Titanic Survival Prediction Report

This report explores the steps involved in preparing, processing, and building models using the Titanic dataset to predict passenger survival.

Data Preprocessing

Before diving into modeling, the dataset needed to be cleaned and processed:

1. **Handling Missing Values**:
- Missing values in the Age column were filled with the median age.
- Missing Embarked values were replaced with the most common embarkation point.
- Missing Fare values in the test set were filled with the median fare.

2. **Categorical Conversion**:
- The Sex column was converted into numerical values (0 for male, 1 for female).
- The Embarked column was transformed into numerical values (0 for Southampton, 1 for Cherbourg, and 2 for Queenstown).

3. **Dropping Irrelevant Columns**:
- Columns such as Name, Ticket, and Cabin were dropped since they are not directly useful for the prediction task.

Simple Rule-Based Model

To start with, a basic rule-based model was created using simple heuristics:

- Females are more likely to survive.
- Passengers in first class are more likely to survive.
- Younger passengers are more likely to survive.

This simple rule-based model achieved an accuracy of **78.8%** on the training dataset, which is quite reasonable for such a basic approach.

Weighted Average Model

Next, I explored a weighted average model where weights were assigned to different features based on domain knowledge and intuition. Initial weights included:

- Pclass: -0.5 (negative weight since higher class passengers had a better chance of survival),
- Sex: 1.0 (being female increased survival chances),
- Age: -0.05 (younger passengers had better chances),
- Fare: 0.003 (higher fare indicated a better chance of survival).

The weighted average model achieved an accuracy of **61.6%** on the training set, which was lower than expected. To improve this, weights were manually adjusted and tested iteratively.

Model Refinement and Feature Engineering

After tuning weights and exploring new features such as family size (SibSp + Parch) and fare per person (Fare/SibSp + Parch), the model's accuracy reached **62.4%**, slightly better than earlier versions.

However, reverting to the original set of features and further adjusting weights yielded better results, improving the model accuracy to **66.6%**. Continued fine-tuning produced a maximum accuracy of **71.5%** on the training dataset.

Final Model and Best Weights

The final model used the following weights:

```
weights_best = {
   'Pclass': -2.6,
   'Sex': 3.8,
   'Age': -0.1,
   'SibSp': -0.05,
   'Parch': -0.05,
   'Fare': 0.11,
   'Embarked': 0.1
}
```

These weights emphasized the importance of Pclass, Sex, and Fare, which are consistent with historical data about survival on the Titanic.

Historical Insights: “Women and Children First”

A well-known historical policy during the Titanic disaster was the “Women and Children First” rule. This policy prioritized women and children for evacuation, significantly influencing survival rates.

To implement this in the model:

1. Females were predicted to survive.
2. Males younger than 14 years old were also predicted to survive.
3. Everyone else was predicted not to survive.

This rule-based model achieved an impressive accuracy of **79.2%** on the training dataset, outperforming some of the more complex models.

Testing and Evaluation

While the model was trained and validated on the training dataset, the test.csv file does not contain the actual Survived values, which means we cannot calculate the accuracy of our model on the test data. This is a common scenario in competitions like Kaggle, where the actual survival labels are hidden, and predictions are submitted to the platform for evaluation.

As a result, while we can confidently evaluate the model’s performance on the training data (with an accuracy of **71.5%** for the weighted average model and **79.2%** for the “Women and Children First” rule-based model), the performance on the test dataset remains unknown without submitting predictions for evaluation.

Conclusion

The Titanic dataset provides valuable insights into feature selection and model building for survival prediction.

In conclusion, the Titanic survival prediction problem offers a fascinating opportunity to explore machine learning techniques on historical data. Through iterative model development, feature engineering, and weight tuning, the model achieved a solid accuracy of **71.5%**, while simpler models, such as the “Women and Children First” rule, performed even better.

The most significant takeaway from this exercise is the importance of simple, interpretable models in some cases, as demonstrated by the success of the “Women and Children First” rule.