CS5011: P2 - Machine Learning

190018035

March 13th 2024

Contents

1 Introduction

For this practical, I was tasked with implementing and evaluating different machine learning models on the Pump It Up: Data Mining the Water Table [1] dataset. This involved experimenting with different preprocessing steps, hyperparameters, and machine learning models to find the best model to predict the status of water pumps in Tanzania. The checklist below details my achievements in the practical for each of the specified parts.

1.1 Project Achievements

- Part 1: Attempted and Fully Working
- Part 2: Attempted and Fully Working

1.2 Usage Instructions

To run the script, navigate to the root directory and run the following command:

```
python3 part1.py <train-input-file> <train-labels-file> <test-input-file> <
    numerical-preprocessing> <categorical-preprocessing> <model-type> <test-
    prediction-output-file>
```

The values for each of the arguments are equivalent to that described in the assignment specification. However, one difference is that the `<numerical-preprocessing>` and `<categorical-preprocessing>` arguments now take in a `Manual` option, which allows features to be preprocessed based on the encoding that are most suitable for the model.

2 Part 1

In this section, I will detail the various preprocessing steps I took to prepare the data for the machine learning models, guided by a deep exploration of the dataset. I will also discuss the design decisions I made while implementing this part, including the parameters of the preprocessing methods and the hyper-parameters of the machine learning models.

2.1 Preprocessing Steps

Removing Irrelevant Features

Removing Single-Value Features

Removing Redundant Features

Replacing Construction Year with Decades

Imputing Missing Values

Fixing Formatting Issues

Limiting High Cardinality Features

Converting Datetime Features

2.2 Design Decisions

2.3 Evaluation

3 Part 2

For this part of the practical, I utilized Optuna [2], an automated hyper-parameter optimization (HPO) tool to find the best hyper-parameters for each of the models. In this section, I will detail my HPO process and analyze the results of the experiments.

3.1 Hyper-Parameter Optimization

3.2 Evaluation

4 Conclusion

References

[1] URL: https://www.drivendata.org/competitions/7/pump-it-up-data-mining-the-water-table/page/23/.

[2] Takuya Akiba et al. "Optuna: A Next-generation Hyperparameter Optimization Framework". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2019.