
LOGIT AD-CLICK
MINI PROJECT

CONTENT

MOTIVATION

- Online advertising is a multi-billion dollar business producing most of the revenue for search engines.
- One area drawing recent attention among both researchers and practitioners is prediction of Click Through Rate.
- Being able to predict this rate means meeting business targets.

3

INTRODUCTION

- Companies are chasing internet marketing to advertise their products on websites and social media platforms.
- Working with advertising data to develop a machine learning algorithm that predicts if a particular user will click on an advertisement.
- The CTR prediction has been used over the past several years in every type of advertisement format, search engine advertisements, contextual advertisements, text advertisements, display banner advertisements, video advertisements etc.

4

CLICK THROUGH RATE (CTR)

- Calculates how frequently people click on your ad.
  - $CTR = \left(\frac{Total\ Measured\ Clicks}{Total\ Measured\ Ad\ Impressions}\right) \times 100$
- **Total measured clicks :** The total amount of clicks on an ad (which were counted by a server)
- **Total measured ad impressions:** Number of times an ad was located on a page (and counted by a server).
- **Higher CTR:** Users are more interested in the specific campaign.
- **Lower CTR:** The ads may not be relevant to the users.

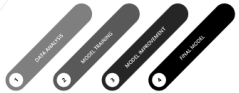
5

LITERATURE REVIEW

- Accurate click-through rate (CTR) prediction can not only improve the advertisement company's reputation and revenue, but also help the advertisers to optimize the advertising performance
- CTR prediction can be made by building the user graph system for the purpose of classifying the advertisement data. The output of this user graph system includes the user's age, gender, and the interest preferences.
- Experiments show that this kind of features has a significant effect on the CTR prediction. ↵

6

HOW DO WE DO IT?



7

DATASET ANALYSIS

- Data analysis is important explore data in meaningful ways.
- Data in itself is merely facts and figures.
- Data analysis organizes, interprets, structures and presents the data into useful information that provides context for the data.
- This context can then be used by decision-makers to take action with the aim of enhancing productivity and business gain.

8

REQUIREMENTS

Python 3.8
- Easy to code
- Python is Portable language
- Large Standard Library

NumPy
- Contains a multi dimensional array
- It can be utilized to perform a number of mathematical operations on arrays.

Pandas
- Provides a really fast and efficient way to manage and explore data.
- Organization and labeling of data.

Matplotlib
- Simple and efficient tools for predictive data analysis.

Seaborn
- Data visualization library based on matplotlib.
- Drawing attractive and informative statistical graphics

Scikit-learn
- Simple and efficient tools for predictive data analysis.
- Accessible to everybody, and reusable in various contexts.

NLTK
- Library to work with human language data
- provides easy to use interfaces such as Word Net

Kaggle
- Kaggle, a subsidiary of Google LLC, is an online community of data scientists and machine learning practitioners.

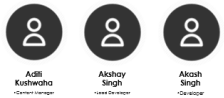
14

Optional Analysis Example

```
import warnings

import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sb

warnings.filterwarnings('ignore')

current_dir = os.getcwd()
file_path = os.path.join(current_dir, r'data_set/train/advertising.csv')

df = pd.read_csv(file_path)
print(df.info())

"""
#1: What age group does the data-set majorly consist of?
"""
sb.distplot(df['age'], bins=20, kde=True,
            hist_kws={"edgecolor":"k"}, color='blue')
print('Oldest person was of:', df['age'].max(), 'Years')
print('Youngest person was of:', df['age'].min(), 'Years')
print('Average age was of:', df['age'].mean(), 'Years')
```

15

TEAM COORDINATION



16

CONCLUSION

- By adding demographic features, we observed notable improvement in the prediction's accuracy.
- The obtained results showed the use value of both machine learning models. ***The Logistic Regression model showed slightly better performance than the decision tree model***, but definitely, both models have shown that they can be very successful in solving classification problems.
- Using the right framework and effective analysis, Click through rate prediction can be a powerful tool to beat the competition in advertising.

20

FUTURE SCOPE

Having full access to the original dataset would be interesting. Explore and manipulate features would undoubtedly bring us exciting and robust insights to optimize model performance. Thus, the feature selecting step would be highly accurate. Deal with imbalanced data is considered a real-world data science problem. Exploring further methods would be interesting. The ***deep neural network*** may be a interesting for the further future study of the CTR prediction.

21

Thank You

23

MOTIVATION

- Online advertising is a multi-billion dollar business producing most of the revenue for search engines.
- One area drawing recent attention among both researchers and practitioners is prediction of Click Through Rate.
- Being able to predict this rate means meeting business targets.

INTRODUCTION

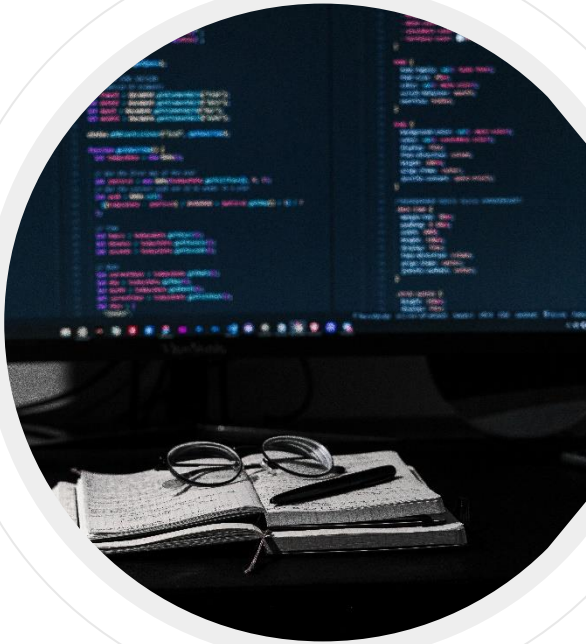
- Companies are chasing internet marketing to advertise their products on websites and social media platforms.
- Working with advertising data to develop a machine learning algorithm that predicts if a particular user will click on an advertisement.
- The CTR prediction has been used over the past several years in every type of advertisement format, search engine advertisements, contextual advertisements, text advertisements, display banner advertisements, video advertisements etc.

CLICK THROUGH RATE (CTR)

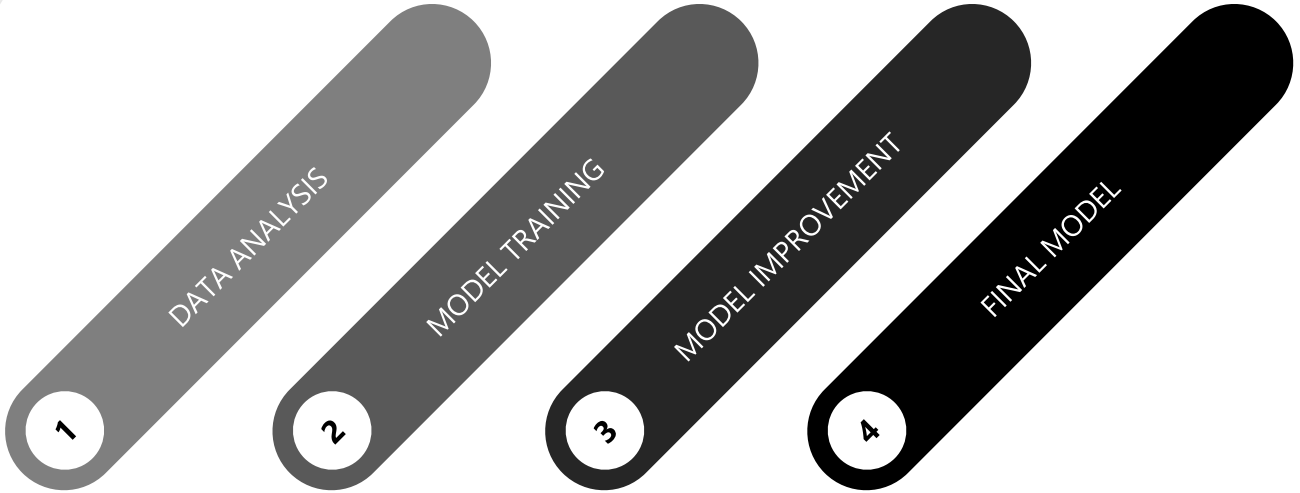
- Calculates how frequently people click on your ad.
  - $CTR = \left(\frac{Total\ Measured\ Clicks}{Totol\ Measured\ Ad\ Impressions}\right) \times 100$
- **Total measured clicks :** The total amount of clicks on an ad (which were counted by a server)
- **Total measured ad impressions:** Number of times an ad was located on a page (and counted by a server).
- **Higher CTR:** Users are more interested in the specific campaign.
- **Lower CTR:** The ads may not be relevant to the users.

LITERATURE REVIEW

- Accurate click-through rate (CTR) prediction can not only improve the advertisement company's reputation and revenue, but also help the advertisers to optimize the advertising performance.
- CTR prediction can be made by building the user graph system for the purpose of classifying the advertisement data. The output of this user graph system includes the user's age, gender, and the interest preferences.
- Experiments show that this kind of features has a significant effect on the CTR prediction. [1]



HOW DO WE DO IT?

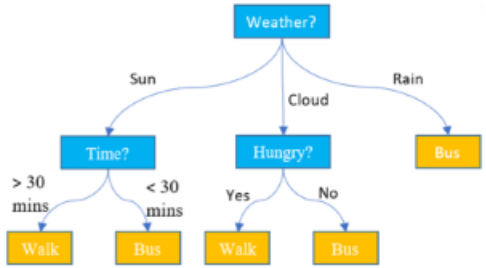


DATASET ANALYSIS

- Data analysis is important explore data in meaningful ways.
- Data in itself is merely facts and figures.
- Data analysis organizes, interprets, structures and presents the data into useful information that provides context for the data.
- This context can then be used by decision-makers to take action with the aim of enhancing productivity and business gain.

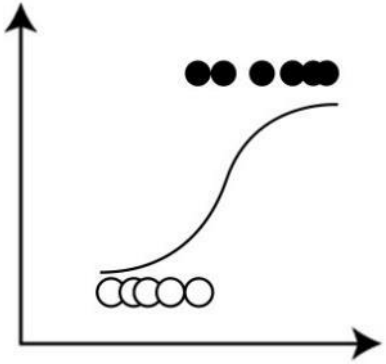
DECISION TREES

- The Decision Tree is one of the most commonly algorithm for analysis and modeling. It is used for classification, prediction, estimation, clustering, data description, and visualization. [2]
- The advantages of Decision Trees, compared to other data mining techniques are simplicity and computation efficiency.
- Disadvantages :
  - Decision-tree learners can create over-complex trees that do not generalize the data well. This is called ***overfitting***.
  - Decision trees can be unstable because small variations in the data might result in a completely different tree being generated. This is called ***variance***.



LOGISTIC REGRESSION

- Logistic Regression is a Machine Learning algorithm which is used for the classification problems, it is a predictive analysis algorithm and based on the concept of probability.[3]
- It can only be used when the target variables fall into discrete categories.
- Few examples:
  - Predicting if an email is spam or not spam
  - Whether a tumor is malignant or benign
  - Whether a mushroom is poisonous or edible.



GRADIENT DESCENT

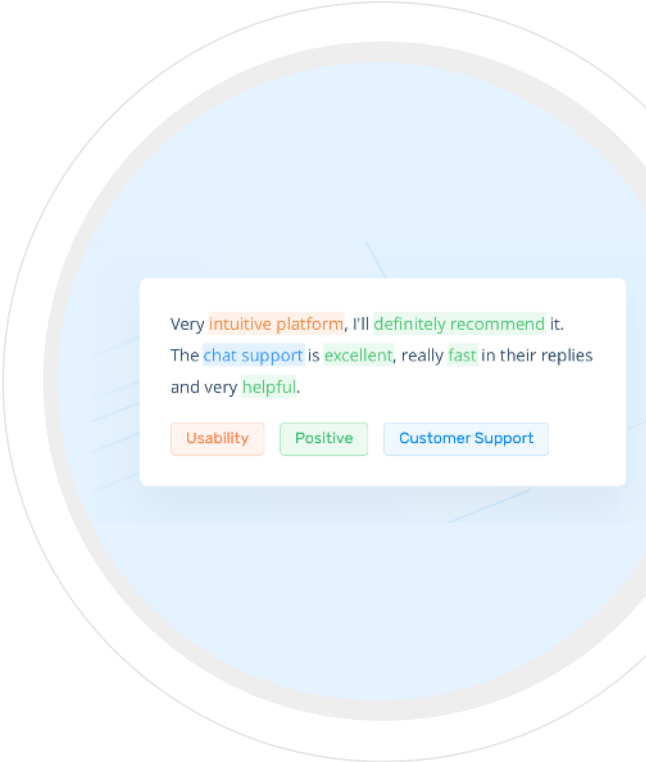
- Gradient Descent is one of the most popular and widely used optimization algorithms. Given a machine learning model with parameters ***(weights and biases)*** and a cost function to evaluate how good a particular model is, our learning problem reduces to that of finding a good set of weights for our model which minimizes the cost function.
- Variants of Gradient Descent
  - Batch Gradient Descent
  - Stochastic Gradient Descent
  - Mini-batch Gradient Descent

Variants of Gradient Descent

- **Batch gradient descent** computes the gradient of the cost function w.r.t to parameter w for **entire training data**.
- **Stochastic gradient descent (SGD)** computes the gradient for each update using a **single training data point** (chosen at random).
- In **mini-batch gradient descent**, we calculate the gradient for each small mini-batch of training data.

NATURAL LANGUAGE PROCESSING

- NLP is a field in machine learning with the ability of a computer to understand, analyze, manipulate, and potentially generate human language.
- Machine learning for NLP helps data analysts turn unstructured text into usable data and insights.
- Examples:
  - Information Extraction (Gmail structures events from emails).
  - Auto-Predict (Google Search predicts user search results).
  - Sentiment Analysis (Hater News gives us the sentiment of the user).



REQUIREMENTS

Python 3.8

- Easy to code
- Python is Portable language
- Large Standard Library

NumPy

- Contains a multi-dimensional array
- It can be utilized to perform a number of mathematical operations on arrays.

Pandas

- Provides a really fast and efficient way to manage and explore data.
- Organization and labeling of data.

Matplotlib

- Simple and efficient tools for predictive data analysis.

Seaborn

- Data visualization library based on matplotlib.
- Drawing attractive and informative statistical graphics.

Scikit-learn

- Simple and efficient tools for predictive data analysis
- Accessible to everybody, and reusable in various contexts.

NLTK

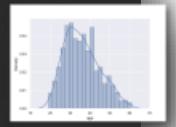
- Library to work with human language data.
- provides easy-to-use interfaces such as Word-Net.

Kaggle

- Kaggle, a subsidiary of Google LLC, is an online community of data scientists and machine learning practitioners. .

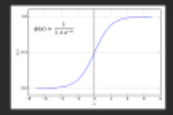
Dataset Analysis Example

```
import warnings

import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sb

warnings.filterwarnings("ignore")

current_dir = os.getcwd()
file_path = os.path.join(current_dir, r"data_set/train/advertising.csv")

df = pd.read_csv(file_path)
print(df.info())

"""
#1: What age group does the data-set majorly consist of?
"""
sb.distplot(df['age'], bins=20, kde=True,
            hist_kws=dict(edgecolor="k", linewidth=1))
print('Oldest person was of:', df['age'].max(), 'Years')
print('Youngest person was of:', df['age'].min(), 'Years')
print('Average age was of:', df['age'].mean(), 'Years')
```



ML Model Example

```python
def sigmoid(z):
    """
    Maps an input to an output of a value between 0 and 1.
    :param z:
    :return: float, [0,1]
    """
    return 1 / (1 + np.exp(-z))


class LogisticRegression:
    """
    Logistic regression is a probabilistic classifier, similar to the Naive Bayes classifier.
    """

    def __init__(self, learning_rate=0.01, max_iter=1000, fit_intercept=False, optimizer='gd', verbose=0):
        self.__learning_rate = learning_rate
        self.__max_iter = max_iter
        self.__fit_intercept = fit_intercept
        self.__optimizer = optimizer
        self.__weights = None
        self.__verbose = verbose

    def __compute_prediction(self, x):
        """
        Compute the prediction y_hat based on current weights
        :param x:
        :return: numpy.ndarray, y_hat of x under weights
        """
        z = np.dot(x, self.__weights)
        predictions = sigmoid(z)
```



Splitting Dataset

```python
def split_data_set(n_rows=1000, train_test_split=0.1):
    """
    Split data-set into training and testing set.
    :return: [x_train, y_train, x_test, y_test]
    """
    current_dir = os.getcwd()
    file_path = os.path.join(current_dir, "data_set", "train", "advertising.csv")

    df = pd.read_csv(file_path, nrows=n_rows)

    x = df.drop(['click', 'ad_topic_line'], axis=1).values
    y = df['click'].values

    n_train = int(n_rows * (1 - train_test_split))

    x_train = x[:n_train]
    y_train = y[:n_train]
    x_test = x[n_train:]
    y_test = y[n_train:]

    enc = OneHotEncoder(handle_unknown='ignore')
    enc.fit(x_train)

    x_train_enc = enc.fit_transform(x_train)
    x_test_enc = enc.transform(x_test)

    return x_train_enc.toarray(), y_train, x_test_enc.toarray(), y_test
```

```
if __name__ == "__main__":
    x_train, y_train, x_test, y_test = split_data_set(n_rows=1000,
 train_test_split=0.3)

    model = LogisticRegression(learning_rate=0.1, max_iter=1000,
 optimizer='gd', fit_intercept=True)

    model.fit(x_train, y_train)
    predictions = model.predict(x_test)

    print("--" * 30, "Logistic Regression (with Gradient Descent)
 Accuracy score:",
          "Training samples: {0}, AUC on testing set: {
 1:.3f}".format(len(x_train), roc_auc_score(y_test, predictions)),
          "--" * 30,
          sep="\n")
```

TEAM CONTRIBUTION

Aditi Kushwaha

- Content Manager

Akshay Singh

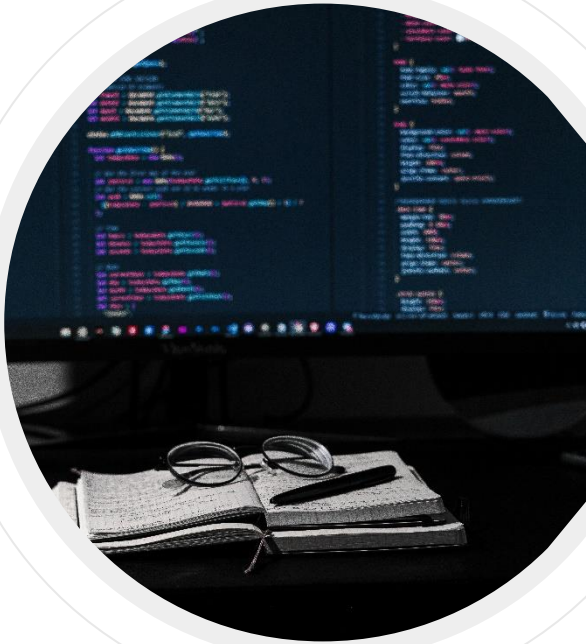
- Lead Developer

Akash Singh

- Developer

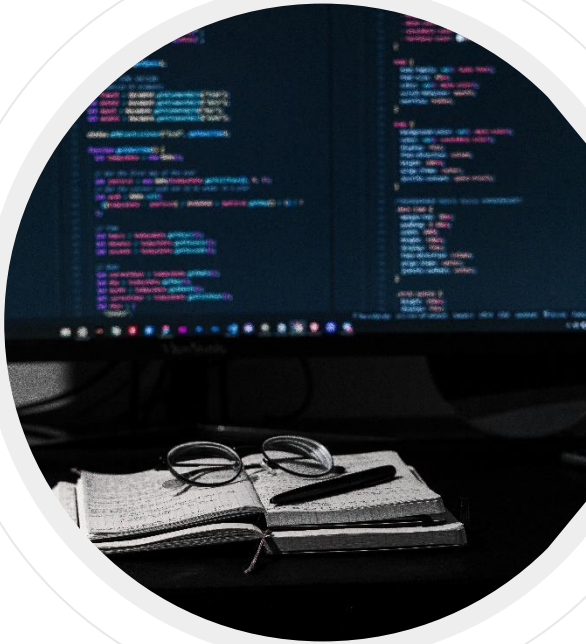
CONCLUSION

- By adding demographic features, we observed notable improvement in the prediction's accuracy.
- The obtained results showed the use value of both machine learning models. ***The Logistic Regression model showed slightly better performance than the decision tree model***, but definitely, both models have shown that they can be very successful in solving classification problems.
- Using the right framework and effective analysis, Click through rate prediction can be a powerful tool to beat the competition in advertising.



FUTURE SCOPE

Having full access to the original dataset would be interesting. Explore and manipulate features would undoubtedly bring us exciting and robust insights to optimize model performance. Thus, the feature selecting step would be highly accurate. Deal with imbalanced data is considered a real-world data science problem. Exploring further methods would be interesting. The ***deep neural network*** may be a interesting for the further future study of the CTR prediction.



REFERENCES

1. SEN ZHANG, QIANG FU, WENDONG XIAO (2017). *Advertisement Click-Through Rate Prediction Based on the Weighted-ELM and Adaboost Algorithm. Scientific Programming*, vol. 2017, Article ID 2938369. https://www.hindawi.com/journals/sp/2017/2938369/ [1].

2. PRASHANT GUPTA (2017). *Decision Trees in Machine Learning*. https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052 [2].

3. AYUSH PANT (2019). *Introduction to Logistic Regression.* https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148 [3].

4. BADREESH SHETTY (2018). *Natural Language Processing(NLP) for Machine Learning.* https://towardsdatascience.com/natural-language-processing-nlp-for-machine-learning-d44498845d5b [4].

5. AURELIEN GERON (2017). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems.*

6. EDWARD LOPER, EWAN KLEIN, AND STEVEN BIRD (2009). *Natural Language Processing with Python.*

Thank You