Data Science Agent – AI@SCALE

Short Description

A virtual assistant with Autogen for data science tasks, from data preprocessing to model training and hyperparameter tuning.

Long Description

Problem: Many fundamental data science tasks, such as predicting sales, prices, machine failures, and fraud, often involve relatively simple machine learning models. The potential for automation in these areas is substantial, allowing for the creation of AI-driven solutions that can expedite the development process. ETL (Extract, Transform, Load) pipelines can also be automated using generic templates. Countless hours are typically spent on writing prototype code for even the most basic tasks can be saved. Also, companies need to build a team of ML engineers, analysts, and data scientists before they can start making data-driven decisions which greatly hinders their progress in their initial years.

Solution: The assistant automates labour-intensive data preprocessing and modelling, making it accessible to a broader audience. Its unique features save time, ensure consistency, and elevate the quality of data analysis and decision-making processes. It can be used as an assistant by people having laymen knowledge of machine learning or data analysis to solve end-to-end ML tasks like classification and regression.

Unique Features:

1. **Complex dataset and tasks:** It can handle complex datasets with multiple columns with various datatypes for tasks such as regression and classification and can solve these tasks in an end-to-end manner.
2. **Data preprocessing:** It takes charge of essential ETL processes, encompassing data loading, missing value handling, redundancy removal, and preprocessing, including scaling, one-hot encoding, and converting categorical data to numerical formats.
3. **Feature Engineering:** It also has the capacity to engineer new features through diverse transformations.
4. **Model training:** The assistant can train multiple models and identify the optimal one on its own. Following model training, it engages in hyperparameter tuning to enhance performance.

Code Evaluation:

Challenge	Model	Score	Score after Hyperparameter tuning	Rank (percentile)
House Price Prediction	Random Forests	0.144	0.148	
	XGBoost	0.142	0.135	
	Gradient Boosting Regressor	0.133	0.130	670 (85.4)
Titanic		-	0.787	1815 (88.3)

Market Scope –

The market scope for our product spans various industries and use cases, here's an overview of the most important ones-

1. **SMEs and Startups**: They generally do not have the resources to employ a large-scale data team, thus our product would be most useful to them making them our primary targets.
2. **Enterprises**: Over **200,000** companies worldwide that employ data scientists. 'Large enterprises and corporations are increasingly adopting machine learning and AI for various purposes, such as customer analytics, predictive maintenance, fraud detection, and more. Our product can cater to their needs for efficient and effective model development.
3. **Data Science and Machine Learning Professionals**: Data scientists, machine learning engineers, and AI researchers are the primary target audience. These professionals across various industries rely on robust tools to streamline their workflow and improve model performance.

Revenue Streams –

The cost to run such a platform at scale will comprise of two major costs-

1. Bill on tokens consumed to write code.
2. Bill on GPU or CPU hours to develop a model.

Our business model will be two-fold-

1. Pay-Per-Use: Charge users based on their usage of the AI assistant's services. You can set pricing based on the number of models trained, datasets processed, or the complexity of hyperparameter tuning tasks.
2. Enterprise Solutions: Target large organizations and offer tailored enterprise solutions that include on-premises or cloud-based deployments of your AI assistant. Enterprises may be willing to pay for customized versions with enhanced security and scalability features.

Prospects -

Develop further to read DB schemas and write PySpark code that works on distributed databases.

Integrate with cloud services for 1 click deployment or CI/CD

Improve individual modules of assistant and make It write better code.