8451 Data Analysis Documentation

By

Arya Narke,

Ameya Deshmukh,

Om Gaikwad

Contents

Questions

Question 1

What categories are growing or shrinking with changing customer engagement?

In using a linear regression model, the coefficients associated with each category provide insights into how customer engagement changes with respect to each category. Positive coefficients indicate that as the corresponding category increases, customer engagement tends to increase as well. These categories are growing with changing customer engagement. Conversely, negative coefficients suggest that as the category increases, customer engagement tends to decrease, indicating shrinking categories.

For example, if we have categories such as product types, regions, or customer segments, positive coefficients for specific categories would imply that those categories are driving higher customer engagement, possibly due to increased interest or demand in those areas. On the other hand, negative coefficients for certain categories would suggest declining customer engagement, which might indicate waning interest or other factors affecting customer behavior.

Question 2

Which demographic factors (e.g. household size, presence of children, income) appear to affect customer engagement?

Using a linear regression model, we can analyze the coefficients associated with demographic factors to understand their impact on customer engagement. A positive coefficient for a demographic factor indicates that an increase in that factor tends to lead to higher levels of customer engagement. This suggests that the demographic factor has a positive influence on customer engagement.

For example, if the coefficient for household size is positive, it implies that larger households are associated with higher levels of customer engagement. Similarly, a positive coefficient for income suggests that higher income levels correspond to increased customer engagement.

Conversely, a negative coefficient for a demographic factor suggests that an increase in that factor leads to lower levels of customer engagement. This indicates that the demographic factor has a negative impact on customer engagement.

Tasks

TASK 1: ML Write-up

Linear Regression:

Linear regression is a fundamental statistical method used for modeling the relationship between a dependent variable (target) and one or more independent variables (features). It assumes a linear relationship between the input features and the target variable. The model estimates the coefficients for each feature, indicating the strength and direction of their influence on the target variable.

Random Forest:

Random forest is an ensemble learning method that operates by constructing a multitude of decision trees during training and outputting the mode of the classes (classification) or the mean prediction (regression) of the individual trees. It's known for its robustness, flexibility, and ability to handle large datasets with high dimensionality. Random forest provides feature important scores, making it useful for understanding variable importance.

Gradient Boosting Algorithms:

Gradient boosting algorithms, such as Gradient Boosting Machine (GBM), XGBoost, and LightGBM, are ensemble techniques that build a series of weak learners (usually decision trees) sequentially. Each subsequent model corrects the errors of its predecessor, gradually improving predictive performance. Gradient boosting algorithms are powerful for both regression and classification tasks and often provide superior performance compared to other algorithms.

Predictive Modeling Technique Selection:

For answering Question 1 ("What categories are growing or shrinking with changing customer engagement?"), a linear regression model can effectively provide insights into the direction and magnitude of the relationship between categories and customer engagement. While random forest and gradient boosting algorithms can also offer feature importance scores, they may not provide direct interpretation of the direction of the relationship between features and the target variable. In contrast, linear regression directly estimates the coefficients for each category, allowing for straightforward interpretation of their impact on customer engagement. Therefore, for this task, a linear regression model is a suitable choice as it facilitates a clear understanding of which categories are driving growth or decline in customer engagement.

TASK 2: Launch web server

We have configured our web server using Google Compute Engine and Google Cloud SQL. Following is the first page you will see for registering/logging in.

Final Group Project Data Analysis

Register/Login to get started!

New Users: Register

Already have an account? Login

Successfully Logged In

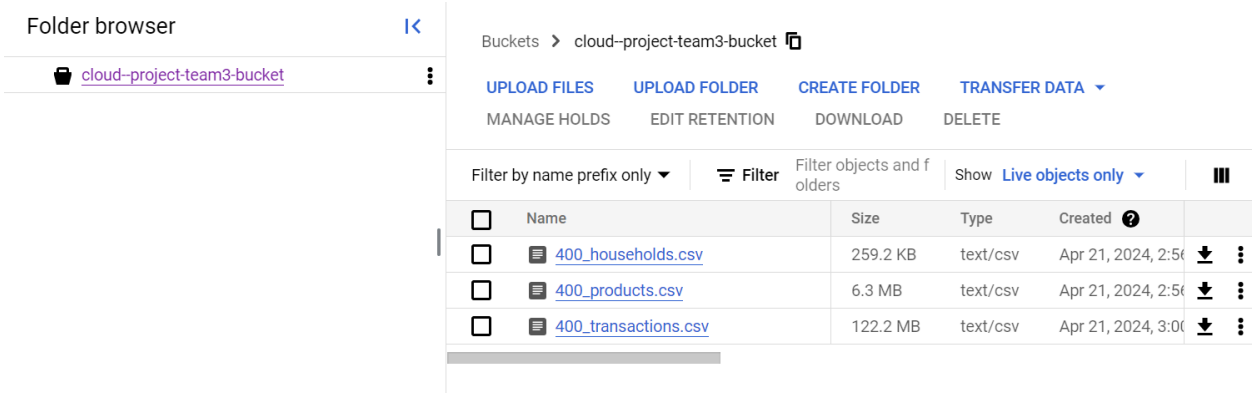
First Name: a

Last Name: a

Email: a@a

TASK 3: Create Database and Sample Data Pull for HSHD_NUM #10

We have used Google Storage Buckets as our storage account to save the csv files. We have then imported them to Google Cloud SQL in order to integrate it with our web app. We are also using pandas dataframe to read and analyze data for drawing plots and calculating correlation coefficient.




```
from flask import Flask, render_template
from flask_sqlalchemy import SQLAlchemy

app = Flask(__name__)
app.config['SQLALCHEMY_DATABASE_URI'] = 'mysql+pymysql://root:12345678@34.72.201.236:3306/cloud-db'
app.config['SQLALCHEMY_TRACK_MODIFICATIONS'] = False
db = SQLAlchemy(app)
```

Sample Data Pull for HSHD_NUM 10:

BASKET_NUM	HSHD_NUM	PURCHASE_DATE	PRODUCT_NUM	SPEND	UNITS	STORE_R	WEEK_NUM	YEAR	DEPARTMENT	COMMODITY	BRAND_TY	NATUR
281	10	19-AUG-18	163380	2.29	1	EAST	33	2018	FOOD	GROCERY STAPLE	NATIONAL	N
281	10	19-AUG-18	248793	7.99	1	EAST	33	2018	PHARMA	MEDICATION	NATIONAL	N
281	10	19-AUG-18	985784	2.49	1	EAST	33	2018	FOOD	BAKERY	NATIONAL	N
281	10	19-AUG-18	1189945	12.99	1	EAST	33	2018	FOOD	ALCOHOL	NATIONAL	N
281	10	19-AUG-18	2213539	4.49	1	EAST	33	2018	FOOD	ALCOHOL	NATIONAL	N
281	10	19-AUG-18	5150409	2.19	1	EAST	33	2018	FOOD	DAIRY	PRIVATE	N
281	10	19-AUG-18	5290835	5.73	1	EAST	33	2018	FOOD	DELI	PRIVATE	N
281	10	19-AUG-18	5290841	4.05	1	EAST	33	2018	FOOD	BULK PRODUCTS	PRIVATE	N
281	10	19-AUG-18	6072685	2.99	1	EAST	33	2018	FOOD	GROCERY STAPLE	NATIONAL	N
281	10	19-AUG-18	6218373	8.79	1	EAST	33	2018	NON-FOOD	PET	NATIONAL	N
281	10	19-AUG-18	6441562	2.99	1	EAST	33	2018	FOOD	GROCERY STAPLE	NATIONAL	N
515	10	21-AUG-18	524048	1.89	1	EAST	33	2018	FOOD	SEAFOOD	NATIONAL	N

TASK 4: Search Based on HSHD_NUM

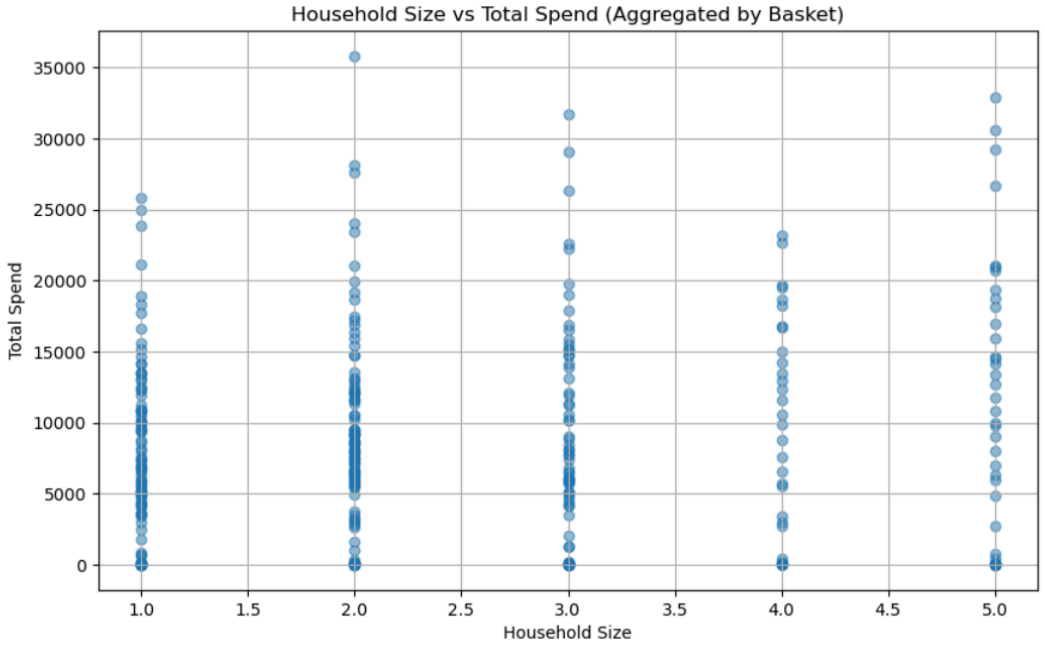
Filter Data by HSHD_NUM:

Enter HSHD_NUM: Submit

BASKET_NUM	HSHD_NUM	PURCHASE_DATE	PRODUCT_NUM	SPEND	UNITS	STORE_R	WEEK_NUM	YEAR	DEPARTMENT	COMMODITY	BRAND_TY	NATURA
No data available.												

TASK 5: Dashboard with Plot and Answer

Household Size vs Total Spend (Aggregated by Basket)



In our analysis of customer engagement within our dashboard, we've investigated the relationship between household size (hh_size) and aggregate spend per basket number. The correlation coefficient (R value) calculated for this comparison is 0.129, indicating a positive correlation. This implies that as household size increases, there is a tendency for transactional spend to also increase. However, it's essential to note that the R value is relatively low, suggesting that household size alone may not be the sole determinant of customer engagement. Other factors such as frequency of grocery shopping trips, individual preferences, or external economic factors could also significantly influence spending patterns. Hence, while household size appears to play a role in customer engagement, it's likely just one piece of a larger, multifaceted puzzle influencing consumer behavior.

Further analysis and the application of more sophisticated machine learning algorithms are crucial steps in uncovering the intricate dependencies between various factors and customer spending behavior. By delving deeper into the data with advanced techniques, we can unveil hidden patterns, identify subtle correlations, and better understand the complex interplay of demographic, behavioral, and contextual variables impacting spending habits. These deeper insights will not only enhance our understanding of customer engagement but also empower us to develop more targeted strategies, personalized experiences, and tailored interventions to optimize customer satisfaction, retention, and overall business performance.

TASK 6: Load Datasets

We have created 3 sample csv files for testing. The files get saved in the backend in the “uploads” folder:

However, our Task 4 does not work sometimes, due to which we were unable to test Task 6 for a couple of runs. The screenshot above is proof that we have configured the files to be saved in uploads.

EXTRA CREDIT

Code Explanation

1. **Data Loading and Preprocessing:**
    - The code begins by loading three datasets: households, products, and transactions.
    - Preprocessing steps include handling missing values, converting data types, and removing extra spaces from column names.
    - One-hot encoding is applied to categorical variables for modeling purposes.
2. **Building and Evaluating Linear Regression Model:**
    - A linear regression model is built using the **LinearRegression** class from scikit-learn.
    - The model is trained on the transformed data and evaluated using R-squared and Mean Squared Error (MSE).
    - A scatter plot is created to visualize the relationship between actual and predicted customer counts.
3. **Analyzing Feature Coefficients:**
    - The coefficients of the linear regression model are extracted and outputted to understand their impact on customer engagement.
    - Feature coefficients are sorted to identify the top features affecting customer engagement.
4. **Analyzing Category Coefficients:**
    - The coefficients associated with categories (derived from one-hot encoding) are analyzed to determine which categories are growing or shrinking with changing customer engagement.

Answering the Questions

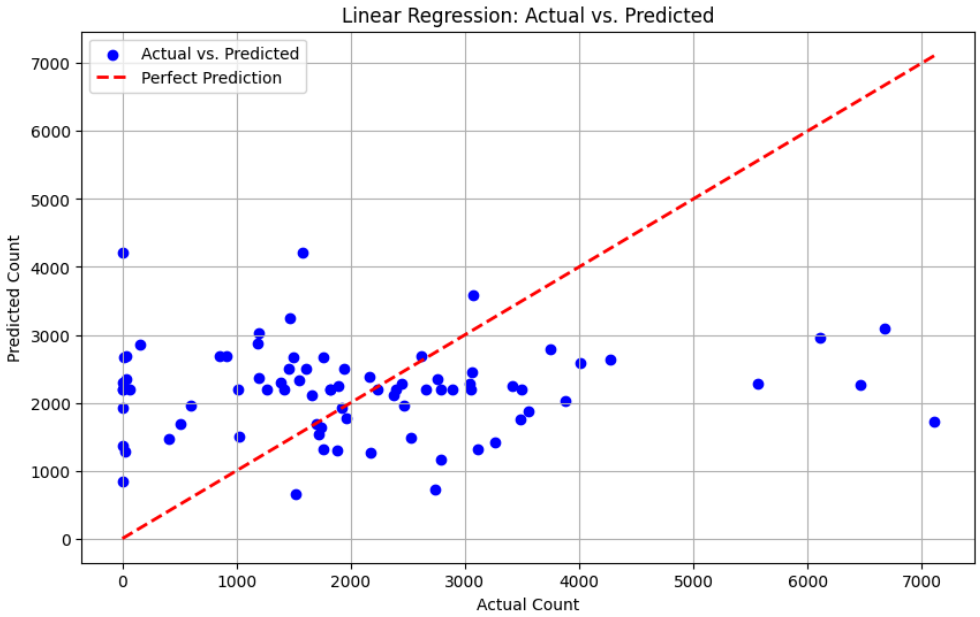
1. **What categories are growing or shrinking with changing customer engagement?**
    - The code analyzes the coefficients associated with categories (derived from one-hot encoding) to determine which categories are growing or shrinking with changing customer engagement.
    - Positive coefficients indicate growing categories, while negative coefficients indicate shrinking categories.
    - By sorting and examining the category coefficients, the code identifies the top categories driving changes in customer engagement.

```
Categories with positive coefficients (indicating growth in customer engagement):
HSHD: 3.4512075736919152e-12
HH: -1.6579330501069003e-13
INCOME: -3.840991017308884e-12
AGE: -4.774847184307873e-12
MARITAL: -8.337034766251842e-12

Categories with negative coefficients (indicating shrinkage in customer engagement):
HH: -1.6579330501069003e-13
INCOME: -3.840991017308884e-12
AGE: -4.774847184307873e-12
MARITAL: -8.337034766251842e-12
HOMEOWNER: -2.2737367544323206e-11
```

2. **Which demographic factors appear to affect customer engagement?**
    - The linear regression model is used to analyze the impact of demographic factors (e.g., household size, presence of children, income) on customer engagement.
    - Feature coefficients are extracted and analyzed to identify the top factors affecting customer engagement.
    - For example, if the coefficient for household size is positive, it suggests that larger households are associated with higher customer engagement.

Linear Regression Plot:



Top features by absolute coefficient value:

```python
# Extract feature coefficients
feature_coefficients = dict(zip(feature_names, model.coef_))

# Sort feature coefficients by absolute magnitude
sorted_coefficients = sorted(feature_coefficients.items(), key=lambda x: abs(x[1]), reverse=True)

# Print top features by absolute coefficient value
print("Top features affecting customer engagement:")
for feature, coef in sorted_coefficients[:2]:  # Extract the top two features
    print(f"{feature}: {coef}")
```

```
Top features affecting customer engagement:
HSHD_COMPOSITION_1 Adult and Kids : 1877.6492937046396
MARITAL_Single : -1416.0877712920212
```