**Student Name – Sachin Shrinivas Kadam**

**Assignment Name – Data Cleaning Techniques**

* Remove non-essential columns from the dataset.
  + Identify the non-essential columns from the modelling point of view. Justify your decisions.
  + Justify the selection or removal of features based on their impact on model performance.

My Analysis & Response as below

When building a predictive model for the dataset (train.csv), we aim to select the most relevant features for predicting survival. Some columns may not contribute significantly to the model’s predictive power and could add unnecessary complexity or noise. Below, I will identify the non-essential columns and provide justifications for their removal or retention, along with the reasoning based on their potential impact on model performance.

* **Understand the train.csv dataset**
  + **PassengerId**: Unique identifier for each passenger (no missing values).
  + **Pclass**: Class of the passenger (no missing values).
  + **Name**: Name of the passenger (no missing values).
  + **Sex**: Gender of the passenger (no missing values).
  + **Age**: Age of the passenger (may have missing values).
  + **SibSp**: Number of siblings or spouses aboard the Titanic (no missing values).
  + **Parch**: Number of parents or children aboard the Titanic (no missing values).
  + **Ticket**: Ticket number (no missing values).
  + **Fare**: Fare for the ticket (no missing values).
  + **Cabin**: Cabin number (may have missing values).
  + **Embarked**: Port of embarkation (may have missing values).
* **Non-Essentials Columns for Model – Justification, and Impact on the Model**

|  |  |  |  |
| --- | --- | --- | --- |
| **Column** | **Justification of Removal** | **Alternative** | **Impact on Model** |
| Cabin | The Cabin column has a significant number of missing values, making it difficult to use without heavy preprocessing. Even if you do extract the deck or section information, it may still not be a very strong predictor for survival when compared to other features. | You can process the Cabin feature to extract the deck (first letter of the cabin number), but even then, it may not provide much predictive power | Some models may benefit from data imputation or feature extraction from this column, removing it will likely simplify the model without sacrificing performance significantly. If processed correctly, it could be useful, but for most purposes, it’s not essential. |
| Ticket | Ticket numbers are unlikely to contain any meaningful patterns that could help in predicting survival. They are unique to each passenger and don't provide generalizable insights | NA | This feature can be safely removed as it adds unnecessary complexity and does not improve the model's performance |
| PassengerId | This column contains a unique identifier for each passenger. While it may be useful for referencing individual passengers, it does not provide any predictive value regarding survival, as it is simply a number that does not correlate with the target variable. |  | Removing this feature will not negatively affect model performance. Including it may confuse the model, as the unique identifier doesn't correlate with survival outcomes |
| Name | The passenger's name, in its raw form, does not offer direct predictive power. However, the title in the name (e.g., "Mr.", "Mrs.") could be useful in some cases because it might correlate with the passenger's gender or social status | Extract titles (e.g., "Mr.", "Mrs.", "Dr.") from the name and use them as a separate feature. | Removing the full name reduces the feature space and avoids introducing irrelevant text data. Extracting titles may provide useful information about survival, such as social status or gender. |

* **Essential Columns to Retain for Model – Justification, and Impact on the Model**

|  |  |  |
| --- | --- | --- |
| **Column** | **Justification to Retain** | **Impact on Model** |
| Sex | Gender played a significant role in survival on the Titanic. Women had a higher survival rate than men, making this a key feature in predicting survival. | This feature is highly valuable and should be retained as it provides a direct signal regarding the likelihood of survival. |
| Pclass | The class of the passenger (1st, 2nd, 3rd) is strongly correlated with survival. Higher-class passengers (1st class) had a higher chance of survival, making this an important predictor. | This feature is essential for survival prediction and should be retained. It provides a strong signal about the likelihood of survival based on social status. |
| Age | Age can have an important impact on survival chances. Children and elderly passengers were more likely to survive | This feature provides crucial information and should be retained. However, you may need to handle missing values (e.g., imputation) to make full use of this feature |
| SibSp | The number of siblings or spouses aboard could affect survival. Passengers traveling with family members may have had a different survival rate compared to those traveling alone. | This feature could be important and should be retained. You may also consider combining it with **Parch** (number of parents/children aboard) to form a family size feature. |
| Parch | The number of parents or children aboard could be important for survival predictions. Family groups may have had different survival rates. | Like **SibSp**, this feature should be retained, and it may be combined with **SibSp** to create a family size feature |
| Fare | The fare paid for the ticket is often correlated with the passenger's class and socio-economic status, which can influence survival chances | This feature provides relevant information regarding the passenger's social status and should be retained |
| Embarked | The port of embarkation (Cherbourg, Queenstown, Southampton) could have an impact on survival due to factors such as passenger demographics or the order in which lifeboats were filled | This feature should be retained as it might provide additional context for predicting survival. |

* **List of Features to Remove and Retain**

|  |  |
| --- | --- |
| **Features to Remove** | **Features to Retain** |
| PassengerId (identifier, non-predictive) | **Survived** (target variable) |
| Name (raw name does not contribute; titles could be extracted) | **Pclass** (passenger class) |
| Ticket (ticket number does not provide predictive value) | **Sex** (gender) |
| Ticket (ticket number does not provide predictive value) | **Age** (age of the passenger) |
|  | **SibSp** (siblings or spouses aboard) |
|  | **Fare** (fare paid) |
|  | **Embarked** (port of embarkation) |
|  | **Parch** (parents or children aboard) |

* **Impact of Feature Selection on Model Performance**
  + **Reducing Dimensionality**: Removing non-essential features like **PassengerId**, **Name**, **Ticket**, and **Cabin** reduces the dimensionality of the dataset, making it easier for the model to focus on important features. This can improve model training times and reduce the risk of overfitting.
  + **Handling Missing Data**: Features like **Age** and **Cabin** often have missing values. Removing features like **Cabin** reduces the need for complex imputation strategies, which can introduce noise if not handled carefully.
  + **Enhancing Performance**: Reducing irrelevant or redundant features often leads to better model performance. Including non-predictive features like **PassengerId** or **Ticket** would only add noise and could degrade model performance, especially for simpler algorithms like logistic regression or decision trees.
  + **Improving Interpretability**: By focusing on features like **Pclass**, **Sex**, **Age**, and **Fare**, the model becomes more interpretable. These features have clear relationships with survival, and keeping them allows for easier explanation of model results.

Conclusion – By Selecting and removing features based on their relevance and impact on model performance is crucial for building an effective predictive model. Removing non-essential columns and retaining those that carry useful signals will improve both the efficiency and accuracy of the model.