ENSF 611

Machine Learning for Software Engineers

Project Proposal

Author:

Steven Duong
(30022492)

Affiliation
Department of Electrical and Software Engineering
University of Calgary
Calgary, Alberta

Lab Block: B01
Date of Report: Mar 24, 2023

1. Why: Question/Topic being investigated

I chose to investigate the Titanic dataset on Kaggle because it offers a real-world problem, well-structured data, an active community, and a wide range of machine learning techniques to explore. This will provide an engaging learning experience and the opportunity to apply machine learning concepts to a historical event with practical implications.

2. How: Plan of attack

In accordance with the established guidelines, I will prepare a Jupyter notebook and submit it for evaluation. By employing Lab 3 (classification & grid search) as a foundational template, I will ensure the seamless integration of essential data preprocessing steps, while incorporating additional techniques tailored specifically to the Titanic dataset.

The following steps will be taken to complete the project:

1. Data acquisition: Download the Titanic dataset, including the training and test sets, from the Kaggle competition page.
2. Data preprocessing & feature engineering: Implement preprocessing methods such as handling missing values through imputation strategies, encoding categorical variables, and normalizing or standardizing numerical attributes.
3. Model selection and training: Utilize Scikit-learn library to experiment with various machine learning algorithms, such as Logistic Regression, Random Forest, Gradient Boosting, and Support Vector Machines. Train the models on the preprocessed training dataset.
4. Model evaluation and hyperparameter tuning: Evaluate the performance of the selected models using appropriate metrics and cross-validation techniques. Fine-tune the hyperparameters of the models to achieve optimal performance.
5. Prediction and submission: Generate predictions on the test dataset using the best-performing model and submit the results to the Kaggle competition leaderboard.

3. What: Dataset, models, framework, components

a. Dataset: The Titanic dataset is available at the Kaggle competition url: https://www.kaggle.com/competitions/titanic/data

b. Models: I will use Scikit-learn classifiers, including Logistic Regression, Random Forest, Gradient Boosting, and Support Vector Machines.

c. Framework: Scikit-learn will be the main machine learning framework for developing my models.