Kaggle: Titanic

Executive summary report for Survival Rate Predictiom

ISSUE / PROBLEM

Based on Titanic data, **fewer than 40% survived**.
Explore and compare **various machine learning models** and find one with **the best performance to predict Titanic survivors** based on passenger data.

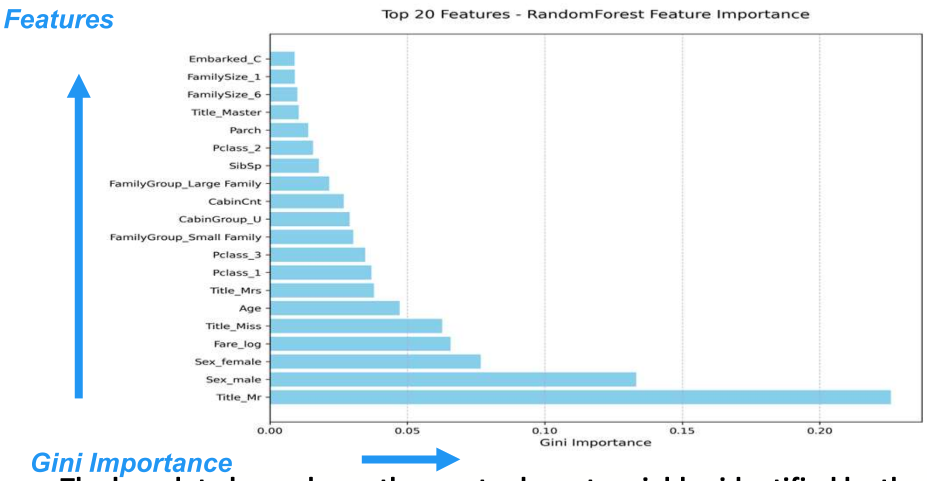
RESPONSE

Since the variable we are seeking to predict is categorical, the team could build either a logistic regression or a tree-based machine learning models or a ensemble models.

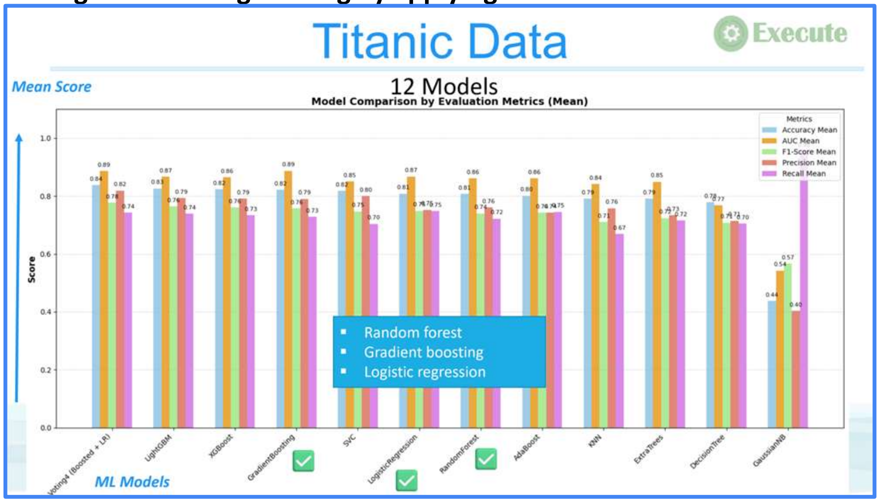
The models perform 82.3% the random forest and get 79.9% Kaggle score.

IMPACT

This model helps predict whether people will survive in Titanic. The result among top 5% (May 2025) people who participated in the competition.



The bar plot above shows the most relevant variables identified by the Random Forest model. This serves as the baseline, using features such as 'male', 'female', etc., which we later extended to 19 features through feature engineering by applying custom rules



- Random forest
- Gradient boosting
- Logistic regression

In our analysis, both Gradient Boosting and Logistic Regression models, along with Random Forest, achieved Kaggle scores exceeding 78%.

INSIGHTS/NEXT STEPS

- Continue refining the Random Forest model to achieve ≥ 80% accuracy
- Expand model testing beyond random forest to include:
  - Gradient boosting: capture complex non-linear relationships
  - Logistic regression: offers interpretable results and serves as a baseline model
- Enhanced feature engineering, e.g., using Fare_group or incorporating feature importance from other models (e.g., SHAP values).
- Implement a Voting Classifier that combines Random Forest, Gradient Boosting, and Logistic Regression for improved overall performance