Titanic

The classification of Titanic passengers

Andreea STROIA
Hritika KATHURIA
Hala ALBAHLOUL

Agenda

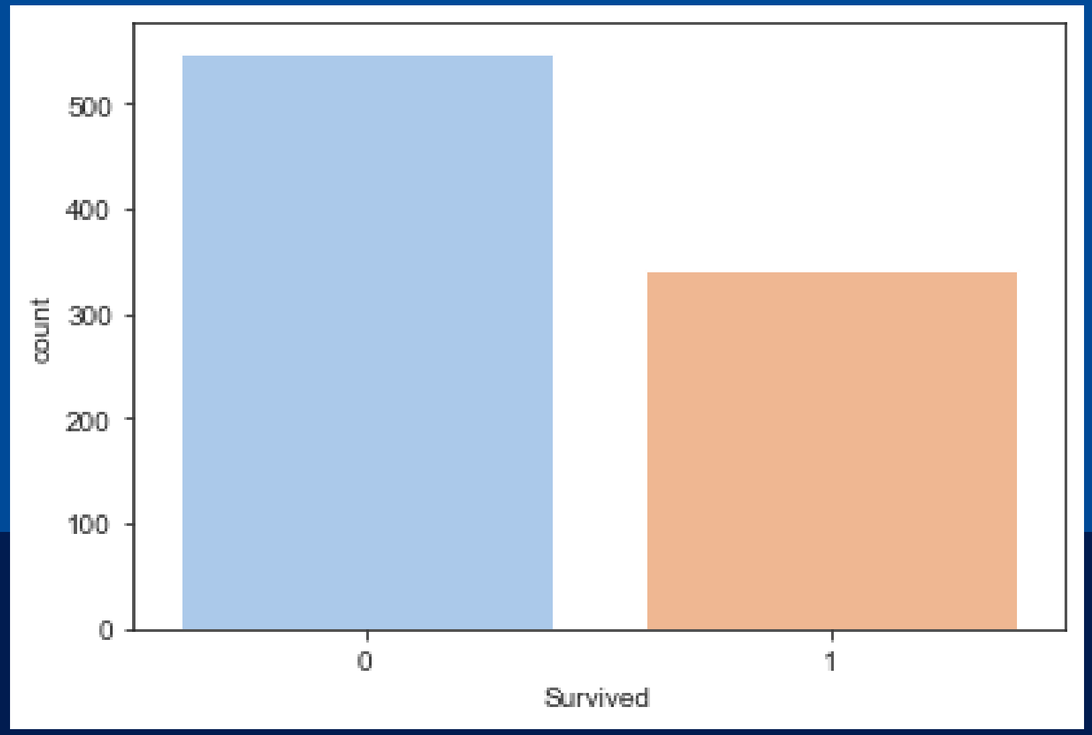
Logistic regression

K-Nearest Neighbors classifier

Discriminant Analysis

Questions

Let's see how we stand by now



We know that:

People who didn't survive ~61.61%
People who survived ~38.38%

Classification

Logistic regression

Modelling Survivability in terms of other variables

Steps..

1. Transforming categorial data

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare	Family	Adult	Sex_female	Sex_male	Embarked_C	Embarked_Q	Embarked_S
0	1	0	3	22.0	1	0	7.2500	2	1	0	1	0	0	1
1	2	1	1	38.0	1	0	71.2833	2	1	1	0	1	0	0

2. Splitting the data

3. Setting the target

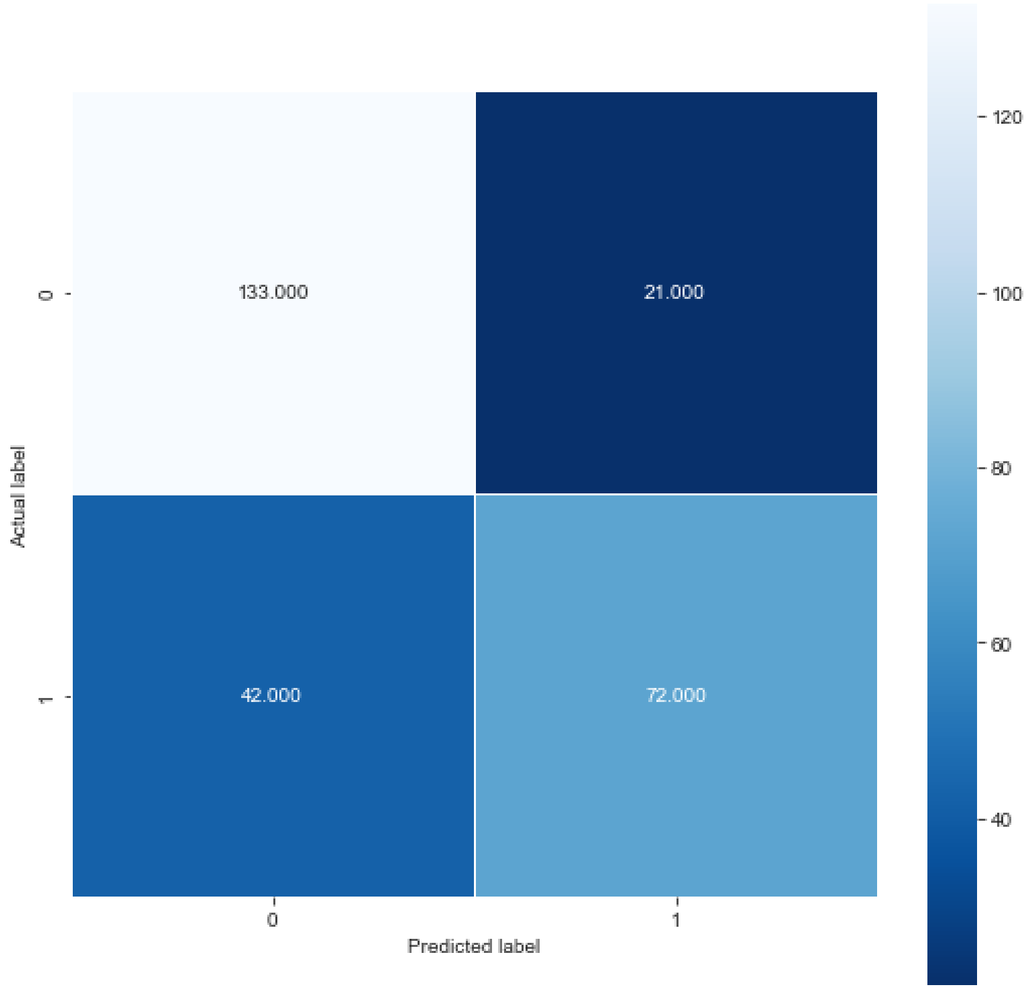
4. Predictions

5. Accuracy of the model

Some insights

	precision	recall	f1-score	support
0	0.76	0.86	0.81	154
1	0.77	0.63	0.70	114
accuracy			0.76	268
macro avg	0.77	0.75	0.75	268
weighted avg	0.77	0.76	0.76	268

- What percent of your predictions was correct?
- From all the positive classes, how many we predicted correctly?
- What percent of positive predictions were correct



KNN

1. Splitting the dataset

```
PredictorColumns = ['Pclass', 'Age', 'Family', 'Fare', 'Sex_female', 'Sex_male', 'Embarked_C', 'Embarked_Q', 'Embarked_S']
TargetColumn = 'Survived'

x = df[PredictorColumns].values
y = df[TargetColumn].values

from sklearn.model_selection import train_test_split

x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.3, random_state = 10)
```

2. Finding k

```
CV accuracy for k=1 is 65

CV accuracy for k=2 is 71

CV accuracy for k=3 is 73

CV accuracy for k=4 is 71

CV accuracy for k=5 is 71

CV accuracy for k=6 is 72

CV accuracy for k=7 is 70

CV accuracy for k=8 is 69

CV accuracy for k=9 is 69

CV accuracy for k=10 is 70
```

```
CV accuracy for k=11 is 69

CV accuracy for k=12 is 71

CV accuracy for k=13 is 69

CV accuracy for k=14 is 70

CV accuracy for k=15 is 69

CV accuracy for k=16 is 69

CV accuracy for k=17 is 70

CV accuracy for k=18 is 69

CV accuracy for k=19 is 69
```

```
optimal_k = final_scores.index(max(final_scores))
print(optimal_k)
```

KNN

3. Applying the model, from k = 1 to k = 19

4. Predicting the test results with the best accuracy

5. Evaluating the model

f1 score	0.71
accuracy score	0.74

Confusion Matrix

	Predicted NO	Predicted Yes
Actual No	121	20
Actual Yes	20	62

What error rate can we expect on the test set?

KNN	0.26
Logistic regression	0.165
Discriminant Analysis	0.17

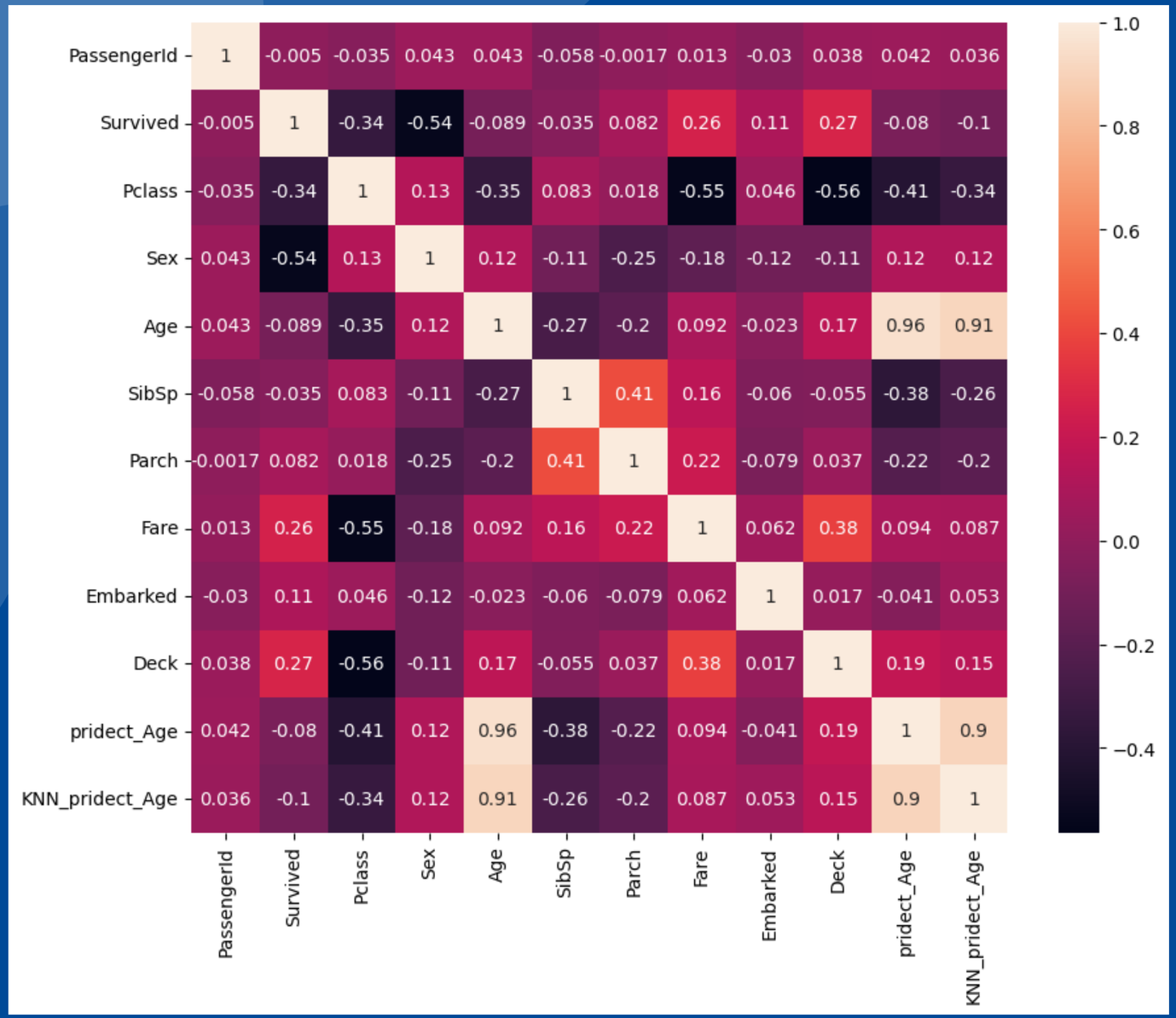
What is the best model?

The best model is Discriminant Analysis with prdecting Age with KNN (0.83)

What are the most important variables to predict the outcome?

Sex, Pclass, Fare, Deck

What are the most important variables to predict the outcome?



Correlation Matrix

Thank you for your attention!