PREDICTION OF CTR OF EMAIL CAMPIGN

Submitted by:

Ujjwal Pratik

Contents

ACKNOWLEDGMENT

I sincerely thanks to the Analytics Vidya for this opportunity. Under their guidance I learned a lot about this project. I had also taken help from YouTube & online videos.

INTRODUCTION

PROBLEM STATEMENT

Most organizations today rely on email campaigns for effective communication with users. Email communication is one of the popular ways to pitch products to users and build trustworthy relationships with them. Email campaigns contain different types of CTA (Call To Action). The **goal** of email campaigns is to maximize the Click Through Rate (CTR). CTR is a measure of success for email campaigns. The higher the click rate, the better your email marketing campaign is. CTR is calculated by the no. of users who clicked on at least one of the CTA divided by the total no. of users the email was delivered **to. CTR** = No. of users who clicked on at least one of the CTA / No. of emails delivered**.** CTR depends on multiple factors like design, content, personalization, etc.

OBJECTIVE

Task at hand is to build a machine learning-based approach to predict the CTR of an email campaign

DATA DESCRIPTION

- There are two dataset training and test data.
- The source of data is taken from Analytics Vidya.
- The information of past email campaigns containing the email attributes like subject and body length, no. of CTA, date and time of an email, type of the audience, whether its a personalized email or not, etc and the target variable indicating the CTR.

METHEDOLOGY

- It gives insights of the dependency of target variables on independent variables using machine learnings techniques to determine the CTR because it gives the best outcome.

METRIC USAGE

a. Linear Regression.
b. Lasso Regularization.
c. Ridge Regularization.
d. Elastic Net Regularization.
e. Random Forest Regressor.

SYSTEM REQUIREMENTS

Hardware and Software Requirements and Tools Used

a) Hardware Requirement:
   i. Intel core i5
   ii. 8 GB Ram

b) Software Requirement:
   i. Python 3.x with packages:
      1. Pandas: Data analysis and manipulation tool
      2. NumPy: Provide support for mathematical functions, random number etc.
      3. Matplotlib: is a low-level graph plotting library in python that serves as a visualization.
      4. Seaborn: is a library mostly used for statistical plotting in python.
      5. Scikit-Learn: is an open-source Python library that has powerful tools for data analysis and data mining.

APPROACH

IDENTIFICATION OF POSSIBLE PROBLEM-SOLVING APPROACHES

- R2 score: is used to evaluate the performance of a linear regression mode.
- Linear Regression: Logistic regression is fast and relatively uncomplicated, and it is convenient for you to interpret the results.
- Lasso: The Lasso is a linear model that estimates sparse coefficients with l1 regularization.
- Ridge: Ridge regression is an extension of linear regression where the loss function is modified to minimize the complexity of the model.
- Elastic Net: is a linear regression model trained with both l1 and l2 -norm regularization of the coefficients.
- Cross-Validation-Score: a model that would just repeat the labels of the samples that it has just seen would have a perfect score but would fail to predict anything useful on yet-unseen data.
- Grid Search CV: This function helps to loop through predefined hyperparameters and fit your estimator (model) on your training set.
- Mean Squared Error: this metric gives an indication of how good a model fits a given dataset.
- Root Mean Squared error: is a frequently used measure of the differences between values (sample or population values) predicted by a model or an estimator and the values observed.
- Z-score: Z-score is also known as standard score gives us an idea of how far a data point is from the mean.
- Label Encoder: Label Encoding refers to converting the labels into numeric form.
- K Nearest Regressor: It observes features of an object and trains a model in the structure of a tree to predict data in the future to produce meaningful continuous output. Continuous output means that the output/result is not discrete, i.e., it is not represented just by a discrete, known set of numbers or values.
- Standard Scaler: Standard Scaler. Standard Scaler helps to get standardized distribution, with a zero mean and standard deviation of one (unit variance).
- Random Forest Regressor: A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

TESTING OF IDENTIFIED APPROACH(Algorithms)

a. Train Test Split
b. Linear Regression
c. Lasso Regularization
d. Ridge Regularization
e. Elastic Net Regularization
f. Grid Search CV
g. Cross Validation
h. Random Forest Regressor

KEY FOR SUCCESS IN SOLVING PROBLEM UNDER CONSIDERATION

- Analysed data for any unique values.
- Extraction information from some columns and made another variable from them.
- Analysed data for distribution.
- Caparison between two variables.
- Checked outliers through z-score method.
- checked skewness present in the dataset.
- Done Standard Scaling.
- Cross validate the r2 score from overfitting.
- Hyper Parameter tuning using Grid Search CV

CONCLUSION

This study shows that it is feasible to predict the CTR based on historical data. R2 Score is 41.64% & Cross val scores is 99.49%

LEARNING OUTCOMES OF THE STUDY IN RESPECT OF DATA SCIENCE

This study gives me opportunity for lots of learning starting from various types of plotting like histograms, boxplot, scatterplot, line chart and many more graphs. These graphs helped me to analyse different aspects of data like outlier, skewness, correlation etc. It also helped me to learn how to apply various model techniques on data and enable predications.

REFRENCES

- Analytics Vidya course videos.
- Google Search
- YouTube
- GitHub
- UCI Machine learning repository