In [3]:

```
pip install pandas scikit-learn matplotlib seaborn
```

```
Requirement already satisfied: pandas in c:\users\sanja\anaconda3\lib\site-packages (2.2.2)
Requirement already satisfied: scikit-learn in c:\users\sanja\anaconda3\lib\site-packages (1.5.1)
Requirement already satisfied: matplotlib in c:\users\sanja\anaconda3\lib\site-packages (3.9.2)
Requirement already satisfied: seaborn in c:\users\sanja\anaconda3\lib\site-packages (0.13.2)
Requirement already satisfied: numpy>=1.26.0 in c:\users\sanja\anaconda3\lib\site-packages (from pandas) (1.26.4)
Requirement already satisfied: python-dateutil>=2.8.2 in c:\users\sanja\anaconda3\lib\site-packages (from pandas) (2.9.0.post0)
Requirement already satisfied: pytz>=2020.1 in c:\users\sanja\anaconda3\lib\site-packages (from pandas) (2024.1)
Requirement already satisfied: tzdata>=2022.7 in c:\users\sanja\anaconda3\lib\site-packages (from pandas) (2023.3)
Requirement already satisfied: scipy>=1.6.0 in c:\users\sanja\anaconda3\lib\site-packages (from scikit-learn) (1.13.1)
Requirement already satisfied: joblib>=1.2.0 in c:\users\sanja\anaconda3\lib\site-packages (from scikit-learn) (1.4.2)
Requirement already satisfied: threadpoolctl>=3.1.0 in c:\users\sanja\anaconda3\lib\site-packages (from scikit-learn) (3.5.0)
Requirement already satisfied: contourpy>=1.0.1 in c:\users\sanja\anaconda3\lib\site-packages (from matplotlib) (1.2.0)
Requirement already satisfied: cycler>=0.10 in c:\users\sanja\anaconda3\lib\site-packages (from matplotlib) (0.11.0)
Requirement already satisfied: fonttools>=4.22.0 in c:\users\sanja\anaconda3\lib\site-packages (from matplotlib) (4.51.0)
Requirement already satisfied: kiwisolver>=1.3.1 in c:\users\sanja\anaconda3\lib\site-packages (from matplotlib) (1.4.4)
Requirement already satisfied: packaging>=20.0 in c:\users\sanja\anaconda3\lib\site-packages (from matplotlib) (24.1)
Requirement already satisfied: pillow>=8 in c:\users\sanja\anaconda3\lib\site-packages (from matplotlib) (10.4.0)
Requirement already satisfied: pyparsing>=2.3.1 in c:\users\sanja\anaconda3\lib\site-packages (from matplotlib) (3.1.2)
Requirement already satisfied: six>=1.5 in c:\users\sanja\anaconda3\lib\site-packages (from python-dateutil>=2.8.2->pandas) (1.
16.0)
Note: you may need to restart the kernel to use updated packages.
```

In [11]:

```
import pandas as pd

# Load cleaned dataset
df = pd.read_csv(r"C:\New Volume\internship\final projects\FINAL PROJECT IN ELEVATE (PLV).csv")

# Check the first 5 rows
df.head()
```

Out[11]:

	customer_unique_id	num_orders	total_spend	First_purchase	last_purchase	AOV	recency_days	tenure_days
0	7c396fd4830fd04220f754e42b4e5bff	4	82.82	04-09-2017 11:26	02-10-2017 10:56	20.705	2896	27
1	708ab75d2a007f0564aedd11139c7708	1	99.33	25-04-2018 22:01	25-04-2018 22:01	99.330	2691	0
2	861eff4711a542e4b93843c6dd7febb0	1	146.87	16-05-2017 15:05	16-05-2017 15:05	146.870	3035	0
3	a8b9d3a27068454b1c98cc67d4e31e6f	1	24.39	26-06-2018 11:01	26-06-2018 11:01	24.390	2629	0
4	af07308b275d755c9edb36a90c618231	1	141.46	24-07-2018 20:41	24-07-2018 20:41	141.460	2601	0

In [12]:

```
df = df.fillna(0)
```

In [13]:

```
df['avg_days_between_orders'] = df.apply(lambda row: row['tenure_days'] / row['num_orders'] if row['num_orders'] > 0 else row[
df['purchase_rate'] = df.apply(lambda row: row['num_orders'] / row['tenure_days'] if row['tenure_days'] > 0 else 0, axis=1)
```

In [14]:

```
X = df[['num_orders', 'recency_days', 'AOV', 'tenure_days', 'avg_days_between_orders', 'purchase_rate']]
y = df['total_spend']
```

In [15]:

```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

In [19]:

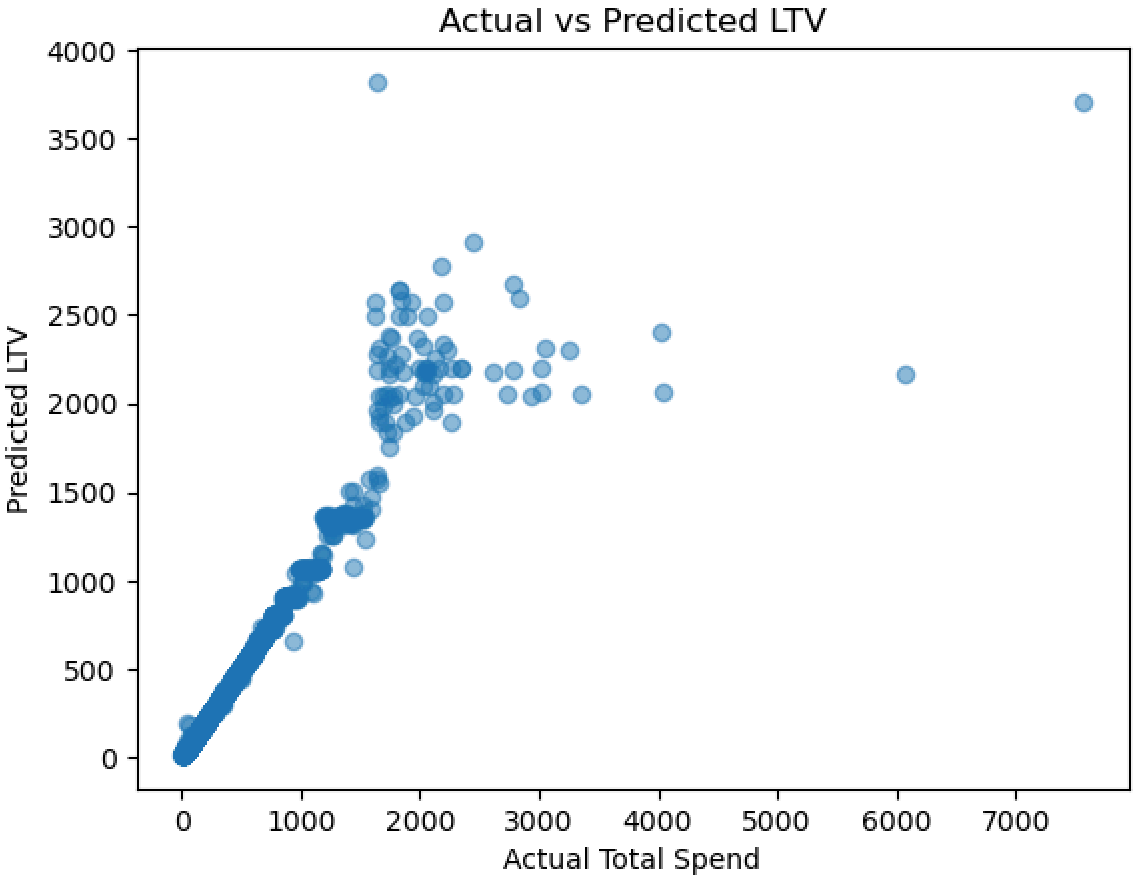
```
import xgboost as xgb
from xgboost import XGBRegressor
from sklearn.metrics import mean_absolute_error, mean_squared_error
import numpy as np

# Initialize model
model = XGBRegressor(n_estimators=100, learning_rate=0.1, max_depth=5, random_state=42)

# Train model
model.fit(X_train, y_train)
```

```
# Predict
y_pred = model.predict(X_test)

# Evaluate
mae = mean_absolute_error(y_test, y_pred)
rmse = np.sqrt(mean_squared_error(y_test, y_pred))

print(f'MAE: {mae}')
print(f'RMSE: {rmse}')
```

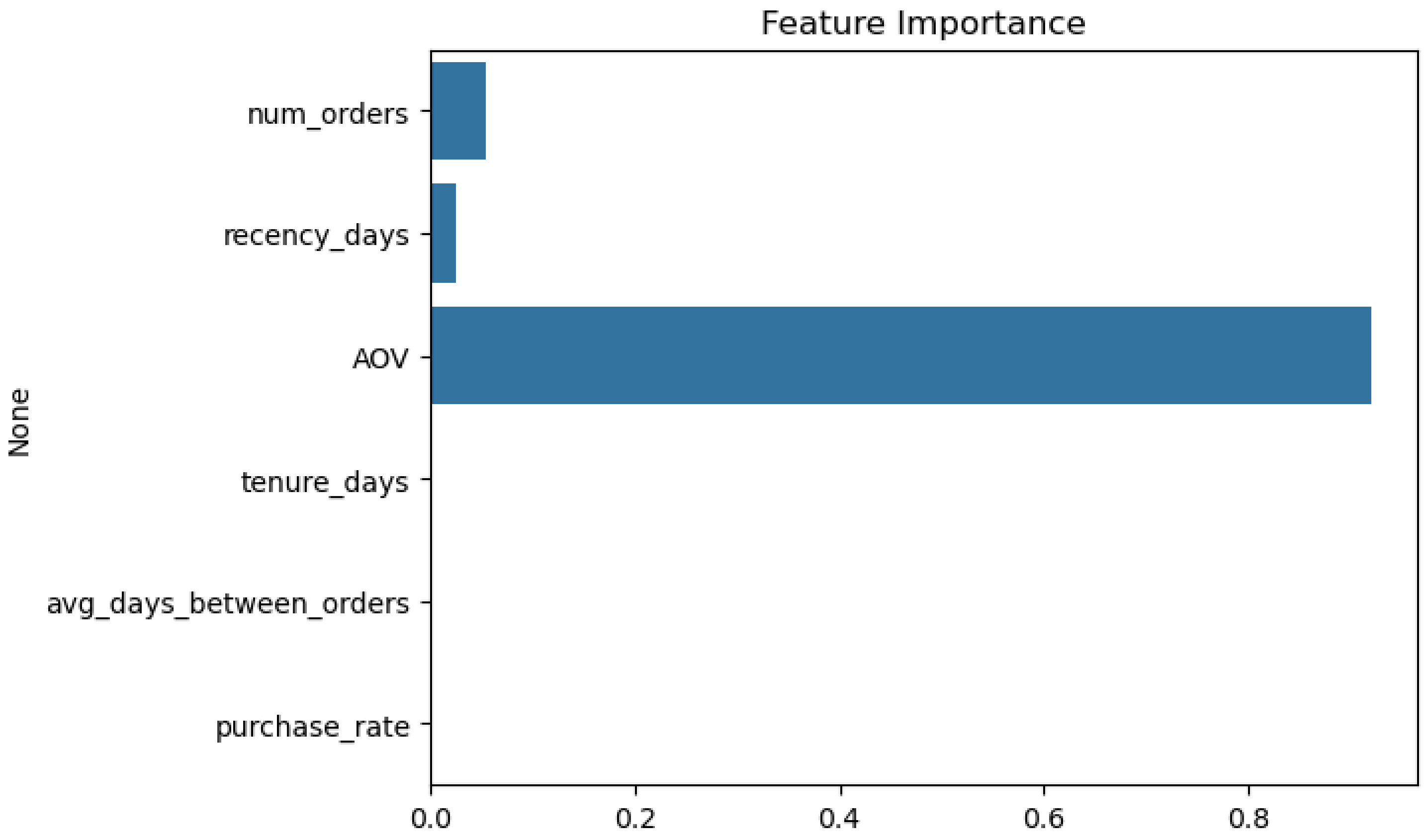
```
MAE: 4.411034032177755
RMSE: 56.225145436323025
```

In [20]:

```
import matplotlib.pyplot as plt

plt.scatter(y_test, y_pred, alpha=0.5)
plt.xlabel('Actual Total Spend')
plt.ylabel('Predicted LTV')
plt.title('Actual vs Predicted LTV')
plt.show()
```



In [21]:
```
import seaborn as sns

feature_importance = model.feature_importances_
sns.barplot(x=feature_importance, y=X.columns)
plt.title('Feature Importance')
plt.show()
```



In [ ]:

In [27]:

```
bins = [0,
        df_test['predicted_ltv'].quantile(0.25),
        df_test['predicted_ltv'].quantile(0.50),
        df_test['predicted_ltv'].quantile(0.75),
        df_test['predicted_ltv'].max() + 1]

labels = ['Low', 'Medium', 'High', 'Very High']

df_test['ltv_segment'] = pd.cut(
    df_test['predicted_ltv'],
    bins=bins,
    labels=labels,
    include_lowest=True
)
# View segments
print(df_test.groupby('ltv_segment').size())
```

```
ltv_segment
Low          5050
Medium       4540
High         4807
Very High    4715
dtype: int64
```

```
C:\Users\sanja\AppData\Local\Temp\ipykernel_24692\575360421.py:16: FutureWarning: The default of observed=False is deprecated a
nd will be changed to True in a future version of pandas. Pass observed=False to retain current behavior or observed=True to ad
opt the future default and silence this warning.
  print(df_test.groupby('ltv_segment').size())
```