**Predicting the Survival of Titanic Passengers**

|  |  |  |
| --- | --- | --- |
|  | **Yongyuan Qu**  yq2361@columbia.edu  **Xiaocong Xuan**  xx2438@columbia.edu  **Wenjun Yang**  wy2431@columbia.edu  **Yuhao Wang**  yw3924@columbia.edu  **Yitong Liu**  yl5549@columbia.edu |  |

1. **Introduction**

The Titanic disaster remains one of history's most poignant maritime tragedies, marked by the substantial loss of life it caused. This project seeks to address the problem of predicting the survival outcomes of passengers based on available demographic and socio-economic data. By analyzing the factors that influenced survival rates, this study aims to shed light on historical socio-economic disparities and inform the development of enhanced safety protocols for contemporary maritime transportation.

1. **Data Description**

The dataset used for this project is derived from the Titanic passenger list, sourced from Kaggle (*https://www.kaggle.com/competitions/titanic/data*). The data is split into two distinct sets:

* Training Set (train.csv): This dataset includes demographic and socioeconomic information for a subset of the Titanic's passengers. It features variables such as passenger class (Pclass), sex, age, number of siblings/spouses aboard (SibSp), number of parents/children aboard (Parch), ticket number, fare, cabin number, and the port of embarkation. Crucially, it also includes the survival status (Survived), where 1 indicates survival and 0 indicates non-survival. This set is used to build and train the machine learning models, providing the 'ground truth' needed for supervised learning.
* Test Set (test.csv): Similar to the training set in structure but without the survival outcome. The purpose of this dataset is to evaluate the performance of the trained models on unseen data. The test set allows for the assessment of how well the predictive model generalizes to new data.

These datasets provide a comprehensive base for developing and testing machine learning models aimed at predicting outcomes from historical events, in this case, the survival of passengers from the Titanic disaster.

1. **Methodology**

The methodology for predicting survival rates involves several machine-learning models and analytic techniques:

a. Data Preprocessing:

* Handling Missing Data: Missing values in 'Age', 'Cabin', and 'Embarked' will be imputed using statistical methods like median imputation for numerical data and mode imputation for categorical data. In addition, rows containing missing values in any of the critical features will be dropped from the dataset.
* Feature Engineering: Features such as 'Title' extracted from passenger names and a binary 'Alone' indicator based on 'SibSp' and 'Parch' will be created to enhance the model's predictive accuracy.

b. Machine Learning Models:

* Logistic Regression: This model will be used to establish a baseline for comparison. It's particularly useful for binary classification problems and will provide insights into the importance of different features in predicting survival.
* Genetic Algorithm: To optimize feature selection, a genetic algorithm will be employed, improving the efficiency and performance of the predictive models by selecting the most relevant features.
* Neural Network: A neural network model will be constructed using TensorFlow's Keras API. The network will feature layers suitable for binary classification tasks, with dropout layers to prevent overfitting and an Adam optimizer for efficient training.

c. Evaluation:

* Model Evaluation Metrics: Models will be evaluated using accuracy, precision, recall, F1-score, and ROC curves. These metrics will help in understanding the effectiveness of the models in classifying the survival outcomes.
* Cross-Validation: To ensure the models do not overfit, k-fold cross-validation will be employed during the training process.

This methodology outlines a comprehensive approach to applying machine learning techniques to historical data, aiming to provide insights into the factors that influenced survival during the Titanic disaster. By leveraging advanced analytics, this project not only aids in historical analysis but also enhances our understanding of data-driven decision-making in safety and crisis management.

1. **Discussion**

Our team initially identified the presence of NA values in the 'Age' and 'Fare' columns. In order to mitigate any potential consequences, the rows containing NA values beneath these columns are eliminated. We then generate a new Data Frame 'X' for clustering purposes, which contains only 'Age' and 'Fare'. Following this, the features are standardized and k-means clustering is executed (Figure 1). The scatter plot (Figure 2) depicting the relationship between passengers' age and fare was generated, as indicated by the graph, using the variables 'Age', 'Fare', and 'Cluster'. Once the data has been visualized, it can be prepared for genetic algorithm processing. To begin, pertinent attributes such as Pclass, Sex, Fare, and Survived are chosen and stored in a Data Frame named 'data'. After that, the 'Sex' column is mapped to numeric values (female = 1, male = 0), and drop rows are populated with NA values.

文本

描述已自动生成

**Figure 1: Feature Selection**

In order to maximize the feature selection procedure and enhance the performance of survival prediction models, we opted for the genetic algorithm. To accomplish this, we initially standardize features in preparation for optimization. The components of the genetic algorithm are subsequently delineated, encompassing toolbox, genetic, and fitness operations. Once these preparations are complete, the evaluation function and the primary function are defined. Observe the result after executing the main function as the final phase.

图表, 散点图

描述已自动生成

**Figure 2: Clustering Result**

The optimal individual vector is evidently [0,1,1,0]. This results in the second and third features being selected, while the initial and final features remain unselected. The 'age' and'sex' attributes are chosen. Furthermore, it was determined that the optimal individual fitness is 6.503512325622854e-17. The fact that the fitness value, which represents the mean accuracy of survival prediction, is relatively low suggests that the accuracy is exceptionally high, essentially fitting the data perfectly.

For neural network model, our tem employ EDA for visualization. Analyzed the distribution of key numerical variables using histograms and bar plots. Utilized count plots to observe the distribution of categorical variables such as 'Sex', 'Pclass', and 'Embarked' with respect to the survival rate (Figure 3). Correlation analysis (Figure 4) helped in understanding the relationship between different features.

The initial dataset contained several features, including passenger class, name, sex, age, number of siblings/spouses aboard, number of parents/children aboard, ticket number, fare, cabin number, and port of embarkation. We performed the following preprocessing steps: Identified and imputed missing values in 'Age', 'Cabin', and 'Embarked' columns using appropriate strategies.

图表, 条形图

描述已自动生成正方形

中度可信度描述已自动生成

**Figure 3: EDA Figure 4: Data Preprocssing**

Extracted titles from names, mapping five titles with numerical values, combined 'SibSp' and 'Parch' features then categorized as 'Single' with binary values to indicate if passengers aloneness. Transformed categorical variables into numerical ones using encoding techniques to make the data suitable for our machine learning algorithms. Constructed a neural network model using TensorFlow's Keras API. Initialized a linear stack of layers and added ‘Dense’ layer type. First layer was specified ‘input\_shape’ to be 7 which is exactly the number of features in the dataset. “activation='relu'” specifies the activation function for the neurons in that layer. ReLU (Rectified Linear Unit) is a common activation function that is used to introduce non-linearity in the model. “kernel\_initializer='he\_normal'” set the initial random weights of the layer according to the He normal initializer, which was based on the number of inputs to the unit. Standardized the inputs to a layer for each mini-batch by ‘BatchNormalization()’. This stabilized the learning process and dramatically reduced the number of training epochs required to train deep networks. Intentionally did not set all ‘Dense’ layer with initialization and bias use specified. This could help other people to fine tune the model with all of above-mentioned terms considered to be hyperparameters in the future. The final layer used “activation='sigmoid'” because this was a binary classification problem (predicting survival, which was either 1 or 0). Sigmoid activation functions output a value between 0 and 1, which was ideal for binary classification. ‘model.compile’ configured the model for training. ‘loss=tf.keras.losses.binary\_crossentropy’ set the loss function, which was appropriate for a binary classification problem. ‘optimizer=tf.keras.optimizers.Adam()’ was used as an optimization algorithm, which was a popular choice that combined the benefits of two other extensions (AdaGrad and RMSProp) of stochastic gradient descent. Fitted the model with batchsize 32 and epoch 50 times. At last, evaluated the performance of the model based on accuracy, precision, recall, and F1-score metrics (Figure 5).

Our model yielded the following results:

**手机屏幕的截图

描述已自动生成**

**Figure 5: Model Result**

The study illustrates the potential of machine learning in historical data analysis. The predictive model offers an interesting lens to understand the survival patterns and socio-economic factors aboard the Titanic. Future work may focus on deploying more sophisticated models and utilizing larger datasets for improved predictions.

1. **Conclusion**

In this comprehensive study, we utilized advanced machine learning techniques to predict the survival outcomes of passengers aboard the Titanic, leveraging a rich dataset that included demographic and socio-economic variables. Our multifaceted approach incorporated logistic regression, genetic algorithms, and neural networks to evaluate the influence of various factors on survival probabilities. The logistic regression model provided a solid baseline, highlighting the significance of attributes such as gender, class, and age—factors historically known to impact survival chances during the tragedy. To enhance the model's efficiency and predictive accuracy, we employed a genetic algorithm for optimal feature selection, which proved crucial in refining our predictions. Further sophistication was introduced through the development of a neural network model using TensorFlow's Keras API, which adeptly handled complex non-linear relationships between the variables, thus offering deeper insights into the survival patterns. Our findings not only demonstrate the power of machine learning in extracting meaningful insights from historical data but also suggest practical applications, such as improving safety protocols in modern maritime transportation and other areas requiring crisis management. By illustrating how socio-economic disparities influenced survival rates, this study also contributes to our understanding of historical social structures and their impacts, providing a window into the past that can inform future research. Future endeavors could focus on integrating more advanced machine learning models and expanding the datasets to refine these predictions further, potentially leading to broader applications in safety standards development and socio-economic research across various domains. This research underscores the potential of machine learning to provide valuable insights into historical events and improve decision-making in contemporary issues related to safety and crisis management.