Kids Model Target Variable 202212 - Sarah's

```
from pyspark.sql import functions as f
from pyspark.sql.functions import *
from pyspark.sql.window import Window
```

```
/usr/lib/spark/python/pyspark/context.py:123: UserWarning: You are passing in an insecure Py4j gateway.  This presents a security risk, and will be completely forbidden in Spark 3.0
  "You are passing in an insecure Py4j gateway.  This "
```

```
## Kids Pack Purchases in Apr 23

pp_recharges  = spark.sql("""SELECT ban AS account_no, rcg_id, subscription_id, left(rcg_req_date, 10) as date, rcg_req_date, input_amount, pps_balance, pps_new_balance, l9_package_type, l9_package_desc, l9_channel_desc, l9_package_id, prepaid_transaction_type, partition_date_key FROM edw.fct_pp_recharges WHERE rcg_status = 'V' AND input_amount > 0 AND method_code = 'SCRD' AND l9_package_desc IS NOT NULL""")

from pyspark.sql.functions import to_date

pp_recharges = pp_recharges.select('account_no', 'rcg_id', 'subscription_id', to_date(pp_recharges.date, 'yyyy.MM.dd').alias('purchase_date'), 'rcg_req_date', 'input_amount', 'pps_balance', 'pps_new_balance', 'l9_package_type', upper(col('l9_package_desc')).alias('l9_package_desc') , 'l9_channel_desc', 'l9_package_id', 'prepaid_transaction_type', 'partition_date_key')

pp_recharges_fil = pp_recharges.filter(col('purchase_date').between('2023-04-01','2023-04-30') & col('l9_package_desc').contains('KIDS'))

target = pp_recharges_fil.select('account_no').dropDuplicates().withColumn('purchased_pack',lit(1))
```

```
target.select(countDistinct('account_no')).show()
```

```
+--------------------------+
|count(DISTINCT account_no)|
+--------------------------+
|                      8297|
+--------------------------+
```

```
merged_table = spark.read.csv("s3://astro-datalake-prod-sandbox/group_data/njoi/model/feature_engineering/merged_table_202304.csv" , header=True, sep=",")
merged_table_target = merged_table.join(target,'account_no','left').fillna(0, subset=['purchased_pack'])
```

```
merged_table_target.groupBy('purchased_pack').agg(count('account_no')).show()
```

```
+--------------+-----------------+
|purchased_pack|count(account_no)|
+--------------+-----------------+
|             1|            12800|
|             0|           303181|
+--------------+-----------------+
```

```
merged_table_target.repartition(1).write.csv("s3://astro-datalake-prod-sandbox/Amirul/03 NJOI-Sooka/3.1 NJOI/Model/Kids_Pack/merged_table_202304", mode="overwrite" , header=True)
```

```
df_11 = spark.read.csv('s3://astro-datalake-prod-sandbox/Amirul/03 NJOI-Sooka/3.1 NJOI/Model/Kids_Pack/merged_table_202304',header=True)
df_11.printSchema()
```

```
root
 |-- account_no: string (nullable = true)
 |-- race: string (nullable = true)
 |-- cust_state: string (nullable = true)
 |-- cust_region: string (nullable = true)
 |-- Recency: string (nullable = true)
 |-- Frequency: string (nullable = true)
 |-- Monetary: string (nullable = true)
 |-- R_Ratio: string (nullable = true)
 |-- F_Ratio: string (nullable = true)
 |-- M_Ratio: string (nullable = true)
 |-- recency_quartile: string (nullable = true)
 |-- frequency_quartile: string (nullable = true)
 |-- monetary_quartile: string (nullable = true)
 |-- RFM_Score: string (nullable = true)
 |-- Recency_Flag: string (nullable = true)
 |-- Frequency_Flag: string (nullable = true)
 |-- Segmentation: string (nullable = true)
```

```
z.show(df_11.limit(5))
```

account_no ▾	race ▾	cust_state ▾	cust_region ▾	Recency ▾	Frequency ▾	Monetary ▾	R_Ratio ▾	F_Ratio ▾	M_Ratio ▾	recency_quartile ▾	frequency_quartile ▾	monetary_quartil
89758399	MALAY	SAR	SARAWAK	273	1	45.0	0.75	0.0	0.01	1	2	2
90385032	MALAY	KED	NORTHERN	0	22	625.0	0.0	0.05	0.08	4	4	4
90695135	MALAY	MEL	SOUTHERN	148	2	31.8	0.41	0.0	0.0	2	2	2
90781672	MALAY	SEL	CENTRAL	0	4	70.0	0.0	0.01	0.01	4	3	2
90943746	CHINESE	SAB	SABAH	0	17	279.0	0.0	0.04	0.04	4	4	4

```
## Kids Pack Purchases in Dec 2022

pp_recharges  = spark.sql("""SELECT ban AS account_no, rcg_id, subscription_id, left(rcg_req_date, 10) as date, rcg_req_date, input_amount, pps_balance, pps_new_balance, l9_package_type, l9_package_desc, l9_channel_desc, l9_package_id, prepaid_transaction_type, partition_date_key FROM edw.fct_pp_recharges WHERE rcg_status = 'V' AND input_amount > 0 AND method_code = 'SCRD' AND l9_package_desc IS NOT NULL""")

from pyspark.sql.functions import to_date

pp_recharges = pp_recharges.select('account_no', 'rcg_id', 'subscription_id', to_date(pp_recharges.date, 'yyyy.MM.dd').alias('purchase_date'), 'rcg_req_date', 'input_amount', 'pps_balance', 'pps_new_balance', 'l9_package_type', upper(col('l9_package_desc')).alias('l9_package_desc') , 'l9_channel_desc', 'l9_package_id', 'prepaid_transaction_type', 'partition_date_key')

pp_recharges_fil = pp_recharges.filter(col('purchase_date').between('2022-11-01','2022-11-30') & col('l9_package_desc').contains('KIDS PACK'))

target = pp_recharges_fil.select('account_no').dropDuplicates().withColumn('purchased_pack',lit(1))

merged_table = spark.read.csv("s3://astro-datalake-prod-sandbox/group_data/njoi/model/feature_engineering/merged_table_202210.csv" , header=True, sep=",")
merged_table_target = merged_table.join(target,'account_no','left').fillna(0, subset=['purchased_pack'])
```

```
merged_table_target.repartition(1).write.csv("s3://astro-datalake-prod-sandbox/Amirul/03 NJOI-Sooka/3.1 NJOI/Model/Kids_Pack/merged_table_202211_target_v2", mode="overwrite" , header=True)
```