August 8, 2025

Forecasting Client Spending Patterns

Data Science Challenge

Jed Jerrel K. Escaran
BS Applied Physics (Instrumentation)
University of the Philippines - Diliman
Class of 2025

Understanding Customer Spending Behavior

Banks keep a record of customer transaction histories and use it to personalize engagement and marketing. However, personalizing for each client may prove expensive and counterproductive.

Understanding customer spending behavior allows the bank to focus on either active or inactive card users, anticipate bank revenues, and offer other products and/or services to customers.

Objective 1:

Personalize marketing strategies given a record of customer spending.

Objective 2:

Predict future spending trend of customers.

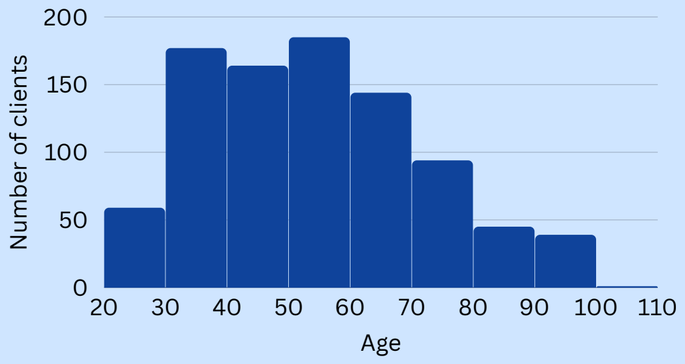
Exploratory Data Analysis

We have data on 908 clients: customer ID, date of birth, gender, occupation, and Half of Month (HOM) transactions (total and subtotal across different categories) from January 1, 2019 to December 15, 2022 (47 HOM periods).

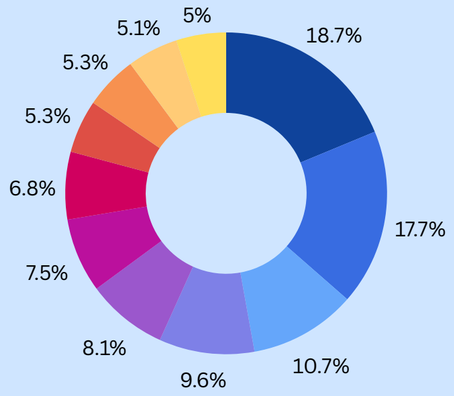
Variable	Data Type	Description
cc_num	str	Credit card number/ customer id
year	int	Year of transaction
month	int	Month of transaction
HOM_tag	int	Half of Month (HOM) tag. HOM=1 refers to days 1 to 15 while HOM=2 refers to days 16 to 30/31
HOM_total	float	Sum of all transactions in one HOM
HOM_{category}	float	Subtotal of transactions per spending type/ category (see list of categories below)
HOM_{category}_freq	int	Frequency of transactions per category within the HOM period
dob	datetime	Date of birth
gender	str	Sex assigned at birth: M/F
job	str	Job description of customer
next_HOM	float	Sum of transactions for the next HOM

Exploratory Data Analysis

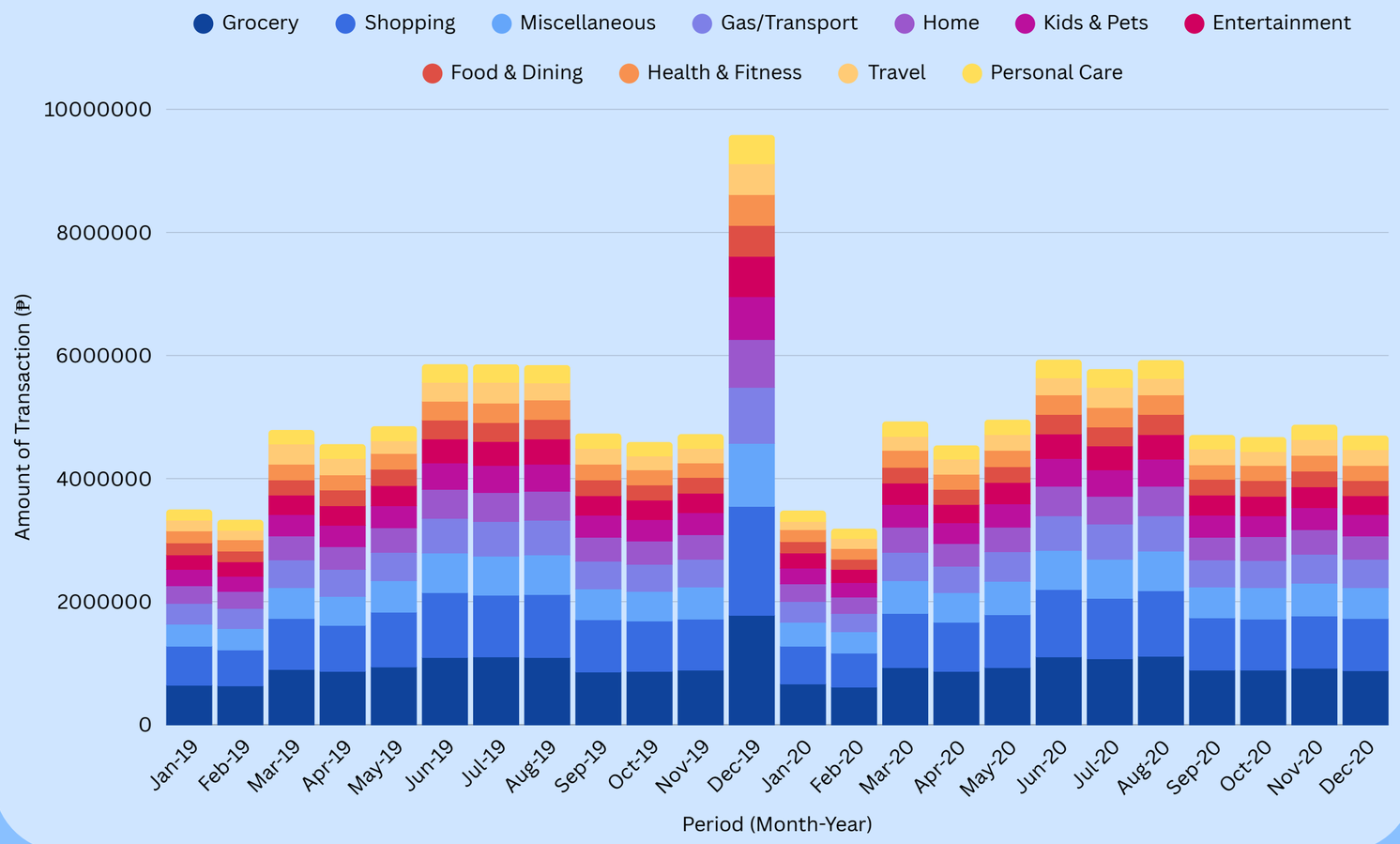
Client Ages



Average Spend Distribution



Monthly Spend by Category



1

Data Cleaning

loading raw data, parsing dates, computing client age, identifying outliers, processing HOM flags

2

Feature Engineering

creating time-series features, clustering clients, encoding categories, combining features

3

Model Training

splitting data 70-30, training then tuning linear regression and random forest regressor models

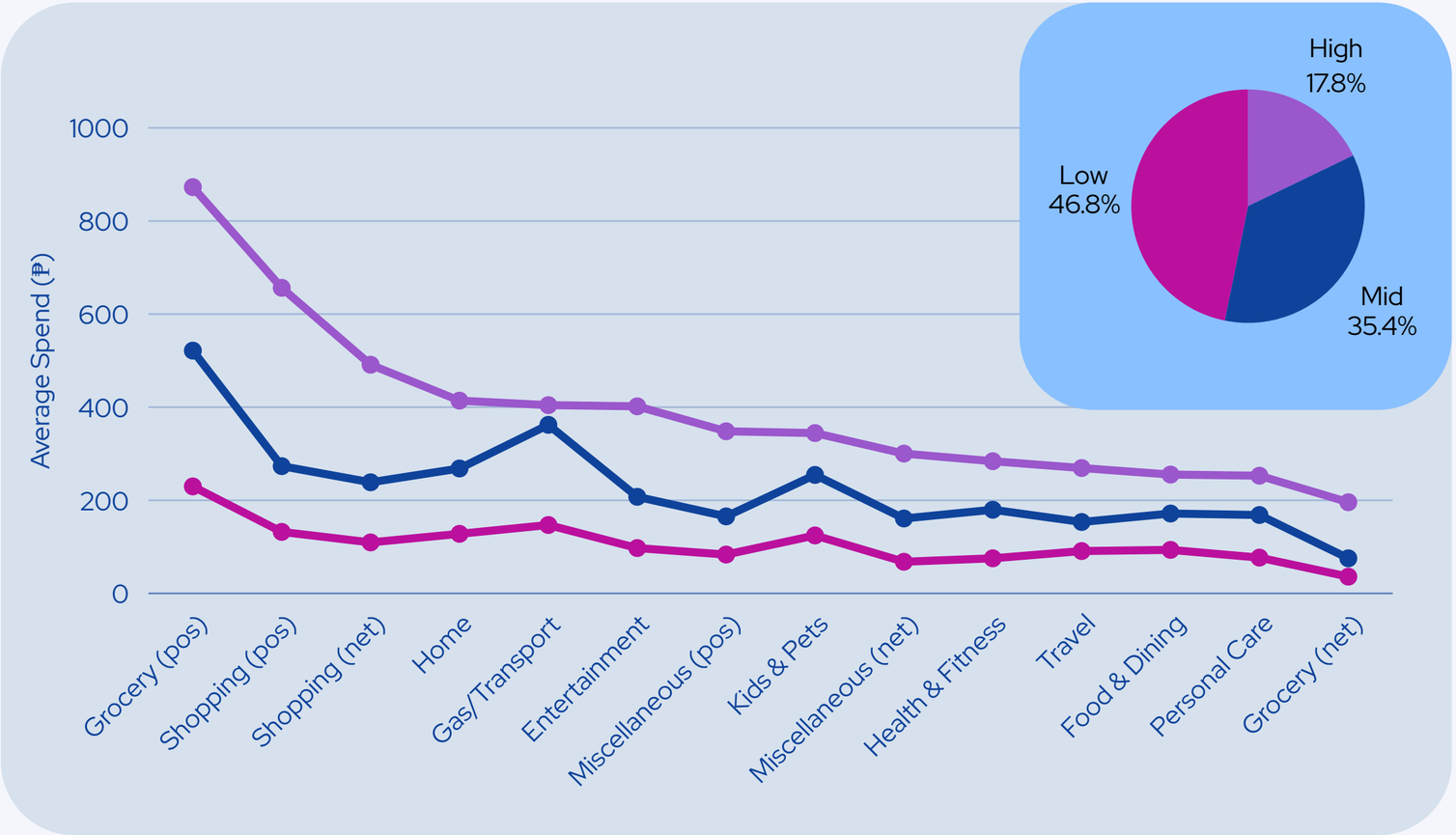
4

Model Evaluation

comparing MAE, MAPE, and R^2 values, actual vs predicted plots, feature importance and PDP plots

Client Clustering

Using K-Means for 908 clients



High Spenders: 162 clients

Age: 40

Gender: Female

Mid Spenders: 321 clients

Age: 55

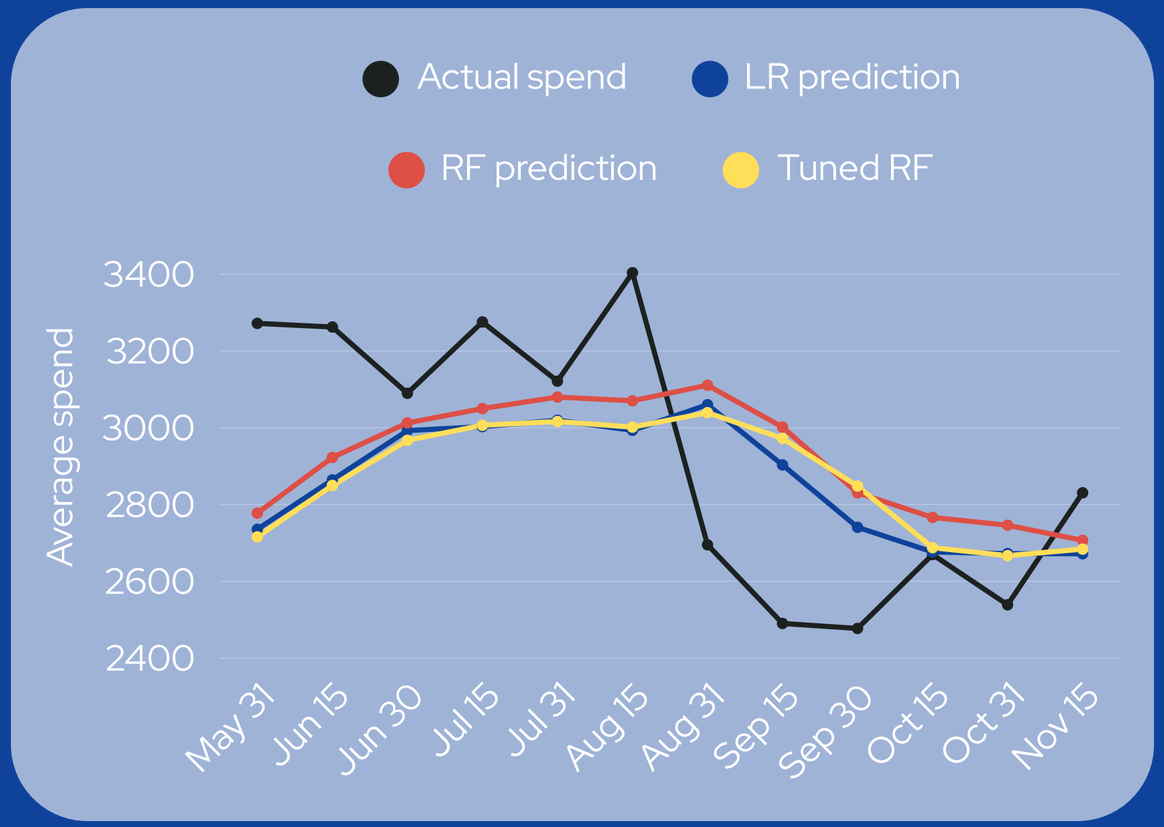
Gender: Female

Low Spenders: 425 clients

Age: 58

Gender: Male

Forecasting Client Spending Patterns



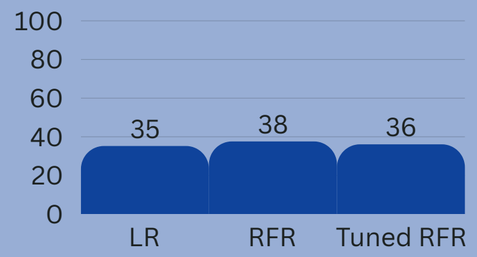
Mean Absolute Error:

Linear Regression
₱ 773.9

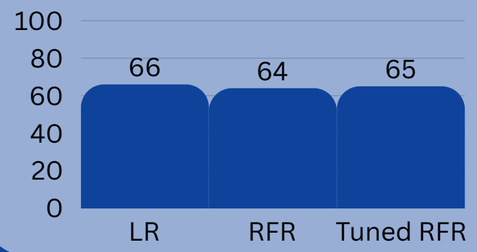
Random Forest Regr.
₱ 805.3

Tuned RF Regressor
₱ 782.7

Mean Absolute Percentage Error (%):



Coefficient of Determination (R^2):



Summary

Using the transaction history of 908 clients, we were able to **form clusters** of clients based on their spending capability. We were also able to **train regression models** to forecast the spending patterns of clients up to one HOM ahead. Predicting past one HOM is possible, but errors will compound.

In this work, the different methods of purchase (POS and NET) were held negligible. This was done since the project did not factor in the method of purchase when training the forecaster.

Key Insights

Outliers

Double December

Top Categories

groceries
shopping
misc.

Client Clustering

for targeted marketing

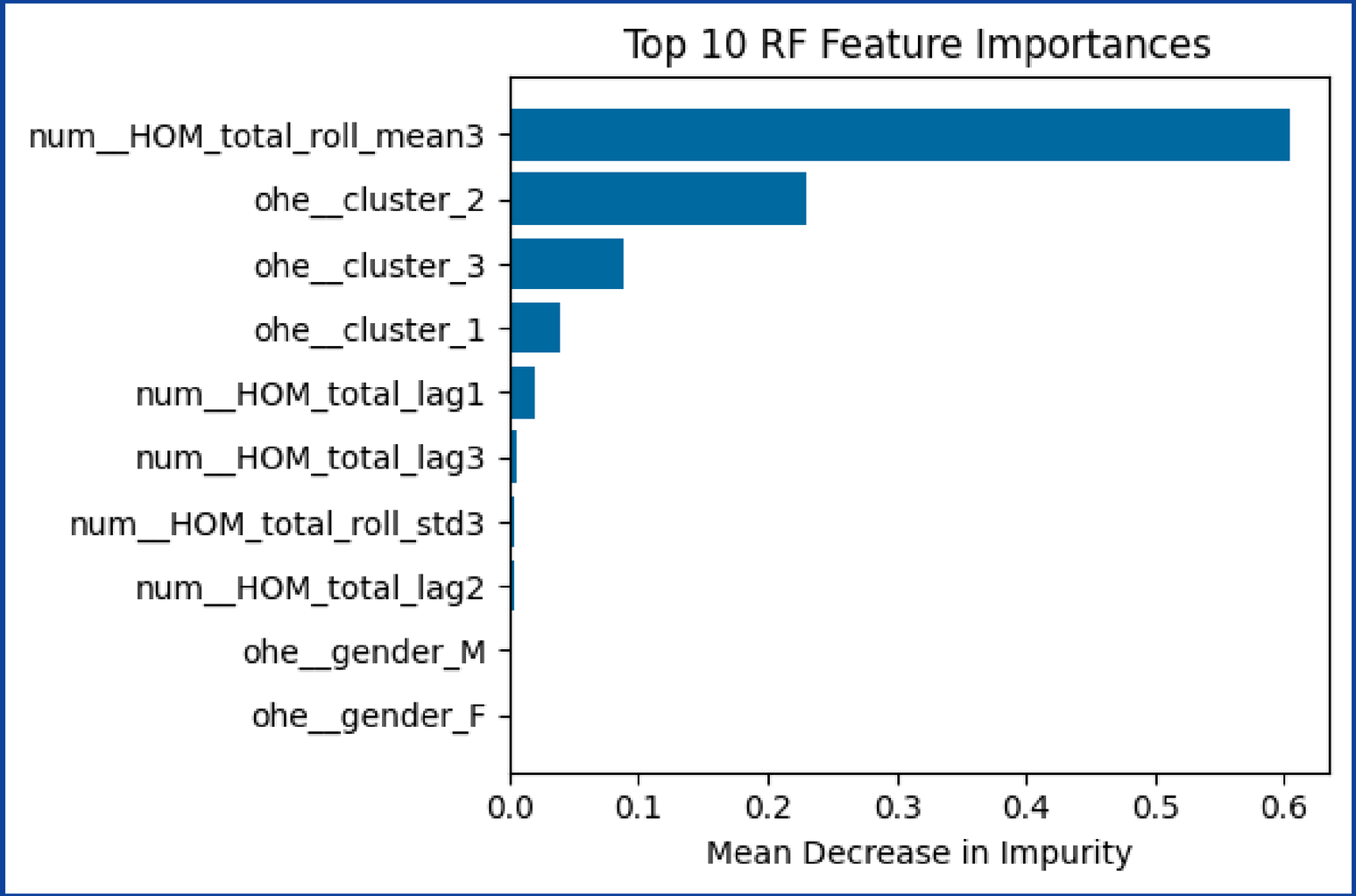
One-HOM Forecast

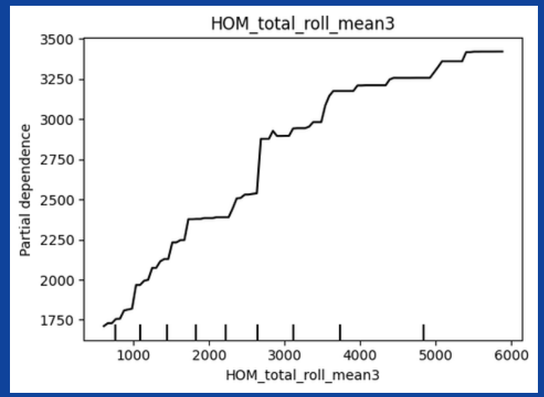
Off by ₱774, on average

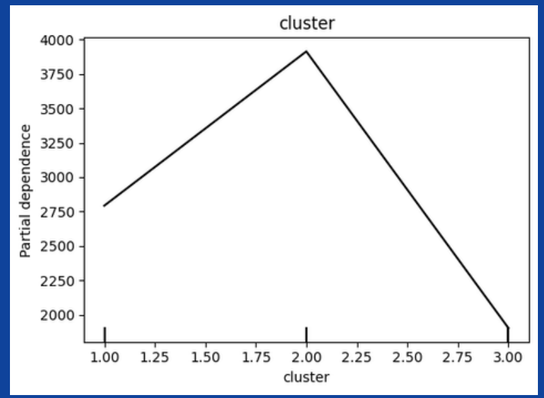
Thank you!

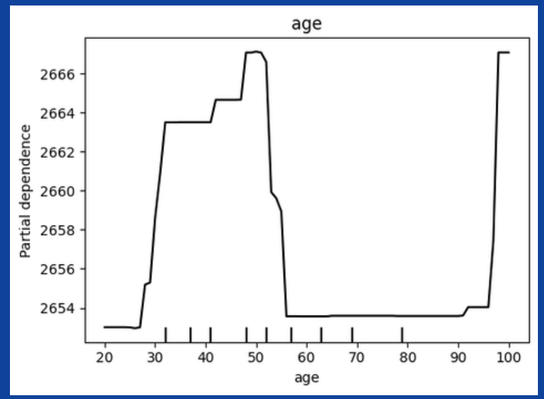
escaran.jed@gmail.com

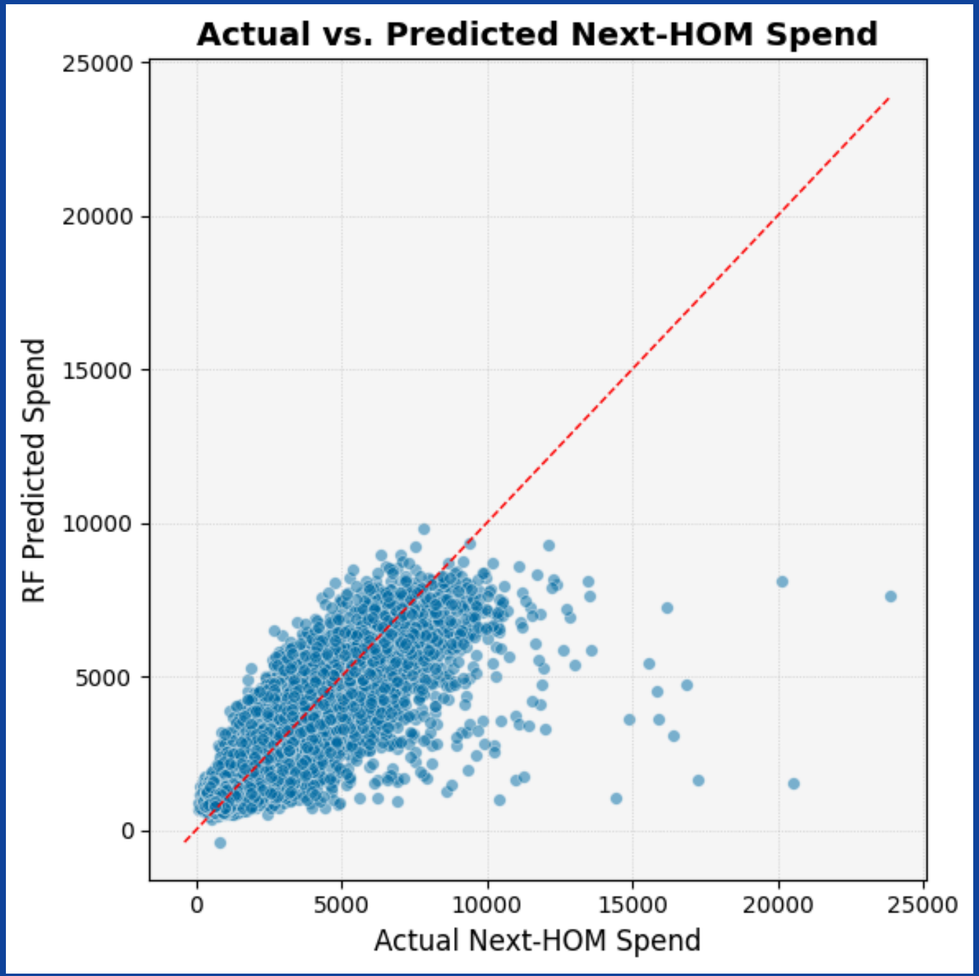
linkedin.com/in/jed-escaran/

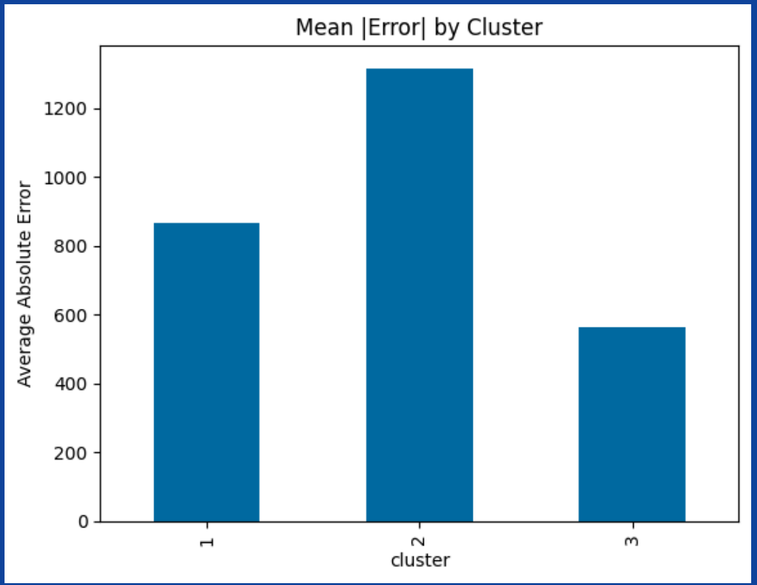

Top 10 RF Feature Importances
num__HOM_total_roll_mean3
ohe__cluster_2
ohe__cluster_3
ohe__cluster_1
num__HOM_total_lag1
num__HOM_total_lag3
num__HOM_total_roll_std3
num__HOM_total_lag2
ohe__gender_M
ohe__gender_F
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Mean Decrease in Impurity

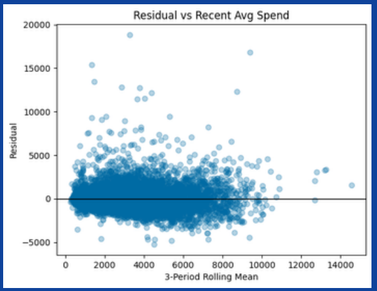

HOM_total_roll_mean3
Partial dependence
3500
3250
3000
2750
2500
2250
2000
1750
1000
2000
3000
4000
5000
6000
HOM_total_roll_mean3

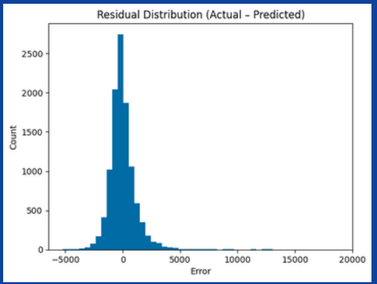

cluster
Partial dependence
4000
3750
3500
3250
3000
2750
2500
2250
2000
1.00
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00
cluster

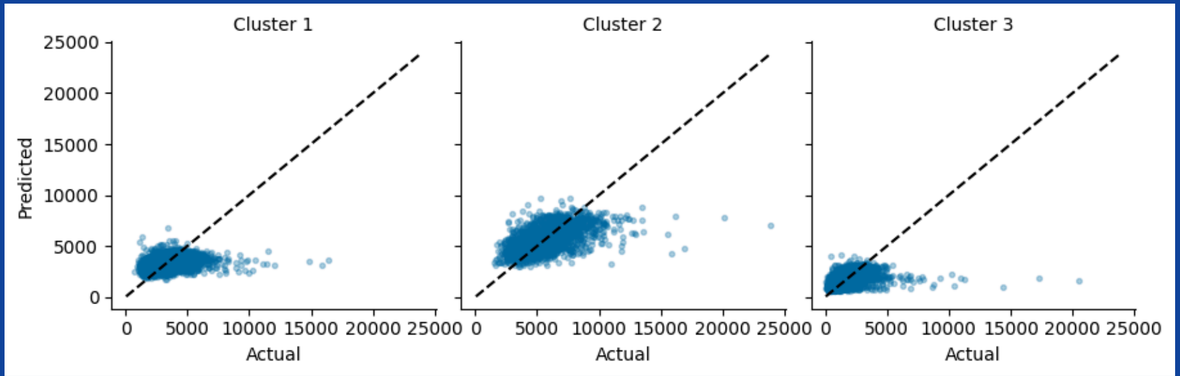

age
Partial dependence
2666
2664
2662
2660
2658
2656
2654
20
30
40
50
60
70
80
90
100
age


Actual vs. Predicted Next-HOM Spend
RF Predicted Spend
Actual Next-HOM Spend
0
5000
10000
15000
20000
25000


Mean |Error| by Cluster
Average Absolute Error
cluster
1
2
3
0
200
400
600
800
1000
1200


Residual vs Recent Avg Spend
Residual
3-Period Rolling Mean


Residual Distribution (Actual – Predicted)
Count
Error


Cluster 1
Cluster 2
Cluster 3
Predicted
Actual
0
5000
10000
15000
20000
25000