中国石油大学（北京）
CHINA UNIVERSITY OF PETROLEUM , BEIJING

PYTHON FOR DATA ANALYSIS

FINAL PROJECT REPORT

Team 58

Full Name	Student ID
Sonet Adhikary (索内特)	2022030089
Nuhu Ibne Shahid (胡歌)	2022030109
Tanha Tabassum (田娜)	2022030110

College of Information Science and Engineering
College of Artificial Intelligence

Date: May 10, 2024

CTR Analysis of Advertisements

Abstract- **this report is shows what actually we coded last few days, pressing our keyboard harder than ever. This exhibits how we opened the large data file, how we used it, how we analyzed it, and how found it that in which factor a user actually cliqued the advertisement. We also made a model by training it on different datas, that can predict the Clique Through Rate. It is a whole Case Study of an advertisement and how user are reacting to that.**

I. Introduction

First of all CTR means Clique Through Rate, that have been using in the ad industry for a long while now. A company if try to sell their product, they have to advertise the product, now user need to see that ad, but why they will clique that ad, and how many users and how often they will clique that, we have to find it out. We also have to analyze it in different perks. We will find it how CTR is used to gauge the performance of keywordsand ads and how it can act as the KPI (Key Performance Indicator) of a business that relies on online advertising. Increasing the accuracy of Advertisement CTR prediction is critical to improve the effectiveness of precision marketing.

II. Problem Statement

The exact problem is for what, for which key factor a user is cliquing the ad. To be specific, from a huge data set, we have to predict the CTR by analysing it and train a module that can predict himself the future.

III. Data Description

Date Range: Seven consecutive days
Platforms: Huawei AppGallery
Ad Types: Video ads
Metrics: Age, city, City Rank, his_app_size

IV. Selection Rationale:

Objective: The purpose of analysis is to optimize ad performance, and understand user engagement.

Relevance: CTR serves as a key performance indicator in advertising, reflecting user engagement, campaign effectiveness, and cost efficiency. It helps advertisers gauge the success of their ad campaigns and optimize strategies to achieve their marketing objectives.

Data Quality: The datasets is masked reliable and accurately represents the performance of the advertisements.

V. Team Member Contributions and Work Distribution

1. **Programmer**: Sonet Adhikary (2022030089)

Contribution: Write, test, debug, and maintain the detailed instructions, building and maintaining data pipelines and infrastructure for processing CTR data. Developing custom analytics tools and dashboards for visualizing and analyzing advertising performance metrics. Integrating third party APIs for data retrieval and analysis.

2. **Data Analyst**: Nuhu Ibne Shahid (2022030109)

Contribution: Analyzing CTR data to identify trends, patterns, and insights. Creating dashboards to track advertising performance metrics. Reviews data to identify key insights into customers and ways the data can be used to solve problems.

3. **Product Manager:** Tanha Tabassum (2022030110)

Contribution: Oversee the development of analytics tools and platforms used to track advertising performance metrics like CTR. Prepared the report and all documentations, formatting and requirements.

VI. Analytical Approach & Methodology

Analytical Approach:

1. Objective Definition:
The objectives of the analysis is to find the factor affecting CTR.

2. Data Collection:
There was two segments, Test Data & Train Data. We basically used Train Data, 20% of datas we used to test, and 80% of it to create module to predict the performance.

3. Exploratory Data Analysis (EDA):
Explore the data to understand its characteristics, patterns, and distributions. Use descriptive statistics, data visualization techniques, and exploratory analyses to uncover insights and identify potential relationships.

4. Hypothesis Formulation:
Based on the insights gained from EDA, formulate hypotheses or assumptions about the relationships between variables or factors in the data. These hypotheses guide the subsequent analysis and testing.

5. Statistical Analysis and Modeling:
Apply statistical techniques and modeling approaches to test hypotheses, make predictions, or uncover hidden patterns in the data.

6. Interpretation and Insight Generation:
Interpret the results of the analysis in the context of the original objectives. Generate actionable insights and recommendations based on the findings, addressing the questions or problems defined at the outset.

7. Validation and Sensitivity Analysis:
Validate the robustness of the analysis by conducting sensitivity analysis, testing assumptions, and assessing the impact of uncertainties or variations in the data. This ensures the reliability and validity of the conclusions drawn.

Methodology:

Methods and Tools:

1. Pandas Library: Created data frame, Analyzed the datas.

2. yaspin library: For Visualization and others.

3. Plotly: For graphing and columns.

4. Scikit-learn package: To train the data to create a module that predict the performances.

Detailed Steps,

1. Data Loading,
First of all, we used pandas library to load and analyse the data from csv file.

We didnt take the whole data at once, we took it chunk by chunk to ensure memory efficiency so that there will be no memory glitches The chunk size was 1000. To make it more efficient we didn't load the all fields of the dataset, we just filtered out necessary fields using a lambda function in usecolls parameters in pandas read_csv function.

2. Data Visualization,
Secondly, we used Plotly library to draw a graph for visualizing the CTR. We sent the data frame (df = pandas.read_csv{}) and two others fields (label-clique, his_app_-size- application size) to a custom function named draw_graph.

```
# import external libns

import plotly.express as px
import plotly.io as pio

def draw_graph(data, x_col, color_col, title, color_map=None):
    pio.templates.default='plotly_white'
    fig = px.box(data,
                 x=x_col,
                 color=color_col,
                 title=title,
                 color_discrete_map=color_map)
    fig.update_traces(quartilemethod="exclusive")
    fig.show()
```

3. Machine Learning,

Thirdly, We trained the data using machine learning algorithm. We used the python's Scikit-learn package to train the data in 20:80 ratio.

We experimented on 20% to determine the accuracy and precision of other 80% of the dataset.

```
train_model.py  train_model
# import external libs

from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, confusion_matrix, precision_score, recall_score, f1_score, ro

def train_model(df,target_col,inp_col):
    df[target_col]=df[target_col].astype(int)

    X = df[[inp_col]]
    y = df[target_col]

    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

    model = RandomForestClassifier()
    model.fit(X_train, y_train)

    y_pred = model.predict(X_test)

    accuracy = accuracy_score(y_test, y_pred)
    conf_matrix = confusion_matrix(y_test, y_pred)
    precision = precision_score(y_test, y_pred)
    recall = recall_score(y_test, y_pred)
    f1 = f1_score(y_test, y_pred)
    roc_auc = roc_auc_score(y_test, y_pred)
    print(f"Accuracy: {accuracy:.4f}")
    print("\nConfusion Matrix:")
    print(conf_matrix)
    print(f"\nPrecision: {precision:.4f}")
    print(f"Recall: {recall:.4f}")
    print(f"F1 Score: {f1:.4f}")
    print(f"\nROC-AUC Score: {roc_auc:.4f}")
```

VII. Analysis and Findings

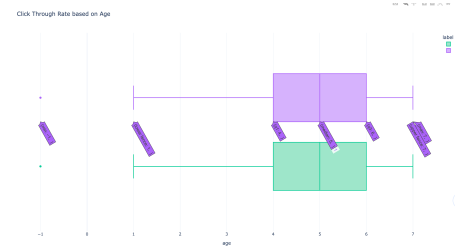
We generated graph on 3 different parameters, Age, City rank and App Size. Below is the function that will draw the graph.

Age:

Here is the code that made the graph based on Age.

```
# drawing graph
with yaspin(text='Drawing Graph ... ',color='yellow'):
    draw_graph(df, inp_col, target_col, "Click Through Rate based on Age", {'Yes': 'blue', 'No': 'red'})
```

The age and the CTR is parallel we can see from the graph. So the advertiser should target their audience based on their age and that is one of the key factors affecting the CTR of an advertisement.

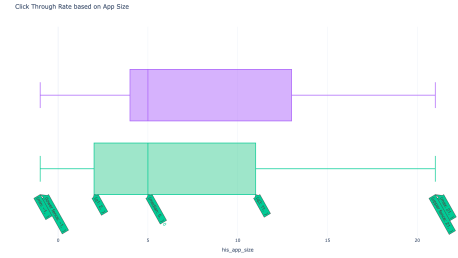


App Size:

Here is the code that made the graph based on App Size.

```
# drawing graph
with yaspin(text='Drawing Graph ... ',color='yellow'):
    draw_graph(df, inp_col, target_col, "Click Through Rate based on App Size", {'Yes': 'blue', 'No': 'red
```

The App Size and the CTR is disproportional we can see from the graph. When the app size increases the CTR decreases. So the advertiser and the developers both should be concerne about this. So app size also is a crucial factor affecting the CTR of an advertisement.

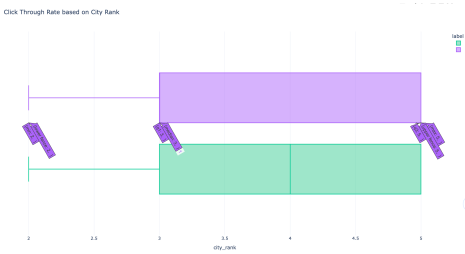


City Rank:

Here is the code that made the graph based on Age.

```
# drawing graph
with yaspin(text='Drawing Graph...',color='yellow'):
    draw_graph(df, inp_col, target_col, 'Click Through Rate based on City Rank', {'Yes': 'blue', 'No': 'g
```

The City Rank and the CTR corelated in the graph. So city rank is a crucial factor too affecting the CTR of an advertisement.



Data Prediction:

We trained a model using scikit-learn to predict the CTR of adverstisements from the given data set.

```
: Training Model with Machine Learning...Accuracy: 0.9654

Confusion Matrix:
[[8091784       0]
 [ 289643       0]]

Precision: 0.6100
Recall: 0.5200
F1 Score: 0.7300

ROC-AUC Score: 0.5000
```

VIII. Discussions and Conclusions:

In conclusion, the analysis of CTR for advertisements based on age, city rank, app size provides valuable insights for advertisers seeking to optimize campaign performance and maximize ROI. By understanding demographic preferences, regional variations, and app-related factors, advertisers can tailor their ad strategies to effectively target specific audience segments and enhance engagement levels.

Key takeaways include the importance of personalized targeting strategies to appeal to different age groups, localization tactics to cater to regional preferences, strategic placement within top-ranking apps to increase visibility, and optimization efforts for apps of varying sizes to improve ad performance and user experience.

Moving forward, advertisers should continue to monitor and analyze CTR data, adapt strategies based on emerging trends and insights, and prioritize data-driven decision-making to achieve advertising objectives and drive business growth in an increasingly competitive digital landscape.

In today's world the internet is the biggest market place to cater a business and advertisement is the key communicator to your audience. CTR plays here a big role for business and we saw its different perks, journey, analysis and findings.

IX. Refferences:

We have used this external python libraries in our code.

1. pandas (data framing)
   https://pandas.pydata.org/

2. yaspin library (for console visualization)
   https://pypi.org/project/yaspin/

3. scikit learn (training ml model)
   https://scikit-learn.org/stable/

4. Plotly (Data Visualization)
   https://plotly.com/

5. Numpy (EDA)
   https://numpy.org/