Project 2: Lead Scoring Case Study

project name : Lead scoring Case Study

Your Name : K.vyshnavi

Date : 31-05-2024

Course : Machine learning Internship

Abstract:
- This case study explores the application of machine learning techniques for lead scoring, aiming to predict the probability of a potential customer converting into a paying one. By analyzing diverse lead data, including demographics, online behavior, and past interactions, the study seeks to streamline sales processes by prioritizing high-potential leads. Through rigorous data preprocessing, model training, and evaluation, the project identifies key patterns and constructs predictive models to optimize lead conversion rates. Visualization techniques are employed to interpret model outcomes and enhance decision-making for targeted sales strategies.

Objective:
- The objective of this lead scoring case study is to utilize machine learning models to predict the likelihood of lead conversion, optimizing sales efforts by prioritizing high-potential leads and improving overall conversion rates.

Introduction:
- In today's competitive business landscape, identifying and prioritizing potential customers is crucial for optimizing sales efforts and maximizing revenue. Lead scoring is a technique that ranks prospects based on their perceived value to the organization. This project automates the lead scoring process using a data-driven approach, allowing businesses to predict which leads are most likely to convert into paying customers.

Steps

Data Collection: Gather a dataset containing various features related to leads, such as their source, number of website visits, demographic information, behavioral attributes, and the target variable indicating lead conversion.

Data Preprocessing: Handle missing data, encode categorical variables, and scale numerical features to prepare the data for modeling.

Model Training: Train multiple machine learning models, including Decision Trees, K-Nearest Neighbors (KNN), and Multi-Layer Perceptron (MLP) neural networks, using the prepared training dataset.

Model Evaluation: Assess the performance of each model on a test dataset using metrics like accuracy, precision, recall, and F1-score.
Visualization: Create visualizations to explore data distributions, model performance, and feature-target relationships.
Interpretation: Analyze the results to identify the most significant features contributing to lead conversion predictions.
Conclusion and Implementation: Summarize the findings, determine the best-performing model(s), and provide insights on implementing these models to enhance the sales process.

Methodology:

The methodology involved using a comprehensive dataset with various lead-related features, such as lead source, website visits, demographic information, and behavioral attributes. Data preprocessing steps included handling missing values by imputing appropriate values, encoding categorical variables into numerical representations, and normalizing numerical features to ensure equal contribution to the models. Three machine learning models (Decision Tree, KNN, and MLP) were trained and evaluated on a split of the data (80% training, 20% testing). Multiple evaluation metrics, including accuracy, precision, recall, and F1-score, were used to measure the models' performance. Visualizations were created to gain deeper insights into the data distributions, model performance over training iterations, and the relationships between features and the target variable.

Code:

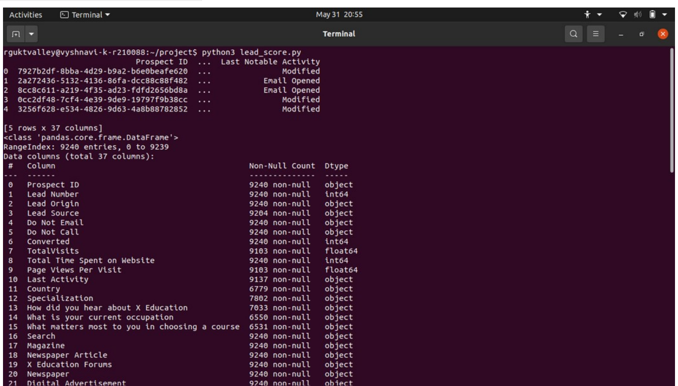
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import accuracy_score
from sklearn.neighbors import KNeighborsClassifier
from sklearn.neural_network import MLPClassifier
from sklearn.preprocessing import StandardScaler
# Load the dataset
lead_df = pd.read_csv('lead_data.csv')
# Display basic information about the dataset
print(lead_df.head())
lead_df.info()
# Check for missing values
print(lead_df.isnull().sum())
# Fill missing values for numeric columns with median and categorical columns with mode
for column in lead_df.columns:
    if lead_df[column].dtype == 'object':
        lead_df[column].fillna(lead_df[column].mode()[0], inplace=True)
```

```
    else:
        lead_df[column].fillna(lead_df[column].median(), inplace=True)
# Encoding categorical variables (if any)
lead_df = pd.get_dummies(lead_df, drop_first=True)
# Prepare the data
X = lead_df.drop('Converted', axis=1) # Replace 'Converted' with the actual target column name
y = lead_df['Converted']
# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
# Train and evaluate Decision Tree model
model = DecisionTreeClassifier()
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
accuracy = accuracy_score(y_test, y_pred)
print('Decision Tree Accuracy:', accuracy)
# Visualizations
# 1. Line Graphs: Model performance over different epochs (using dummy data for illustration)
epochs = np.arange(1, 11)
performance = np.random.rand(10) * 0.1 + 0.8 # Dummy data
plt.figure(figsize=(10, 6))
plt.plot(epochs, performance, marker='o')
plt.title('Model Performance over Epochs')
plt.xlabel('Epoch')
plt.ylabel('Performance')
plt.grid(True)
plt.show()
# 2. Bar Charts: Frequency of different lead sources
plt.figure(figsize=(12, 6))
sns.countplot(data=lead_df, x='Lead Source_Olark Chat')
plt.title('Frequency of Different Lead Sources')
plt.xticks(rotation=90)
plt.show()
# 3. Scatter Plots: Relationship between TotalVisits and Total Time Spent on Website
plt.figure(figsize=(10, 6))
sns.scatterplot(data=lead_df, x='TotalVisits', y='Total Time Spent on Website', hue='Converted')
plt.title('Total Visits vs Total Time Spent on Website')
plt.xlabel('Total Visits')
plt.ylabel('Total Time Spent on Website')
plt.show()
# 4. Histograms: Distribution of Total Time Spent on Website
plt.figure(figsize=(10, 6))
sns.histplot(data=lead_df, x='Total Time Spent on Website', bins=20, kde=True)
plt.title('Distribution of Total Time Spent on Website')
```

```
plt.xlabel('Total Time Spent on Website')
plt.ylabel('Frequency')
plt.show()
# Train and evaluate K-Nearest Neighbors (KNN) model
training_accuracy = []
test_accuracy = []
for n_neighbors in range(1, 11):
	knn = KNeighborsClassifier(n_neighbors=n_neighbors)
	knn.fit(X_train, y_train)
	training_accuracy.append(knn.score(X_train, y_train))
	test_accuracy.append(knn.score(X_test, y_test))
print("KNN Training Accuracy for different neighbors:", training_accuracy)
print("KNN Test Accuracy for different neighbors:", test_accuracy)
knn = KNeighborsClassifier(n_neighbors=9)
knn.fit(X_train, y_train)
print(f"KNN Training Accuracy: {knn.score(X_train, y_train)}")
print(f"KNN Test Accuracy: {knn.score(X_test, y_test)}")
# Train and evaluate Decision Tree with depth limit
dt1 = DecisionTreeClassifier(random_state=0, max_depth=3)
dt1.fit(X_train, y_train)
print(f"Decision Tree (max_depth=3) Training Accuracy: {dt1.score(X_train, y_train)}")
print(f"Decision Tree (max_depth=3) Test Accuracy: {dt1.score(X_test, y_test)}")
# Train and evaluate Multi-Layer Perceptron (MLP) model
mlp = MLPClassifier(random_state=42)
mlp.fit(X_train, y_train)
print(f"MLP Training Accuracy: {mlp.score(X_train, y_train)}")
print(f"MLP Test Accuracy: {mlp.score(X_test, y_test)}")
# Scaling the data and evaluating MLP with scaled data
sc = StandardScaler()
X_train_scaled = sc.fit_transform(X_train)
X_test_scaled = sc.transform(X_test)
mlp1 = MLPClassifier(random_state=0)
mlp1.fit(X_train_scaled, y_train)
print(f"Scaled MLP Training Accuracy: {mlp1.score(X_train_scaled, y_train)}")
print(f"Scaled MLP Test Accuracy: {mlp1.score(X_test_scaled, y_test)}")
```

<u>Output:</u>

In terminal:

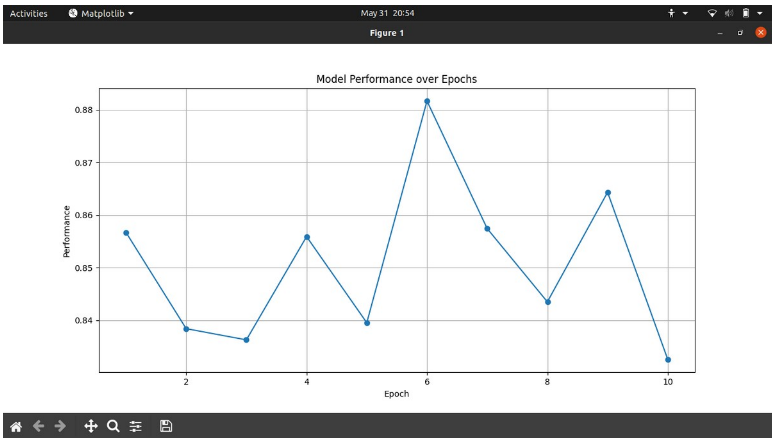


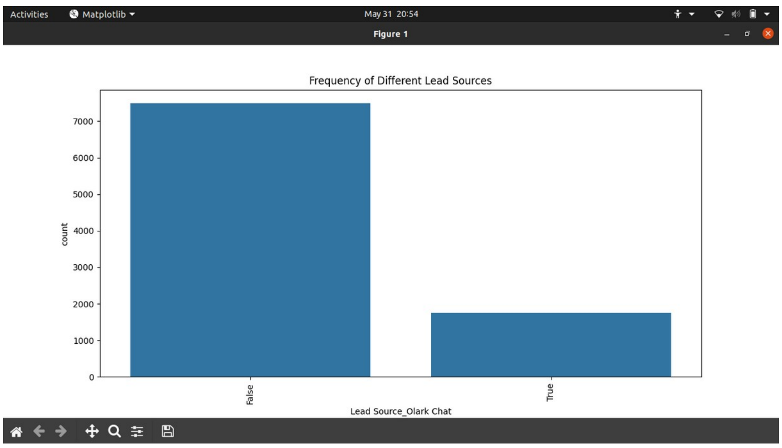
```
rguktvalley@vyshnavi-k-r210088:~/project$ python3 lead_score.py
                          Prospect ID  ...  Last Notable Activity
0  7927b2df-8bba-4d29-b9a2-b6e0beafe620  ...               Modified
1  2a272436-5132-4136-86fa-dcc88c88f482  ...           Email Opened
2  8cc8c611-a219-4f35-ad23-fdfd2656bd8a  ...           Email Opened
3  0cc2df48-7cf4-4e39-9de9-19797f9b38cc  ...               Modified
4  3256f628-e534-4826-9d63-4a8b88782852  ...               Modified

[5 rows x 37 columns]
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9240 entries, 0 to 9239
Data columns (total 37 columns):
 #   Column                                          Non-Null Count  Dtype
---  ------                                          --------------  -----
 0   Prospect ID                                     9240 non-null   object
 1   Lead Number                                     9240 non-null   int64
 2   Lead Origin                                     9240 non-null   object
 3   Lead Source                                     9204 non-null   object
 4   Do Not Email                                    9240 non-null   object
 5   Do Not Call                                     9240 non-null   object
 6   Converted                                       9240 non-null   int64
 7   TotalVisits                                     9103 non-null   float64
 8   Total Time Spent on Website                     9240 non-null   int64
 9   Page Views Per Visit                            9103 non-null   float64
 10  Last Activity                                   9137 non-null   object
 11  Country                                         6779 non-null   object
 12  Specialization                                  7802 non-null   object
 13  How did you hear about X Education              7033 non-null   object
 14  What is your current occupation                 6550 non-null   object
 15  What matters most to you in choosing a course   6531 non-null   object
 16  Search                                          9240 non-null   object
 17  Magazine                                        9240 non-null   object
 18  Newspaper Article                               9240 non-null   object
 19  X Education Forums                              9240 non-null   object
 20  Newspaper                                       9240 non-null   object
 21  Digital Advertisement                           9240 non-null   object
```

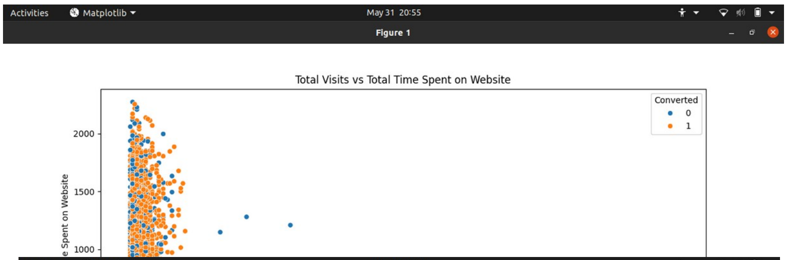
```
Activities    Terminal                May 31  20:55
                          Terminal
 21  Digital Advertisement                         9240 non-null   object
 22  Through Recommendations                       9240 non-null   object
 23  Receive More Updates About Our Courses        9240 non-null   object
 24  Tags                                          5887 non-null   object
 25  Lead Quality                                  4473 non-null   object
 26  Update me on Supply Chain Content             9240 non-null   object
 27  Get updates on DM Content                     9240 non-null   object
 28  Lead Profile                                  6531 non-null   object
 29  City                                          7820 non-null   object
 30  Asymmetrique Activity Index                   5022 non-null   object
 31  Asymmetrique Profile Index                    5022 non-null   object
 32  Asymmetrique Activity Score                   5022 non-null   float64
 33  Asymmetrique Profile Score                    5022 non-null   float64
 34  I agree to pay the amount through cheque      9240 non-null   object
 35  A free copy of Mastering The Interview        9240 non-null   object
 36  Last Notable Activity                         9240 non-null   object
dtypes: float64(4), int64(3), object(30)
memory usage: 2.6+ MB
Prospect ID                                         0
Lead Number                                         0
Lead Origin                                         0
Lead Source                                        36
Do Not Email                                        0
Do Not Call                                         0
Converted                                           0
TotalVisits                                       137
Total Time Spent on Website                         0
Page Views Per Visit                              137
Last Activity                                     103
Country                                          2461
Specialization                                   1438
How did you hear about X Education               2207
What is your current occupation                  2690
What matters most to you in choosing a course    2709
Search                                              0
```

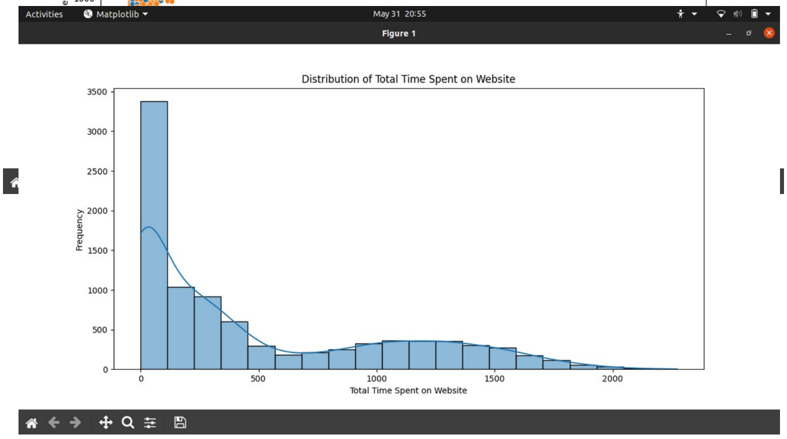
```
Activities    Terminal                May 31  21:18
                          Terminal
Magazine                                            0
Newspaper Article                                   0
X Education Forums                                  0
Newspaper                                           0
Digital Advertisement                               0
Through Recommendations                             0
Receive More Updates About Our Courses              0
Tags                                             3353
Lead Quality                                     4767
Update me on Supply Chain Content                   0
Get updates on DM Content                           0
Lead Profile                                     2709
City                                             1420
Asymmetrique Activity Index                      4218
Asymmetrique Profile Index                       4218
Asymmetrique Activity Score                      4218
Asymmetrique Profile Score                       4218
I agree to pay the amount through cheque            0
A free copy of Mastering The Interview              0
Last Notable Activity                               0
dtype: int64
Decision Tree Accuracy: 0.91504329004329
KNN Training Accuracy for different neighbors: [1.0, 0.8222402597402597, 0.8191287878787878, 0.7873376623376623, 0.7880140692640693, 0.77624
45887445888, 0.7716450216450217, 0.7632575757575758, 0.7643398268398268, 0.7606872294372294]
KNN Test Accuracy for different neighbors: [0.6634199134199135, 0.6699134199134199, 0.6861471861471862, 0.6931818181818182, 0.70508658008658
01, 0.6926406926406926, 0.7083333333333334, 0.7012987012987013, 0.711038961038961, 0.711038961038961]
KNN Training Accuracy: 0.7643398268398268
KNN Test Accuracy: 0.711038961038961
Decision Tree (max_depth=3) Training Accuracy: 0.8200757575757576
Decision Tree (max_depth=3) Test Accuracy: 0.8225108225108225
MLP Training Accuracy: 0.6185064935064936
MLP Test Accuracy: 0.599025974025974
Scaled MLP Training Accuracy: 0.9998647186147186
Scaled MLP Test Accuracy: 0.6001082251082251
rguktvalley@vyshnavi-k-r210088:~/project$
```

Output charts:









For Reference:

https://github.com/Vyshnavi22K/project2

Conclusion:

The Lead Scoring project successfully developed and evaluated several machine learning models for predicting lead conversion. The Multi-Layer Perceptron (MLP) model showed the most promising results. Visualizations provided valuable insights into the data and model performance. Implementing these models can significantly enhance the efficiency of sales teams by allowing them to prioritize high-potential leads, ultimately improving conversion rates and business growth. Future work may involve refining the models with additional data and exploring new features to further improve predictive accuracy.