Predictive Analytics: Titanic Passenger Survivability

Data Science Fundamentals: Learning Module Practical Quiz

Business Understanding

The Titanic, encountered a significant tragedy during its inaugural voyage in 1912.

You are a data analyst and are asked to delve into comprehensive passenger data, which includes details like names, ages, and ticket prices, to develop predictive models to assess survival probabilities.

Using historical insights with data analysis and predictive modeling to showcase the potential of data science.

Data Understanding

- Train.csv: use this as your input to build your model
  - The values in the second column (**"Survived"**) can be used to determine whether each passenger survived or not:
    - if it's a "1", the passenger survived.
    - if it's a "0", the passenger died.
- Test.csv: use this to test your model post-build
  - Note that **test.csv** does not have a **"Survived"** column.

Data Prep

Explore the data and provide descriptive statistics values.

Are all features relevant? If not, which ones were dropped

- I dropped Ticket, Name, and Cabin since they won't be needed to build the model.

Do any of the features need to be transformed to quantitative values? Which fields and what method was used?

- I used Embarked and Sex fields as Categorial, the rest I used as Numeric.

Are any fields nulls? How was this handled?

- Age: 177 null values (~ 20%) - replaced the null values with median

-Embarked: 2 null values (<1%) – replaced them with constant

Modeling

Did you create any new features?

Which feature is your predictor (target)? Which features are your input variables (i.e X)

- I have Pclass, Sex, Age, SibSp, Parch, Fare, and Embarked as my variables.

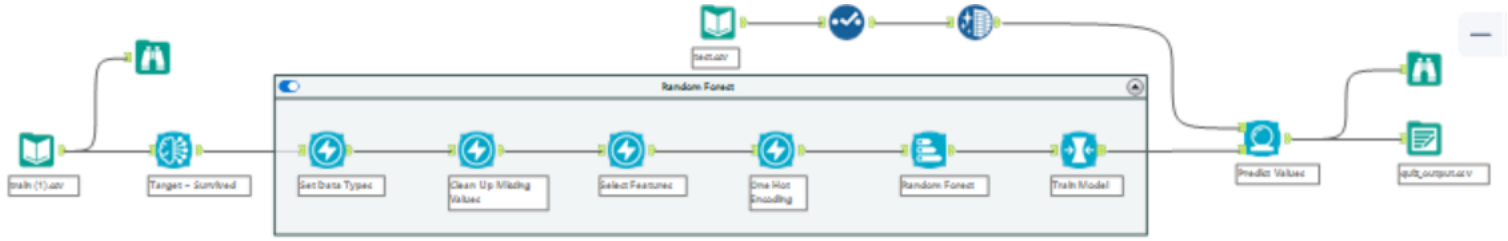
Did you use unsupervised or supervised learning? Did you try multiple algorithms?

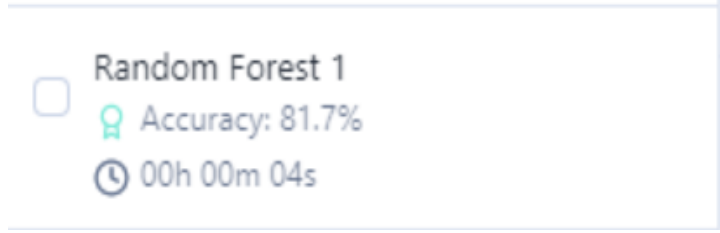
- I used supervised learning. I have tried decision tree, logistic regression, random forest, and XGBoost.

Accuracy and precision of your chosen model?

- From the models that I selected to run, I decided to choose Random Forest as it gave me the highest accuracy (~81,7 %)

Add a picture of your flow and a snapshot of your accuracy %





Evaluation

What new fields were added within the test data once your model was run?

- I have 3 new fields: Survived_predicted, Survived_0 and Survived_1

What do they mean?

- Survived_predicted: 0 means “Survived”, 1 means “Die”
- Survived_0 means – chance of not survival
- Survived_1 means – chance of survival

Are you able to evaluate your model? Are their constraints of missing data that limits evaluation?

- I am able to evaluate the model after a data cleansing process, to ensure the data type and handle the null values.

Insights (points won’t be deducted on this one)

What insights did you discover? Was there a pattern between who survives?

- Female has higher chance of surviving than male
- Passengers with higher fare tend to have higher chance of surviving