A Project Report *on*

Enhancing Survival Prediction on the Titanic Dataset

GitHub Repository: https://github.com/roshan2429/CS584_Machine-Learning

by

Mr. Roshan Dattatray Hyalij (A20547441)

Under the guidance of

Prof. Oleksandr Narykov

Department of Computer Engineering

ILLINOIS INSTITUTE OF TECHNOLOHY

2023-2024

ILLINOIS TECH

Acknowledgements

We take this opportunity to express our deepest sense of gratitude and sincere thanks to those who have helped us in completing this task. We express our sincere thanks to our guide Prof. Oleksandr Narykov, who has given us valuable suggestions, excellent guidance, continuous encouragement and taken interest in the completion of this work. His kind help and constant inspiration will always help us in our future also.

ILLINOIS TECH

Contents

Chapter 1

1.1 Introduction

The Titanic dataset is a renowned collection of data that depicts the individual passenger profiles on board the RMS Titanic during its fatal first voyage in 1912. Its historical relevance has resonance in the field of predictive modelling, providing a special chance to investigate the variables that affected survival rates during the disaster.

This dataset contains a wide range of information, including demographics, socioeconomic statuses, cabin assignments, and boarding sites. These parameters are important indicators for comprehending the complex relationship between passenger features and survival probability.

1.2 Objective

The principal goal of this research project is to improve survival prediction accuracy by utilizing a variety of sophisticated machine learning techniques. The goal is to extract useful information from the Titanic dataset by utilizing a range of methods, from ensemble modelling to data preparation, thus improving the survival result prediction accuracy.

The emphasis is on using these approaches to extract latent patterns and dependencies from the dataset and to construct strong predictive models. The goal of this investigation is to gain priceless knowledge about the parameters that determine survival and illuminate the critical elements that shaped the outcomes of passengers during this historic incident.

Chapter 2

2.1 Scope

The objective of this study is to increase the accuracy of survival predictions by utilizing machine learning techniques on the Titanic dataset. Standard computing hardware and related software tools are needed, such as an integrated development environment (IDE) and Python 3.x with the required libraries. The aim is to investigate diverse techniques for improving predictive models and gaining insights into factors that impact survival rates.

2.2 Hardware and Software Requirements

1) Hardware Requirements:

Processor: Recommended Intel Core i5 or AMD Ryzen 5 (or equivalent) for faster computation.
Memory (RAM): Minimum 8GB RAM, preferably 16GB for larger datasets and smoother performance.
Storage: Adequate storage space (at least 256GB SSD recommended) for dataset storage and model training.
GPU (Optional): A dedicated GPU (NVIDIA GeForce or AMD Radeon) with CUDA support can significantly expedite model training for complex algorithms.

2) Software Requirements:

Operating System: Compatible with Windows, macOS, or Linux distributions.
Python Environment: Python 3.x installed, preferably managed via Anaconda for streamlined package management.
Integrated Development Environment (IDE): Any preferred IDE (e.g., Jupyter Notebook, PyCharm, VSCode) for coding and experimentation.
Libraries: Essential Python libraries such as Pandas, NumPy, Scikit-learn, Matplotlib, and Seaborn for data manipulation, analysis, and visualization.
Report Generation: Tools like Microsoft Word, LaTeX, or Google Docs for report writing.

Chapter 3

1) Data Exploration and Preprocessing:

Types of Data and Overview: The Titanic dataset includes both category (like "Sex," "Embarked," and "Age") and numeric (like "Fare," and so on. The mean, median, and standard deviation are examples of summary statistics that provide information about the major patterns and variances within numerical data.

Managing Missing data: It was essential to identify any missing data. Imputation methods that replaced 'Age' with a mean or median preserved the integrity of the data. To maintain data quality, columns with a high percentage of missing values (such "Cabin") were removed.

Categorical Variables: 'Sex' and 'Embarked' were examples of the categories that were converted into numerical representation utilizing techniques such as one-hot encoding. The smooth incorporation of categorical data into machine learning models was made possible by this translation.

2) Model training and cross validation:

RandomForestClassifier Training:
Five-fold cross-validation was used to train the RandomForestClassifier model on the dataset. The model performed fairly well in predicting survival outcomes, as evidenced by the average accuracy across folds of about 82%.
Cross-Validation Metrics: Individual fold accuracies ranged from 79% to 85%, and the model showed consistent performance across folds. This variation in accuracy between folds sheds light on the model's stability and generalizability.

3) Validation and Model Evaluation:

Data Splitting and Training:
Upon splitting the dataset into training (80%) and validation (20%) sets, the RandomForestClassifier achieved an accuracy of 85% on the training set, indicating a good capture of underlying patterns within the data.
Classification Report Insights:
The classification report detailed precision, recall, accuracy metrics, and F1-score. The model exhibited a precision of 80%, recall of 85%, and an F1-score of 82%, showcasing a balanced performance in predicting both survival and non-survival cases.

4) Feature Importance and Model Refinement:

Feature Importance Analysis: 'Age', 'Fare', and 'Sex' were found to be the three most influential characteristics in the RandomForestClassifier feature importance analysis. The two most important factors influencing the likelihood of survival are "age" and "sex."

Retraining with Selected Features: The accuracy increased by 3% to 83% after the model was retrained using these particular features. This improvement showed how important feature selection is for improving model performance.

.

5) Regularization and Alternative Models:

Regularization with Logistic Regression: The accuracy of the model was kept at about 82% by applying regularization approaches to Logistic Regression, which produced only slight alterations. Regularization, however, significantly reduced worries about overfitting.

Testing Different Models: Support Vector Machines (SVM) and Gradient Boosting Classifier, two alternative models, demonstrated accuracies of 80% and 77%, respectively. Despite performing somewhat worse than the RandomForestClassifier, these models offered a variety of strategies and trade-offs in terms of accuracy.

6) Feature Engineering and Model Comparison Experiment:

Creation of the 'FamilySize' Feature: The addition of this feature had a big influence on model comparison. With a 2% accuracy gain over Logistic Regression, RandomForestClassifier highlights the significance of designed features in predictive modelling.

7) Hyperparameter Tuning and Ensemble Methods Experiment:

GridSearchCV for Hyperparameter Tuning:
The RandomForestClassifier's ideal hyperparameters were found via GridSearchCV optimization, which improved the model's robustness and accuracy by 2% when compared to the default settings.

Formation and Execution of Ensembles:
Combining RandomForestClassifier with Logistic Regression, the Voting Classifier ensemble achieved an accuracy of 83%, which was a slight improvement above the separate models.

Limitations:

Data Limitations: It was difficult to do a more thorough study due to the incompleteness of the dataset, particularly in categories like "Cabin."

Further Investigations: In order to increase forecast accuracy even more, future research may concentrate on investigating new data sources or utilizing more advanced methods.

Chapter 5

Conclusion:

In summary, careful examination and analysis of the Titanic dataset, together with feature engineering and strategic modelling approaches, provided insightful information about survival prediction. The project shows the potential for data-driven insights in decision-making processes and advances our understanding of the complexities of predictive modelling. It also identifies areas that warrant further research.