Titanic survival prediction model

Project overview

This data analysis project aims to predict the survival of passengers on the Titanic using machine learning techniques. The 1912 Titanic disaster is one of the most notorious maritime tragedies in history, resulting in over 1,500 fatalities. Based on the classic dataset provided by Kaggle, the project explores the key factors influencing passenger survival through systematic data analysis and modeling, and constructs a predictive model.

The project adopts CRISP-DM (cross-industry data mining standard process) methodology, which includes six complete stages: business understanding, data understanding, data preparation, modeling, evaluation and deployment. We pay special attention to feature engineering and model interpretation, not only to pursue prediction accuracy, but also to understand the story and rules behind the data.

Through this project, we have achieved the following goals:

A survival prediction model with an accuracy of about 83% was established

Identify key influencing factors such as gender and cabin class

The principle of "women and children first" in maritime rescue has been verified

The nonlinear effect of family size on survival was found

Provides the basic framework for model deployment.

data set

Using the Titanic data set provided by Kaggle, which includes:

Training set: information about 891 passengers

Test set: Information on 418 passengers

Field description

The data set contains the following 12 original features:

Passenger information:

PassengerId: Unique identifier of the passenger

Name: Passenger's name (including title)

Sex: (male/female)

Age: age (partially missing)

Ticket Information:

Pclass: Cabin class (1/2/3 class)
Ticket: a ticket office

Fare： the price of a ticket

Cabin: Cabin number (large number missing)
Embarked: Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)
Family member information:
SibSp: Number of siblings/spouses on the same ship
Parch: Number of parents/children on the same boat

target variable ：

Survived: Survival status (0= dead, 1= alive)
Data quality analysis
The original data has the following problems:
Missing values: age (about 20%), cabin number (about 77%), port of embarkation (0.2%), ticket price (a small number in the test)
Data distribution: ticket price is seriously right skewed, and age is approximately normal distribution
Exception: The fare for individual passengers is extremely high
Technology stack
Python 3.x
Main libraries:
Pandas/numpy: data processing
Scikit-learn: Machine learning model
Matplotlib/seaborn: Data visualization

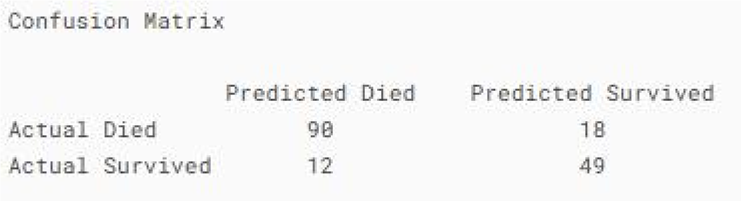
Random forest selection criteria
It can handle mixed type features
Insensitive to outliers
Provide feature importance measures
Suitable for medium sized data sets

Model evaluation analysis

precision ：

Validation set accuracy: 83.24%
Better than baseline model

confusion matrix ：



```
Confusion Matrix

                  Predicted Died    Predicted Survived
Actual Died             90                  18
Actual Survived         12                  49
```

Type I error (false positive): 18/108=16.7%

Type II error (false negative): 12/61=19.7%

Feature importance analysis

Gender (importance 0.28):

The strongest predictor confirms the principle of "women first"

The female survival rate was 74.2 percent, compared with 18.9 percent for men

Ticket price (0.16):

It is highly related to the class of accommodation

The survival rate was 62.6 percent in first class and 24.2 percent in third class

Age (0.12):

The survival rate of children (0-12 years) was 59.0%, showing a U-shaped relationship

Future improvement direction

Model level:

Try gradient boosting trees (XGBoost/LightGBM)

Add the Stacking integration method

feature engineering ：

Extract the pattern information from the check office

Create cabin deck features (from the first letter of Cabin)

comprehensibility ：

Introduce SHAP value analysis

Build LIME local interpretation

System expansion:

Develop the Streamlit interface

Deploy a RESTful API service

The complete code and data set of this project have been open source on

GitHub, welcome to contribute and improve!