Analytic Solutions Portfolio applying classic methods

Bolaños. Julian (A00365918), Mosquera. Giovanni (A00365672), Rodríguez. Juan (A00365843), Zorrilla. Juan (A00365972), Posso. Breyner, Facultad de Ingeniería

I. Abstract

Is a well known fact that all websites collect a bunch of information from its users, and even more, e-commerce sites. This data is commonly used for marketing and statistic purposes. Colombian e-commerce are just beginners in this practices and in this article we are going to analyze a colombian case of study.

This article describes the process of development, building, training and deployment of a series of analytics solutions proposed to an e-commerce based on a marketing study. Along all this case of study the analytics team applied the CRISP-DM (Cross Industry Standard Process for Data Mining) methodology. In a first stage was defined a list of business objectives according to e-commerce interests, following with the definition of the analytics problems to solve, metrics and a brief analysis of previous studies.

In the next steps was made an EDA (Exploratory Data Analysis) using the marketing study brought by the commerce, based on that analysis were defined the modeling techniques to be applied and the evaluation protocols. Finally the models were trained in order to will full-fill the analytics and business objectives. After the evaluation of all the models we go on the final stage: Deployment. All the models were deployed on a web page were those could be consumed as services by any e-commerce employee.

Keywords: E-commerce, sales, expenses, regression, Linear Regression, K-Means, clustering, Random Forest, Flask, classification, metrics, confusion matrix, Logistic Regression, accuracy.

II. Introducción

LÍMPIK es un comercio en línea que se centra principalmente en la venta de productos de mercado, la canasta familiar, alcohol y bisutería. Hace un par de meses, LÍMPIK decidió empezar a vender productos por catálogo y en una sede que abrieron en la ciudad de Cali.

El comercio tiene interés en aplicar inteligencia artificial y analítica en su negocio. Teniendo como principal objetivo, ampliar el entendimiento de sus clientes actuales y nuevos. Para ello han realizado estudios de mercado, pensado en hacer un perfilamiento de clientes, al igual que identificar cuáles de sus medios de venta es más viable, entre otras ideas. Para ello, se ha contactado con nosotros y en este proyecto se buscará generar un portafolio de soluciones de analítica que estén en línea con los objetivos del comercio y den una solución satisfactoria a los mismos.

III. Marco Teórico

Machine Learning: El aprendizaje automático o aprendizaje automatizado o aprendizaje de máquinas (del inglés, machine learning) es el subcampo de las ciencias de la computación y una rama de la inteligencia artificial, cuyo objetivo es desarrollar técnicas que permitan que las computadoras aprendan.

Regresión Logística: En estadística, la regresión logística es un tipo de análisis de regresión utilizado para predecir el resultado de una variable categórica en función de las variables independientes o predictoras. Es útil para modelar la probabilidad de un evento ocurriendo en función de otros factores.

K Means: K-means es un método de agrupamiento, que tiene como objetivo la partición de un conjunto de n observaciones en k grupos en el que cada observación pertenece al grupo cuyo valor medio es más cercano. Es un método utilizado en minería de datos.

Random Forest: Random forest también conocidos en castellano como '"Bosques Aleatorios"' es una combinación de árboles predictores tal que cada árbol depende de los valores de un vector aleatorio probado independientemente y con la misma distribución para cada uno de estos.

E Commerce: El comercio electrónico también conocido como ecommerce, comercio por Internet o comercio en línea consiste en la compra y venta de productos o de servicios a través de internet, tales como redes sociales y otras páginas web.

IV. Antecedentes

Tras realizar un estudio riguroso sobre los antecedentes de nuestro problema encontramos diversos casos de estudio previos al nuestro, entre estos se encuentran proyectos como "Data Prep, Visual EDa and Statistical Hypothesis", o como "Marketing Analytics, Classification, and EDA". En estos se realizaron trabajos de clasificación, regresión, segmentación y planteamiento de hipótesis emplenado tests como: Point-Biserial Correlation test, Chi cuadrado y Spearman Rank Correlation test y modelos clásicos como: Random Forest, Decision Tree y KNN (K-Nearest Neighbors), los cuales podríamos tomar de base para nuestro contexto. Además se va a brindar el valor agregado de llegar al despliegue, etapa a la cual no llegan los proyectos basados sobre este mismo conjunto de datos.

V. Metodología

A. Preguntas de Interés

Una vez se realizó el contacto con el comercio se manifestaron las siguiente preguntas y elementos de interés para el desarrollo del proyecto:

1) ¿Es posible identificar cuál de mis medios de venta (local físico, catálogo y tienda en línea) está siendo más frecuentado por los clientes, para así invertir más en este?
2) Como negocio quisiéramos conocer mejor a nuestros clientes actuales y futuros: ¿Qué compran?, ¿Existen características en común entre ellos?, ¿Podemos definir perfiles de cliente?
3) ¿Cómo podemos hacer recomendaciones más adecuadas a cada tipo de cliente que tenemos?
4) ¿Son los niños o jóvenes en un hogar, un factor determinante en cuanto a la cantidad de dinero gastado en nuestro comercio?

B. Tipo de problema de analítica

En base a nuestro contexto y las preguntas de interés, podemos determinar los objetivos de analitica que se definirán para el proyecto que se desea llevar a cabo, estos objetivos son los siguientes:

- Elaborar modelos de regresión para predecir cuántas compras realizará un cliente por medio del catálogo, página web y tienda física.
- Construir un modelo de clasificación de clientes acorde a los artículos que adquiere para generar perfiles de compradores
- Elaborar un modelo de clasificación para identificar los usuarios que podrían aceptar una campaña de marketing/oferta a futuro, en base al número de campañas aceptadas y diferentes características del comprador.
- Elaborar un modelo de clustering para clasificar a un cliente del comercio, en base al número de niños y jóvenes que viven en su hogar

Como se menciona en los objetivos, se desean implementar modelos de regresión, clasificación y clustering de clientes, todos estos pertenecientes al tipo de analítica **predictiva** debido a que como su nombre lo indica tratan de predecir acontecimientos o comportamientos a partir de relaciones entre diferentes factores que permiten valorar riesgos o probabilidades asociadas sobre la base de un conjunto de condiciones.

C. Datos Recolectados

En nuestro dataset encontramos datos de clientes de una campaña de marketing realizada por LIMPIK para recolectar información pertinente acerca de sus clientes, dentro de este dataset podemos encontrar la siguiente información:

- Año de nacimiento
- Nivel académico
- Estado civil
- Ingresos anuales
- Número de infantes y jóvenes que hay en el hogar
- Si el cliente aceptó la oferta en alguna de las anteriores campañas
- Si el cliente realizó quejas en los últimos dos años
- Fecha de inscripción del cliente a la empresa
- Cantidad gastada de productos (pesqueros, cárnicos, frutas, dulces, vinos, productos oro) en los últimos dos meses
- Número de compras hecho con descuento
- Número de compras hechas usando el catálogo
- Número de compras realizadas directamente en las tiendas
- Número de compras realizadas a través del sitio web de la empresa
- Número de visitas al sitio web de la empresa en el último mes
- Número de días desde la última compra

D. Análisis exploratorio de los datos (EDA)

Se realizó un primer vistazo sobre con cuánta información contamos, cuántas variables de interés, su tipo de dato entre Cualitativo y Cuantitativo, estadísticas básicas sobre el set como lo son la media, mediana, desviación estándar, etc.

Luego se realizaró una visualización básica de los datos para encontrar irregularidades en los estos, las gráficas más tentativas serán Histogramas, Gráficos de Violín, Gráficos de caja, Gráficos de barras, Diagrama de torta, Diagrama de correlación y Diagramas de puntos.

Una vez identificados los datos que nos generan ruido dentro de la información se realizó a una limpieza de los mismos. Se buscaron outliers y duplicados, una vez encontrados se vio que estos presentaban un patrón en partícular y eran de vital importancia para realizar predicciones fieles a la realidad.

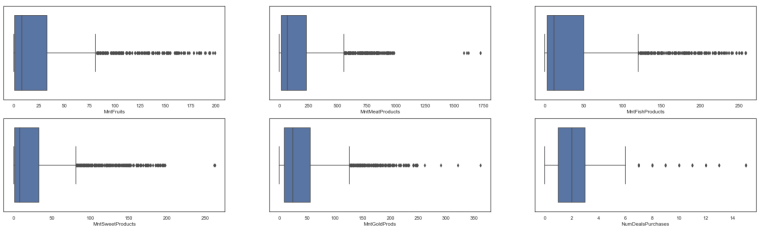

Fig. 1. Diagramas de caja que presentan el patrón..

Se siguió la identificación y limpieza a los datos con fallos estructurales o con poco sentido. Se corrigieron los casos en los que se presentaron fallos sobre el tipo de dato.

Ahora, finalmente se relalizó una búsqueda de datos faltantes o nulos, haciendo uso de las distribuciones de cada variable. Se reemplazaron estos registros teniendo en cuenta ciertas restricciones del set de datos.

En último lugar se validó la información para saber si esta era congruente y los pasos anteriores fueron aplicados satisfactoriamente. Todos estos pasos fueron reportados para poder llevar un control de la información. Finalmente, se realizaron de nuevo los gráficos del primer paso para ver visualmente los cambios realizados en las representaciones y se extrajó la información estadística básica del set de datos.

E. Modelamiento

Para avanzar con el desarrollo de un portafolio de soluciones de analítica para LÍMPIK, se pasó a la siguiente etapa, el modelamiento de nuestra propuesta para solucionar los problemas y las preguntas de interés del comercio.

1) Técnicas de Modelamiento: Para dar cumplimiento a cada uno de los objetivos, se plantea la elaboración de diferentes modelos de predicción que nos faciliten llegar a una respuesta correcta a los objetivos de negocio. Por cada objetivo de analítica se presentan el o los posibles modelos a emplear:

1) Random Forest, Regresión Linea, Árboles de decisión, XG-Boost
2) K-Means

3) Regresión Logística, Árboles de decisión, SVM (Support Vector Machines)
4) K-Means

2) Protocolos de Evaluación para Modelos: Para los modelos de regresión se plantea el uso de las siguiente métricas: MSE, RMSE, Coeficiente R^2 y la gráfica de residuos.

Para el modelo de K-Means, se usarán estimadores como el método de codo, silueta y el índice Calinski-Harabasz, para encontrar la mejor definición de clusters.

En el último modelo, para la clasificación se emplearán la exactitud, precisión, TPR, FPR, puntaje f1 y la matriz de confusión.

3) Construcción de Modelos: Para este apartado se empleó la librería de Python Sklearn de Scikit Learn, la cual permitía manejar todos los modelos anteriormente mencionados, dando la oportunidad de modificar los hiperparámetros de los mismos. En esta etapa se evaluaron los modelos anteriormente mencionados, sin embargo, se tomaron únicamente como resultados finales los mejores modelos.

A continuación se listan los mejores modelo con sus respectivos hiperparámetros:

TABLE I
MODELOS FINALES PARA CUMPLIR OBJETIVOS DE ANALÍTICA

Objetivo	Modelo	Hiperparámetros	Predicción
1	Random Forest	default	Número de compras de un cliente en tienda física
1	Random Forest	default	Número de compras de un cliente en catálogo
1	Random Forest	default	Número de compras de un cliente en página web
2	K-Means	$k = 3$, k-means++	Perfil de comprador de un cliente en base a los productos que adquiere
3	Regresión Logística	default	Probabilidad de aceptación para campañas de marketing
4	K-Means	$k = 3$, k-means++	Perfil de comprador de un cliente en base a los menores en casa

F. Evaluación

1) Resultados y Métricas: A continuación se listan los resultados obtenidos y las métricas relacionados a estos para cada uno de los modelos presentados anteriormente:

Modelo de Regresión Logística (Objetivo de analítica 2):

- Se obtuvo un perfíl de comprador el cual es casi seguro (92% de probabilidad) que no aceptaría una campaña de marketing.
- Se obtuvo un perfíl de comprador que posiblemente acepte campañas (65% de probabilidad).

A continuación se puede evidenciar la matriz de confusión para el conjunto de datos de prueba. En general el modelo de regresión logística presentó una exactitud del 0.875.

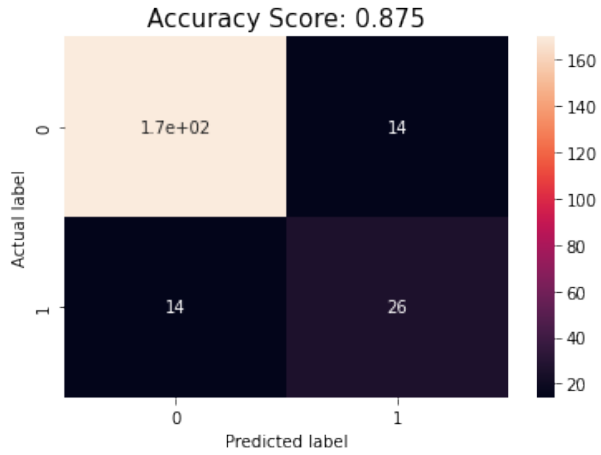

Fig. 2. Mátriz de confusión de regresión logística sobre datos de prueba. *0: No acepta campaña. 1: Acepta campaña.*

Modelos Random Forest (Objetivo de analítica 1):

- Se lograron generar 3 algoritmos de regresión para cada uno de los medios de venta del comercio.
- El modelo que mejor describe las compras de los clientes en los diferentes medio de venta es el modelo de predicción para la compras por la página web ($R^2 = 0.768$), mientras que el que peor los describe es el de compras en la tienda física ($R^2 = 0.689$)
- Por el bajo ajuste de los modelos al intentar predecir la cantidad de compras que un cliente realizará por cualquier medio, no se realizará una predicción muy efectiva.

En las siguientes imágenes se presenta la gráfica de residuos para los tres modelos, junto con su R^2, MSE y $RMSE$.

```
Validation
R^2: 0.7481
MSE: 1.8577
RMSE: 1.3630
```

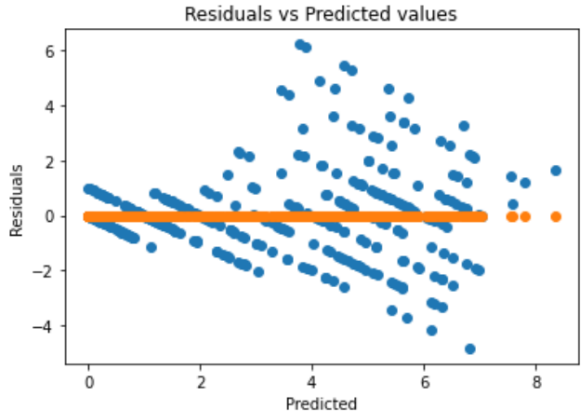

Fig. 3. Gráfica de residuos y métricas de Random Forest para predecir compras por catálogo.

```
Validation
R^2: 0.7680
MSE: 1.5381
RMSE: 1.2402
```

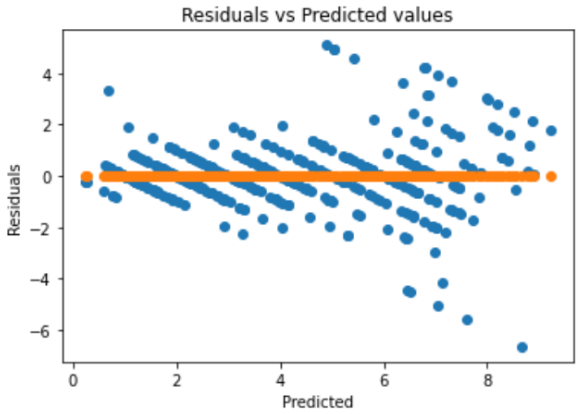

Fig. 4. Gráfica de residuos y métricas de Random Forest para predecir compras por web.

```
Validation
R^2: 0.6897
MSE: 3.3261
RMSE: 1.8238
```

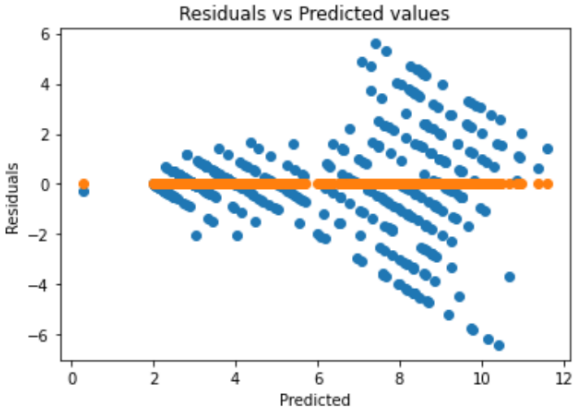

Fig. 5. Gráfica de residuos y métricas de Random Forest para predecir compras por tienda física.

Modelos K-Means (Objetivos de analítica 2 y 4):

- Se obtuvo que la cantidad de jóvenes y niños por hogar, no es un factor determinante en el consumo, sin embargo, sí se pueden describir perfiles de usuario en base al número de niños y jóvenes dentro de un hogar. Perfiles definidos: 3 (Objetivo 4)

- Definimos 3 perfiles de compradores diferentes, cada uno con sus respectivas características. En base a factores como los productos que consumían, edad, tiempo que llevan en el comercio, ingresos, entre otros. (Objetivo 2)

A continuación se listan los perfiles de compradores definidos por medio de los dos modelos usados anteriormente.

Perfil 0

El pefil 0 es un cliente promedio, sus ingresos están entre 40.000 y 60.000 dólares. Adquiere en diferentes cantidades los productos del comercio, pero sin presentar un enfoque en particular. Es un perfil que prefiere realizar sus compras por la página web del comercio, en lugar del catálogo y de la tienda física. Predomina con 1 o 2 jovenes en su hogar.

Perfil 1

Este comprador presenta ingresos bajos, entre 0 y 40.000 dólares. No es un comprador que gaste mucho su dinero en los productos del comercio, es quien menos productos adquiere dentro de la población de clientes. Además de que a pesar de que frecuenta mucho la página web de LÍMPIK, prefiere realizar sus compras en el punto físico. Predomina en que no hay ningun joven en el hogar.

Perfil 2

Podemos identificar a estos compradores como una persona que cuenta con altos ingresos, más de 60.000 mil dólares. También este comprador se caracteriza por sus constantes gastos en productos como alcohol (vino) y diferentes productos del comercio como carnes, pescado y accesorios de Oro. Igualmente, este perfil prefiere realizar sus comprar por internet. La cantidad de jovenes en el hogar está entre 1 y 2.

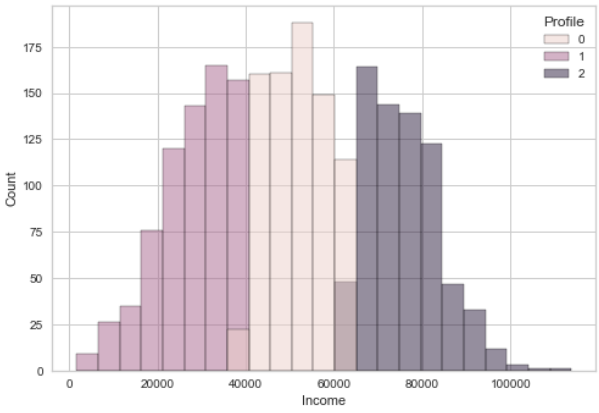

Fig. 6. Pefiles de comprador separados por nivel de ingresos.

G. Despliegue

Pasando a la última etapa del proyecto, se elaboró un plan de despliegué simple. Con este plan se quería facilitar el acceso de los empleados del comercio a los modelos generados en las etapas anteriores. Para ello se planteó generar una página web en donde quién deseara hacer uso de los modelos y sus predicciones, debería ingresar el listado de datos necesarios para la predicción de cada modelo. Para el desarrollo de la página web se hizo uso Flask y Dash para elaborar un despliegue sencillo y efectivo. Finalmente, para el despliegue se elaboraron los siguientes pasos:

1) Usando joblib, se extrajo cada uno de los modelos generados en los notebooks.
2) Usando un script de python con Dash se elaboraron formularios para ingresar la información de cada modelo
3) Se ejecutó el script de python usando Flask para correr la página web

VI. Resultados y Análisis

Como ya se mencionó en la etapa de *Resultados y métricas* de la metodología, se obtuvieron 6 modelos, de los cuales, los modelos de regresión empleados para el objetivo de analítica 1,

si bien realizarán predicciones y harán generalizaciones correctas para la mayoría de los casos, estos no se acoplan de la mejor manera a los conjuntos de datos.

Por parte de los modelos de clustering K-Means, encontramos que realizan una segmentación adecuada de los clientes del comercio, permitiéndonos definir perfiles de clientes adecuados para la clasificación de los mismos.

Finalmente, con lo que respecta a los modelos construídos, el modelo de regresión logística pra el objetivo de analítica 3, presentó un ajuste muy bueno para realizar predicciones ante clientes que rechazarían una campaña, sin embargo, no se presenta un buen ajuste y generalización para predecir clientes que sí aceptarían una campaña.

Si analizamos detenidamente estos modelos, podemos entender que si bien todos están en capacidades de realizar predicciones adecuadas y acertadas para la mayoría de los casos, no resultan siendo muy óptimos y confiables ante una posible decisión de la empresa, ya sea por ejemplo, disminuir el número de tiendas físicas por sus bajas ventas o empezar una nueva campaña de mercado específica. Ambas decisiones involucran una inversión muy alta de capital por parte del comercio, decisiones que, contando con modelos sin mucha precisión como los propuestos, no resultan fáciles de tomar.

Sin embargo, con el trabajo hecho, se ha logrado de manera muy efectiva empezar a describir de una manera más amplia y clara a los clientes de LÍMPIK. Esto permite definir unas bases sobre el comportamiento de los mismos y algunas decisiones que estos tomarán a futuro. Todo esto nos ayuda a entender que para un trabajo a futuro o en una siguiente iteración, se tendrá más información y mayor conocimiento de los clientes a manejar y por ende, se podrán definir un mayor número de objetivos de negocio con los que emplear técnicas de analítica más avanzadas que porporcionen resultados de mayor calidad.

Como último aspecto, se puede llegar a que se han cumplido cumplido los objetivos de analítica, sin embargo, no se han logrado generar soluciones lo suficientemente robustas para dar cumplimiento a todas las preguntas de intéres o preguntas del negocio.

VII. Conclusiones

A manera de cierre de este informe, se logró seguir un flujo de analítica basado en la metodología CRISP-DM para el desarrollo de una serie de soluciones de analítica para dar solución a un conjunto de objetivos de negocio definidos con un cliente. A lo largo de todo este tiempo de desarrollo se realizaron constantes iteraciones y mejoras sobre las soluciones propuestas, esto con la finalidad de experimentar y realizar diferentes ajustes empleando diferentes modelos que podían entregar una mejor solución a los objetivos definidos. Si bien los modelos obtenidos no cumplen de manera óptima y confiable los objetivos de negocio, se ha logrado plasmar un procedimiento claro y organizado del desarrollo de los mismos.

Finalmente, como ya hemos mencionado en los resultados, se han dejado unos cimientos y bases adecuadas para futuras iteraciones, esto debido a que se logró un entendimiento mayor de los clientes y su información. Dicho entendimiento abre las puertas a oportunidades de mejora, como por ejemplo la aplicación de ingeniería de atributos en la etapa de análisis para extraer una mayor cantidad de información reelevante de los clientes. Otra clara oportunidad de mejora sería incluir la construcción de modelos que apliquen técnicas de machine learning más avanzadas (Ej. Deep Learning).

References

[1] Comunidad de Wikipedia, Machine Learning, Wikipedia, 19 ago 2022 https://es.wikipedia.org/wiki/Aprendizaje_automtico

[2] Comunidad de Wikipedia, Regresión Logística, Wikipedia, 13 jun 2022 https://es.wikipedia.org/wiki/Regresin_logstica

[3] Comunidad de Wikipedia, K Means, Wikipedia, 14 jun 2022 https://es.wikipedia.org/wiki/K-medias

[4] Comunidad de Wikipedia, Random Forest, Wikipedia, 23 dic 2021 https://es.wikipedia.org/wiki/Random_forest

[5] Comunidad de Wikipedia, E Commerce, Wikipedia, 11 nov 2022 https://es.wikipedia.org/wiki/Comercio_electrnico