National University of Computer and Emerging Sciences

Task 1

Explore Other Testing Needs

Group Members

Name	Roll Number	Contact
Hafiz Abdullah	19L-2288	l192288@lhr.nu.edu.pk
Muhammad Arslan Ashraf	19L-2303	l192303@lhr.nu.edu.pk
Muhammad Usman Arshad	19L-2356	l192356@lhr.nu.edu.pk

CS4036 Software Testing
Spring 2023
Department of Computer Science
FAST-NU, Lahore, Pakistan

Selected Topic: Testing and Test Automation in AI and ML

Introduction

The utilization of machine learning (ML) models is gaining significant popularity across numerous domains such as healthcare, finance, biology, and more. With the availability of advanced software packages, training complex models has become convenient, resulting in exceptional predictive accuracies. However, these modern software packages have led to a higher level of abstraction, making it increasingly challenging to test and automate the evaluation of AI-based algorithms and ML models.

But, there is a powerful tool called deepchecks that can be incredibly helpful in addressing these challenges. It is a comprehensive tool for testing, validating, and monitoring machine learning models and data at every stage of the model's lifecycle. With this tool, users can easily detect issues related to data quality, distributions, and the model's performance without expending significant effort. Furthermore, deepchecks is the only package that supports checks for both tabular and image datasets.

Deepchecks Library

Deepchecks is a package that provides a framework of data, models, checks, suites, and conditions for configurable and extendable ML testing. It proposes a uniform architecture for testing models and data, together with an application programming interface (API). It also allows the user to concatenate numerous checks into a single test command to be executed later. Furthermore, The Deepchecks library includes several built-in checks and suites that may be used to validate various points along the machine learning development process.

Integrations

Deepchecks can easily be integrated with many additional tools such as TensorFlow, scikit learn, Spark & Databricks, Pytest, H2O, HuggingFace Transformers, Airflow, CML, JUnit, wandb, and more.

Deepchecks' Suites

Deepchecks includes predefined suites for the following scenarios:

- Data Validation: Deepchecks plays a vital role in ensuring the integrity of the data. It helps detect issues such as duplicate samples, problems with string or categorical features, significant outliers, and inconsistent labels.
- Train-Test Validation: Before training the model, Deepchecks verifies that the data splits are balanced, such as ensuring balanced classes, checking for

significant distribution drift between features or labels in each dataset, and identifying potential data leakage that could impact the model's integrity.

- Model Evaluation: With a trained model, Deepchecks assesses its performance by examining various performance metrics and comparing them to benchmarks. It provides a clear understanding of the model's performance and identifies areas where it may underperform, offering valuable insights to enhance its effectiveness.

Checks

A check is designed to examine a particular element of the data or model. These checks can address various common issues like data leakage and concept drift. Currently, the Deepchecks library comprises 62 checks. Among them, 42 checks are applicable to classification and regression models trained on tabular data, while the remaining checks are tailored for computer vision models trained on image data.
Each check in Deepchecks can produce two types of results. Firstly, there is a visual result that is intended for display purposes. This can provide a clear representation of the check's outcome. Secondly, there is a return value that can be utilized to validate and verify the expected results of the check. This allows users to programmatically assess and confirm the correctness of the check's outcomes.

Condition

In Deepchecks, a condition refers to a function that can be included in a Check. This function evaluates the return value of the Check and provides a pass, fail, or warning result. Its purpose is to validate and assess the Check's outcome. Many checks in Deepchecks already have pre-defined conditions that can be added to the check. Additionally, these conditions often allow users to adjust their parameters to customize their behavior and tuning.

Conclusion

In conclusion, deepchecks emerges as a remarkable tool for thoroughly assessing the performance of AI and ML models. Its implementation empowers data scientists and ML engineers to validate the accuracy, robustness, and fairness of their models. Moreover, the automation of testing tasks facilitated by deepchecks not only saves valuable time but also mitigates the potential risks associated with human errors. As a result, deepchecks contributes significantly to enhancing the reliability and efficiency of AI and ML model evaluation processes.