Customer Segmentation Classification

Introduction

In this Project, based on an automobile company is planning to enter new markets with its existing products. After doing extensive market research, they conclude that the behavior of the new market is like the behavior of the existing market. In the current market, the sales team has categorized all customers into 4 segments (A, B, C, D). They plan to use the same strategy for new markets and have identified 2,627 new potential customers. We must help the manager to predict the right group of new clients.

Data

The dataset is Customer Segmentation from Kaggle Website, it contains 8068 rows and 11 columns.

Algorithms

Feature Engineering:

- Applied a Standard Scaler method to scaling data before applying the model.
- Splitting data into train data and test data, 60% of data is used to train models during the learning process.
- Checking missing values

Models

Logistic Regression (LR), Random Forest Classifier (RFC), Decision Tree Classifier (DTC), Gaussian Naive Bayes Classifier, KNeighbors(KNN), XGBClassifier.
XGBClassifier gives higher accuracy.

Tools

These are the technologies and libraries that we will be using for this project:

- Technologies: Python, Jupyter Notebook.
- Libraries: NumPy, Pandas, Matplotlib, Seaborn, plotly, Scikit-learn.

Model Evaluation and Selection

Logistic Regression:

	precision	recall	f1-score	support
A	0.43	0.51	0.47	324
B	0.44	0.24	0.31	304
C	0.52	0.63	0.57	336
D	0.67	0.68	0.68	369
accuracy			0.53	1333
macro avg	0.52	0.52	0.51	1333
weighted avg	0.52	0.53	0.52	1333

KNeighbors:

	precision	recall	f1-score	support
A	0.29	0.25	0.27	324
B	0.32	0.03	0.05	304
C	0.27	0.78	0.40	336
D	0.89	0.13	0.23	369
accuracy			0.30	1333
macro avg	0.44	0.30	0.24	1333
weighted avg	0.46	0.30	0.24	1333

Decision Tree Classifier:

	precision	recall	f1-score	support
A	0.40	0.52	0.45	324
B	0.00	0.00	0.00	304
C	0.47	0.66	0.55	336
D	0.59	0.71	0.65	369
accuracy			0.49	1333
macro avg	0.37	0.47	0.41	1333
weighted avg	0.38	0.49	0.43	1333

Random Forest Classifier:

	precision	recall	f1-score	support
A	0.48	0.51	0.49	324
B	0.42	0.30	0.35	304
C	0.56	0.60	0.58	336
D	0.64	0.71	0.68	369
accuracy			0.54	1333
macro avg	0.52	0.53	0.52	1333
weighted avg	0.53	0.54	0.53	1333

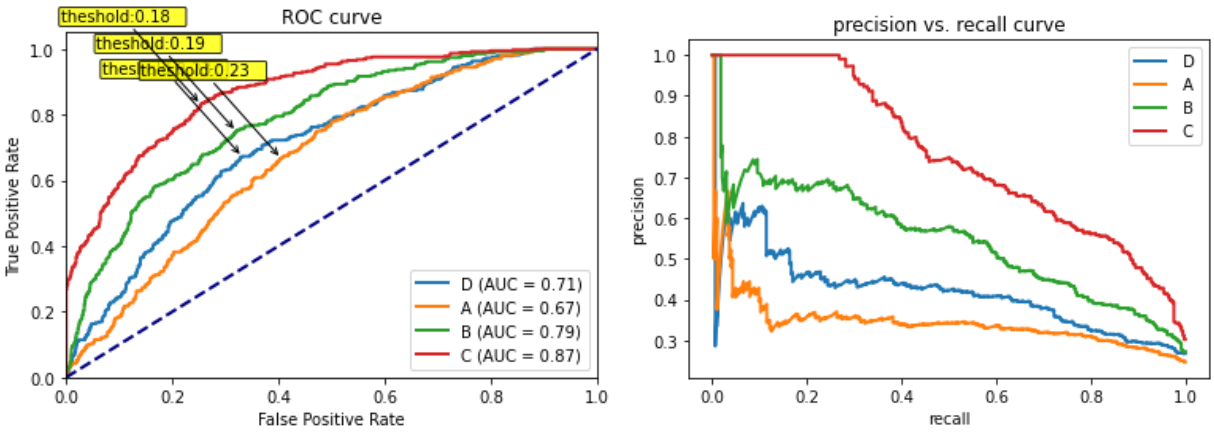
XGB Classifier:

	precision	recall	f1-score	support
A	0.45	0.45	0.45	324
B	0.35	0.30	0.32	304
C	0.52	0.57	0.55	336
D	0.64	0.65	0.65	369
accuracy			0.50	1333
macro avg	0.49	0.49	0.49	1333
weighted avg	0.50	0.50	0.50	1333

Gaussian Naive Bayes

	precision	recall	f1-score	support
A	0.46	0.42	0.44	324
B	0.33	0.14	0.20	304
C	0.43	0.67	0.52	336
D	0.64	0.67	0.65	369
accuracy			0.49	1333
macro avg	0.46	0.47	0.45	1333
weighted avg	0.47	0.49	0.46	1333

The following figure shown the Curve:



Conclusion

Summary of data modeling It can be found that all the performance are generally not very good, However Refer to the accuracy list above, XGBClassifier seems to be the best approach. However, would suggest having a better sampling again for better data modeling.