Universidad de Navarra

Máster Oficial Big Data Science

Predicción de Contratación de Productos:Un Enfoque Analítico para la Estrategia Empresarial

Año académico 2023/2024

Autores: Alfonso Andrés Lafuente, Mario Lamas Herrera, Mario Pérez Vicente, Jaime Vila Sagaseta de Ilúrdoz.

Tutor: Matías Ávila

1. INTRODUCCIÓN

Nos enfrentamos a un reto significativo: predecir cuáles de los 25 productos distintos que ofrece una empresa serán contratados por sus clientes el próximo mes. Cada cliente puede tener un máximo de un producto activo por tipología y los contratos son mensuales, renovables indefinidamente. Disponemos de una base de datos extensa que incluye tanto las características de los clientes como su historial mensual de productos activos. Nuestra tarea es anticipar no sólo los productos que se mantendrán, sino específicamente los nuevos contratos que se añadirán, lo que plantea un desafío complejo y multifacético.

Para abordar este problema, hemos colaborado de manera efectiva utilizando GitHub, organizando nuestro trabajo en dos Jupyter notebooks principalmente. El primer archivo se dedica al análisis y la información de los datos, permitiéndonos explorar y comprender en profundidad las características y patrones presentes en el conjunto de datos. El segundo archivo se enfoca en la limpieza y el modelado, donde realizamos los cambios necesarios en el dataset y aplicamos los modelos de machine learning. Esta estructura nos ha permitido trabajar de manera eficiente y colaborativa, asegurando que cada paso del proceso esté meticulosamente documentado y que nuestras contribuciones se integren de manera coherente.

Con este enfoque, buscamos no solo la precisión en nuestras predicciones, sino también generar insights profundos y accionables que puedan transformar la estrategia de productos de la empresa y mejorar significativamente la experiencia del cliente.

2. EDA (Análisis Exploratorio de los Datos)

El primer paso para resolver el problema que se plantea es comprender los datos que disponemos y estudiarlos para plantear una estrategia adecuada para resolver el problema eficientemente. El dataset cuenta con 635.000 observaciones con 48 atributos cuya descripción se adjunta en el anexo y que describen el comportamiento de compra histórico de cada usuario para una serie de 25 productos.

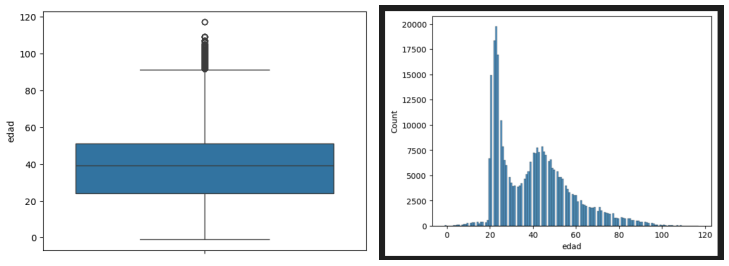
2.1 Valores Faltantes

A excepción de las columnas ***'cod_persona','mes','edad','num_antiguedad' y 23 de los 25 productos***, se observó que el resto contaban con un número considerable de valores nulos cuyo procesamiento se expone en el apartado de Limpieza de datos.

2.2 Análisis descriptivo

Con el fin de asegurar la calidad y comprender los datos, se calcularon algunos estadísticos como: 'media','desviación estándar','percentiles'...

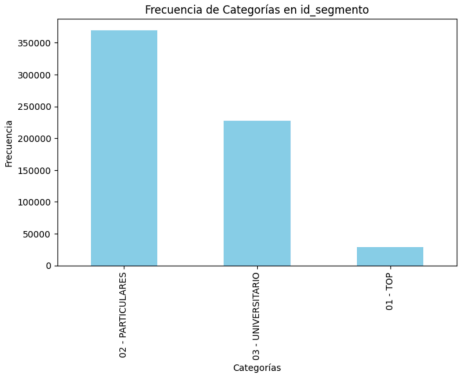
Para entender de manera visual la distribución de los datos numéricos generamos histogramas y diagrama de cajas.



Una de las conclusiones de este análisis fue que las edades se distribuyen principalmente en 2 grupos, menores (<18 años), jóvenes (18-30 años) y adultos (+30 años).

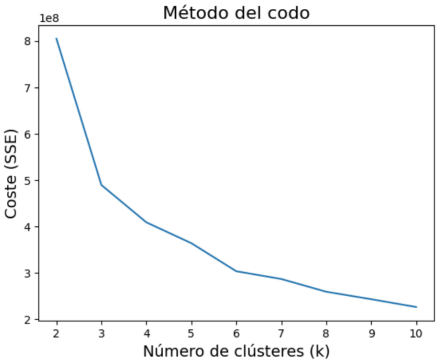
Posteriormente repetimos el análisis para las variables categóricas con información como el número de categorías, la categoría más frecuente …

Los gráficos de barras nos permitieron entender de forma visual esta información.

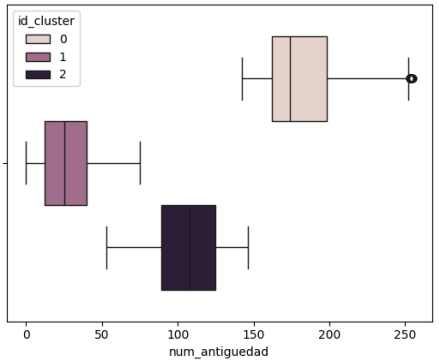


2.3 Clustering

Debido a la limitada información que disponíamos, decidimos hacer un clustering para comprender las agrupaciones existentes en nuestros datos y tratar de caracterizarlos. El paso inicial fue definir el número de clusters, para ello empleamos 3 métodos, el método del codo y silhouette, llegando a la conclusión que para favorecer la explicabilidad manteniendo el rendimiento, el número óptimo de clusters eran 3.



Debido al elevado tiempo de cálculo, se realizó este proceso sobre una muestra aleatoria de los datos. A continuación, procedimos a caracterizar los clusters resultantes.



La variable que mostraba mayor varianza en este análisis es el **número de antiguedad** y el **id_segmento** donde se aprecia una diferencia considerable entre las 3 agrupaciones.

2.4 Análisis de productos por id_segmento y edad

Con las conclusiones obtenidas del clustering, decidimos comprobar qué productos compran los usuarios en función del segmento al que pertenecen, las conclusiones fueron las siguientes:

- **PARTICULARES:**

  Este grupo de clientes no compran el producto 2.

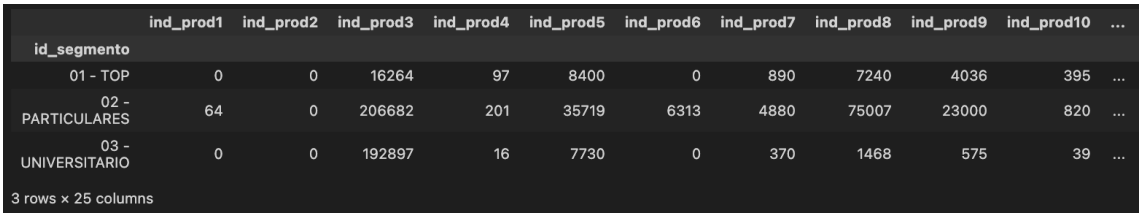
- **UNIVERSITARIOS:**

  Este grupo de clientes no compran los productos: 1,2,6,15 y 21.

- **TOP:**

  Este grupo de clientes no compran los productos: 1,2 y 6.

Resumen:



id_segmento	ind_prod1	ind_prod2	ind_prod3	ind_prod4	ind_prod5	ind_prod6	ind_prod7	ind_prod8	ind_prod9	ind_prod10	...
01 - TOP	0	0	16264	97	8400	0	890	7240	4036	395	...
02 - PARTICULARES	64	0	206682	201	35719	6313	4880	75007	23000	820	...
03 - UNIVERSITARIO	0	0	192897	16	7730	0	370	1468	575	39	...

3 rows × 25 columns

Finalmente, decidimos realizar el mismo análisis empleando la categorización de edad expuesta anteriormente.Creamos una nueva variable llamada “edad_dividida” para agrupar diferentes edades en los siguientes grupos de edad: “menor”, “adolescente”, “universitario” y “adulto”. Al comparar los diferentes grupos, se observan patrones claros en la compra de productos.

Los adultos son los mayores compradores en términos absolutos, con productos como ind_prod3, ind_prod13 e ind_prod24 liderando las ventas. Además, la compra de productos en este grupo es amplia, cubriendo prácticamente todos los productos disponibles.

Los adolescentes presentan un patrón de compra muy limitado, concentrándose casi exclusivamente en ind_prod6, con muy pocas compras en otros productos. Este grupo no muestra interés significativo en la mayoría de los productos.

Por último, los menores tienen un volumen de compras insignificante, con la mayoría de los productos con 0 compras en este grupo, lo que refleja una baja participación en el mercado.

Estos datos subrayan la importancia de adaptar las estrategias de marketing según la combinación de edad y segmento para maximizar la efectividad.

Tras estudiar la información, decidimos plantear la resolución del problema, realizando por cada tipo de id_segmento, definimos un modelo de clasificación para predecir cada uno de los productos que los clientes de dicho segmento han comprado, para los producto que no han sido comprados, no vamos a recomendarlos si el cliente pertenece a este segmento.

3. LIMPIEZA Y PREPROCESAMIENTO DE DATOS

Con las conclusiones obtenidas durante el proceso de exploración y caracterización de los datos, procedemos a la limpieza de la información para asegurar la calidad y maximizar el rendimiento del modelo.

Para llevar a cabo la limpieza de los datos, se reemplazan los valores erróneos como ‘ NA’ y aquellos que representan datos faltantes con “-1”, estos posteriormente serán procesados en función de la columna a la que pertenecen.

El mapeo de variables categóricas a valores numéricos se realiza mediante la definición de los siguientes diccionarios, que representan un LabelEncoder:

```
emp_dict = {'N':0,-1:-1,'A':1,'B':2,'F':3,'S':4}
inreaderall_dict = {'N':0,-1:-1,'S':1}
sexo_dict = {'V':0,'H':1,-1:-1}
tiprel_dict = {'A':0,-1:-1,'I':1,'P':2,'N':3,'R':4}
indresi_dict = {'N':0,-1:-1,'S':1}
indext_dict = {'N':0,-1:-1,'S':1}
conyuemp_dict = {'N':0,-1:-1,'S':1}
segmento_dict = {-1:4,'01 - TOP':1,'02 - PARTICULARES':2,'03 - UNIVERSITARIO':3}
```

El preprocesamiento de datos continúa convirtiendo columnas clave a tipos de datos numéricos más eficientes para optimizar el uso de memoria. Además, se reemplazan valores anómalos en la columna “xti_rel”, cambiando el 99 por 2 para mantener la coherencia de los datos. Se utilizan los diccionarios anteriores para mapear variables categóricas a valores numéricos, asegurando que las columnas sean adecuadas para el modelado.

Para aquellas variables de tipo fecha se extrajo información como el mes, año y día, posteriormente generamos el diccionario con la selección de productos comentada en el apartado 2.4.

Conocíamos que la información de productos positivos era reducida, para explorar esta hipótesis, creamos una clasificación para observar las columnas que solo contenían 0, aquellas con menos de un 0.1% de valores positivos, aquellas con entre un 0.1% y 0.5% de valores positivo y finalmente aquellos con más de un 0.5% de valores positivos.

Al principio del trabajo pensamos en que al haber varias observaciones del mismo individuo en distintas fechas era buena idea hacer series temporales. Tras hacer una prueba con redes neuronales (tratándolo como series temporales) y otra con una clasificación (usando solamente las últimas fechas) vimos que la diferencia era insignificante, por lo que decidimos quedarnos con la idea de clasificación al ser más sencilla.

Decidimos entonces añadir columnas con la información de las compras del N periodo pasado, en este caso 1. De esta forma conservamos algo de

información histórica sin aumentar considerablemente el tamaño del dataset con técnicas como por ejemplo la generación de ventana.

La idea principal era realizar el proceso completo para diferentes N periodos, por ejemplo de 1 a 5 y hacer una media de las predicciones de los diferentes modelos. Sin embargo, supone un coste computacional que no podemos afrontar. Consideramos igualmente que de esta forma mejoraría la robustez de la solución.

Para ejecutar lo anterior se crean nuevas columnas en el conjunto de datos reducido mediante la función “shift”, la cual desplaza los valores de las columnas hacia abajo en un número específico de filas (en este caso, de 1 a 17). Esto se hace para capturar el estado anterior de ciertas variables. También se generan condiciones (“DIFF_CONDS”) para identificar cambios en el ID del usuario entre registros consecutivos. Luego, estas nuevas columnas se utilizan para calcular diferencias y establecer valores específicos en función de estas condiciones. Esto ayuda a crear características temporales que reflejan cambios en el comportamiento de los usuarios.

4. MODELAJE Y EVALUACIÓN

4.1 Modelaje

Trás realizar una investigación sobre el estado del arte de los modelos de clasificación, concluimos que Catboost mostraba el mejor rendimiento para este tipo de tareas, empleando la selección de productos por segmento, entrenamos un modelo de Catboost para cada producto que aplicaba en cada segmento y guardamos dichos modelos y sus correspondientes métricas.

Decidimos dividir las muestras en 80% datos de entrenamiento y 20% de test ya que consideramos que es la distribución adecuada para las dimensiones de datos que manejamos.

Para corregir el desbalance de clases, calculamos los pesos de cada clase y se los pasamos al modelo.

Obteniendo las métricas medias por tipo de segmento (se muestran las categorías como números) que se muestran a continuación:

```
Valores medios por segmento:  4
Accuracy medio:  0.9638532290306554
Recall medio:  0.908732481166843
Valores medios por segmento:  2
Accuracy medio:  0.9608985321362405
Recall medio:  0.7816858722866797
Valores medios por segmento:  3
Accuracy medio:  0.9577446655524896
Recall medio:  0.8262330727975903
Valores medios por segmento:  1
Accuracy medio:  0.9592054389649982
Recall medio:  0.793933084125231
```

Diccionario para mapear id_segmento:

```
segmento_dict = {-1:4,'01 - TOP':1,'02 - PARTICULARES':2,'03 - UNIVERSITARIO':3}
```

Con estas métricas, evaluamos el rendimiento de predicción empleando el penúltimo periodo como variable para predecir(X) y el último periodo como etiqueta (y).

Como se puede observar en las métricas, los resultados son positivos teniendo en cuenta que se están prediciendo 25 variables.

5. RECOMENDACIONES FINALES

Para poder generar la predicción sobre el próximo periodo, teniendo en cuenta que para el entrenamiento hemos empleado los productos del penúltimo periodo, debemos sustituir estos por los del último periodo, de esta manera obtenemos las probabilidades de que el cliente compre dicho producto en el próximo periodo y para aquellos producto que por clasificación de segmento hemos decidido no recomendar, retornamos un cero.

Empleando la librería Pandas, creamos un dataframe con los resultados para facilitar el procesamiento final de la solución que se divide en los siguientes pasos:

- Añadimos a nuestro set de datos de predicciones la información de los productos para el periodo anterior para en caso de que el usuario ya contará con el producto, no recomendarselo.

- A continuación, creamos una columna **'predicted'** donde ordenamos las predicciones de mayor a menor siempre y cuando superen un umbral de 0.5 y las sustituimos por el nombre de la columna a la que pertenecen para finalmente guardar las predicciones en un archivo **predicciones.csv.**

- Finalmente, para los valores de los productos sustituimos la probabilidad por 1 en caso de que supere el umbral o por 0 en caso contrario y terminamos guardando este archivo con el nombre **soluciones.csv.**

ANEXO

A) Descripción de las variables

Nombre Variable	Descripción
cod_persona	Customer code
mes	The table is partitioned for this column
pais	Customer's Country residence
sexo	Customer's sex
edad	Customer’s age
fecha1	The date in which the customer became as the first holder...
xti_empleado	Employee index: A active, B ex employed, F filial, N not...
xti_nuevo_cliente	New customer Index. 1 if the customer registered in the...
num_antiguedad	Customer seniority (in months)
xti_rel	1 (First/Primary), 99 (Primary customer during the month...
fec_ult_cli_1t	Last date as primary customer (if he isn't at the end of...
xti_rel_1mes	Customer type at the beginning of the month ,1 (First/Pr...
tip_rel_1mes	Customer relation type at the beginning of the month, A ...
indresi	Residence index (S (Yes) or N (No) if the residence coun...
indext	Foreigner index (S (Yes) or N (No) if the customer's bir...
des_canal	Channel used by the customer to join
xti_extra	Deceased index. N/S
tip_dom	Addres type. 1, primary address

cod_provincia	Province code (customer's address)
xti_actividad_cliente	Activity index (1, active customer; 0, inactive customer)
imp_renta	Gross income of the household
id_segmento	Segmentation: 01 - VIP, 02 - Individuals 03 - college gr...
mean_engagement	Mean customer engagement
ind_prod1	1 (customer uses the product this month) 0 (the customer...
ind_prod2	1 (customer uses the product this month) 0 (the customer...
…	
ind_prod25	1 (customer uses the product this month) 0 (the customer...

B) Métricas de los diferentes modelos

Resultados del Segmento 4

Segmento	Producto	Accuracy	Recall	Matriz de Confusión
4	ind_prod3	0.8410645575032065	0.904480722473081	[[2661, 937], [550, 5208]]
4	ind_prod5	0.8399957246686618	0.8231404958677686	[[7361, 1390], [107, 498]]
4	ind_prod6	0.999786233433091	1.0	[[9278, 2], [0, 76]]
4	ind_prod7	0.9763787943565626	0.6752136752136753	[[9056, 183], [38, 79]]
4	ind_prod8	0.8760153911928175	0.938953488372093	[[7227, 1097], [63, 969]]
4	ind_prod10	0.9927319367250962	0.35135135135135137	[[9275, 44], [24, 13]]
4	ind_prod12	0.8892689183411714	0.7294117647058823	[[8072, 944], [92, 248]]
4	ind_prod13	0.8367892261650278	0.84	[[7346, 1435], [92, 483]]
4	ind_prod16	0.9790508764429243	0.15294117647058825	[[9147, 124], [72, 13]]
4	ind_prod18	0.8654339461308251	0.7236842105263158	[[7822, 1154], [105, 275]]
4	ind_prod19	0.8665027789653699	0.7822349570200573	[[7834, 1173], [76, 273]]
4	ind_prod24	0.8420265070542967	0.8693918245264207	[[7006, 1347], [131, 872]]
4	ind_prod25	0.8575245831551945	0.8322981366459627	[[7621, 1252], [81, 402]]

Resultados del Segmento 2

Producto	Accuracy	Recall	Matriz de Confusión
ind_prod1	0.9998931167165456	0.0	[[9355, 0], [1, 0]]
ind_prod3	0.8410645575032065	0.904480722473081	[[2661, 937], [550, 5208]]
ind_prod4	0.9991449337323642	0.0	[[9348, 2], [6, 0]]
ind_prod5	0.8399957246686618	0.8231404958677686	[[7361, 1390], [107, 498]]
ind_prod6	0.999786233433091	1.0	[[9278, 2], [0, 76]]
ind_prod7	0.9763787943565626	0.6752136752136753	[[9056, 183], [38, 79]]
ind_prod8	0.8760153911928175	0.938953488372093	[[7227, 1097], [63, 969]]
ind_prod9	0.9011329628046174	0.7287671232876712	[[8165, 826], [99, 266]]
ind_prod10	0.9927319367250962	0.35135135135135137	[[9275, 44], [24, 13]]
ind_prod11	0.9965797349294571	0.0	[[9324, 11], [21, 0]]
ind_prod12	0.8892689183411714	0.7294117647058823	[[8072, 944], [92, 248]]
ind_prod13	0.8367892261650278	0.84	[[7346, 1435], [92, 483]]
ind_prod14	0.9512612227447628	0.3706293706293706	[[8847, 366], [90, 53]]
ind_prod15	0.9890979050876443	0.125	[[9248, 60], [42, 6]]
ind_prod16	0.9790508764429243	0.15294117647058825	[[9147, 124], [72, 13]]
ind_prod17	0.9972210346301839	0.5217391304347826	[[9318, 15], [11, 12]]
ind_prod18	0.8654339461308251	0.7236842105263158	[[7822, 1154], [105, 275]]
ind_prod19	0.8665027789653699	0.7822349570200573	[[7834, 1173], [76, 273]]
ind_prod20	0.9197306541256948	0.6022099447513812	[[8496, 679], [72, 109]]
ind_prod21	0.993587002992732	0.041666666666666664	[[9295, 37], [23, 1]]
ind_prod22	0.8601966652415562	0.7529691211401425	[[7731, 1204], [104, 317]]
ind_prod23	0.8527148353997435	0.7548387096774194	[[7627, 1264], [114, 351]]
ind_prod24	0.8420265070542967	0.8693918245264207	[[7006, 1347], [131, 872]]
ind_prod25	0.8575245831551945	0.8322981366459627	[[7621, 1252], [81, 402]]

Segmento: 3

Producto	Accuracy	Recall	Matriz de Confusión
ind_prod3	0.8410645575032065	0.904480722473081	[[2661, 937], [550, 5208]]
ind_prod4	0.9991449337323642	0.0	[[9348, 2], [6, 0]]
ind_prod5	0.8399957246686618	0.8231404958677686	[[7361, 1390], [107, 498]]
ind_prod7	0.9763787943565626	0.6752136752136753	[[9056, 183], [38, 79]]
ind_prod8	0.8760153911928175	0.938953488372093	[[7227, 1097], [63, 969]]
ind_prod9	0.9011329628046174	0.7287671232876712	[[8165, 826], [99, 266]]
ind_prod10	0.9927319367250962	0.35135135135135137	[[9275, 44], [24, 13]]
ind_prod11	0.9965797349294571	0.0	[[9324, 11], [21, 0]]
ind_prod12	0.8892689183411714	0.7294117647058823	[[8072, 944], [92, 248]]
ind_prod13	0.8367892261650278	0.84	[[7346, 1435], [92, 483]]
ind_prod14	0.9512612227447628	0.3706293706293706	[[8847, 366], [90, 53]]
ind_prod16	0.9790508764429243	0.15294117647058825	[[9147, 124], [72, 13]]
ind_prod17	0.9972210346301839	0.5217391304347826	[[9318, 15], [11, 12]]
ind_prod18	0.8654339461308251	0.7236842105263158	[[7822, 1154], [105, 275]]
ind_prod19	0.8665027789653699	0.7822349570200573	[[7834, 1173], [76, 273]]
ind_prod20	0.9197306541256948	0.6022099447513812	[[8496, 679], [72, 109]]
ind_prod22	0.8601966652415562	0.7529691211401425	[[7731, 1204], [104, 317]]
ind_prod23	0.8527148353997435	0.7548387096774194	[[7627, 1264], [114, 351]]
ind_prod24	0.8420265070542967	0.8693918245264207	[[7006, 1347], [131, 872]]
ind_prod25	0.8575245831551945	0.8322981366459627	[[7621, 1252], [81, 402]]

Resumen de Segmento 1

Producto	Accuracy	Recall	Matriz de Confusión
ind_prod3	0.8410645575032065	0.904480722473081	[[2661, 937], [550, 5208]]
ind_prod4	0.9991449337323642	0.0	[[9348, 2], [6, 0]]
ind_prod5	0.8399957246686618	0.8231404958677686	[[7361, 1390], [107, 498]]
ind_prod7	0.9763787943565626	0.6752136752136753	[[9056, 183], [38, 79]]
ind_prod8	0.8760153911928175	0.938953488372093	[[7227, 1097], [63, 969]]
ind_prod9	0.9011329628046174	0.7287671232876712	[[8165, 826], [99, 266]]
ind_prod10	0.9927319367250962	0.35135135135135137	[[9275, 44], [24, 13]]
ind_prod11	0.9965797349294571	0.0	[[9324, 11], [21, 0]]
ind_prod12	0.8892689183411714	0.7294117647058823	[[8072, 944], [92, 248]]
ind_prod13	0.8367892261650278	0.84	[[7346, 1435], [92, 483]]
ind_prod14	0.9512612227447628	0.3706293706293706	[[8847, 366], [90, 53]]
ind_prod15	0.9890979050876443	0.125	[[9248, 60], [42, 6]]
ind_prod16	0.9790508764429243	0.15294117647058825	[[9147, 124], [72, 13]]
ind_prod17	0.9972210346301839	0.5217391304347826	[[9318, 15], [11, 12]]
ind_prod18	0.8654339461308251	0.7236842105263158	[[7822, 1154], [105, 275]]
ind_prod19	0.8665027789653699	0.7822349570200573	[[7834, 1173], [76, 273]]
ind_prod20	0.9197306541256948	0.6022099447513812	[[8496, 679], [72, 109]]
ind_prod21	0.993587002992732	0.041666666666666664	[[9295, 37], [23, 1]]
ind_prod22	0.8601966652415562	0.7529691211401425	[[7731, 1204], [104, 317]]
ind_prod23	0.8527148353997435	0.7548387096774194	[[7627, 1264], [114, 351]]
ind_prod24	0.8420265070542967	0.8693918245264207	[[7006, 1347], [131, 872]]
ind_prod25	0.8575245831551945	0.8322981366459627	[[7621, 1252], [81, 402]]