

Desafío Itaú

Proyecto Semestral - Predicción de compra de productos

Curso: IN7615 - Aprendizaje automático con Redes Neuronales
Integrantes: Felipe Jorquera Díaz
Constanza Peña Soto

Contenidos



1. Contexto y metodología
2. Desarrollo del proyecto
3. Resultados y discusión
4. Conclusiones

1. Contexto y metodología

1.1. Banco Itaú

Banco Itaú Corpbanca es una institución financiera con base en Chile que pertenece al holding de Banco Itaú (Brasil) y que posee presencia en Perú, Colombia y Panamá.

Dentro de sus productos principales se encuentran cuentas bancarias, inversiones, seguros y créditos de todo tipo.



1.2. Desafío Itaú

Se busca generar modelos predictivos que permitan predecir la compra de productos bancarios, para orientar mejor las campañas.

Para ello se trabajará usando 4 datasets de interés:

1. **Datos demográficos.**
2. **Datos transaccionales de 16 productos bancarios.**
3. **Datos de campañas.**
4. **Datos de comunicaciones enviadas.**

Dentro del desafío se encuentra lograr un entendimiento y manejo de los datos, y obtener el mejor modelo para realizar predicciones sobre 5 de estos 16 productos.








1.3. Metodología de trabajo



La metodología de trabajo utilizada fue CRISP-DM, la cual se basa en un **entendimiento** íntegro tanto del **problema de negocio** como de los **datos** para así **preprocesar** la información que servirá de “input” a los **modelos** y finalmente **evaluar** su desempeño.

1.4. Herramientas

Las herramientas utilizadas en el proyecto fueron:

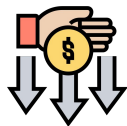
Descripción	Detalle
Lenguaje de programación	 Python
Environment	 Anaconda
Versionamiento	 Github
IDLEs	 Spyder,  Jupyter
Dashboards	 PowerBI
Otros	 MS Excel

2. Desarrollo del proyecto

2.1. Entendimiento del problema de negocio

El objetivo de Itaú es mejorar la orientación de sus campañas de compra de productos mediante la segmentación de sus clientes. **Si sabe que clientes utilizarán o no los productos del banco, entonces las campañas podrían ser focalizadas** con tal de aumentar la eficacia y obtener otros beneficios asociados a las operaciones, costos, métricas, etc.

Beneficios de la focalización de campañas



Reducción de costos asociados a la generación y despliegue de campañas.



Mejora de KPIs al focalizar clientes propensos tanto a las compras como a las campañas.



Aumento del uso de productos y fidelización de clientes.

2.2. Entendimiento del problema de analítica

El objetivo es mejorar la orientación de sus campañas de compra de productos. Para ello, se debe conocer que clientes comprarán un producto en los siguientes meses.

En base a lo anterior, se propone un **modelo de clasificación que prediga con cuál probabilidad un cliente comprará un producto específico en los siguientes 3 meses.**

Además, como métrica de desempeño se escoge el “**Recall**”, ya que el objetivo es orientar de la mejor forma posible las campañas en el segmento de clientes objetivo, sin importar los “Falsos Positivos”. Sin embargo, también se considerarán otras métricas.

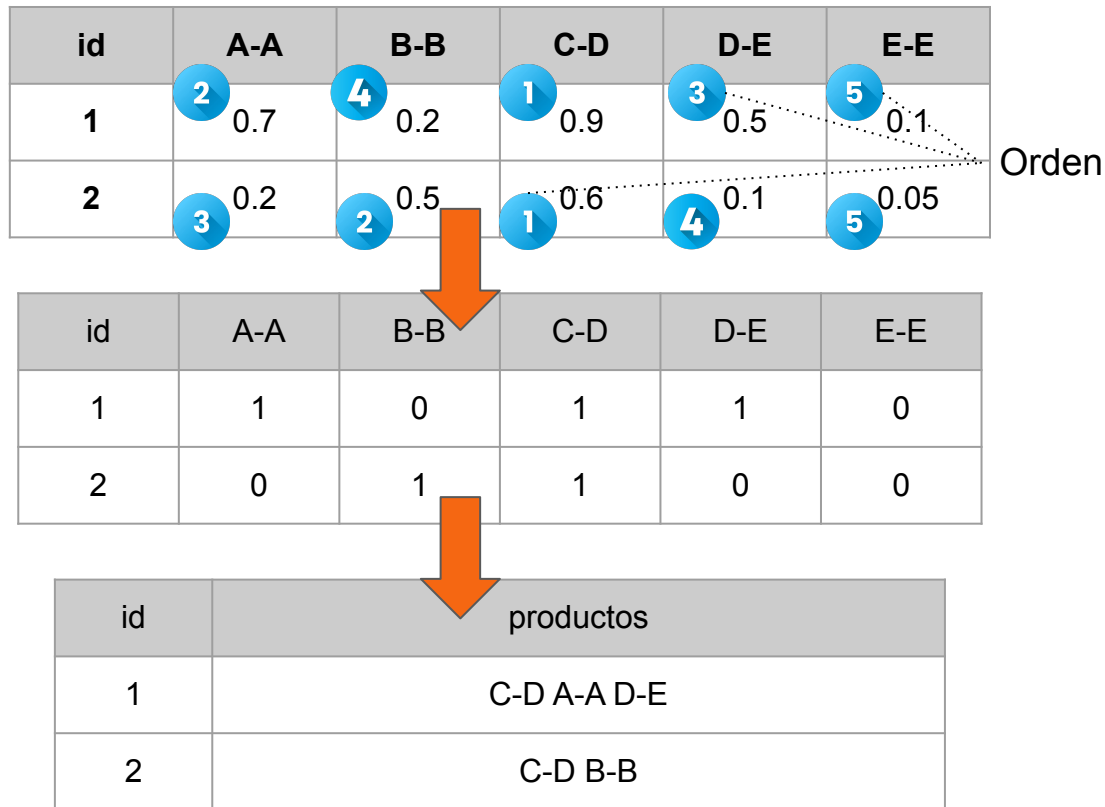
2.2. Entendimiento del problema de analítica

Para generar la variable objetivo, en primer lugar tomamos un mes como base para predecir. **La variable objetivo estará definida por si un cliente compró (1) o no compró (0) el producto dentro de los siguientes 3 meses.**

2022-04	2022-05	2022-06	2022-07	2022-08	2022-09
Base	Predicción	Predicción	Predicción		
	Base	Predicción	Predicción	Predicción	
		Base	Predicción	Predicción	Predicción

2.2. Entendimiento del problema de analítica

Además, para generar el “output” requerido, se utilizará la probabilidad de compra que proviene de cada modelo y se ordenarán (de mayor a menor). Luego, se aproxima cada probabilidad para decidir si finalmente se compra cada producto.



2.3. Entendimiento de los datos

Los datos proveídos poseen diversa información de todos los clientes (anonimizados) y su relación con el banco.



Transacciones

Información de uso de productos, montos y flujo de dinero.



Comunicaciones

Información sobre las interacciones del banco con el consumidor, el producto asociado y la respuesta del cliente.



Campañas

Información de campañas aplicadas a un cliente, su canal de comunicación y el resultado de dicha acción.



Consumidores

Información sociodemográfica y de caracterización de los clientes.

2.4. Análisis de los datos



PRODUCTO TIPO: A-A

Cantidad de clientes que
usaron A-A:

10429

Promedio de transacciones
realizadas por periodo

1,00

Maximo de transacciones
realizadas en un periodo

3,00

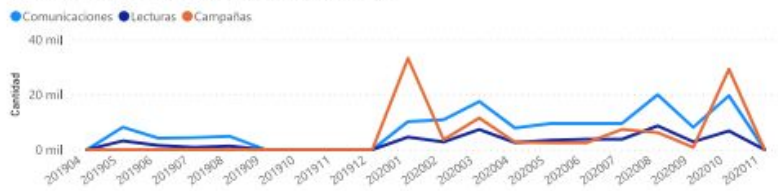
Std, Desv de las
transacciones promedio

0,06

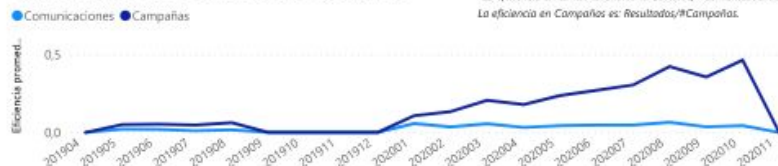
Relación entre transacciones totales de un cliente y la cantidad de transacciones realizadas en producto tipo



Cantidad de comunicaciones, lecturas y campañas



Eficiencia promedio de comunicaciones y campañas



El análisis de los datos se puede visualizar en el dashboard disponible en el siguiente link: [LINK](#)

2.5. Preprocesamiento

1. Se crean múltiples variables en cada dataset que muestran el comportamiento histórico de un cliente con los productos, con los tipos de productos, con las campañas, entre otros.
2. En base a la definición del problema analítico y de negocio es que se genera la variable objetivo para cada mes y para cada producto-tipo de interés entre 2019-01 y 2020-04, al realizar los cruces necesarios entre todas las bases de datos disponibles.
3. Con tal de obtener un mayor poder diferenciador de los modelos, el dataset se separó en los 5 producto-tipo de interés.

2.5. Preprocesamiento

4. Una vez separados, obtenemos un dataset por producto, con todos los clientes en todos los periodos, donde se crea un *semi-target* que toma el **valor 1** es que compró ese producto ese mes y **0** si no lo hizo.
5. Finalmente, para obtener el *target* se escoge si ese cliente compró el producto en los siguientes 3 meses.
6. Los productos con mayor cantidad de “Clase positiva” son:
 1. C-D: 105.187 registros
 2. D-E: 90.142 registros
 3. E-E: 39.343 registros
 4. B-B: 31.249 registros
 5. A-A: 9.159 registros

2.6. Entrenamiento y Testeo

Los datos de entrenamiento y testeo fueron generados por un periodo de corte, en este caso: 2020-02.

Entrenamiento	Testeo	Predicción
2019-03 → 2020-01	2020-02 → 2020-04	2020-07

Notar que existe una pérdida de datos de los meses 2020-05 y 2020-06 debido a que el modelo predice los siguientes 3 meses pero que en este caso, dichos meses no tienen la variable objetivo, lo que hace imposible testear.

2.6. Entrenamiento y Testeo

Dado el gran desbalance de las clases, se utiliza el mecanismo de UnderSampling para realizar el entrenamiento para cada producto-tipo. De esa forma se evita en primera instancia el sesgo hacia la Clase 0 (No compra).

Para el set de testeo, se decidió ocupar el dataset de testeo completo con tal de conocer la robustez del modelo a la hora de clasificar en data imbalanceada.

2.7. Modelos utilizados

Para el problema ya descrito es que se ocupan los siguientes modelos de clasificación (aprendizaje supervisado):

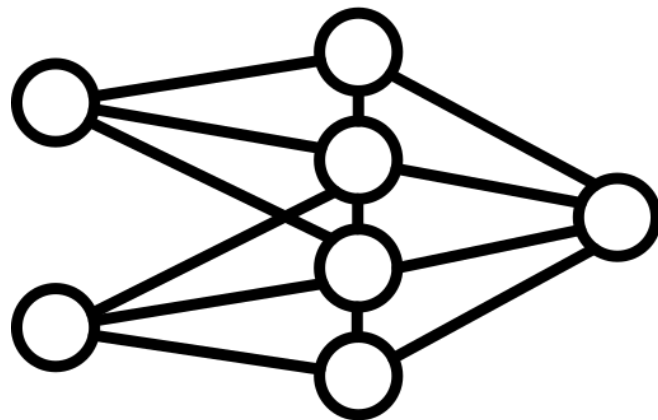
- Naive Bayes
- Regresión Logística
- Random Forest
- XGBoost
- Red Neuronal

Tal como se mencionó anteriormente, se crean 5 modelos de cada tipo (uno para cada producto-tipo).

2.8. Red Neuronal

Se prueban los resultados de distintas arquitecturas de redes:

- Modelos de 3 capas:
 - Con 4, 3, 3 neuronas
 - Con 10, 8, 8 neuronas
- Modelos de 2 capas:
 - Con 5 y 5 neuronas

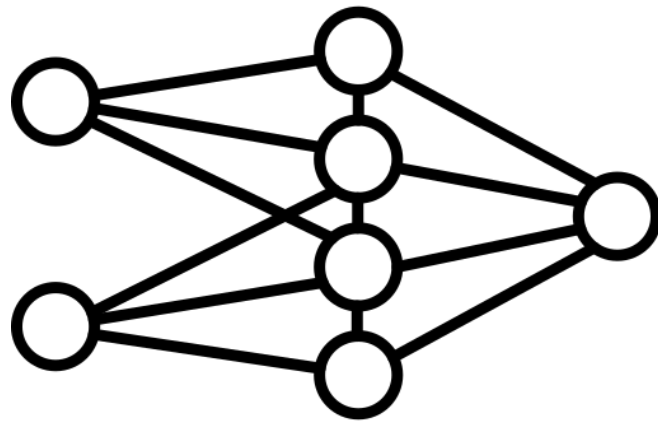


Dentro de ello se prueban distintos learning rate para el modelo, usando ADAM como optimizador.

2.8. Red Neuronal

La red que entregó un mejor desempeño fue:

- Modelo de 3 capas:
 - Primera capa: 10 neuronas
 - Segunda capa: 8 neuronas
 - Tercera capa: 8 neuronas



Con un learning rate de 0.1 y utilizando el optimizador Adams.

3. Resultados y discusión

3.1. Resultados: *Precision*

Precision de cada modelo por cada producto.

Producto	Naive Bayes	Logit	Random Forest	XGBoost	ANN
A-A	0.07	0.05	0.06	0.04	0.04
B-B	0.11	0.10	0.10	0.09	0.10
C-D	0.69	0.86	0.90	0.83	0.85
D-E	0.58	0.56	0.56	0.56	0.56
E-E	0.31	0.28	0.31	0.30	0.30

3.2. Resultados: *Recall*

Recall de cada modelo por cada producto.

Producto	Naive Bayes	Logit	Random Forest	XGBoost	ANN
A-A	0.36	0.50	0.40	0.51	0.61
B-B	0.50	0.67	0.81	0.63	0.66
C-D	0.83	0.82	0.81	0.84	0.84
D-E	0.59	0.67	0.68	0.65	0.65
E-E	0.50	0.52	0.50	0.50	0.50

3.3. Discusión de las métricas

Sobre la métrica Precision, los productos C-D y D-E son los únicos tipos en donde se poseen buena precisión, ya que reduce la cantidad de “Falsos Positivos”. El modelo Naive Bayes es el modelo con peor performance en ese sentido, pero aunque los otros 4 modelos son mejores, no marcan una gran diferencia y son todos similares entre sí.

Para el Recall, el modelo **Random Forest** domina en esta métrica ya que logra detectar la mayoría de las “Clase positiva” presentes en el testeo, especialmente aquellos productos más frecuentes: B-B y C-D. La **red neuronal** posee un gran recall en todos los productos, pero aún así se ve superada por el Random Forest.

3.4. Discusión de los modelos

Random Forest es un modelo ensamblado que es muy bueno para detectar patrones dentro de los datos lo que le da un poder diferenciador muy potente, a diferencia de los modelos lineales. Esta versatilidad le otorga una gran ventaja, resultando así el modelo con mayor robustez de predicción.

Muy de cerca le sigue la **Red Neuronal**, cuya construcción le permite adaptarse lo suficiente a cualquier problema y generar un modelo consistente. Sin embargo, la dificultad de ejecución, construcción y la falta de explicabilidad de sus resultados, lo hace un modelo muy costoso tanto en recursos computacionales como en tiempo.

Los otros modelos (Naive Bayes, XGBoost y Logit) tuvieron un gran performance, especialmente en recall, pero son claramente opacados por el rendimiento del Random Forest.

3.5. Discusión de los resultados

Variables más importantes:

- El porcentaje de uso que le da entre sus productos.
- El número de transacciones que realiza con ese producto ese periodo en el periodo actual y el anterior.
- El total de transacciones del mes.

Las transacciones están relacionadas al uso de sus productos por parte del cliente, y al uso histórico que le ha dado a los productos de interés. Esto coincide con lo esperado, pensando en que los clientes que han usado los productos los volverán a usar.

4. Conclusiones

Conclusiones

El presente proyecto es un gran acercamiento a lo que son las labores diarias de un Data Scientist y es de gran utilidad conocer de primera mano la dificultad e importancia de su trabajo en la organización.

Dentro de las principales habilidades de un científico de datos no solo se encuentran las habilidades técnicas de programación, sino que también matemáticas, de negocio y de comunicación. En este proyecto, aunque se contaba con habilidades técnicas ya trabajadas, el conocimiento sobre el negocio y el cómo traducir sus necesidades a un problema analítico fue fundamental para el modelamiento y el inicio del plan de trabajo.

Conclusiones

Además, el conocimiento sobre el trasfondo de los modelos de Machine Learning fue de alta importancia para lograr obtener las mejores métricas de desempeño posibles y obtener un trabajo de mayor calidad. Dentro de estos modelos el más difícil de implementar fueron las Redes Neuronales, ya que su capacidad de adaptarse es tan grande que requiere un enorme trabajo y dominio en el tema para poder alcanzar su máximo potencial.

A métodos generales, este modelo resulta altamente motivante para seguir incursionando en la analítica avanzada y seguir aportando a las organizaciones a llevar sus decisiones al siguiente nivel.

Desafío Itaú

Proyecto Semestral - Predicción de compra de productos

Curso: IN7615 - Aprendizaje automático con Redes Neuronales
Integrantes: Felipe Jorquera Díaz
Constanza Peña Soto