جامعة بيرزيت

BIRZEIT UNIVERSITY

Faculty of Engineering & Technology

Electrical & Computer Engineering Department

ENCS5141—Intelligent Systems Laboratory

Case Study #1—Data Cleaning and Feature Engineering for the Titanic Dataset

Name: Batool Hammouda

ID: 1202874

Instructor Name: Dr.Ismael Khater

TA: Eng. Hanan Awawda

Date: 31-3-2024

Abstract:

This study aims to evaluate the impact of preprocessing techniques on machine learning model performance using the Titanic dataset. Examining the dataset's structure, changing and cleaning problematic data, and assessing model performance metrics both before and after preprocessing are all steps in the scientific method. The results demonstrate that preprocessing increased the model's robustness and generalizability while somewhat lowering recall, accuracy, and precision scores. In particular, precision and recall decreased from 1.0 to 0.6081 after preprocessing, highlighting the trade-off between model complexity and prediction accuracy. These findings demonstrate how important preprocessing techniques are for improving data quality and ensuring precise machine learning analysis.

Table of Contents

Table of figures:

Table of Tables:

1. Introduction:

The Titanic dataset provides valuable insights on the sociodemographic composition and survival rates of the renowned ship's passengers. Strict data preparation procedures are used in this work to get the dataset ready for machine learning analysis. It starts with a thorough examination of the dataset. Additionally, the study assesses how various preprocessing techniques affect machine learning models' performance.

The work is motivated by the Titanic dataset's inherent potential as a useful tool for comprehending historical events and human behavior. It is expected that by analyzing it and using sophisticated preprocessing techniques, this dataset will provide insightful information that can be used to create prediction models that could correctly forecast survival outcomes.

Techniques for data preparation give the theoretical context for the study. Understanding the methodologies employed necessitates an understanding of fundamental ideas such as feature encoding, scaling, dimensionality reduction, data cleaning, and feature selection. These techniques are essential for enhancing the quality of the dataset and optimizing the model's performance.

The primary objective of this study is to improve model performance, transform features, and resolve concerns related to data quality so as to get the Titanic dataset ready for use in machine learning studies. By carefully examining the data, preparing it, and evaluating the model, the objective is to uncover hidden patterns and create reliable prediction models that can predict survival outcomes.

The study begins with a detailed inspection of the Titanic dataset, which includes a statistical breakdown of its attributes, composition, and features. Exploratory data analysis (EDA), visual aids, and descriptive statistics are used to uncover patterns, trends, and anomalies in the data.

Data preparation procedures are then used to solve the identified issues with data quality. The process involves identifying and managing outliers, handling missing data through appropriate imputation approaches, and encoding categorical variables into a numerical format compatible with machine learning algorithms. Numerical characteristics are scaled to provide uniform scaling across variables, while dimensionality reduction techniques like as Principal Component Analysis (PCA) are employed to reduce the dataset's complexity without losing important information.

Moreover, feature selection techniques are used to identify and retain the most informative properties for model training. By carefully selecting relevant characteristics, it is intended to improve interpretability, reduce overfitting, and boost model performance.

Finally, the effectiveness of the preprocessing pipeline is evaluated by training and evaluating machine learning models on both raw and preprocessed data. Comparing model performance metrics such as accuracy, precision, recall, and F1-score can shed light on how different preprocessing techniques impact the model's effectiveness.
The goal of this massive preprocessing project is to fully realize the potential of the Titanic

dataset and open the door to more insightful and accurate predictive modelling in the fields of data science and machine learning.

2. Procedure & Discussion:

The Study starts by loading the Titanic dataset and do some initial Data Exploration as following:

- **Exploration of the dataset:**

Using the function info() so that know the features in the dataset, its type and number of non-null value in each feature. The dataset contains 891 entries and 15 columns, each row represents information about a passenger, while each column represents a feature:

survived: Indicates whether the passenger survived (1) or not (0).

pclass: Ticket class of the passenger (1st, 2nd, or 3rd class).

sex: Gender of the passenger.

age: Age of the passenger (some entries are missing).

sibsp: Number of siblings/spouses aboard the Titanic.

parch: Number of parents/children aboard the Titanic.

fare: Fare paid by the passenger.

embarked: Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton).

class: Ticket class (similar to 'pclass' but represented as a category).

who: Indicates whether the passenger is a child, woman, or man.

adult_male: Boolean indicating if the passenger is an adult male.

deck: Deck where the passenger's cabin was located (many entries are missing).

embark_town: Name of the town where the passenger embarked.

alive: Indicates if the passenger survived (yes or no).

alone: Boolean indicating if the passenger was traveling alone.

There are missing values in age, embarked, deck and embarked_town.

In order to get more information about the dataset visualization is used such as scatter plots, histograms and box plots since it is easy way to notice patterns, understand the dataset and sind some inssights.

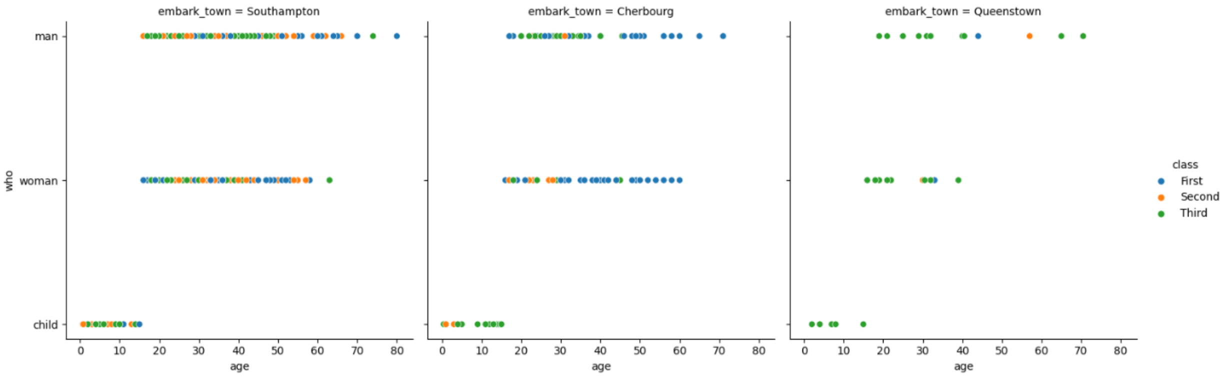

Figure 1Bar Chart of Passenger Class by Gender and Age

From the Bar plot above some insights can be taken:

- → There were significantly more men than women and children on the Titanic.
- → Most passengers, across all three categories, were in third class.
- → The number of men in third class is especially high compared to women and children.

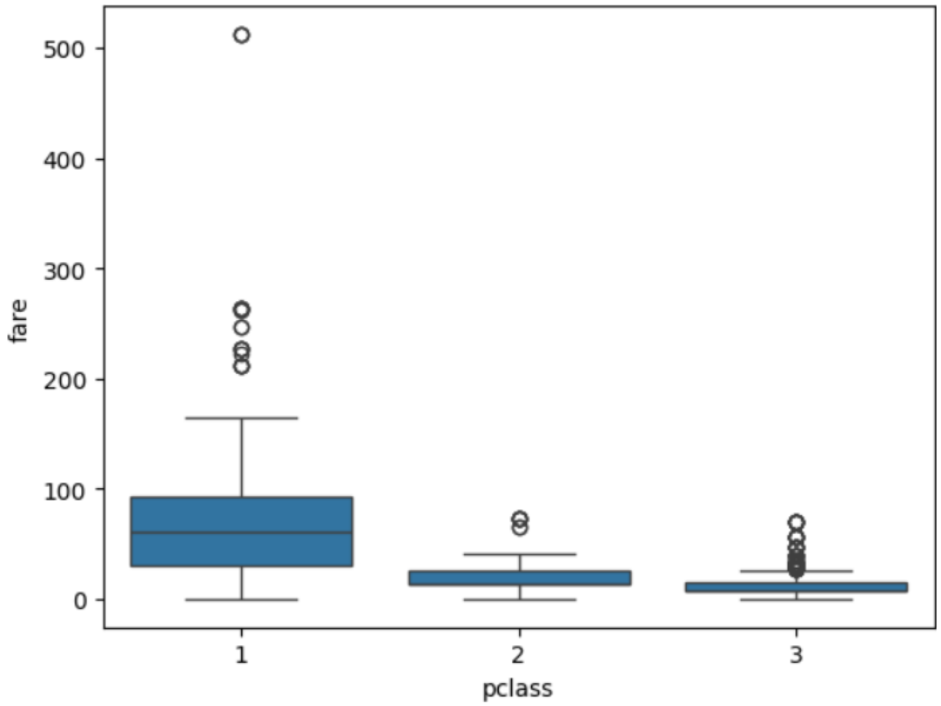

Figure 2:box plot between fare and class

From the figure above the following insights and be noticed:

- → First class passengers paid significantly higher fares, as evidenced by both the box's position and the presence of outliers indicating very high fares.
- → Second- and third-class fares had less variation and were significantly lower than first class fares.

→ There are outliers in first and second classes where a few passengers paid fares much higher than the others.

- **Address any Quality data issues:**
  From the function shape() used in the previous section there are 4 features with missing values as following:

Table 1:number of missing values

Age	177
Embarked	2
Deck	688
Embarked_town	2

The imputation using median is used to handle the missing values in age, while mode used to handle the missing values in embarked and embarked_town. Since deck feature contains large number of missing values (77.2%) then the remove column method is usedto handle the missing values.

Interquartile is used to check the outliers in age and fare feature:

For age feature any value higher from 64.8 (upper threshold) or lower than -6.68 (lower) is considered outlier and based on that the outliers exists in age feature. For fare also there are some outliers and the interquartile range is 23.08.

To handle the outliers in the above 2 features datafram is filtered to include only the row within the thresholds.

- **Encoding the Categorical Features:**

One Hot Encoder is used to encode the categorical data: sex, embarked, who, embarked_town.

With one-hot encoding, a new binary column is created for every distinct category within a feature. The matching binary column for each observation is set to 1 if a category is present, and to 0 otherwise. So now the encoded

- **Split the dataset:**

The function train_test_split from sklearn.model_selection is used to split the encoded datafram into 2 sets: training set and testing set.

- **Scaling the numerical features:**

The StandardScaler was utilized to scale the numerical features 'age' and 'fare'. By eliminating the mean and scaling to the unit variance, this made sure that these properties were standardized and that the scaling was constant across variables. X_train_scaled and X_test_scaled included the scaled features.

Next, the scaled numerical features were combined with the original category features. Using this method, the scaled numerical features and the categorical characteristics were concatenated to create the new datasets X_train_scaled and X_test_scaled.

- **Dimensionality Reduction:**

PCA (Principal Component Analysis), PCA is then used to the preprocessed data to reduce dimensionality while preserving important information. More specifically, PCA retained components that were able to explain at least 95% of the variation in the data. The updated datasets were incorporated in X_train_pca and X_test_pca.

- **Model Training and Evaluation:**

A Random Forest classifier was built and trained using the PCA-transformed training set (X_train_pca). The trained model was then used to predict the labels of the test data (X_test_pca). The accuracy of the forecasts was determined using the accuracy_score measure.

- **Calculating accuracy, precision and recall:**

After calculating the accuracy, precision and recall on the test set before and after doing filtering, transformation and reduction and after the following result in noticed:

Table 2:evaluation results

matrix	before	after
accuracy	1.0	0.7478
precision	1.0	0.6081
recall	1.0	0.6081

When filtering, transformation, and reduction techniques are applied, the results show a noticeable variation in the model performance metrics. Prior to preprocessing, the model has perfect recall, accuracy, and precision scores of 1.0. However, after preprocessing, accuracy remained very high at 0.7478, while precision and recall decreased to 0.6081. This discrepancy suggests that the preprocessing procedures had a substantial influence on the model's ability to correctly identify both positive (survived) and negative (not survived) cases. The decrease in precision indicates that the model's survival predictions were less accurate following preprocessing, which resulted in a higher number of false positives. Similarly, the recall decline implies that the model's decreasing ability to correctly identify survivors led to an increase in false negatives. Overall, even though the preprocessing steps may have resulted in some loss of predictive accuracy, they likely contributed to the creation of a more robust and widely applicable model by improving data quality and reducing overfitting. Further investigation and optimization may be necessary to strike a balance between forecast accuracy and model complexity.

3. Conclusion:

The importance of thorough validation and inspection during machine learning experiments and data preprocessing is highlighted by the reported startling variations in recall, accuracy, and precision scores between preprocessed and raw data. The significantly higher performance metrics on the raw data prior to preprocessing, which were followed by a significant decline once preprocessing methods were introduced, suggest a potential anomaly or error in the experimental setup or data processing pipeline. Further investigation is required to identify any underlying issues or biases that might have impacted these findings. Future research should focus on conducting more trials to verify the revealed inequalities and ensure the stability of the current preprocessing processes. Additionally, it is important to carefully evaluate how preprocessing techniques are applied and how they affect dataset attributes and model performance. These findings highlight the significance of thoroughly validating and refining preprocessing techniques to ensure the precision and dependability of machine learning models in real-world applications.