

AUTOML FOR IMPROVEMENTS IN MACHINE LEARNING SYSTEMS

AUTOMATED DATA PREPROCESSING

TEAM MEMBERS: AMIRTHA KRISHNAN T, SACHIN S K, YEKANTHAVASAN A
20UCS011 20UCS133 20UCS175

TEAM MENTOR: KUMARAKRISHNAN S, ASSOCIATE PROFESSOR

INTRODUCTION

1. AutoML (Automated Machine Learning)

- Automation of machine learning processes for model development, reducing the need for manual intervention.
- Aims to make machine learning accessible by automating tasks like feature engineering, hyperparameter tuning, and model selection.

2. ML (Machine Learning)

- Subfield of artificial intelligence focusing on algorithms that learn patterns from data.
- Involves creating models capable of making predictions or decisions without explicit programming, widely used across various applications.

3. Amelioration

- Refers to the act of improvement or enhancement.
- In the context of machine learning, it signifies refining processes or enhancing system performance to achieve overall optimization.

MODULES

1. Data Preprocessing Module

- Responsible for cleaning and preparing raw data for analysis.
- Includes tasks such as handling missing values, scaling features, and encoding categorical variables.

2. AutoML Core Module

- The central module orchestrating the entire AutoML process.
- Integrates sub-modules for hyperparameter tuning, feature engineering, and model selection.

3. Categorical Variable Encoding Standardization Module

- Ensures consistent encoding of categorical variables for regression and classification tasks.
- May employ methods such as one-hot encoding, label encoding, or target encoding.

4. User Interface (UI) Module

- Provides a user-friendly interface for practitioners to interact with the AutoML system.
- Allows users to input data, set parameters, and visualize results.

5. Report Generation Module

- Facilitates the deployment of AutoML-generated models into production.
- Include options for model versioning, scalability, and monitoring.

FEATURES

1. Imbalanced Data Management

Automatically balances the distribution of different classes in your data. This ensures that the model doesn't favor one class over others, leading to better overall performance.

2. Hyperparameter Tuning

Efficiently explores a wide range of settings for your model. This optimization process helps find the best combination of hyperparameters, enhancing the model's performance.

3. Target Encoding

Takes care of encoding categorical variables for regression and classification tasks. This is important for properly handling non-numerical data in your machine learning model.

4. Feature Selection

Automatically identifies and selects relevant features, improving the accuracy of your model while preventing it from learning noise and overfitting.

FEATURES

5. Time-Series Handling

Streamlines the preparation of time-series data, extracts useful features, and selects appropriate models for forecasting tasks. This ensures accurate predictions in scenarios where data evolves over time.

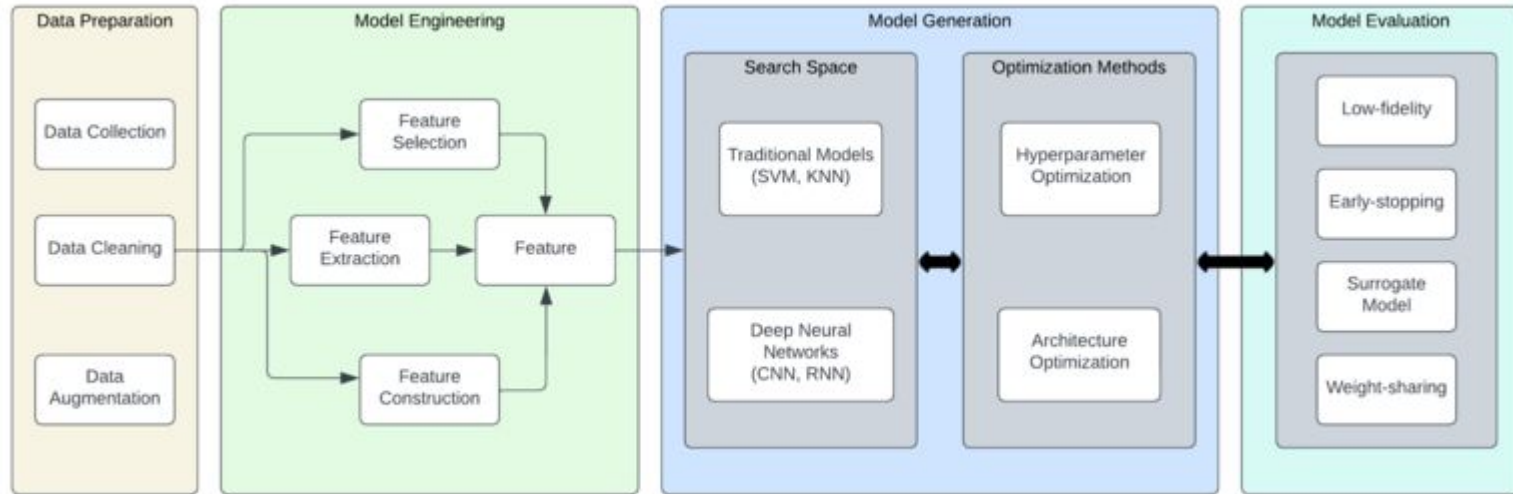
6. Feature Engineering

Automates the process of creating useful features from your data. This not only improves model accuracy but also reduces the need for manual feature engineering, saving time and effort.

7. Automated Functionality

Enables people without extensive expertise to easily build and deploy machine learning models. This means you can use the system even if you're not a machine learning expert, reducing the reliance on manual coding.

ARCHITECTURE DIAGRAM





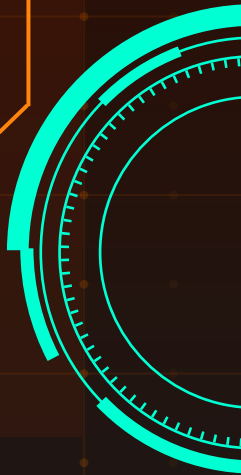
LITERATURE SURVEY:

S.No.	Title of the Paper	Author/s	Year	Advantages
1	DataAssist: A Machine Learning Approach to Data Cleaning and Preparation	Kartikay Goyle, Quin Xie, V. Goyle	2023	Improved Data Quality and Reliability (DataAssist)
2	Investigating Data Repair steps for EHR Big Data	Suraj Juddoo	2022	Efficient Data Repair (Investigating Data Repair)
3	Benchmarking AutoML algorithms on a collection of synthetic classification problems	Pedro Ribeiro, P. Orzechowski, Joost B. Wagenaar, J. H. Moore	2022	Versatile Benchmarking (Benchmarking AutoML algorithms)
4	REIN: A Comprehensive Benchmark Framework for Data Cleaning Methods in ML Pipelines	Mohamed Abdelaal, Christian Hammacher, Harald Schoening	2023	Comprehensive Benchmarking Framework (REIN)
5	Data Cleaning and AutoML: Would an Optimizer Choose to Clean?	Felix Neutatz, Binger Chen, Yazan Alkhatib, Jingwen Ye, Ziawasch Abedjan	2022	Improved model performance
6	AutoCure: Automated Tabular Data Curation Technique for ML Pipelines	Mohamed Abdelaal, Rashmi Koparde, Harald Schoening	2023	Automated Data Curation (AutoCure)
7	Detecting errors in databases with bidirectional recurrent neural networks	Severin Holzer,Kurt Stockinger	2022	Error Detection with Neural Networks (Detecting errors in databases)
8	DiffPrep: Differentiable Data Preprocessing Pipeline Search for Learning over Tabular Data	Peng Li, Zhiyi Chen, Xu Chu, Kexin Rong	2023	Differentiable Data Preprocessing (DiffPrep)
9	DataVinci: Learning Syntactic and Semantic String Repairs	Mukul Singh, José Cambronero, Sumit Gulwani, Vu Le, Carina Negreanu, Gust Verbruggen	2023	Semantic String Repairs (DataVinci)
10	Automated Data Cleaning Can Hurt Fairness in Machine Learning-based Decision Making	Shubha Guha; Falaah Arif Khan; Julia Stoyanovich; Sebastian Schelter	2023	Fairness in Data Cleaning

PROBLEM IDENTIFIED



- In modern machine learning, the complex interplay of data preprocessing, hyperparameter tuning, and feature engineering poses a significant challenge, burdening practitioners with labor-intensive tasks.
 - This system autonomously tackles issues like class imbalances, hyperparameter optimization, categorical variable encoding standardization, and revolutionizes feature selection, streamlining the model development process.
 - Through rigorous comparative analyses, our research underscores the transformative potential of this AutoML system, offering not only a solution to current challenges but a visionary path toward democratizing and empowering both ML experts and novices.
- 
- 

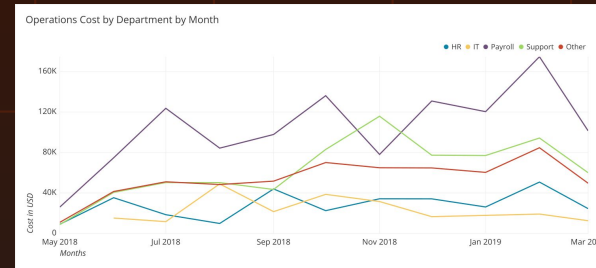
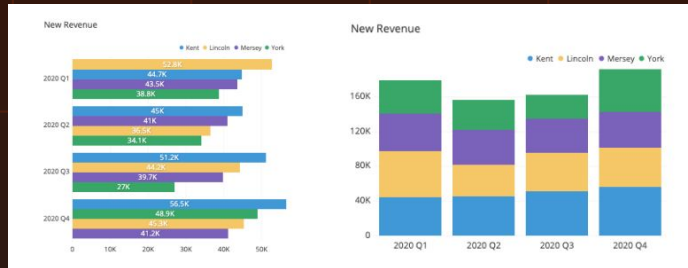


PROBLEM IDENTIFICATION

- Imbalanced class distribution
- Suboptimal Hyperparameter tuning
- Manual Categorical Variable Encoding
- Feature Selection Challenges
- Time-Series Preprocessing Complexity
- Need for Advanced Features Engineering
- Accessibility Barriers for Non-Experts

DATA VISUALIZATION

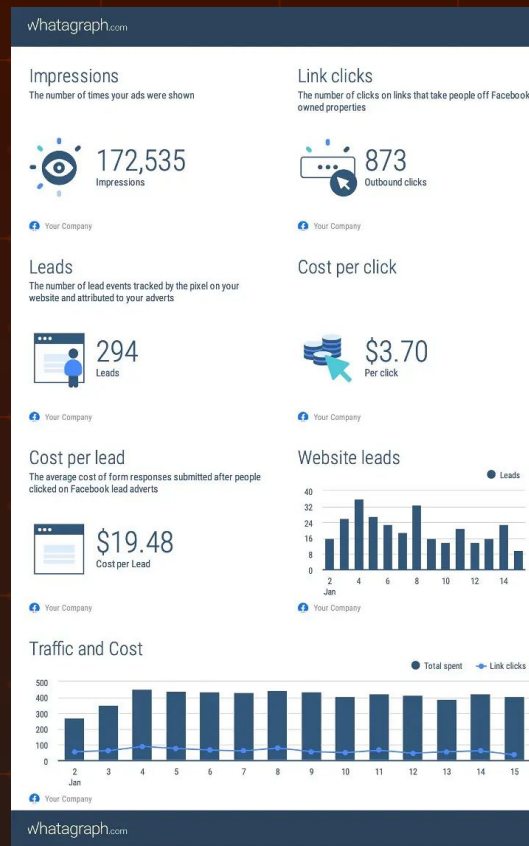
It will allow users to interactively explore and understand their datasets. Implement this using popular JavaScript libraries like D3.js and Plotly to create informative charts and graphs. Users can customize visualizations, and gain insights into their data's distribution and patterns. This feature helps users make informed decisions during the model-building process.



REPORT GENERATION

Once a model is trained and tested, users can request detailed reports summarizing key performance metrics, feature importance, and insights. We'll utilize Python libraries like Pandas and ReportLab to generate PDF or HTML reports.

These reports will include visualizations, model evaluation results, and suggestions for further optimization. Users can easily share these reports with stakeholders for informed decision-making.



REFERENCES:

- [1] K. Goyle, Q. Xie, & V. Goyle, "DataAssist: A Machine Learning Approach to Data Cleaning and Preparation," eprint arXiv:2307.07119, 2023.
- [2] S. Juddoo, "Investigating Data Repair steps for EHR Big Data," in International Conference on Next Generation Computing Applications, 2022.
- [3] P. Ribeiro, P. Orzechowski, J. B. Wagenaar, & J. H. Moore, "Benchmarking AutoML algorithms on a collection of synthetic classification problems," eprint arXiv:2212.02704, 2022.
- [4] M. Abdelaal, C. Hammacher, & H. Schoening, "REIN: A Comprehensive Benchmark Framework for Data Cleaning Methods in ML Pipelines," eprint arXiv:2302.04702, 2023.

THANK YOU