Trustworthy AI Lab x GES UCLA Hackathon

Enhancing CTR Predictions through a Data Clean Room

Team Name: **BitBuilders**

Team Leader: Arnav Sonavane

Members:
Manas Sewatkar
Aarian Thakur
Harshal More
Devdatta Talele

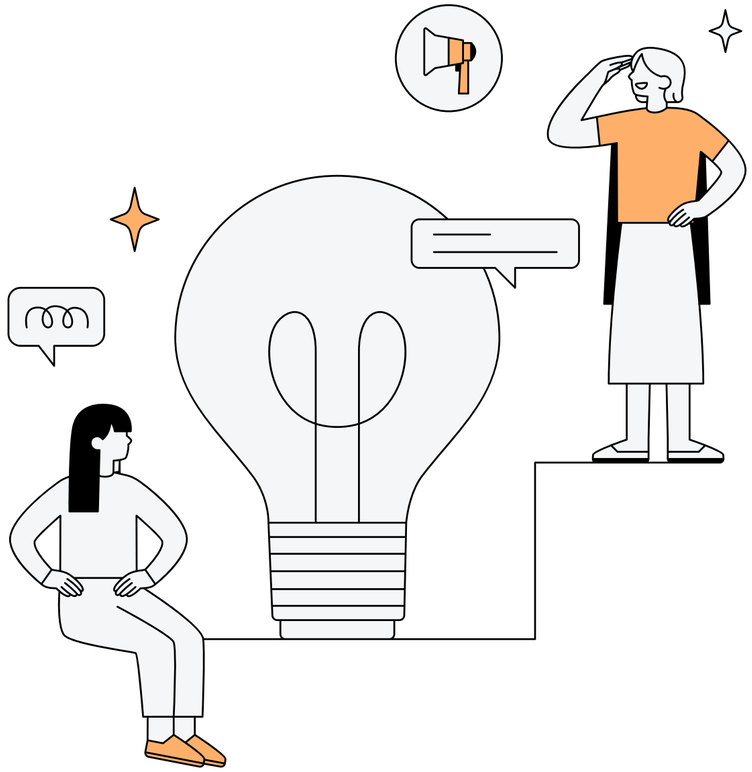
Overview

Highlighting the main theme of this event of utilizing Generative AI to empower “Data Collaboration Intelligence”. We developed a data sharing platform that allows private data sharing, predictive analytics and model building among different data parties, in essence, the “Data Clean Room” to enhance Click-through Rate (CTR) predictions using privacy-preserving synthetic data.

In both of the Parts, i.e., Implementation and Evaluation, we used various technologies including Azure Confidential Cloud VM, imbibing TDVM was well as TEE (Trusted Execution Environments),specifically Intel® TDX and SGX, GAN neural network, Gen AI, and Machine Learning Models. Similarly, several concepts like Hashing, Trusted Computing, Synthetic Data Fidelity and Synthetic Data Utility, Verification of Quote and Return Key.

Mission

Developing a secure confidential platform for secure data sharing and improved CTR prediction through collaboration. By building a secure environment and anonymizing data, teams can explore the potential of this approach without compromising user privacy. The results will provide valuable insights into the feasibility and effectiveness of privacy-preserving data collaboration for advertising optimization.



Challenges

01.

Create synthetic datasets that accurately reflect the real-world distributions of user behavior data.

02.

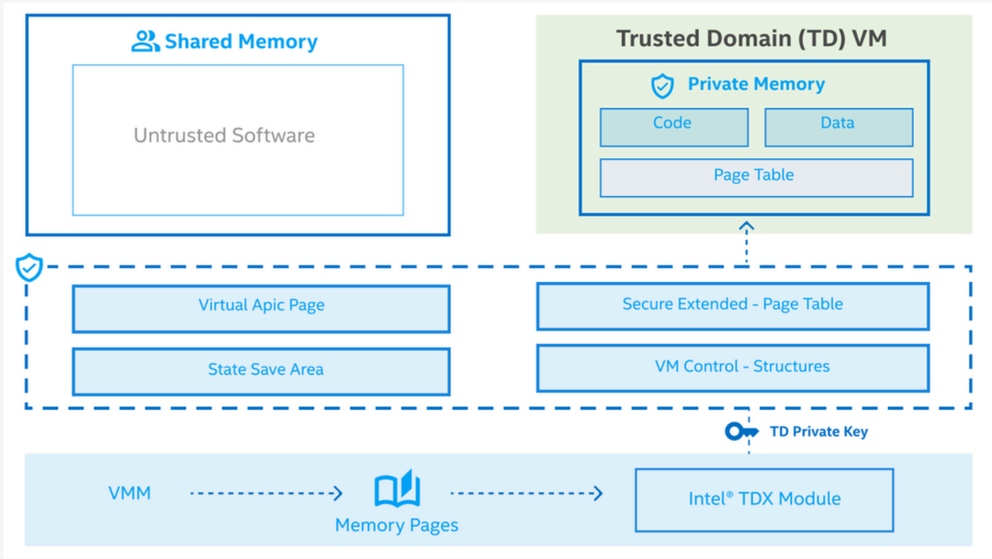
Publishers: Enhance predictive accuracy for user engagement with news content by integrating user interaction patterns from the advertisement dataset. This aims at better content personalization and placement strategies.

03.

Advertisers: Refine prediction models for advertisements by using synthesized data from news content interactions to understand the most effective contexts for ad placements.

Data Clean Room

Data Clean Rooms provide a secure environment for collaborative data analysis without revealing private user information. They allow multiple parties to combine data for better insights while maintaining privacy. By anonymizing data within controlled spaces, Data Clean Rooms reduce privacy risks and ensure compliance with regulations. This approach enables richer data analysis and more accurate models while respecting user trust.



Data Security and IP Protection

Protect apps and data from attack, tampering, or theft.

Privacy and Compliance

Strengthen data confidentiality and regulatory compliance.

Data Sovereignty and Control

Prohibit access by cloud providers or other tenants. Add safeguards to data sovereignty and governance.

Confidential AI

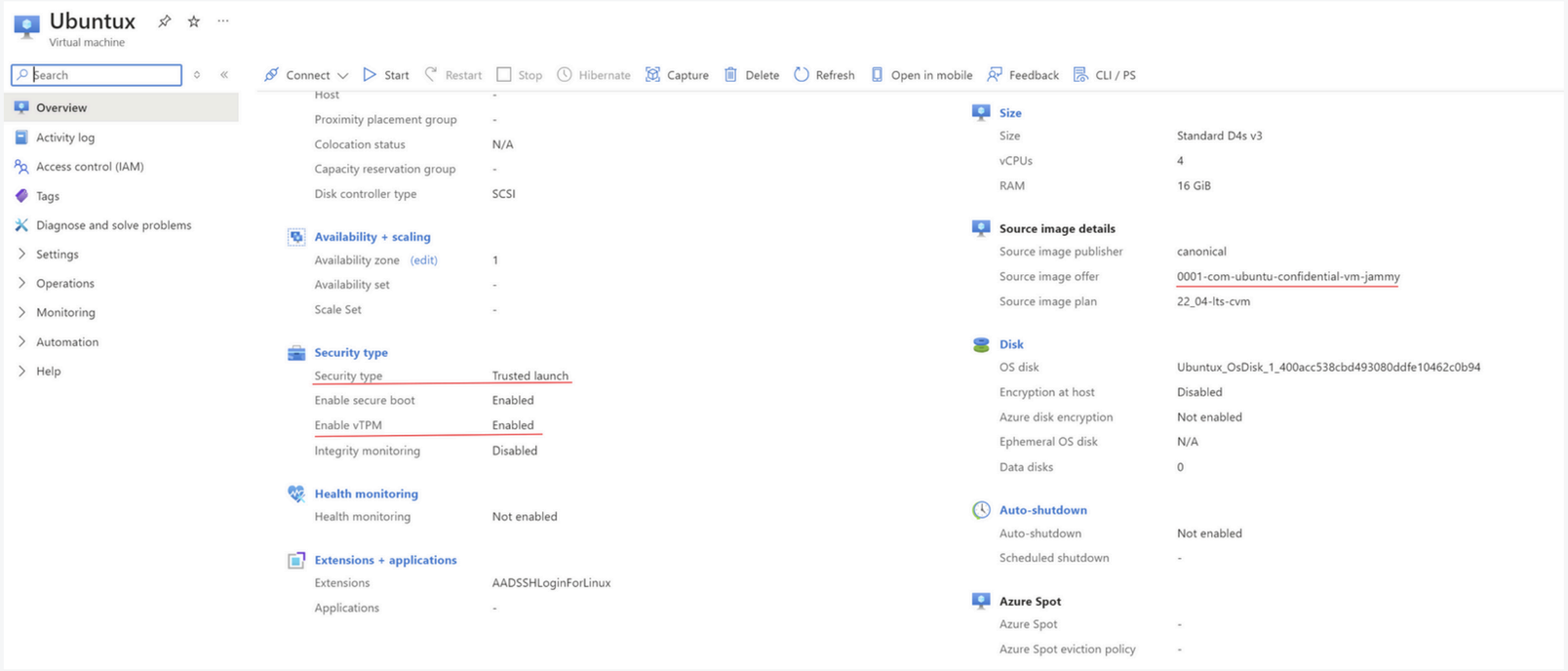
Safeguard your AI data and models by providing robust isolation, integrity, and confidentiality.

Confidential computing leverages Trusted Execution Environments (TEEs) like Intel® TDX and SGX. These enclaves create secure spaces within processors, protecting data confidentiality and integrity even if the system is compromised. Think of it as having locked vaults inside your computer for processing sensitive information. Confidential computing is ideal for Virtual Machines (VMs), enabling secure data processing on shared hardware.

Data Clean Room

Implementation of Task 1 & Task 2

Our secure Data Clean Room (DCR) uses Microsoft Azure Confidential VMs (Cloud VM) for enhanced data protection. These C VMs utilize Trusted Execution Environment (TEE) technology, specifically Intel® TDX and SGX , to create a secure enclave for processing data and cryptographically isolate and protect your data confidentiality and integrity.



Data Clean Room

These C VM’s have virtual Trusted Platform Modules (vTPM) built-in, and also combining the use tpm2-tools we can setup enables Remote Attestation (which creates the keys),required for verifying that the environment is trustable.

For additional security, our programs are containerized using Docker. Containerization isolates applications from the underlying system and other applications, minimizing potential vulnerabilities.

While this approach focuses on data-in-use security, we acknowledge the importance of protecting data at rest and in transit. To achieve this, we plan to implement Azure storage encryption for data at rest and secure transfer protocols like FTPS for data transfers

```
harshal@Ubuntux:~$ ls
SEAL     ak.priv  data_clean_room  pcr_values  pcrs.txt     quote.msg  snap
ak.ctx   ak.pub   home             pcrs.bin    primary.ctx  quote.sig
harshal@Ubuntux:~$
```

Dockerfile

data_clean_room > Dockerfile > ...

You, 2 days ago | 1 author (You)

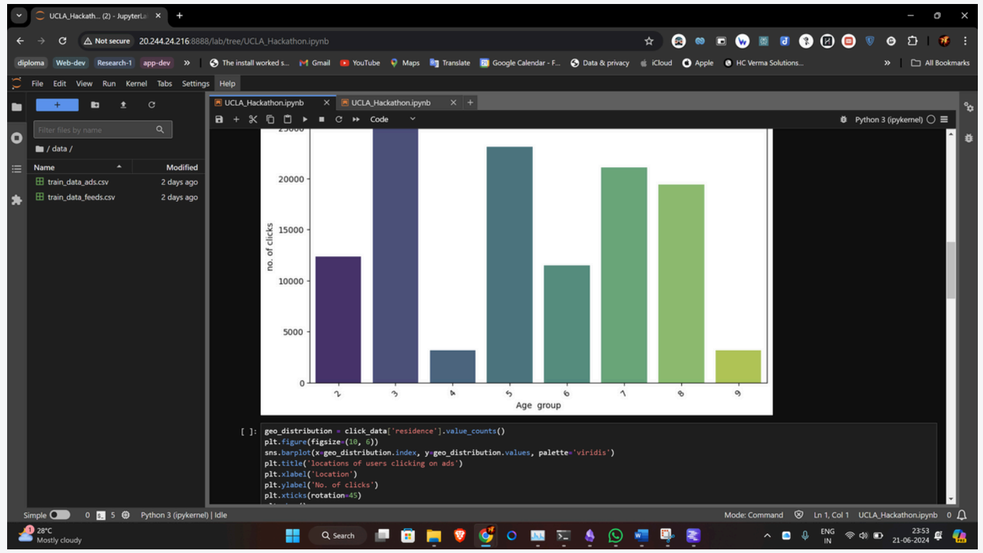
```
FROM ubuntu:22.04

# Set arguments and environment variables
ARG DEBIAN_FRONTEND=noninteractive
ARG PCCS_VERSION="1.21.100.3-jammy1"

# Install essential tools and dependencies
RUN apt-get update && apt-get install -y \
    wget \
    gnupg \
    ca-certificates \
    curl \
    vim \
    git \
    python3.9 \
    python3-pip \
    tpm2-tools \
    && rm -rf /var/lib/apt/lists/*

# Install Python packages
RUN pip3 install pandas numpy jupyterlab

# Configure JupyterLab
RUN jupyter lab --generate-config && \
    echo "c.NotebookApp.ip = '0.0.0.0'" >> /root/.jupyter/jupyter_notebook_config.py && \
    echo "c.NotebookApp.open_browser = False" >> /root/.jupyter/jupyter_notebook_config.py && \
    echo "c.NotebookApp.allow_root = True" >> /root/.jupyter/jupyter_notebook_config.py

# Download and setup Intel SGX dependencies
RUN wget https://download.01.org/intel-sgx/sgx-dcap/1.21/linux/distro/ubuntu22.04-server/sgx_debian_local_repo.tgz && \
    tar zxvf sgx_debian_local_repo.tgz && \
    rm -f sgx_debian_local_repo.tgz

SHELL ["/bin/bash", "-o", "pipefail", "-c"]
RUN echo "deb [trusted=yes arch=amd64] file:/sgx_debian_local_repo jammy main" | \
    tee /etc/apt/sources.list.d/sgx_debian_local_repo.list

RUN mkdir -p /etc/apt/keyrings
RUN set -o pipefail && curl -fsSL https://deb.nodesource.com/gpgkey/nodesource-repo.gpg.key | \
    gpg --dearmor -o /etc/apt/keyrings/nodesource.gpg
RUN set -o pipefail && echo "deb [signed-by=/etc/apt/keyrings/nodesource.gpg] https://deb.nodesource.com/node_18.x nodistro main" | \
    tee /etc/apt/sources.list.d/nodesource.list

# Install SGX dependencies
RUN apt-get update && apt-get install -y \
    nodejs=18.18.2-1nodesource1 \
    cracklib-runtime \
    sgx-dcap-pccs=${PCCS_VERSION} \
    sgx-pck-id-retrieval-tool=${PCCS_VERSION} \
    && rm -rf /var/lib/apt/lists/*
```

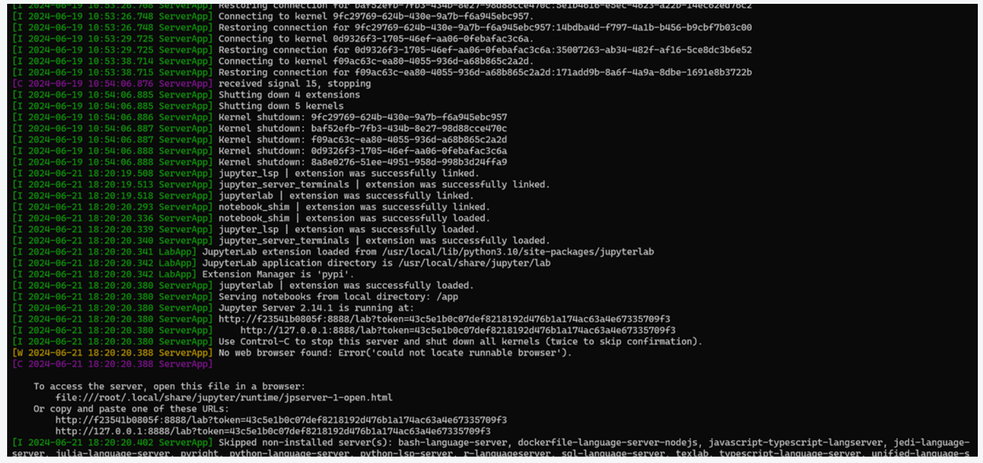
Data Clean Room

The Docker_Image is a comprehensive Docker setup designed to create a secure data clean room environment. It utilizes an Ubuntu 22.04 base and essential tools such as wget, curl, vim, git, python3.9, pip, and tpm2-tools. For data manipulation, it includes Python libraries pandas and numpy, alongside a configured JupyterLab for remote access.

The image incorporates Intel SGX packages like sgx-dcap-pccs and sgx-pck-id-retrieval-tool to establish secure enclaves. It installs Node.js for running the PCCS server and sets up its configurations and SSL keys. A new user, "ubuntu," is added with the necessary permissions.

TPM initialization is performed by creating TPM Endorsement and Attestation Keys and reading PCR values for integrity checks. The Docker image ensures both JupyterLab and the PCCS server run concurrently, providing a robust and secure environment for data handling and analysis.



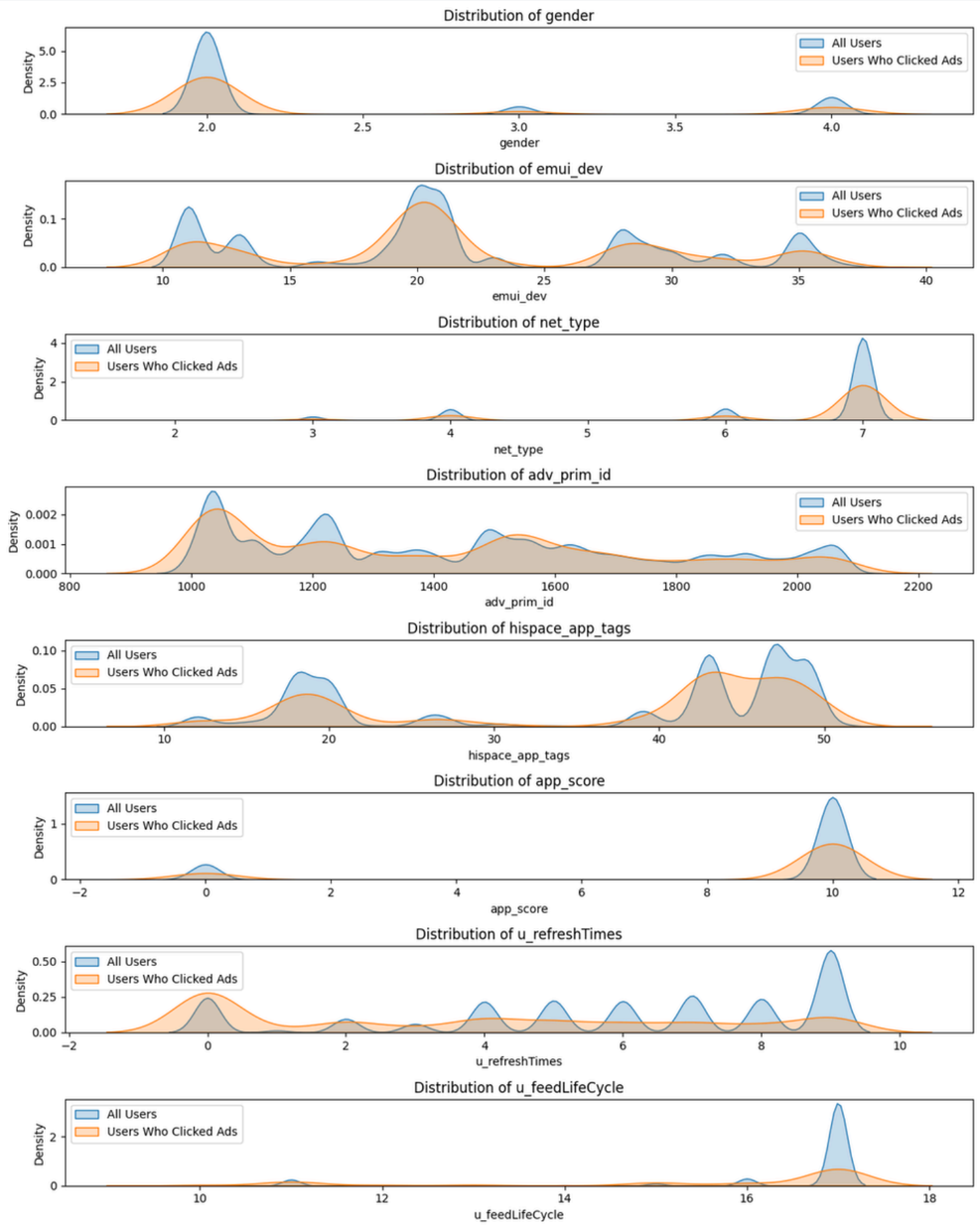


Data Evaluation

Implementation of Task 1

We aim to derive insights about potential customers by analyzing various aspects of user behavior and demographics. The following insights can help ad agencies better understand their potential customers and refine their advertising strategies to increase engagement and conversion rates.

Some of the key questions we explored include: Age Group Distribution, Geographic Distribution, Device Usage, Content Preferences. These insights can help ad agencies better understand their potential customers and refine their advertising strategies to increase engagement and conversion rates.

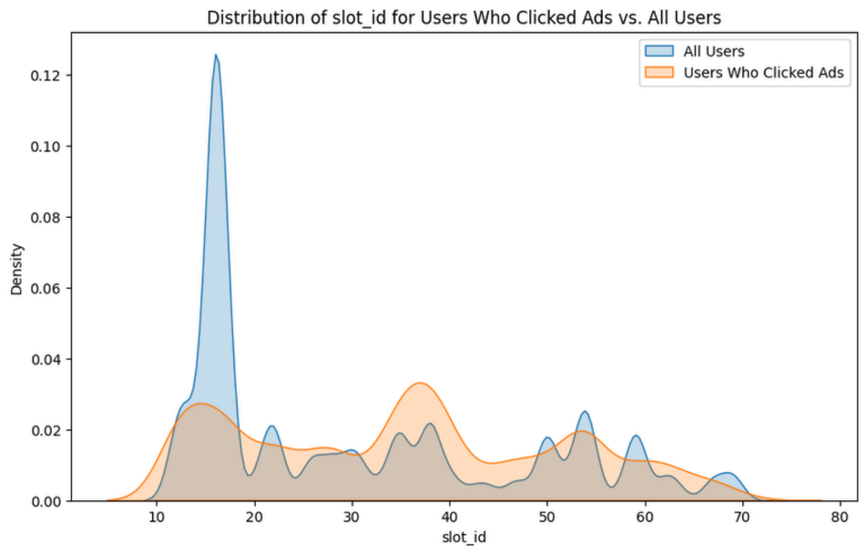


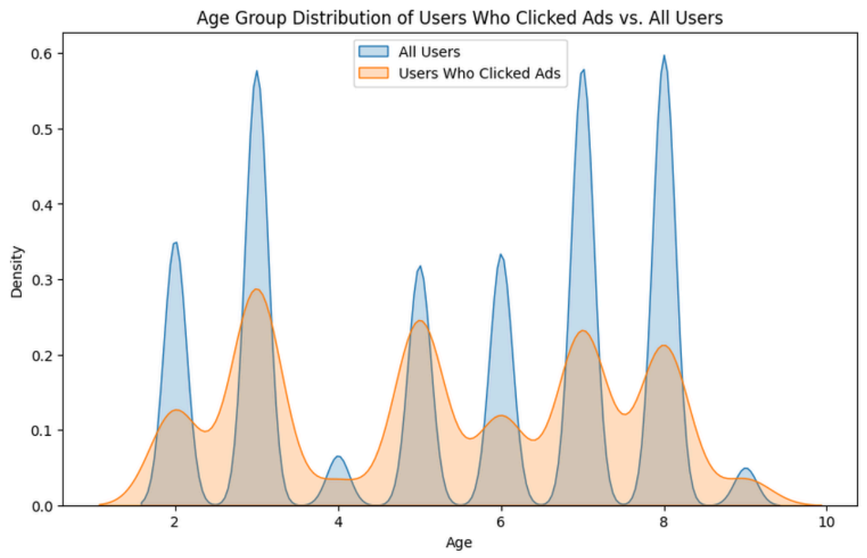
Data Evaluation

Implementation of Task 1

We aim to derive insights about potential customers by analyzing various aspects of user behavior and demographics. The following insights can help ad agencies better understand their potential customers and refine their advertising strategies to increase engagement and conversion rates.

Some of the key questions we explored include: Age Group Distribution, Geographic Distribution, Device Usage, Content Preferences. These insights can help ad agencies better understand their potential customers and refine their advertising strategies to increase engagement and conversion rates.

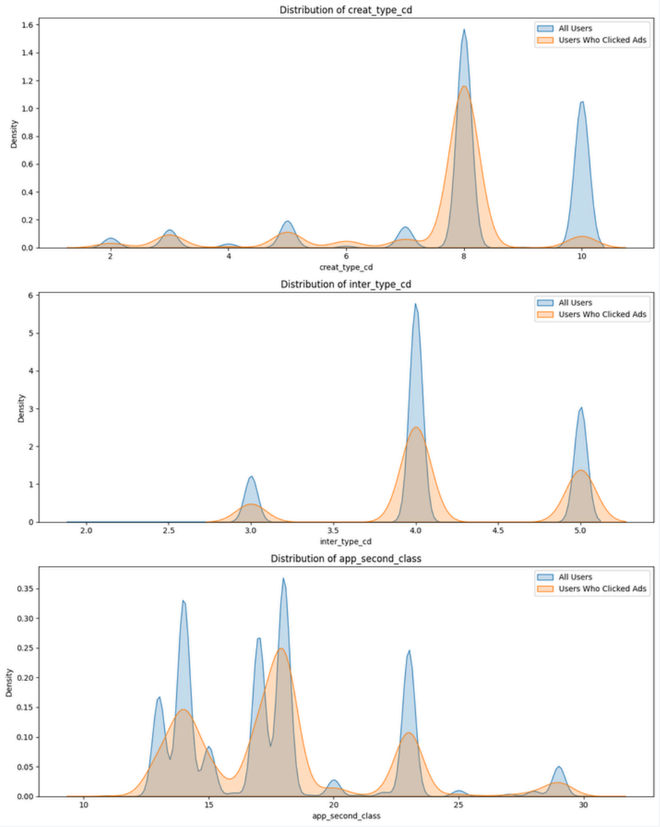


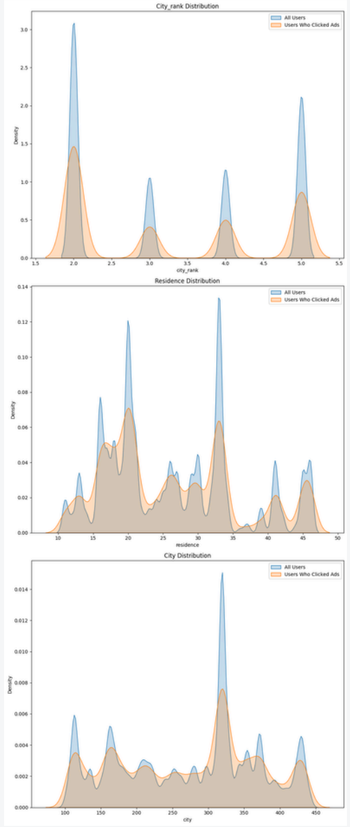


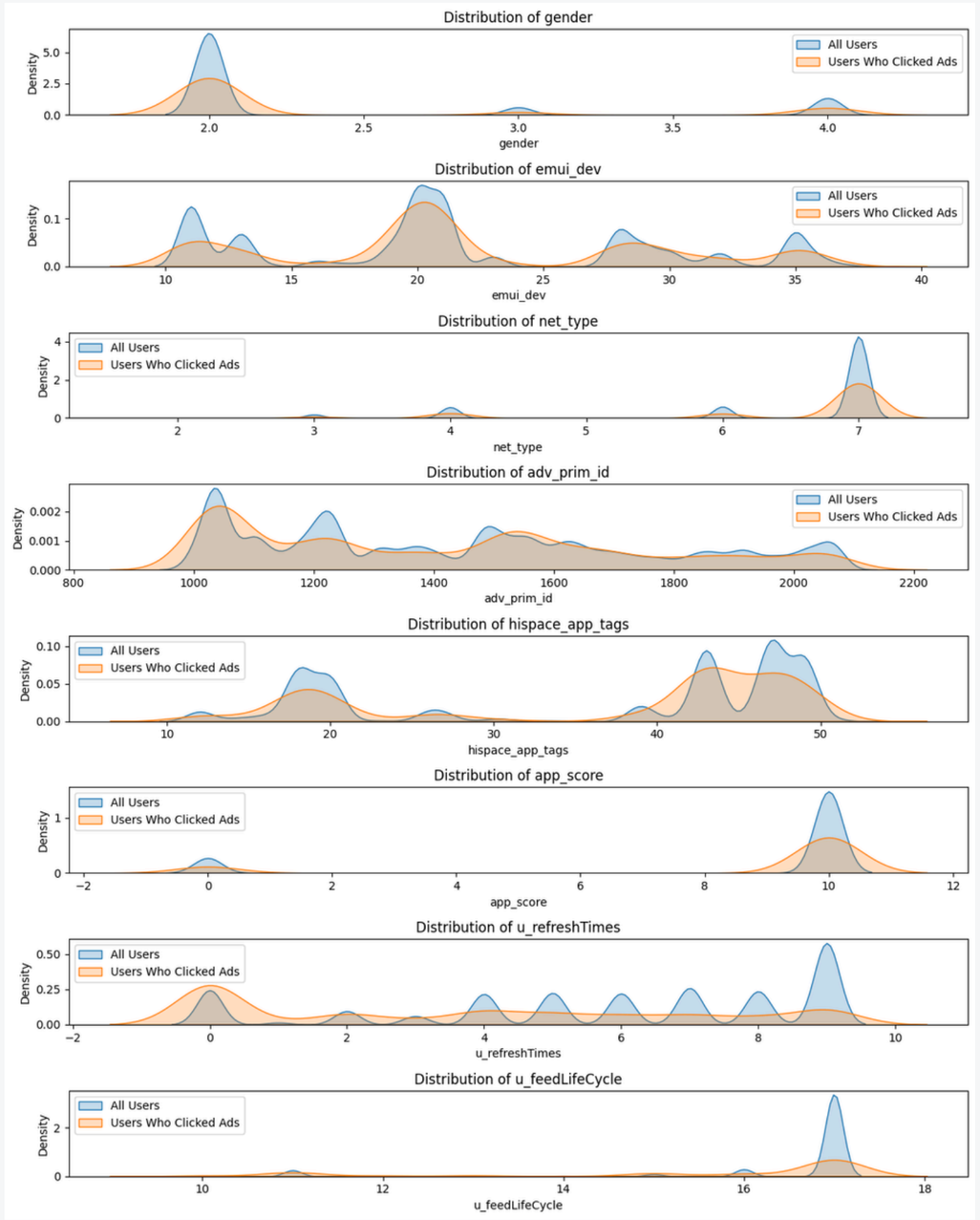
Data Evaluation

Implementation of Task 1

We aim to derive insights about potential customers by analyzing various aspects of user behavior and demographics. The following insights can help ad agencies better understand their potential customers and refine their advertising strategies to increase engagement and conversion rates.



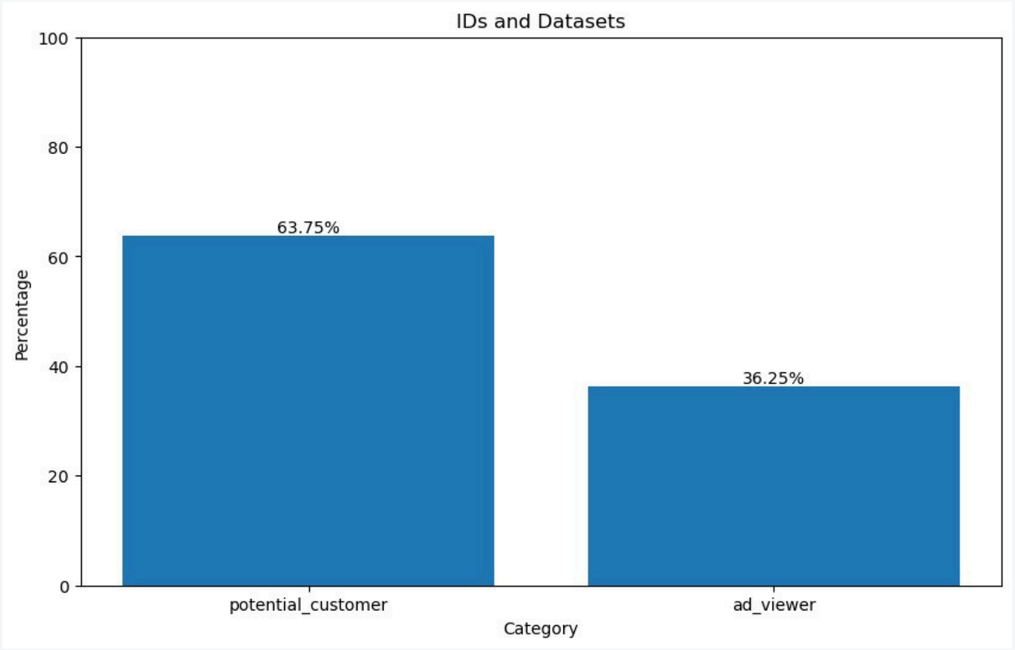




Data Evaluation

Implementation of Task 2 & 3

We aimed to train a predictive model with the following parameters - Identifying Potential Customers, Incorporating Diverse Attributes, Probability Prediction. which utilizes a comprehensive set of attributes, including audience demographics, news content attributes, advertisement attributes, and device attributes, to enhance the prediction accuracy and predicts the probability that a given audience member will become a potential customer based on the aforementioned attributes, enabling more targeted and effective advertising strategies, including features like String to Integer Encoding, Minimum Distance Calculation, Dask Utilization, Privacy Risk Evaluation and performed Metrics Calculation alongwith Benchmarking and Validation.

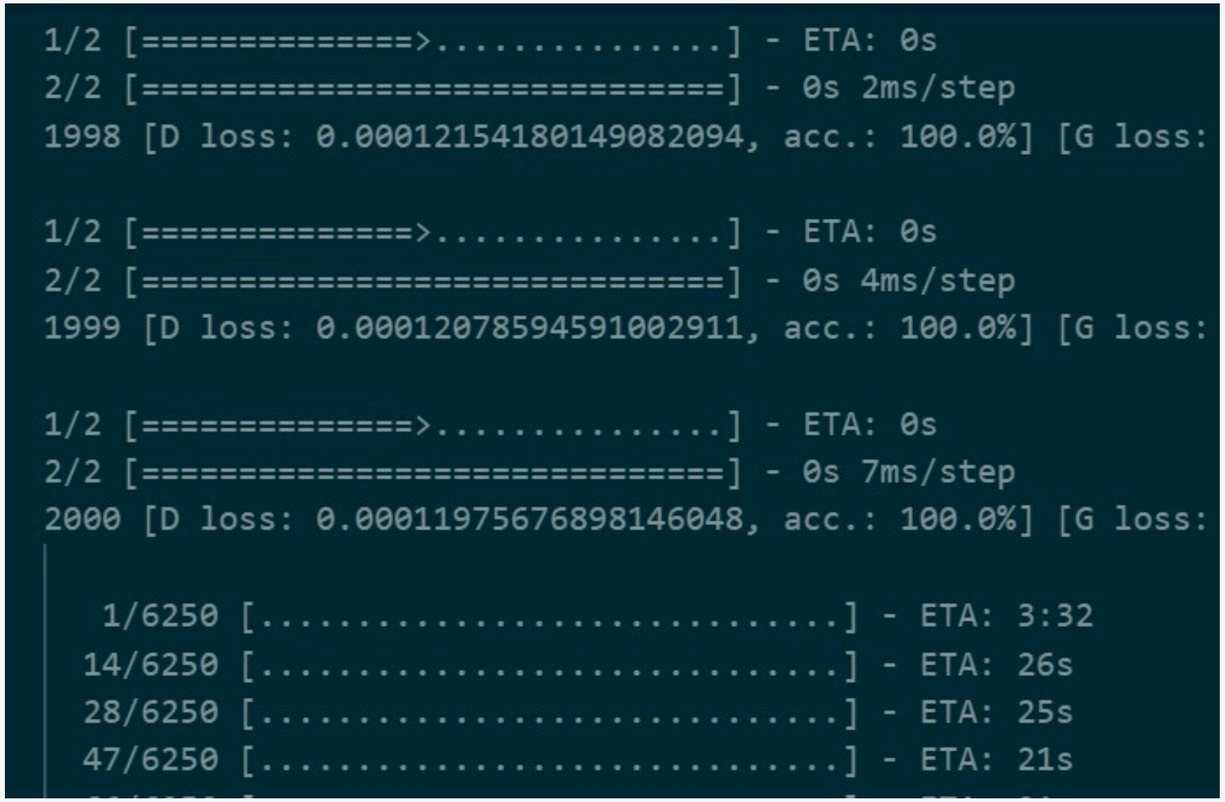



```
(env) manas@ganu3010:/media/manas/Ubuntu/UCLA_Hackathon$ python test.py
Data merged
Data split into training and testing sets:
        Training set: (12800, 26)
        Testing set: (3200, 26)
Model:
VotingClassifier(estimators=[('rf', RandomForestClassifier()),
                             ('gbm', GradientBoostingClassifier()),
                             ('hgbm', HistGradientBoostingClassifier())])
Model fitting done.
Features:
['age', 'inter_type_cd', 'slot_id', 'hispace_app_tags', 'u_newsCatInterestsST', 'u_refreshTimes']
Accuracy:  0.7071875
              precision    recall  f1-score   support

           0       0.70      0.73      0.71      1611
           1       0.71      0.69      0.70      1589

    accuracy                           0.71      3200
   macro avg       0.71      0.71      0.71      3200
weighted avg       0.71      0.71      0.71      3200

(env) manas@ganu3010:/media/manas/Ubuntu/UCLA_Hackathon$
```

Data Evaluation

Implementation of Task 2 & 3

We aimed to train a predictive model with the following parameters - Identifying Potential Customers, Incorporating Diverse Attributes, Probability Prediction. which utilizes a comprehensive set of attributes, including audience demographics, news content attributes, advertisement attributes, and device attributes, to enhance the prediction accuracy and predicts the probability that a given audience member will become a potential customer based on the aforementioned attributes, enabling more targeted and effective advertising strategies, including features like String to Integer Encoding, Minimum Distance Calculation, Dask Utilization, Privacy Risk Evaluation and performed Metrics Calculation alongwith Benchmarking and Validation.



Considerations on Efficiency and Scalability

Using Cloud Confidential VM and Containerization with latest TEE technologies like Intel TDX & SGX ensures secure scalability and effiecient resource management.
Using Synthetic Data created by training GAN (and other) models helps in secure data transaction as well as can be generate according on one's needs.

Thank you !

Team Name: **BitBuilders**