گزارش نهایی پروژه ی یادگیری ماشین

جهاد دانشگاهی تهران

عنوان پروژه :

تحلیل رفتار مشتریان یک فروشگاه با استفاده از :

RFM -

KMeans Clustering -

- و مدل های طبقه بندی

نام : آزاد زیرک

تاریخ : 15 / 09 / 1404

1/ مقدمه :

هدف این پروژه تحلیل رفتار مشتریان یک فروشگاه و دسته بندی آن ها بر اساس ویژگی های خریدشان است.

1.1/ تحلیل دیتای ایجاد شده جهت استخراج ویژگی های رفتاری مشتریان

1.2/ مدل های یادگیری ماشین شامل:

KMeans Algorithm for Clustering -

Random Forest and Logistic Regression Models for -Predicting Customer Behavior

خروجی نهایی شامل خوشه‌بندی مشتریان، مقایسه مدل‌های پیش‌بینی و تحلیل نتایج است.

2/ آماده سازی و پاک سازی داده ها

در ابتدا داده ها بررسی و مراحل زیر انجام شد:

- حذف داده های ناقص و تکراری

- تبدیل انواع داده ها به فرمت مناسب

- ساخت ویژگی های دیتا فریم جدید شامل:

- (مدت زمان از آخرین خرید) Recency

- (تعداد خریدها) Frequency

- (مجموع مبلغ خریدها) Monetary

در این مرحله یک فایل داده ی تمیز شده نیز ذخیره شد (برای ارائه در پوشه ی نهایی).

3/ ساخت ویژگی های دیتا فریم جدید

پس از پاکسازی برای هر مشتری 3 ویژگی دیتا ست جدید شامل مدت زمان از آخرین خرید و تعداد خریدها و مجموع مبلغ خریدها محاسبه شد.

این ویژگی ها پایه اصلی خوشه بندی و مدل های طبقه بندی بودند.

4/ خوشه بندی با الگوریتم ☐ **KMeans** ☐

در این بخش با استفاده از الگوریتم خوشه بندی توانستیم مشتریان را بر اساس ویژگی های دیتا فریم جدید در خوشه های جداگانه قرار دهیم.

مراحل انجام کار :

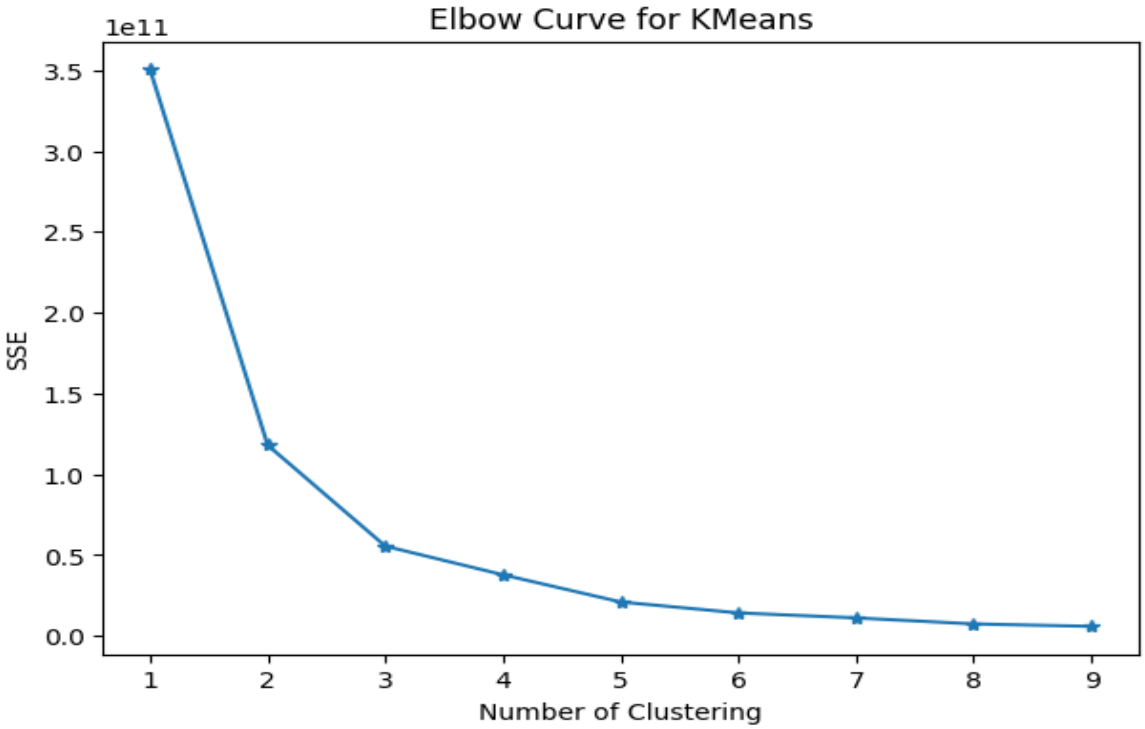
- Standardize data with StandardScaler

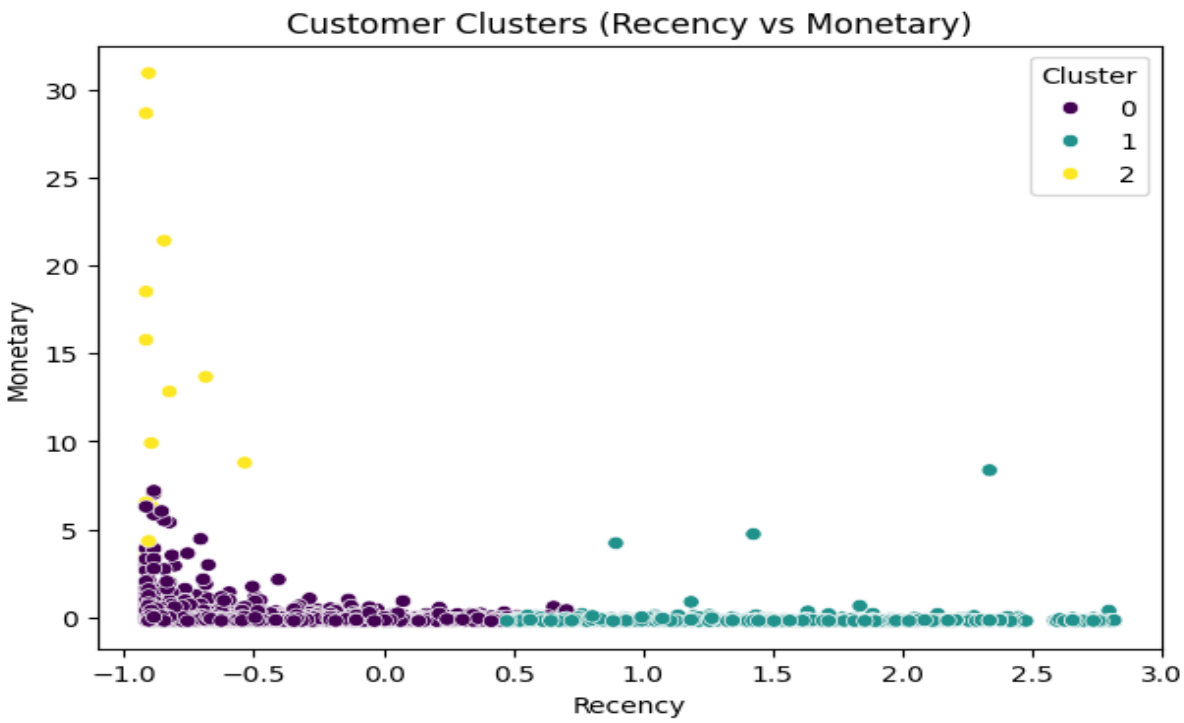
- Test the number of clusters with Elbow Method

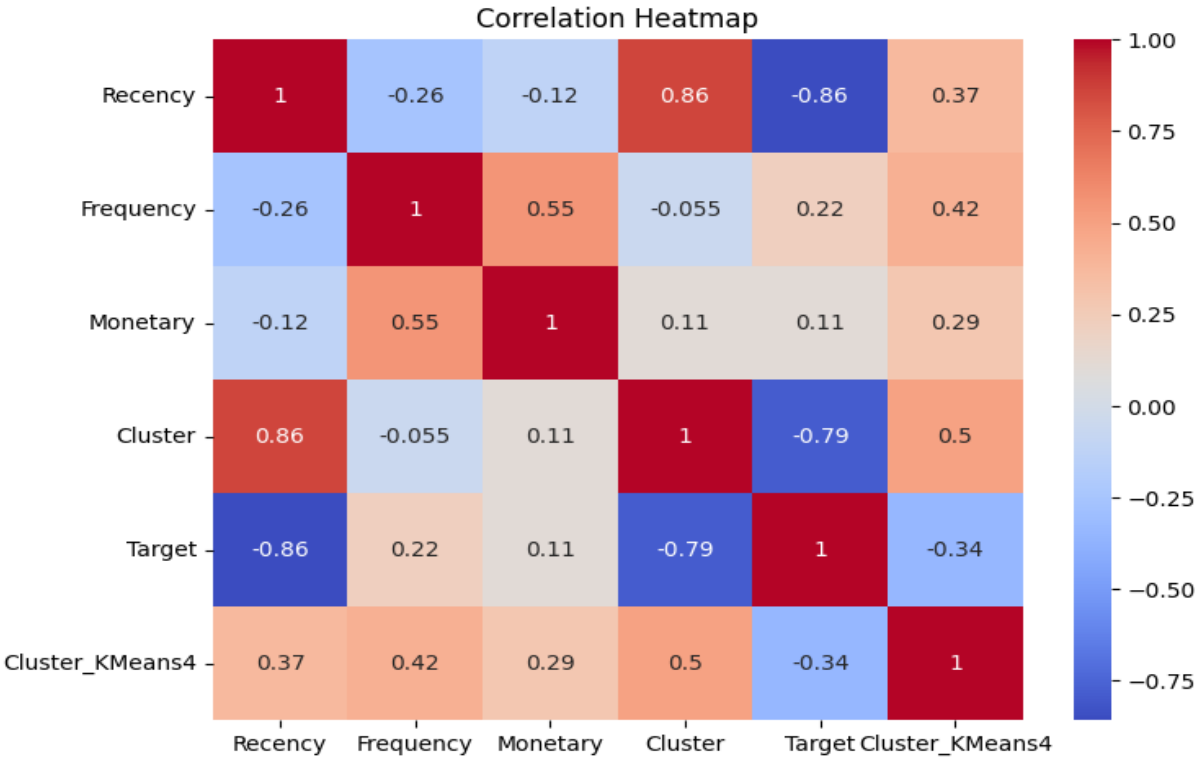
- Choosing the best value of k = 4

- Running KMeans and assigning a cluster to each customer

- Examining the average RFM in each cluster and analyzing the nature of the clusters




Customer Clusters (Recency vs Monetary)
Cluster
0
1
2
Monetary
Recency
30
25
20
15
10
5
0
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0


Correlation Heatmap
Recency
Frequency
Monetary
Cluster
Target
Cluster_KMeans4
1
-0.26
-0.12
0.86
-0.86
0.37
-0.26
1
0.55
-0.055
0.22
0.42
-0.12
0.55
1
0.11
0.11
0.29
0.86
-0.055
0.11
1
-0.79
0.5
-0.86
0.22
0.11
-0.79
1
-0.34
0.37
0.42
0.29
0.5
-0.34
1
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

5/ مدلسازی طبقه بندی

برای پیش بینی رفتار مشتریان یا احتمال تعلق به یک گروه دو مدل مورد استفاده قرار گرفت:

5.1/ Logistic Regression

پارامترهای پایه استفاده شد. نتایج مدل :

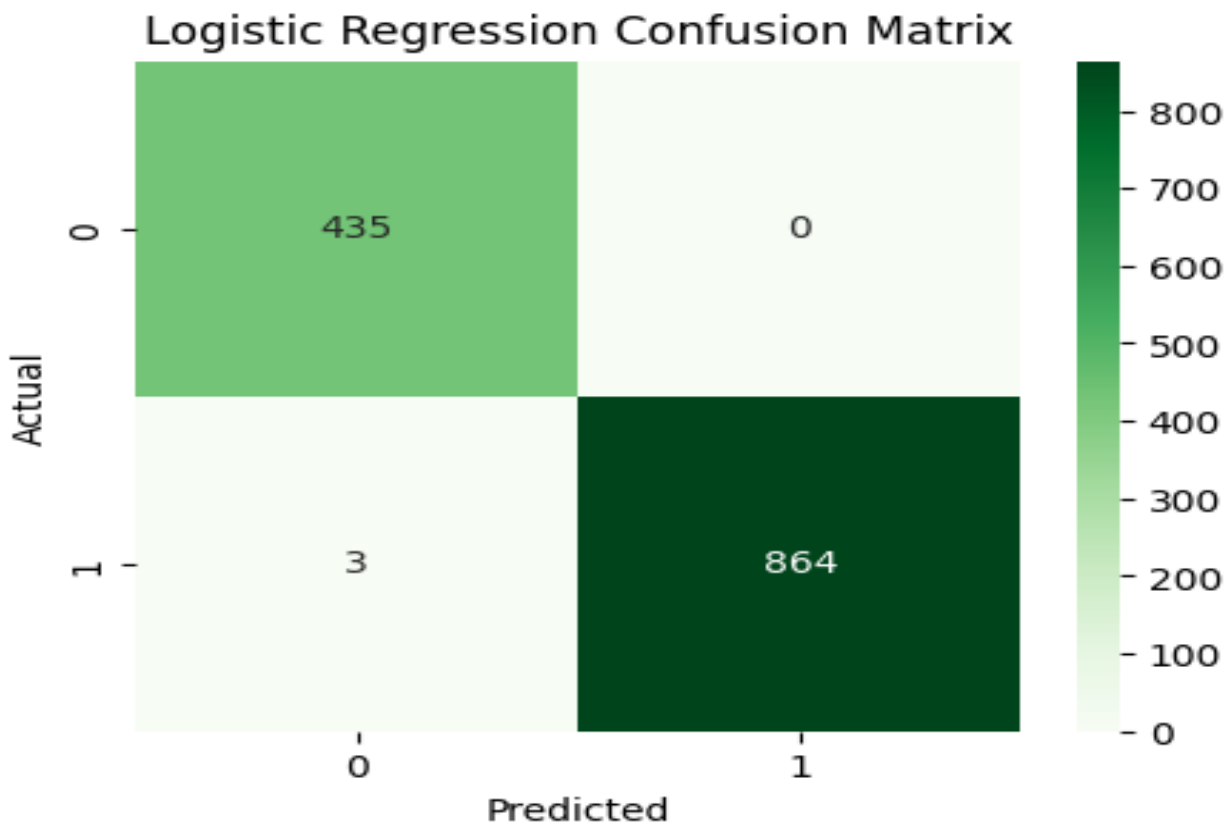
Accuracy: ≈ 0.998

Precision: ≈ 1.000

Recall: ≈ 0.997

F1 Score: ≈ 0.998

Confusion Matrix:



Random Forest /5.2

این مدل با 200 درخت و پارامترهای پیش فرض اجرا شد. نتایج :

Accuracy: 1.00

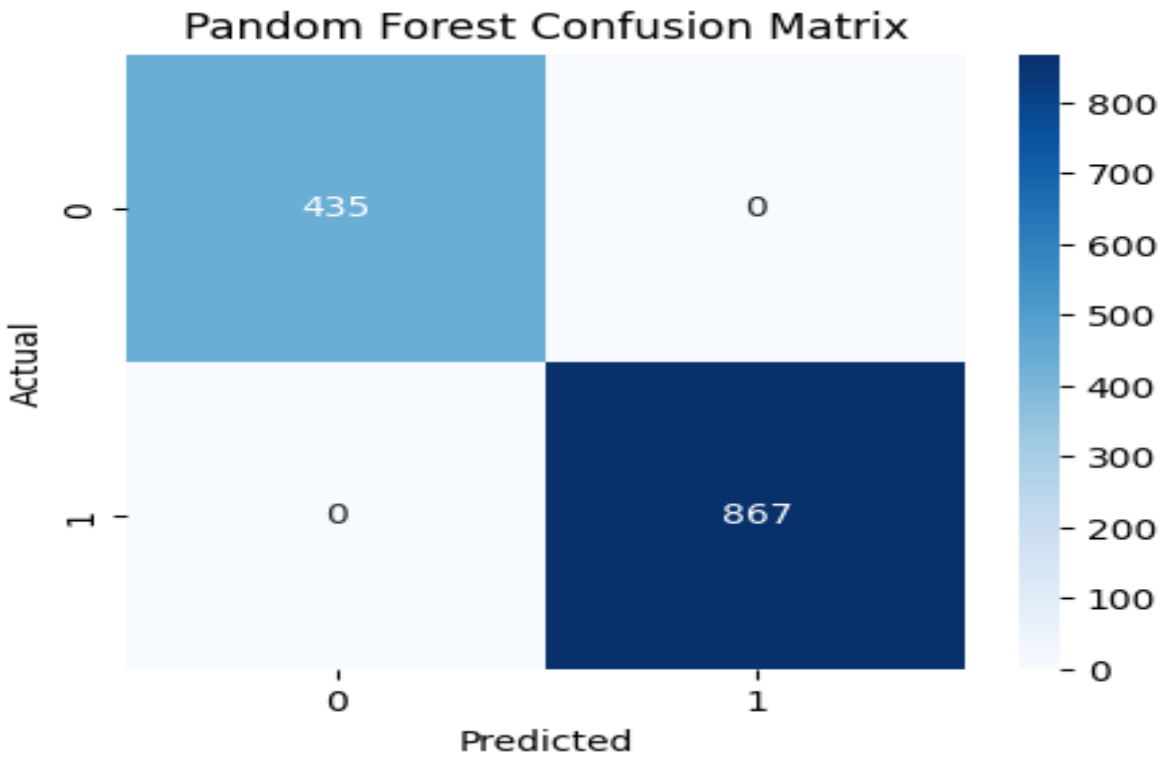
Precision: 1.00

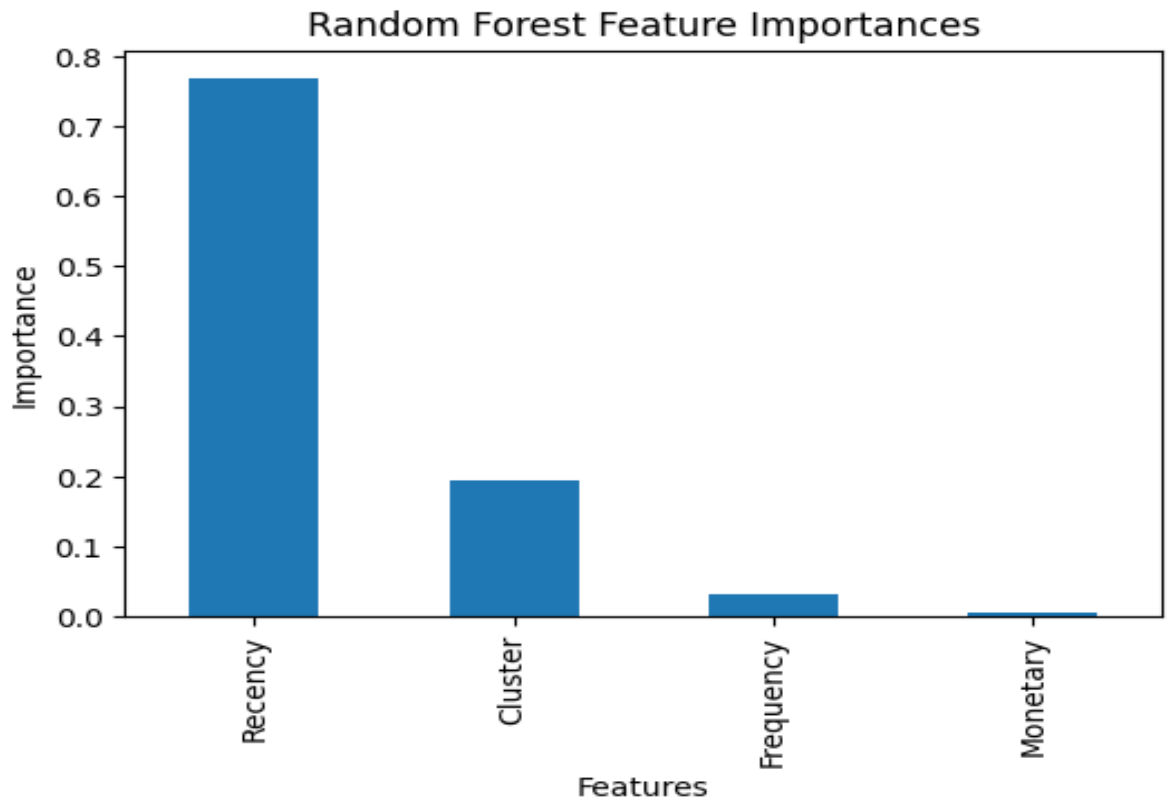
Recall: 1.00

F1 Score: 1.00

AUC: 1.00

Perfect: کاملاً Confusion Matrix


Pandom Forest Confusion Matrix
435
0
0
867
Actual
0
1
Predicted
0
1
800
700
600
500
400
300
200
100
0


Random Forest Feature Importances
Importance
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Recency
Cluster
Frequency
Monetary
Features

همچنین نمودار اهمیت ویژگی ها نیز بررسی شد.

6/ مقایسه مدل ها و انتخاب مدل نهایی

دو مدل زیر اجرا و ارزیابی شدند:

Random Forest –

Logistic Regression –

نتایج نشان می دهد که:

کارایی مدل شماره یک (جنگل تصادفی) بیشتر است.

مدل شماره دو (رگرسیون لجستیک) عملکرد بسیار خوب دارد اما مدل شماره یک (جنگل تصادفی) در معیارهای ارزیابی به صورت کامل و بدون خطا عمل کرد.

خطاهای مدل شماره دئ (رگرسیون لجستیک) تنها مربوط به 3 نمونه اشتباه طبقه بندی شده بود.

بنابراین :

مدل نهایی پیشنهادی :

Random Forest

به دلیل :

توانایی بالا در یادگیری الگوهای غیرخطی

عملکرد پایدار

دقت حداکثری در داده های تست

7/ نتیجه گیری نهایی

در این پروژه دو دستاورد اصلی به دست آمد:

یک/ خوشه بندی مشتریان

با استفاده از ویژگی های خوشه بندی با الگوریتم موردنظر و روش تولید دیتا فریم جدید مشتریان در 4 گروه مشخص تقسیم شدند.

این خوشه ها به تصمیم گیری های بازاریابی کمک می کند.

دو/ پیش بینی رفتار مشتریان

دو مدل طبقه بندی شده اجرا شد و مدل جنگل تصادفی بهترین عملکرد را نشان داد.

این مدل برای پیش بینی احتمال رفتار آینده ی مشتریان و تحلیل سودآوری می تواند بسیار موثر باشد.