

INFORMATICS INSTITUTE OF TECHNOLOGY

In Collaboration with

UNIVERSITY OF WESTMINSTER, UK



SoftesT: A Novel Approach to Software Effort Estimation by Integrating Ensemble Models and Explainable AI

A dissertation by

Ms. Sujana Tharmalingam

W1866994 / 20210070

Supervised by

Ms. Abishani Muthalagan

April 2025

Submitted in partial fulfilment of the requirements for the BEng (Hons) Software Engineering degree at the University of Westminster

ABSTRACT

Problem: Software effort estimation is a crucial task that makes a project successful by completing it in estimated time and budget. Traditional method often struggles because of the agile project nature and lack accuracy making it difficult for the project managers to estimate the accurate time needed to complete the project as a result project can get delayed and exceed budget. This project aims to tackle these challenges by building a more accurate and interpretable model that combines advanced machine learning techniques with explainability driven feature selection method.

Methodology: The proposed solution involves a novel approach using china dataset. Initially, data preprocessing is performed to handle skewness, outliers and scale features. We then used SHAP (Explainable AI) values, derived from initial models, to select the most impactful features. Our core model is a heterogeneous ensemble, combining Linear Regression, Decision Tree, Random Forest, XGBoost, and CatBoost. Also experimentally developed and integrated a custom neural network SHAP-Guided attention mechanism. Optuna is used to fine-tune the hyperparameters for the neural network component.

Results: Testing on the China dataset resulted in excellent performance for the final ensemble model (combining the base learners trained on SHAP-selected features), achieving an R^2 of approximately 0.99 and a Mean Magnitude of Relative Error (MMRE) around 1.14%. This indicates high accuracy and reliability on the test data. However, the evaluation also revealed that the experimental SHAP-guided attention neural network component performed poorly (negative R^2), likely due to data limitations and complexity challenges, and it was therefore excluded from the final successful ensemble prediction.

Subject Descriptors: Computing methodologies -> Machine learning -> Ensemble methods -> Boosting algorithms;

Explainable AI -> Feature Importance -> SHAP -> Software Effort Estimation

-> Natural Language Processing-> Feature Selection ->Attention Mechanism