Assignment 5 - Logistic regression

INSTRUCTIONAL DETAILS

Logistic regression is a statistical model that is used for classification tasks. It is a type of supervised learning algorithm that takes input data and learns to predict a binary outcome (a value of 0 or 1) based on the features of the data.

Here is an example of how you might implement logistic regression in Python using the scikit-learn library:

```
from sklearn.linear_model import LogisticRegression

# Load the data
X = # Features
y = # Target labels (0 or 1)

# Split the data into training and test sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)

# Create a logistic regression model
logistic_regression_model = LogisticRegression()

# Train the model on the training data
logistic_regression_model.fit(X_train, y_train)

# Make predictions on the test data
predictions = logistic_regression_model.predict(X_test)

# Calculate the accuracy of the model
accuracy = logistic_regression_model.score(X_test, y_test)
```

This example assumes that you have the features (X) and target labels (y) for your data, and that you have split the data into training and test sets using the train_test_split function from scikit-learn. The LogisticRegression model is then trained on the training data using the fit method, and the model's accuracy is calculated using the score method, which compares the model's predictions to the actual labels in the test set.

To interpret the output of a logistic regression model, you can first apply a threshold to the predicted probabilities. For example, you might choose a threshold of 0.5, which means that any prediction with a probability greater than 0.5 will be classified as belonging to the positive class, and any prediction with a probability less than 0.5 will be classified as belonging to the negative class.

You can then evaluate the performance of the model using various metrics, such as accuracy, precision, recall, and the F1 score. These metrics can help you understand how well the model is able to predict the correct class for a given input.

It's also important to consider the context in which the model is being used, as well as the consequences of incorrect predictions. For example, if you are using the model to predict the likelihood of a medical condition, you might want to prioritize high recall (the ability to correctly identify all cases of the condition) over high precision (the ability to correctly classify only the positive cases).

Overall, interpreting the output of a logistic regression model requires considering both the predicted probabilities and the evaluation metrics, as well as the specific context in which the model is being used.

BACKGROUND

Logistic regression is a statistical technique used to predict the probability of an outcome belonging to a particular class. It is a useful tool for businesses because it allows them to understand the factors that influence the likelihood of a particular outcome.

For example, a business might use logistic regression to understand the factors that influence the likelihood of a customer making a purchase. This might include factors such as the customer's age, income, and location, as well as factors related to the product or service being offered, such as price and features. Understanding these factors can inform marketing and sales strategies and help the business improve its conversion rate.

Logistic regression can also be used to predict the probability of an outcome occurring in the future. For example, a business might use logistic regression to predict the probability of a customer making a purchase based on their past behavior and other relevant factors.

Overall, logistic regression is an important tool for businesses because it helps them understand the factors that influence the likelihood of a particular outcome and predict the probability of that outcome occurring in the future. This can inform decision-making and help businesses achieve their goals.

RESEARCH QUESTION

Data analysis can be a valuable tool in lead scoring, which is the process of assigning a numeric value to a lead based on its likelihood of becoming a customer. Here are a few steps that organizations can follow to use data analysis in lead scoring:

Identify relevant data: The first step in lead scoring is to identify the data that is most relevant to predicting the likelihood of a lead becoming a customer. This may include data such as the lead's demographics, their behavior on the organization's website or social media channels, and their interactions with the organization's sales and marketing efforts.

Collect and clean the data: The next step is to collect and clean the data that has been identified as relevant. This may involve gathering data from a variety of sources, such as customer databases, website analytics, and social media analytics.

Analyze the data: Once you have collected and cleaned the data, you can use data analysis techniques such as visualization and statistical analysis to identify trends and patterns in the data.

Set scoring criteria: Based on the trends and patterns identified in the data, you can set criteria for assigning a score to each lead. This may involve assigning different weights to different factors based on their importance in predicting the likelihood of a lead becoming a customer.

Monitor and adjust: Finally, it is important to regularly monitor the lead scoring system and make adjustments as needed. Data analysis can be used to track the performance of the lead scoring system and identify areas for improvement.

REQUIREMENTS FOR SUBMISSION

GitHub
<XXXX>
Write-up
<XXXX>
Syntax
<XXXX>

Data
<XXXX>

FORMATTING

See “Assignment 1 - Descriptives” for a detailed list of assignment formatting guidelines. Also, assignment formatting guidelines can be found in the course document cache.

DATASET DETAILS (all data sets can be found here or here)

leadscoring.csv

DATASET FIELDS

organization_id
data_category
purchased
num_employees
first_call_duration_mins
industry
dm_familiarity
dm_perception
dm_awareness
existing_customer
current_spend
company_age_months
trial