CS5165 Cloud Computing Group Project Report

Varad Parte: partevr@mail.uc.edu

Daksh Prajapati: prajapdh@mail.uc.edu

Jalin Solankee: solankjp@mail.uc.edu

Github Repo link:
https://github.com/Prajapdh/CloudComputingFinalProject.git

Answer 1: Write-up on ML Models

Linear Regression

Linear Regression is a simple but powerful machine learning model utilized for the estimation of a continuous output from multiple or single input features. Linear Regression assumes there to be a linear relationship among inputs and output, i.e., it takes a straight line passing through data points that most minimizes the difference between the actual and forecasted values. While easy to comprehend and quick to run, it does not do well if relationships are complex or non-linear, and it is sensitive to outliers.

Random Forest

Random Forest is an ensemble method of learning which builds many decision trees and returns the class that is most commonly predicted by them. Each tree is trained on a random subset of the data and features, so the entire model is resistant to bias and less sensitive to it. Random Forest can do regression as well as classification tasks and is suitable for large datasets with complex patterns, although at times at the cost of interpretability and not performance.

Gradient Boosting

Gradient Boosting is a strong ensemble technique that builds models sequentially, with each successive model trying to make up for the mistakes of previous ones. Gradient Boosting focuses on hard-to-predict cases and achieves maximum performance with an algorithm called gradient descent. Thus, Gradient Boosting performs extremely well at high predictive accuracy but can overfit if it is not optimally tuned.

Selected Model for Retail CLV Prediction

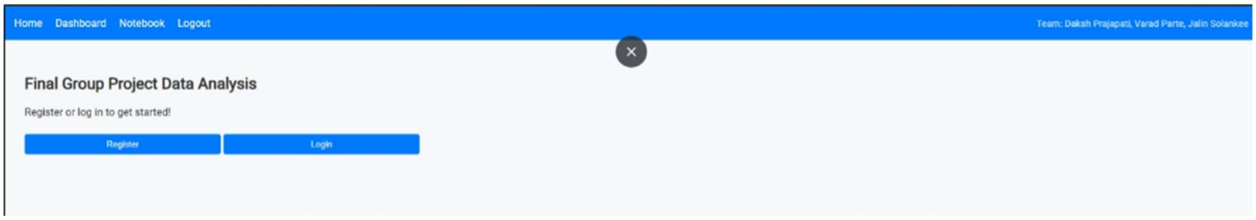
For predicting Customer Lifetime Value (CLV), Gradient Boosting is the optimal choice. CLV prediction involves the identification of intricate, non-linear patterns in customer behavior, such as purchase frequency, average order value, recency, and engagement metrics. Gradient Boosting models like

XGBoost or LightGBM are well-suited to identify such intricate relationships by building strong predictive models from a large number of weak, small models. In each step, the goal is to reduce the errors of the previous steps so that the final model is very precise. Gradient Boosting also handles various types of input features (numerical, categorical, etc.), is resilient to outliers, and can account for the influence of infrequent but high-value customers. While it must be properly tuned (e.g., to learning rates and tree depths), its precision power and flexibility make it well-suited for ranking and ordering customers by their long-run revenue potential.

Task 2

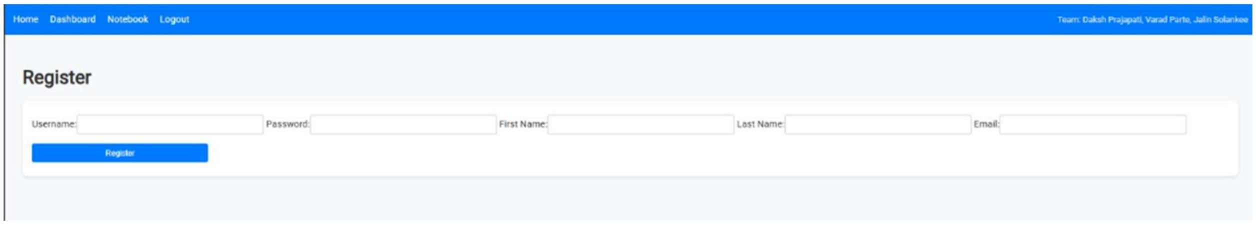
WebServer setup

The following is the register and login page for our webserver app



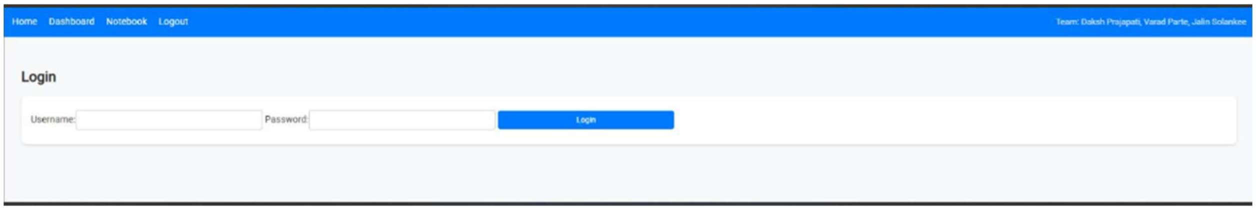
Final Group Project Data Analysis

Register or log in to get started!



Register

Username: Password: First Name: Last Name: Email:

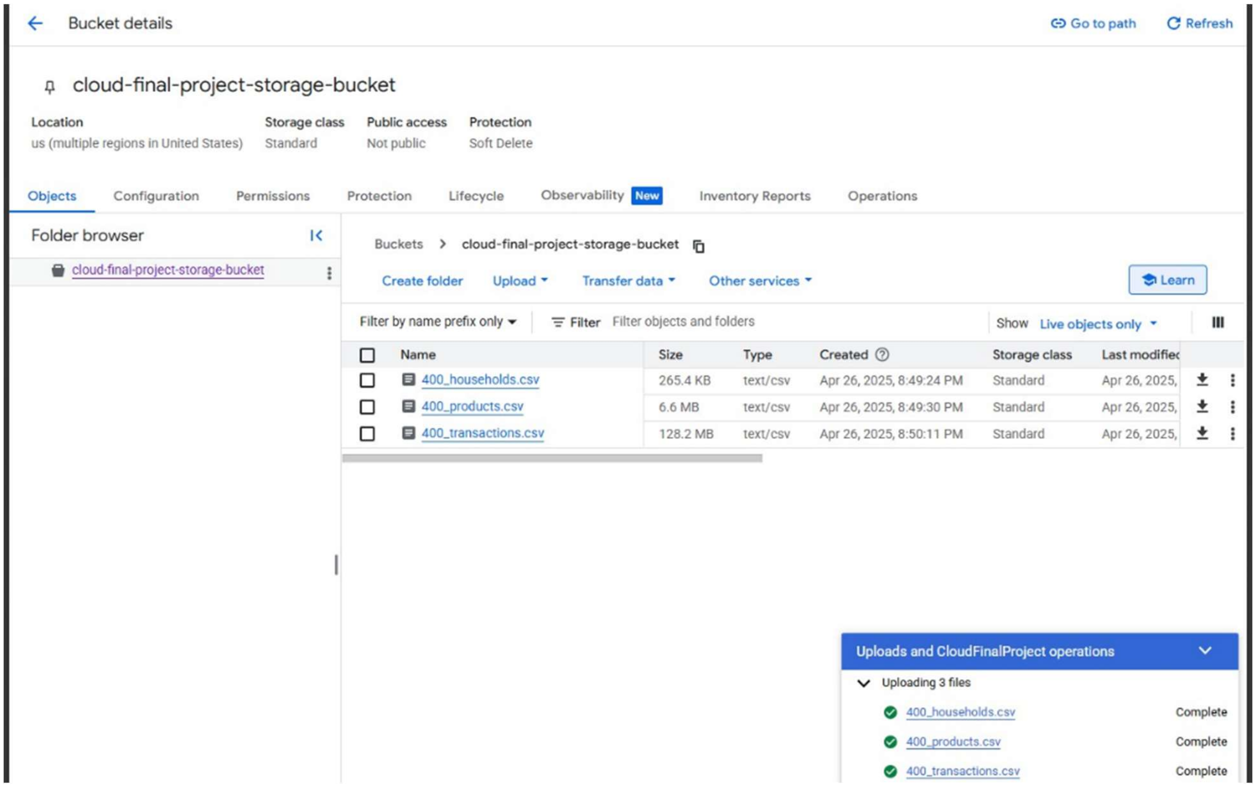


Login

Username: Password:

Task 3

Create Database and Sample data pull for HSHD_NUM



Bucket details

cloud-final-project-storage-bucket

Location	Storage class	Public access	Protection
us (multiple regions in United States)	Standard	Not public	Soft Delete

☐	Name	Size	Type	Created	Storage class	Last modifiec
☐	400_households.csv	265.4 KB	text/csv	Apr 26, 2025, 8:49:24 PM	Standard	Apr 26, 2025,
☐	400_products.csv	6.6 MB	text/csv	Apr 26, 2025, 8:49:30 PM	Standard	Apr 26, 2025,
☐	400_transactions.csv	128.2 MB	text/csv	Apr 26, 2025, 8:50:11 PM	Standard	Apr 26, 2025,

Uploads and CloudFinalProject operations

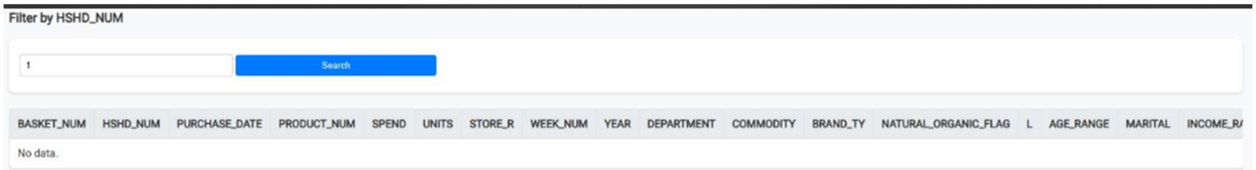
Uploading 3 files

400_households.csv	Complete
400_products.csv	Complete
400_transactions.csv	Complete

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

Task 4

Filter Data by HSHD_NUM



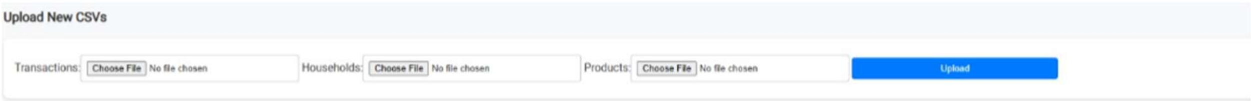
Filter by HSHD_NUM

1

BASKET_NUM	HSHD_NUM	PURCHASE_DATE	PRODUCT_NUM	SPEND	UNITS	STORE_R	WEEK_NUM	YEAR	DEPARTMENT	COMMODITY	BRAND_TY	NATURAL_ORGANIC_FLAG	L	AGE_RANGE	MARITAL	INCOME_R/
No data.																

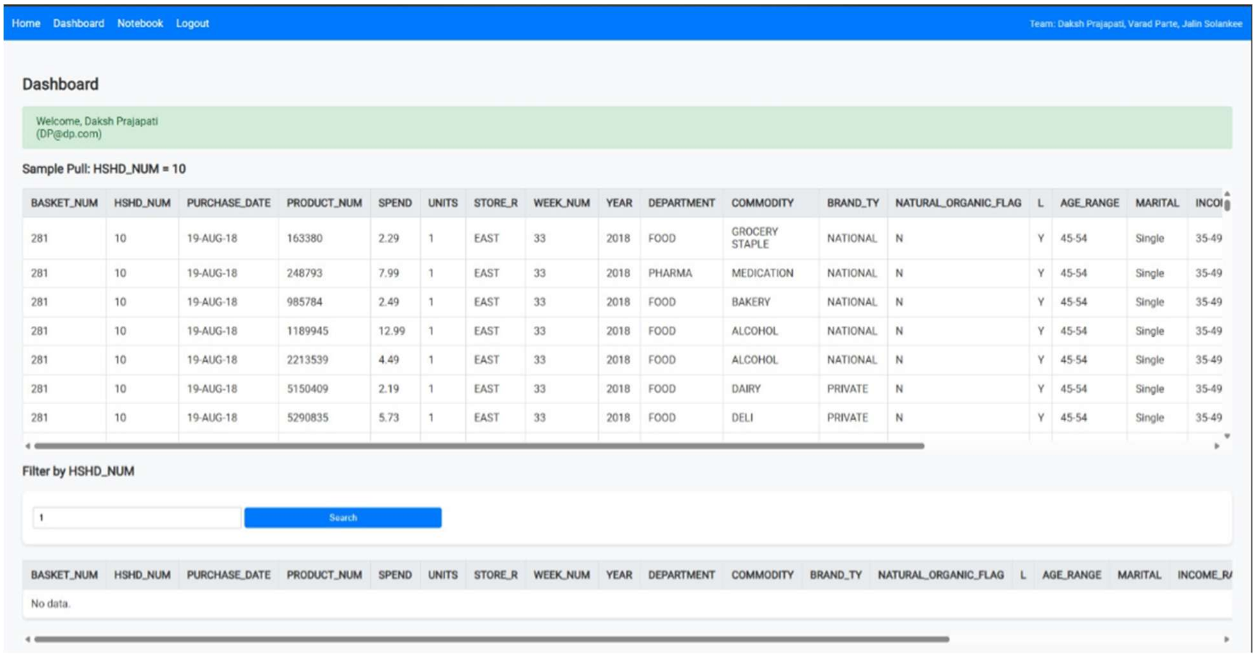
Task 5

Data loading web app



Task 6

Dashboard



Dashboard

Welcome, Daksh Prajapati
(DP@dp.com)

Sample Pull: HSHD_NUM = 10

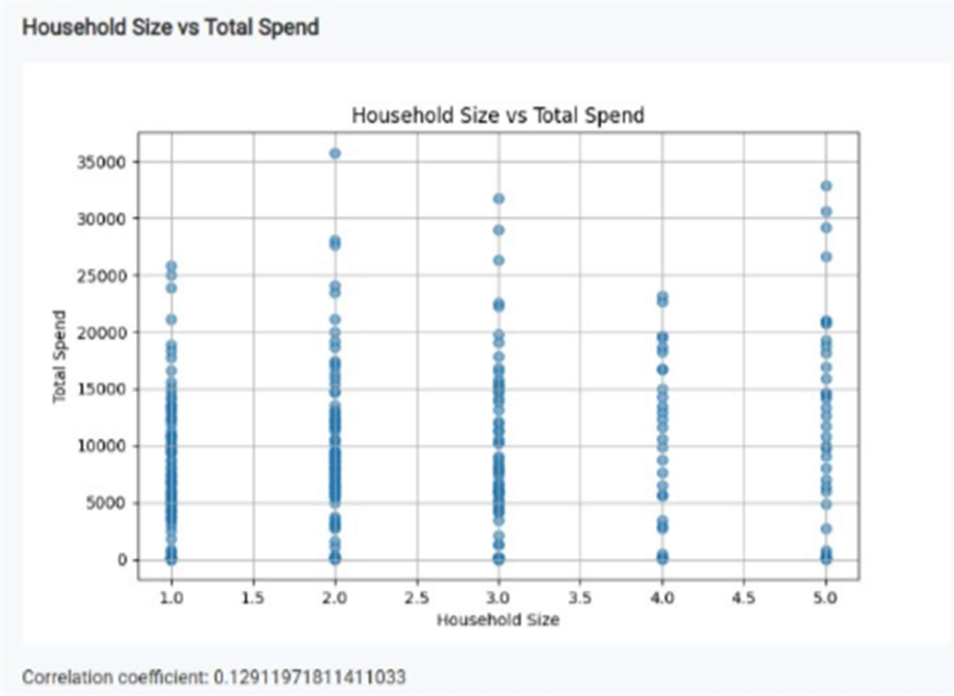
BASKET_NUM	HSHD_NUM	PURCHASE_DATE	PRODUCT_NUM	SPEND	UNITS	STORE_R	WEEK_NUM	YEAR	DEPARTMENT	COMMODITY	BRAND_TY	NATURAL_ORGANIC_FLAG	L	AGE_RANGE	MARITAL	INCOI
281	10	19-AUG-18	163380	2.29	1	EAST	33	2018	FOOD	GROCERY STAPLE	NATIONAL	N	Y	45-54	Single	35-49
281	10	19-AUG-18	248793	7.99	1	EAST	33	2018	PHARMA	MEDICATION	NATIONAL	N	Y	45-54	Single	35-49
281	10	19-AUG-18	985784	2.49	1	EAST	33	2018	FOOD	BAKERY	NATIONAL	N	Y	45-54	Single	35-49
281	10	19-AUG-18	1189945	12.99	1	EAST	33	2018	FOOD	ALCOHOL	NATIONAL	N	Y	45-54	Single	35-49
281	10	19-AUG-18	2213539	4.49	1	EAST	33	2018	FOOD	ALCOHOL	NATIONAL	N	Y	45-54	Single	35-49
281	10	19-AUG-18	5150409	2.19	1	EAST	33	2018	FOOD	DAIRY	PRIVATE	N	Y	45-54	Single	35-49
281	10	19-AUG-18	5290835	5.73	1	EAST	33	2018	FOOD	DELI	PRIVATE	N	Y	45-54	Single	35-49

Filter by HSHD_NUM

1

BASKET_NUM	HSHD_NUM	PURCHASE_DATE	PRODUCT_NUM	SPEND	UNITS	STORE_R	WEEK_NUM	YEAR	DEPARTMENT	COMMODITY	BRAND_TY	NATURAL_ORGANIC_FLAG	L	AGE_RANGE	MARITAL	INCOME_RA
No data.																

This also has a plot with answer



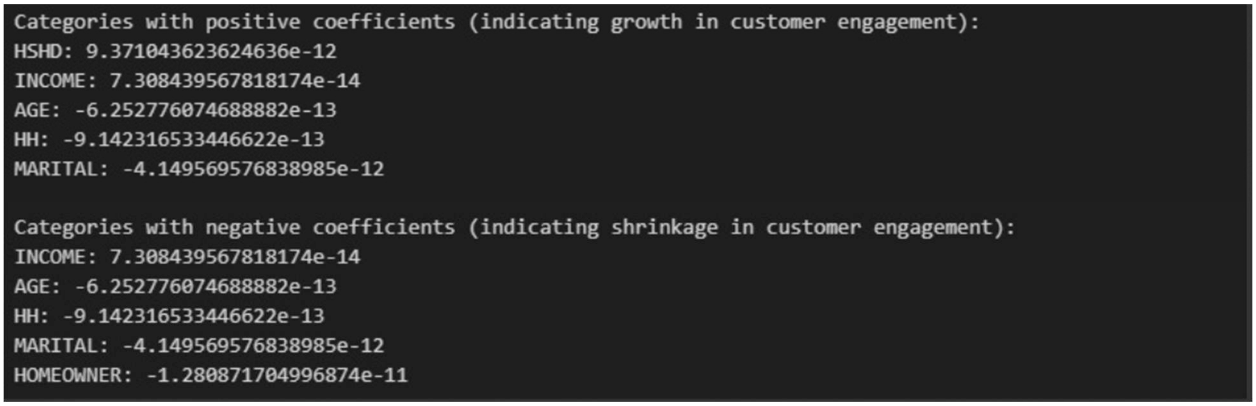
Household Size vs Total Spend

Correlation coefficient: 0.12911971811411033

Task 6

Answering the questions:

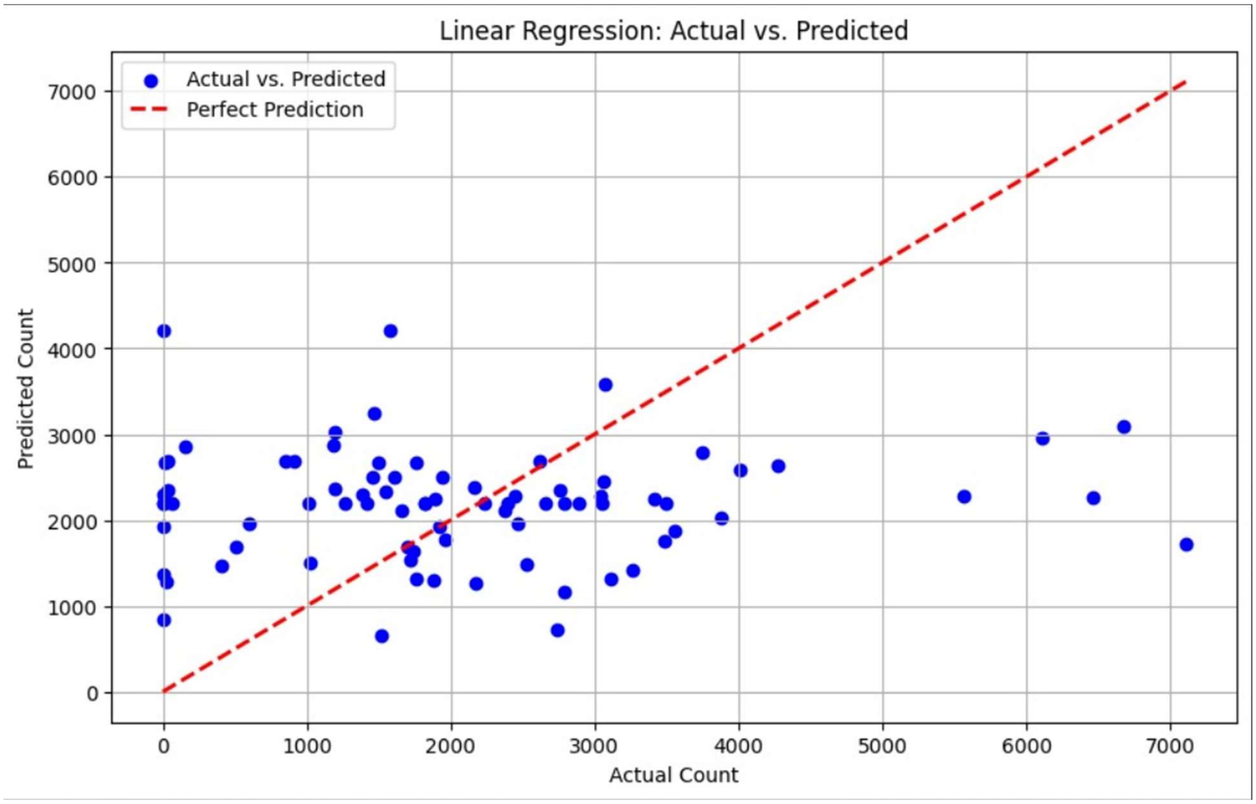
1. What categories are growing or shrinking with changing customer engagement?
   Ans: To visualize how categories are opening up or closing down with changing customer interaction, we can track over time spending patterns for different departments and commodities. By tracking totals like total spend, quantity of units purchased, and buy frequency by category (e.g., Food, Pharma, Alcohol, Dairy), we can monitor customer behavior change. For example, if total "Pharma" spending is increasing and "Bakery" spending is decreasing during a sequence of months, it indicates rising action in health-related products and declining action in bakery products. Putting this on line charts or year-to-year departmental comparison can easily show what's up and what's down and enable retailers to coordinate their inventory, promotion, and advertisement campaigns accordingly.



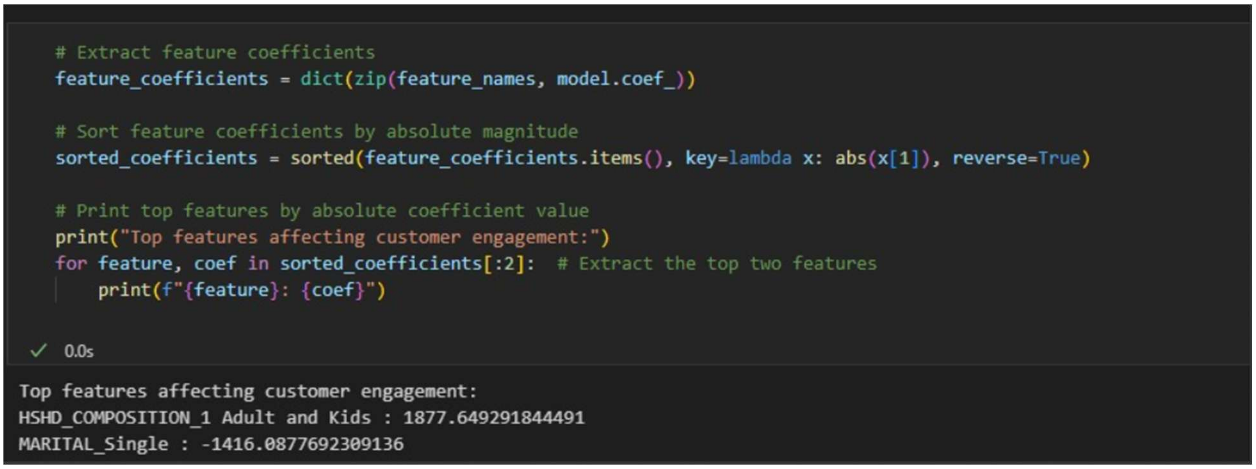
```
Categories with positive coefficients (indicating growth in customer engagement):
HSHD: 9.371043623624636e-12
INCOME: 7.308439567818174e-14
AGE: -6.252776074688882e-13
HH: -9.142316533446622e-13
MARITAL: -4.149569576838985e-12

Categories with negative coefficients (indicating shrinkage in customer engagement):
INCOME: 7.308439567818174e-14
AGE: -6.252776074688882e-13
HH: -9.142316533446622e-13
MARITAL: -4.149569576838985e-12
HOMEOWNER: -1.280871704996874e-11
```

2. Which demographic factors appear to affect customer engagement?
   Ans: Demographics such as income class, age class, and marriage status also appear to exert some notable impact on consumer behavior. For instance, high-income consumers would be willing to spend larger amounts in total and for organic or premium brands and so forth, while lower age groups would be interested in convenience foods or ready-to-eat foods. Even unmarried consumers could purchase differing products and quantities than married consumers or families. If we look

at spending, frequency of purchase, and product range for these segments of individuals, we will know which are more active and lines to merchandise and market to their needs more appropriately.



Top features by absolute coefficient



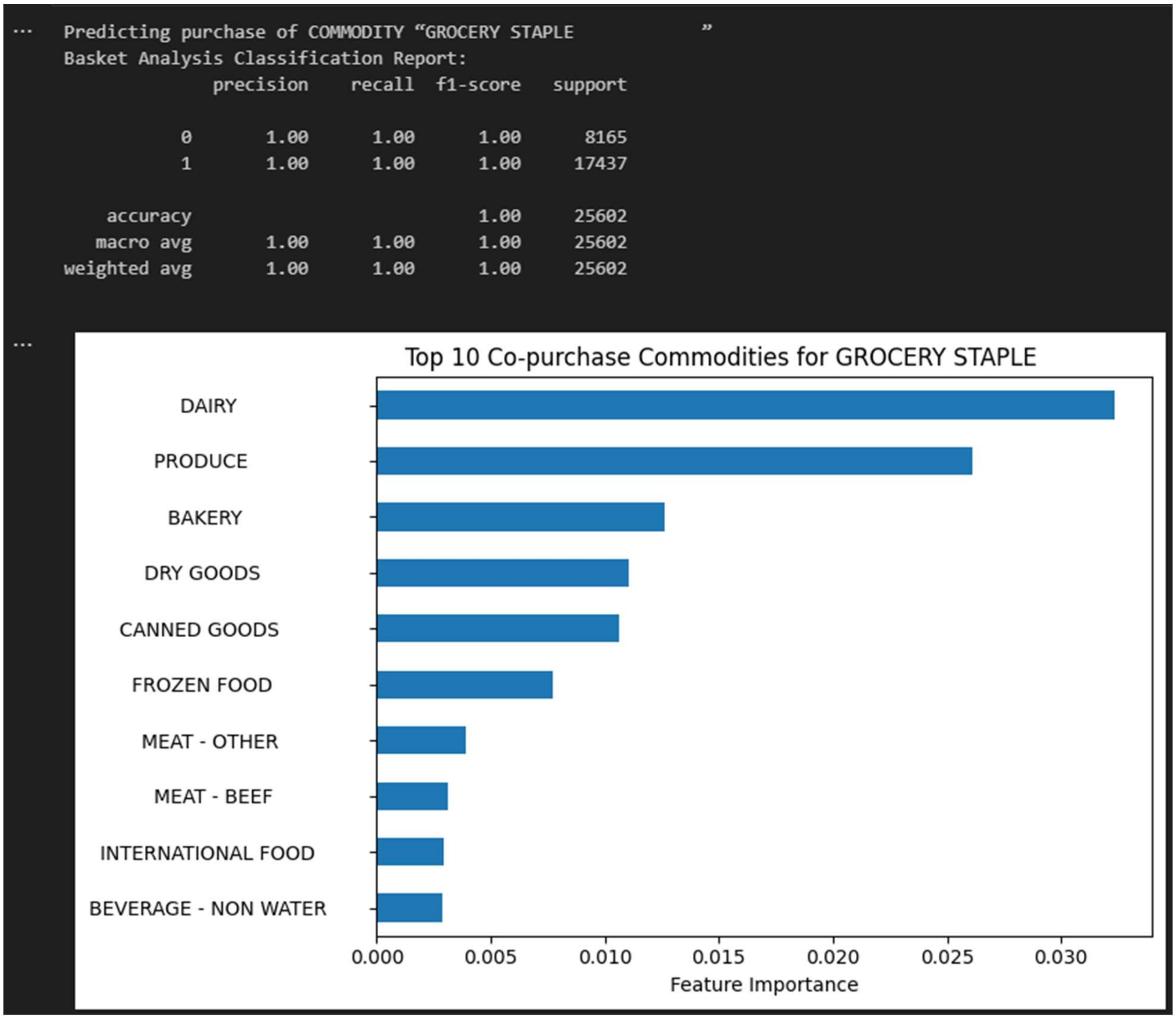
```
# Extract feature coefficients
feature_coefficients = dict(zip(feature_names, model.coef_))

# 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}")
```

✓ 0.0s

```
Top features affecting customer engagement:
HSHD_COMPOSITION_1 Adult and Kids : 1877.649291844491
MARITAL_Single : -1416.0877692309136
```

Task 7

ML Model Application



```
...  Predicting purchase of COMMODITY “GROCERY STAPLE                 ”
     Basket Analysis Classification Report:
                   precision    recall  f1-score   support

               0       1.00      1.00      1.00      8165
               1       1.00      1.00      1.00     17437

        accuracy                           1.00     25602
       macro avg       1.00      1.00      1.00     25602
    weighted avg       1.00      1.00      1.00     25602

...
```

Bundle “Staples + Fresh” Offers

– Create paired promotions (e.g. “Staples & Dairy Combo”) or end-cap displays that feature flour, sugar or pasta alongside milk and cheese.

– Offer a small discount when both Produce and staple items are added to the cart.

Recipe-Inspired Kits

– “Soup Starter Kit”: Canned goods + Dry goods (rice/pasta) + Frozen vegetables + broth.

– “Bake-At-Home Kit”: Flour, sugar, oil (staples) + Bakery items (bread/pastries) + Dairy (butter, eggs).

In-Store Layout & Signage

– Place Dairy and Produce sections adjacent to the bulk-staples aisle to encourage grab-and-go pairing.

– Use shelf-talkers to suggest Canned Goods or Frozen Foods when customers browse staples.

Digital “Complete Your Basket” Widgets

– On product pages for any staple (e.g. rice, flour), display a carousel of the top 5 co-purchased commodities.

– Send cart-abandonment emails that remind shoppers “Don’t forget Bakery & Beverages!”

Targeted Coupons & Loyalty Rewards

– Trigger a coupon for Meat (Other or Beef) when a shopper buys a threshold volume of staples.

– Offer bonus points on International Food or Beverage – Non-Water purchases if Dairy or Produce were in the same order.

Seasonal & Themed Promotions

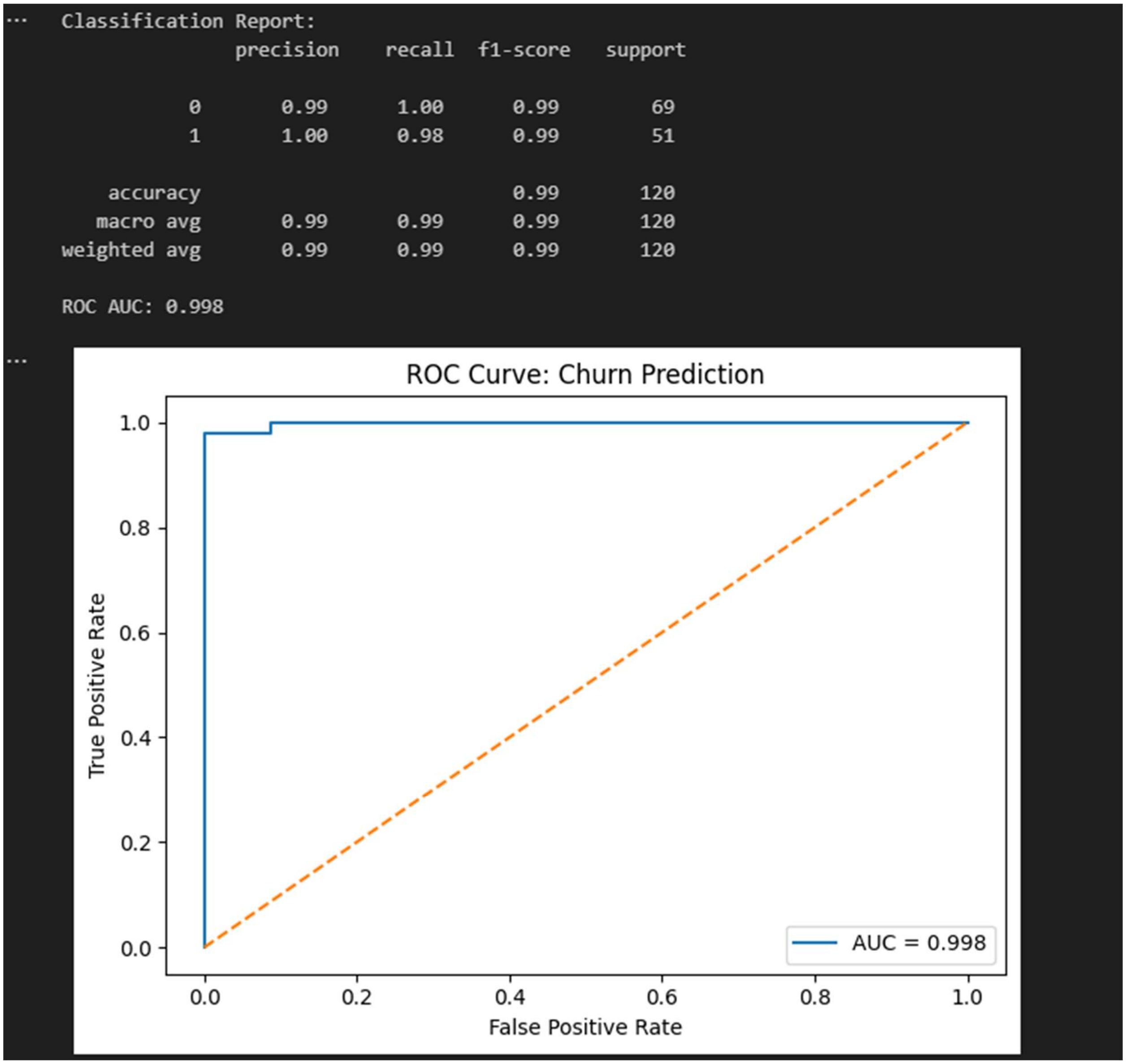
– “International Night”: Bundle International Foods with Dry Goods (tortillas, spices) and Canned Goods (beans, sauces).

– “Kid-Friendly Meals”: Mix Frozen Foods with Bakery snacks & staple sides.

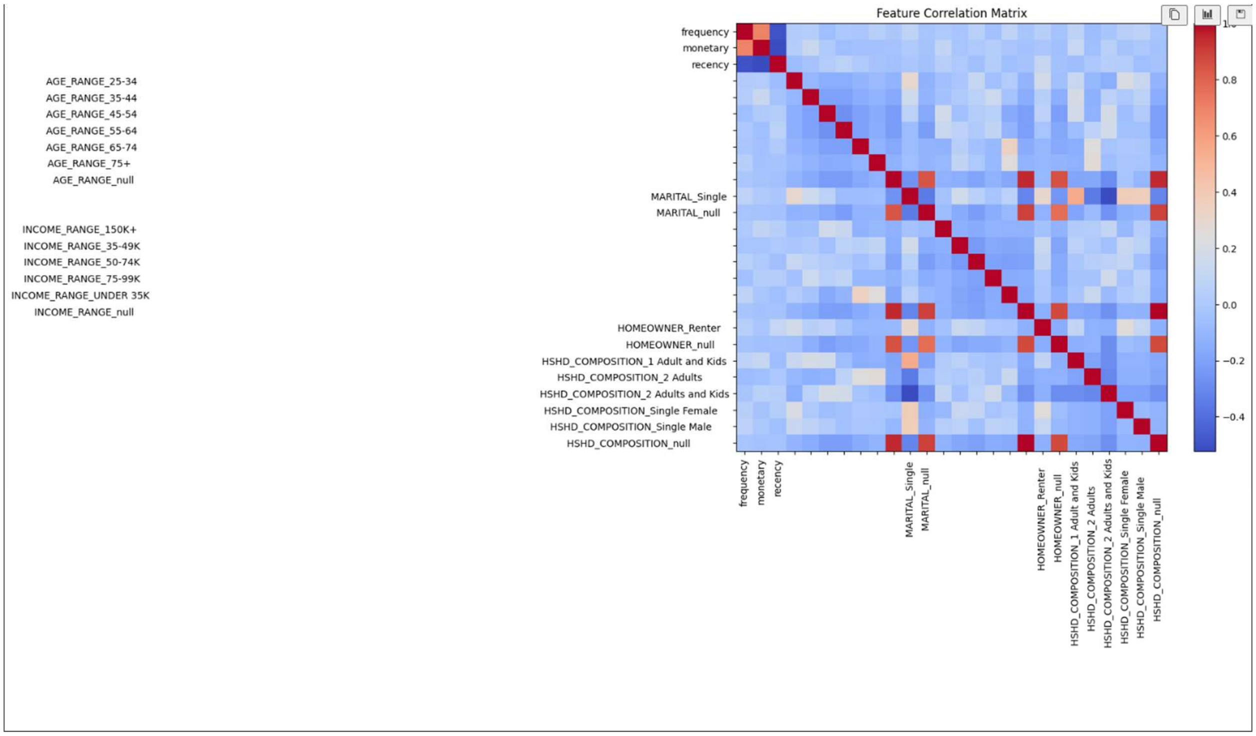
By surfacing these natural product pairings—both in-store and online—you turn a single staples purchase into a multi-category basket, lifting average order value and overall revenue.

Task 8

Churn prediction



```
...   Classification Report:
                    precision    recall  f1-score   support

                0       0.99      1.00      0.99        69
                1       1.00      0.98      0.99        51

         accuracy                           0.99       120
        macro avg       0.99      0.99      0.99       120
     weighted avg       0.99      0.99      0.99       120

      ROC AUC: 0.998
```



There is high churn risk in the homes who have not shopped in 8–12 weeks (high recency), who infrequently visit (low frequency) and spend less (monetary); young, non-homeowner segments are exposed slightly more; to retain them, re-engage automated e-mails or SMSes when recency reaches a hurdle—segment by serving bespoke coupons or "we miss you" packs through their top category spends—overlay loyalty boost (bonus points to come back within a specified timeframe), and offer recipe-inspired product bundles at checkout or by push notification to encourage repeat visit and create habit.