Final Paper

By :

Aryan Mainkar

Abstract :

This study integrates customer demographic data and sales transactions to construct a predictive model for individual customer spending behavior. Utilizing a Random Forest Regressor, we delve into the intricate interplay of age, gender, and payment method on total spending per customer. The research aims to offer detailed insights for businesses aiming to refine marketing strategies based on a granular understanding of customer characteristics.

Introduction :

Understanding and predicting customer behavior is essential for businesses striving to optimize marketing strategies and enhance customer engagement. This study seeks to address this challenge by leveraging machine learning techniques to predict individual customer spending patterns. The primary objective is to develop a predictive model that enables businesses to tailor marketing strategies to individual customer profiles, thereby fostering personalized engagement and revenue growth.

Problem Statement :

Effectively predicting customer spending and identifying the determinants of spending patterns is crucial for businesses seeking to optimize their marketing efforts. This study seeks to unravel these complexities by merging and preprocessing datasets and employing a sophisticated Random Forest Regressor. The overarching goal is to provide businesses with actionable insights to enhance customer engagement and optimize marketing efforts.

Proposed Methodology :

1. Data Exploration

   1.1. Loading and Initial Exploration :

   Two datasets, encompassing customer demographics and sales transactions, are loaded. This initial exploration provides a holistic view of customer characteristics and sales dynamics.The first dataset, labeled df1, comprises information about customer demographics.

   This dataset includes the following columns:

   customer_id: A unique identifier for each customer.
   gender: The gender of the customer.
   age: The age of the customer.
   payment_method: The preferred payment method used by the customer.

   1.2. Exploratory Data Analysis (EDA) :

   EDA uncovers nuanced insights into the age distribution, gender representation, prevalent payment methods, and the sales distribution by category. This initial analysis informs subsequent modelling decisions.

The second dataset, labeled df2, contains detailed information about sales transactions. It includes the following columns:

invoice_no: A unique identifier for each transaction.
customer_id: The customer associated with the transaction.
category: The category of the purchased item.
quantity: The quantity of items purchased.
price: The price of the items.
invoice_date: The date of the transaction.
shopping_mall: The shopping mall where the transaction took place.

```
First few rows of df1:
  customer_id  gender   age payment_method
0     C241288  Female  28.0    Credit Card
1     C111565    Male  21.0     Debit Card
2     C266599    Male  20.0           Cash
3     C988172  Female  66.0    Credit Card
4     C189076  Female  53.0           Cash

Summary of df1:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 99457 entries, 0 to 99456
Data columns (total 4 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   customer_id     99457 non-null  object
 1   gender          99457 non-null  object
 2   age             99338 non-null  float64
 3   payment_method  99457 non-null  object
dtypes: float64(1), object(3)
memory usage: 3.0+ MB
None

Statistical summary of df1:
                age
count  99338.000000
mean      43.425859
std       14.989400
min       18.000000
25%       30.000000
50%       43.000000
75%       56.000000
max       69.000000
```

```
Summary of df2:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 99457 entries, 0 to 99456
Data columns (total 7 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   invoice_no     99457 non-null  object
 1   customer_id    99457 non-null  object
 2   category       99457 non-null  object
 3   quantity       99457 non-null  int64
 4   price          99457 non-null  float64
 5   invoice_date   99457 non-null  object
 6   shopping_mall  99457 non-null  object
dtypes: float64(1), int64(1), object(5)
memory usage: 5.3+ MB
None

Statistical summary of df2:
           quantity         price
count  99457.000000  99457.000000
mean       3.003429    689.256321
std        1.413025    941.184567
min        1.000000      5.230000
25%        2.000000     45.450000
50%        3.000000    203.300000
75%        4.000000   1200.320000
max        5.000000   5250.000000
```

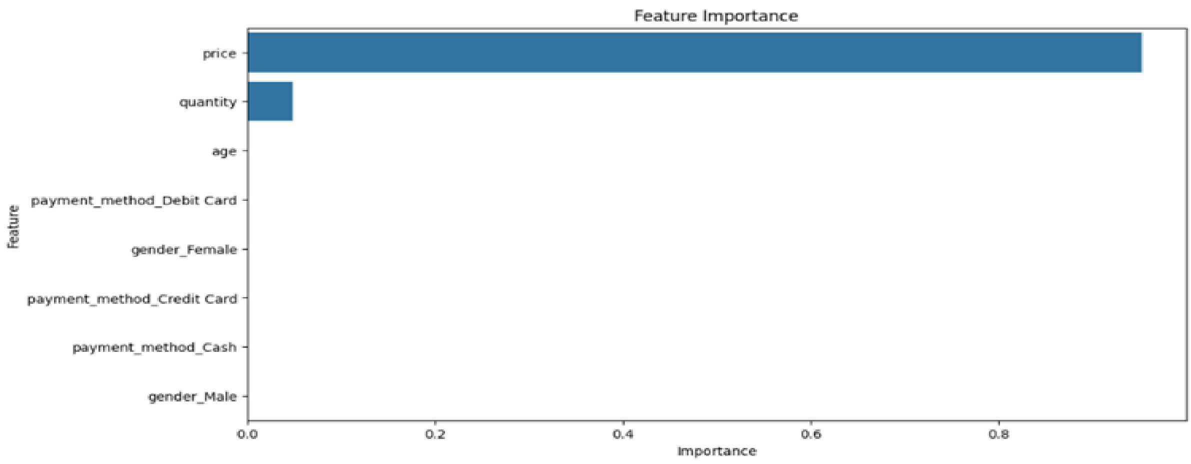
Data Preprocessing :

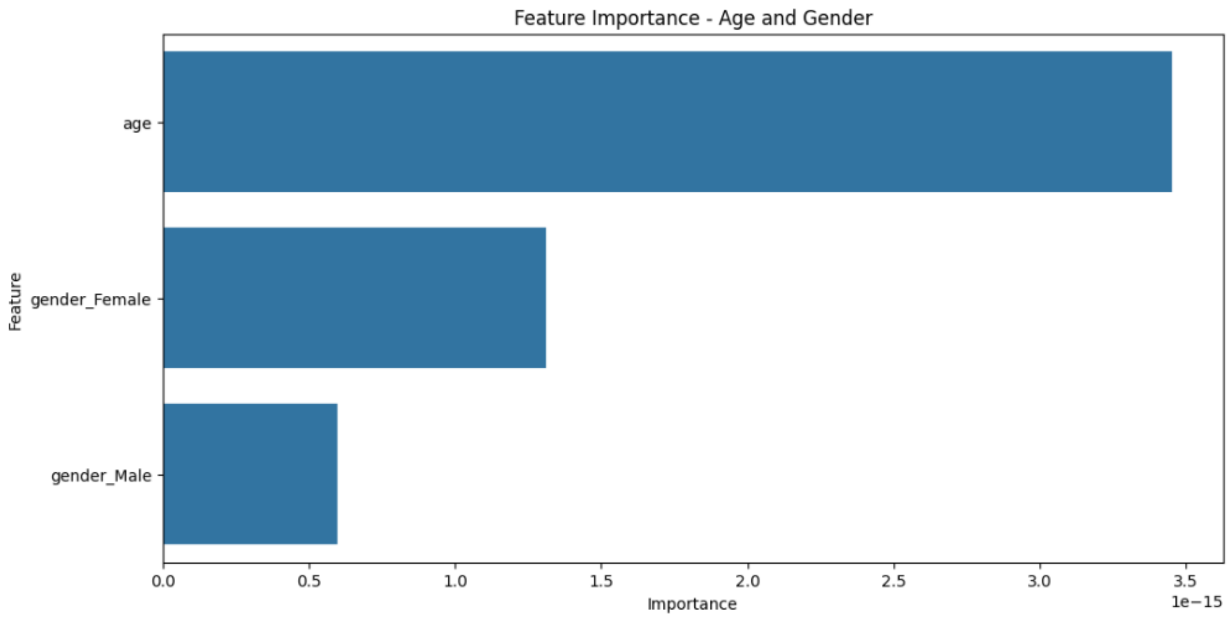
1. Merging Datasets :

Integration of customer and sales datasets based on customer ID establishes the foundation for a comprehensive analysis. This step enables the synthesis of individual customer profiles with transactional data.

2 . Feature Engineering :

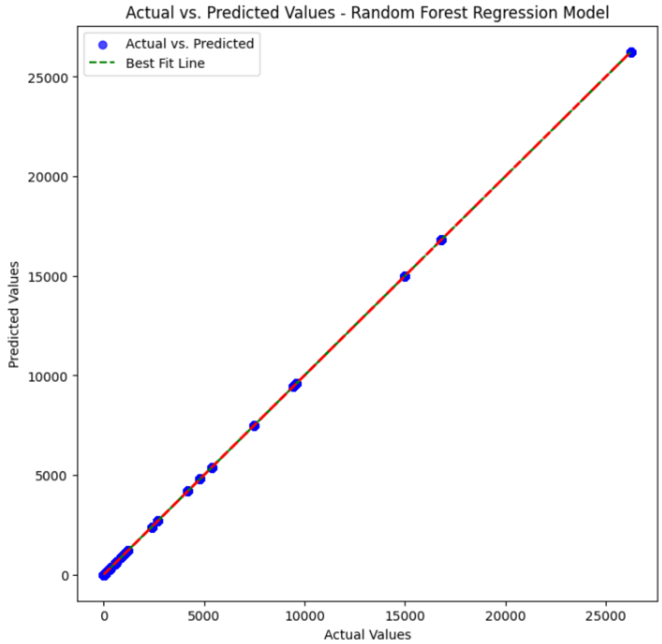
The introduction of a new feature, total spending per customer, serves to enhance the predictive capacity of the model. This feature encapsulates the cumulative spending of each customer, providing a more comprehensive metric for analysis.





Modelling :

1. Random Forest Regression :
   The study employs a Random Forest Regressor, a powerful ensemble learning method, to predict total spending accurately. The model's adaptability to non-linear relationships and feature interactions makes it well-suited for this predictive task.
2. Model Evaluation :
   Quantitative metrics, including Mean Squared Error (MSE), R-squared, and feature importance, are employed to assess model performance. These metrics provide a comprehensive view of the model's predictive capabilities and the factors influencing customer spending.



Analysis and Results :

The Random Forest Regressor demonstrates exceptional performance, accurately predicting total spending. The analysis reveals the significance of gender and age, with female customers and older age groups exhibiting higher spending. The visualizations, including the 3D plot of age, gender, and spending, provide a comprehensive view of the model's predictive capabilities.

```
Insights:
         Feature    Importance
0            age  3.454343e-15
3  gender_Female  1.310763e-15
4    gender_Male  5.978938e-16

Total Spending by Gender:
gender_categorical
Female    $150207136.02
Male      $101298658.23
Name: total_spending_per_customer, dtype: object

Total Spending by Age Group:
age_group
18-25    $33680374.15 Spending
26-35    $47826744.49 Spending
36-50    $74410409.30 Spending
51+      $91193246.77 Spending
Name: total_spending_per_customer, dtype: object
```

```
  customer_id   age payment_method invoice_no  category  quantity    price
0     C241288  28.0    Credit Card    I138884  Clothing       5.0  1500.40
1     C111565  21.0     Debit Card    I317333     Shoes       3.0  1800.51
2     C266599  20.0           Cash    I127801  Clothing       1.0   300.08
3     C988172  66.0    Credit Card    I173702     Shoes       5.0  3000.85
4     C189076  53.0           Cash    I337046     Books       4.0    60.60

  invoice_date   shopping_mall gender_categorical  \
0   05-08-2022          Kanyon             Female
1   12-12-2021  Forum Istanbul               Male
2   09-11-2021       Metrocity               Male
3   16-05-2021    Metropol AVM             Female
4   24-10-2021          Kanyon             Female

   total_spending_per_customer  gender_Female  gender_Male
0                      7502.00            1.0          0.0
1                      5401.53            0.0          1.0
2                       300.08            0.0          1.0
3                     15004.25            1.0          0.0
4                       242.40            1.0          0.0
Mean Squared Error: 6.348410112892832e-22
R-squared: 1.0
```

Conclusions :

The study underscores the efficacy of the Random Forest Regressor in accurately predicting customer spending. The research translates these insights into actionable recommendations for businesses, emphasizing the importance of tailoring marketing strategies to specific customer segments. The integration of demographic data with transactional information proves instrumental in crafting targeted and effective engagement strategies.

Lessons Learned :

1. Data Quality Impact:

The study highlights the pivotal role of meticulous data preprocessing, including handling missing values and encoding categorical variables. The quality of input data significantly influences the robustness of predictive models.

2. Feature Importance Insights:

Through feature importance analysis, age and gender emerge as critical factors influencing customer spending. Understanding these drivers is essential for crafting nuanced and effective marketing strategies. Using the Random Forest model's feature importance analysis, we observed that both age and gender play pivotal roles in shaping customer spending patterns. The Random Forest model identified these variables as highly influential in predicting total spending per customer:

3. Visualization Enhancement:

Utilizing advanced visualization techniques, such as 3D plots and interactive graphs, facilitates a deeper understanding of the model's complexities. Effective communication of insights is crucial for stakeholders to grasp the intricacies of the model and its implications. The use of 3D plots allows us to visualize relationships between age, gender, and spending with enhanced clarity. This approach facilitates a more immersive exploration of the data, aiding stakeholders in grasping the intricate dynamics of customer spending patterns.

4. Model Evaluation Metrics:

The inclusion of comprehensive model evaluation metrics, including MSE and R-squared, provides a thorough assessment of the model's accuracy and explanatory power. These metrics offer quantitative insights into the model's performance, helping stakeholders gauge the accuracy of predictions and the extent to which the model explains the variability in customer spending.

Bibliography :

- Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825-2830.
- McKinney, W. (2010). Data structures for statistical computing in Python. In Proceedings of the 9th Python in Science Conference (Vol. 445, pp. 51-56).
- Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with python. In 9th Python in Science Conference (Vol. 57).