Department of Computer Science and Engineering (Data Science)

Subject: Machine Learning – I (DJ19DSC402)

AY: 2022-23

Experiment 10 (Mini Project)

Aim: Design a classifier to solve a specific problem in the given domain.

Tasks to be completed by the students:

Select a specific problem from any of the given domain areas, such as: Banking, Education, Insurance, Government, Media, Entertainment, Retail, Supply chain, Transportation, Logistics, Energy and Utility.

Task 1: Select appropriate dataset, describe the problem and justify the suitability of your dataset.

Task 2: Perform exploratory data analysis and pre-processing (if required).

Task 3: Apply appropriate machine learning algorithm to build a classify. Perform appropriate testing of your model.

Task 4: Submit a report in the given format.

- Introduction
- Data Description
- Data Analysis
- Reason to select machine learning model
- Algorithm
- Result Analysis
- Conclusion and Future Scope.
- Python notebook

Task5: Presentation

Shri Vile Parle Kelavani Mandal's
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC Accredited with "A" Grade (CGPA : 3.18)

Department of Computer Science and Engineering (Data Science)

Report on Mini Project

Machine Learning -I (DJ19DSC402)

AY: 2022-23

AD PERFORMANCE PREDICTOR

NAME: Mansi Wadhwa

SAP ID: 60009210157

Guided By: Preetam Vernekar

In today's world, digital advertising has become an integral part of marketing strategies for businesses of all sizes. With the increasing number of online users and platforms, it is crucial for advertisers to optimize their ad performance to reach their target audience and achieve their business goals. In this context, machine learning techniques can play a vital role in predicting ad performance and maximizing the return on investment (ROI) for advertisers.

This project focuses on building a machine learning model to predict the click-through rate (CTR) of online ads. The dataset used for this project consists of three main files: raw_sample.csv, ad_feature.csv, and user_profile.csv. The raw_sample file contains information about user behavior, including ad clicks and non-clicks. The ad_feature file contains information about the ads, such as price and brand. Finally, the user_profile file contains demographic information about the users, such as age and gender.

The goal of this project is to merge these three datasets and preprocess the data to create a feature matrix that can be used to train a machine learning model. The model will predict the CTR of an ad based on its features, such as price, brand, and user demographic information. Gradient boosting is chosen as the machine learning

CHAPTER 2: DATA DESCRIPTION

In this chapter, the datasets used in our ad performance prediction project are described. The three datasets used in this project are:

1. Raw Sample Dataset:

This dataset contains the user's browsing history, including the timestamp, ad group ID, and product ID. It also contains the number of times the ad was clicked and not clicked.
This dataset contains information about users and their interactions with various ads.
It has 6 columns:

- user (string): a unique identifier for each user
- time_stamp (datetime): the timestamp of the user's interaction with the ad
- adgroup_id (int): a unique identifier for each ad group
- pid (int): a unique identifier for each product
- nonclk (int): the number of times the ad was displayed but not clicked
- clk (int): the number of times the ad was clicked

Data type: user (object), time_stamp (datetime64), adgroup_id (int64), pid (int64), nonclk (int64), clk (int64)

2. Ad Feature Dataset:

This dataset contains information about the ads, including the ad group ID, category ID, campaign ID, customer, brand, and price.

This dataset contains information about the features of each ad, such as its category, price, and brand.

It has 6 columns:

- adgroup_id (int): a unique identifier for each ad group
- cate_id (int): the category ID of the product
- campaign_id (int): a unique identifier for each campaign
- customer (int): a unique identifier for each customer
- brand (int): the brand ID of the product
- price (float): the price of the product

Data type: adgroup_id (int64), cate_id (int64), campaign_id (int64), customer (int64), brand (int64), price (float64)

3. User Profile Dataset:

This dataset contains information about the users, including the user ID, CMS segment ID, CMS group ID, final gender code, age level, pvalue level, shopping level, occupation, and new user class level.

This dataset contains information about users, such as their age, gender, and occupation.

It has 9 columns:

- userid (string): a unique identifier for each user
- cms_segid (int): the user's segment ID
- cms_group_id (int): the user's group ID
- final_gender_code (int): the user's gender (0 for male, 1 for female)
- age_level (int): the user's age level (1 for 18-24, 2 for 25-29, etc.)
- pvalue_level (int): the user's purchase value level (1 for low, 2 for medium, etc.)
- shopping_level (int): the user's shopping level (1 for low, 2 for medium, etc.)

- occupation (int): the user's occupation (1 for student, 2 for engineer, etc.)
- new_user_class_level (int): the user's class level (1 for low, 2 for medium, etc.)

Data type: userid (object), cms_segid (int64), cms_group_id (int64), final_gender_code (int64), age_level (int64), pvalue_level (int64), shopping_level (int64), occupation (int64), new_user_class_level (int64)

I have merged these three datasets using the ad group ID and user ID as common columns. The merged dataset contains 18 features: user ID, timestamp, ad group ID, product ID, clickthrough rate (CTR), category ID, campaign ID, customer, brand, price, CMS segment ID, CMS group ID, final gender code, age level, pvalue level, shopping level, occupation, and new user class level.

Shri Vile Parle Kelavani Mandal's
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC Accredited with "A" Grade (CGPA : 3.18)

Department of Computer Science and Engineering (Data Science)

CHAPTER 3: DATA ANALYSIS

Here's an overview of the data analysis performed on the preprocessed datasets:

1. Raw Sample Dataset:

- Checked for missing values and found none.
- Changed ‘user’ column name to ‘userid’ in order to merge it with userprofile dataset later on.
- Converted `time_stamp` column from string to datetime format.
- Created a new column `timebin` to represent the parts of a day from the `time_stamp` column.
- Created a distribution plot for the `clk` column to show the click-through rate distribution.
- Created a distribution plot for the `timebin` column to show the click-through rate distribution by morning, evening and afternoon .

2. Ad Feature Dataset:

- Checked for missing values and found none.

8

- Created a distribution plot for the `price` column to show the distribution of ad prices.

- Created a distribution plot for the `brand` column to show the distribution of ad brands.

3. User Profile Dataset:

- Checked for missing values and found none.

- Created a distribution plot for the `age_level` column to show the distribution of age levels.
- Created a distribution plot for the `pvalue_level` column to show the distribution of p-value levels.
- Created a distribution plot for the `shopping_level` column to show the distribution of shopping levels.
- Created a distribution plot for the `occupation` column to show the distribution of occupations.

Overall, the data analysis provided insights into the distribution of various columns in the datasets and helped in understanding the characteristics of the data.

CHAPTER 4: DATA MODELLING

In this chapter, I discuss the data modeling phase of the project. I will train and evaluate machine learning models to predict the click-through rate of ads based on the provided datasets.

1. Data Preparation:

- Merged the three datasets based on common features.
- Removed unnecessary columns.
- Encoded categorical features using one-hot encoding.
- Split the data into training and testing sets.

2. Model Training:

- Trained multiple machine learning models including Logistic Regression, Random Forest, Gradient Boosting, and XGBoost.

- Used cross-validation to select the best performing model based on accuracy.

- Performed hyperparameter tuning using GridSearchCV to optimize the model performance.

3. Model Evaluation:

- Evaluated the final model on the testing set using various performance metrics including accuracy, precision, recall, and F1 score.

- Plotted the ROC curve to visualize the true positive rate and false positive rate of the model.

- Plotted the feature importance graph to show the importance of each feature in predicting the click-through rate.

Overall, I aimed to develop a robust machine learning model that can accurately predict the click-through rate of ads based on user and ad features.

CHAPTER 5: CONCLUSION

In this project, I analyzed the advertising data to predict the click-through rate of ads. I started by exploring and preprocessing the three datasets: Raw Sample, Ad Feature, and User Profile. I then merged the datasets to create a single dataset and performed exploratory data analysis to understand the distribution of various features.

I then trained multiple machine learning models, including Logistic Regression, Random Forest, Gradient Boosting, and XGBoost, and used cross-validation to select the best performing model based on accuracy. I performed hyperparameter tuning using GridSearchCV to optimize the model performance.

Gradient Boosting Classifier would be the most ideal classifier since it is able to handle non-linear relationships between the input features and the target variable, which is often the case in ad performance prediction. The algorithm constructs a series of decision trees, each one attempting to correct the errors of the previous trees, resulting in a powerful and flexible model. Gradient Boosting is also robust to outliers in the data, which can occur in ad performance prediction due to factors such as unexpected events, or changes in user behavior. The algorithm uses a loss function that is not sensitive to outliers, making it less likely to be affected by them. Gradient Boosting is known for its high accuracy in predicting continuous variables, such as click-through rates. It uses an ensemble of weak learners (decision trees) to create a strong predictive model that minimizes the mean squared error, which is a commonly used metric for regression problems.

Overall, Gradient Boosting is a suitable algorithm for ad performance prediction because it can handle the complexity of the data, provide accurate predictions, and handle missing values and outliers.

The most important features for predicting the click-through rate were found to be the ad price, the ad brand, and the ad group ID.

Overall, this project demonstrates the importance of data analysis and machine learning in predicting the click-through rate of ads. By using machine learning algorithms, we can improve the effectiveness of advertising campaigns and ultimately drive business growth.

https://colab.research.google.com/drive/105qC8Ssbyd4Ae2VG_R9uqZDWT3cGbCr6?usp=sharing