning Rate
max_depth	Tree Depth
nrounds	Boosting Rounds
subsample	Subsample Ratio

Accuracy: 81.13%

xgBoosting


Learning Curves
Accuracy
0.52
0.51
0.50
0.25
0.50
0.75
1.00
Training Set Size
Learning Curves
Accuracy
0.90
0.88
0.86
0.84
0.82
0.25
0.50
0.75
1.00
Training Set Size
Learning Curves
Accuracy
0.95
0.90
0.85
0.80
0.25
0.50
0.75
1.00
Training Set Size

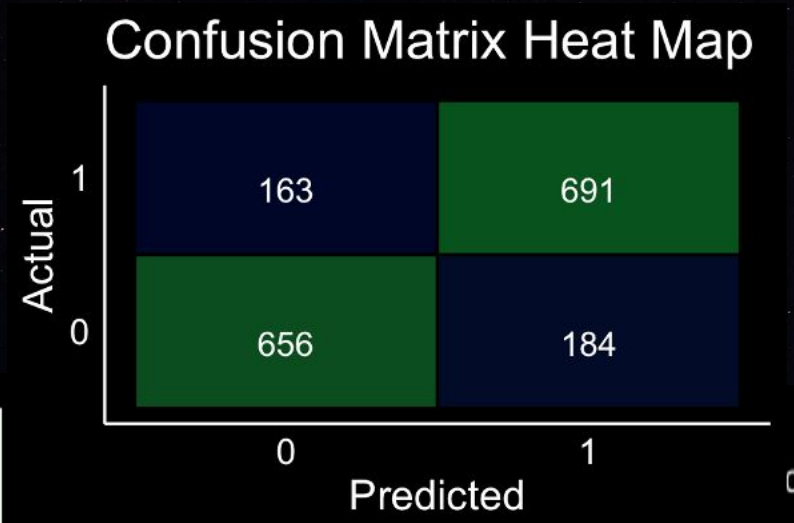


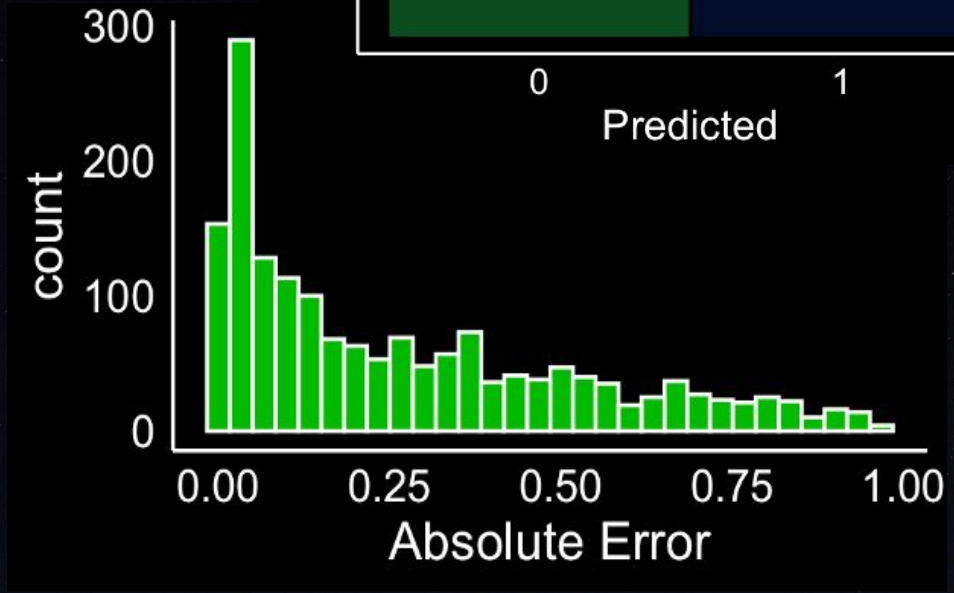
xgBoosting

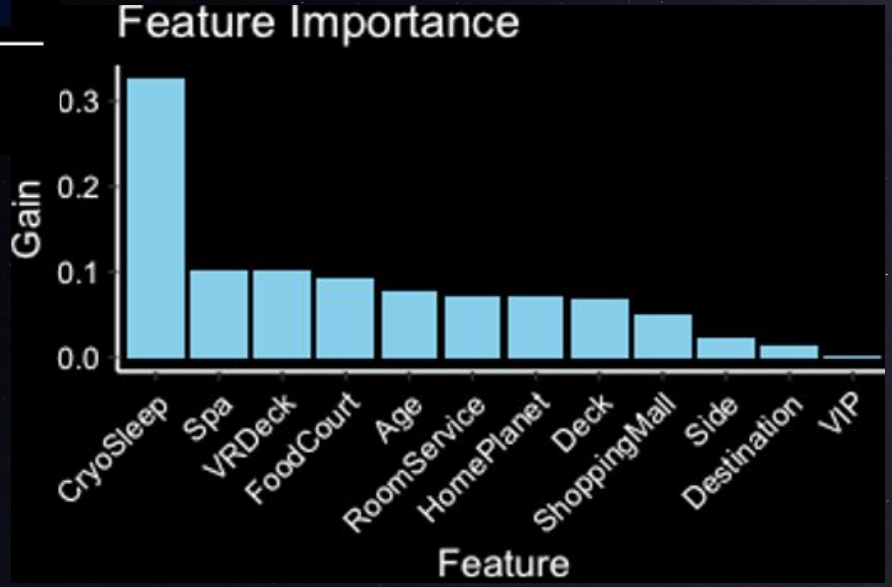
eta	0.01
max_depth	7
nrounds	500
subsample	0.8

Accuracy: 81.13%

xgBoosting


Confusion Matrix Heat Map
Actual
1
0
163
691
656
184
0
1
Predicted

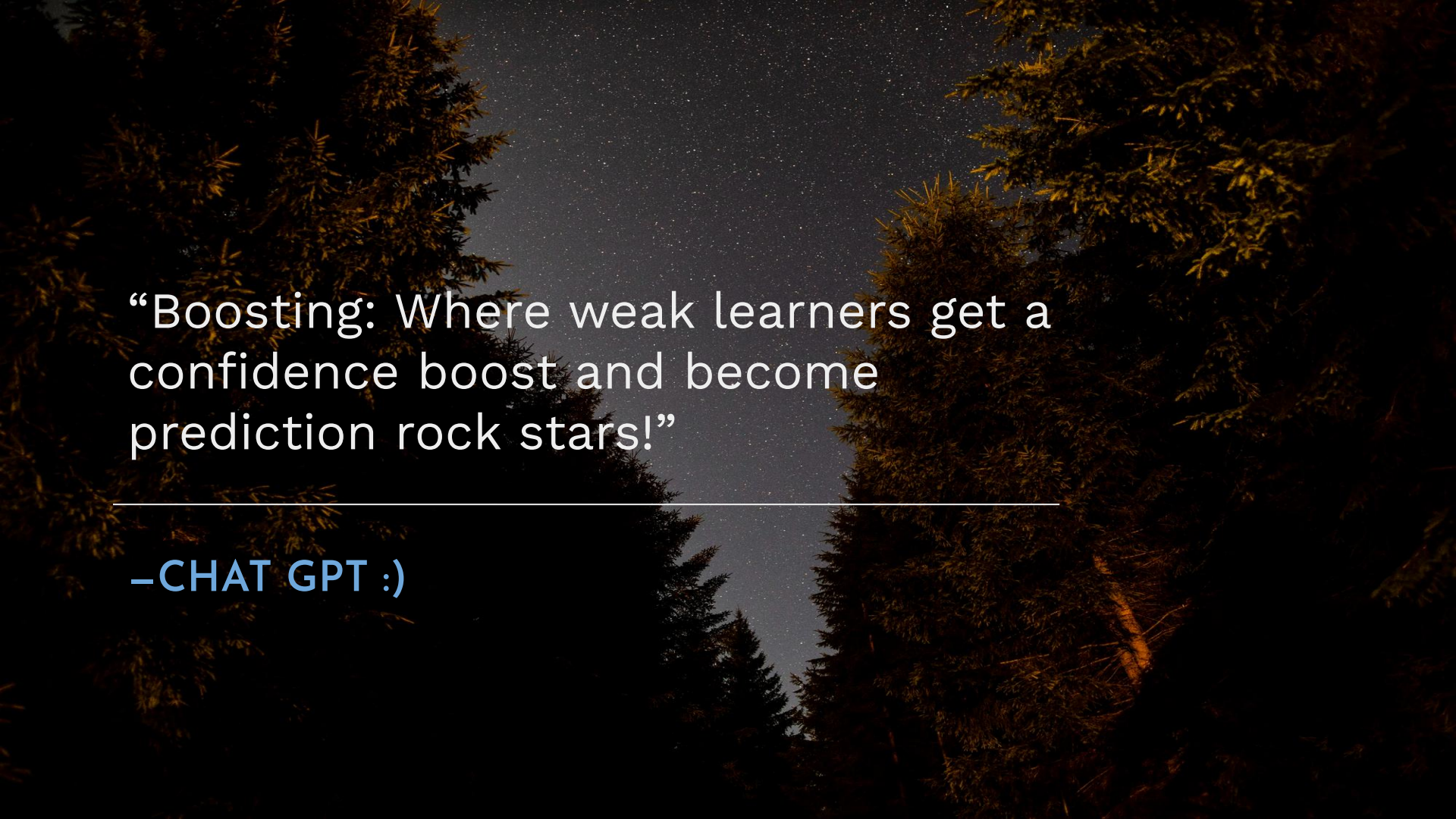

300
200
100
0
count
0.00
0.25
0.50
0.75
1.00
Absolute Error


Feature Importance
Gain
0.3
0.2
0.1
0.0
CryoSleep
Spa
VRDeck
FoodCourt
Age
RoomService
HomePlanet
Deck
ShoppingMall
Side
Destination
VIP
Feature


BEST
PREDICTION

ACCURACY RATE COMPARISON

Logistic Regression	78.06%
KNN	78.88%
Decision Tree	78.99% [Random Forest]
xgBoosting	81.13%


“Boosting: Where weak learners get a confidence boost and become prediction rock stars!”
–CHAT GPT :)

CONCLUSION


THANK YOU
Do you have any questions?