Entrega 1 - Proyecto
Entregan: Julian Bolaños, Giovanni Mosquera, Juan Rodríguez y Juan Zorrilla

Contexto del problema

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.

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?

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 un modelo 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
- En base a los perfiles de compradores, construir un modelo para generar sugerencias de productos que puedan ser afines al perfil de un comprador

- Elaborar un modelo de regresión para predecir el dinero gastado por 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 lineal, modelos de clasificación y modelos de sugerencias de productos, 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.

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 y segmentación, los cuales podríamos tomar de base para nuestro contexto y además brindar el valor agregado de llegar al despliegue, etapa a la cual no llegan los proyectos basados sobre este mismo dataset.

Métricas a utilizar para medir el progreso

Para la medición del progreso y satisfacer nuestros objetivos indicados anteriormente.

Tenemos las siguientes métricas para el criterio de éxito:

- ☐ Para los modelos de regresión se sugiere emplear una precisión del 95%.
- ☐ Para el modelo de clasificación se propone un porcentaje de precisión y asignación del 90%.
- ☐ Para el motor de sugerencias se plantea un porcentaje de precisión del 90%.

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

Este dataset fue extraído de kaggle: link

Análisis exploratorio de los datos

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

Luego se realizará 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 procederá a una limpieza de los mismos. Se buscarán outliers y duplicados, una vez encontrados se decidirá si son eliminados o reemplazados por algún estadístico. Se sigue la identificación y limpieza a los datos con fallos estructurales o con poco sentido.. Si existen fallos en cuanto al tipo de dato, también serán corregidos en esta etapa.

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

En último lugar se validará la información para saber si es congruente y los pasos anteriores fueron aplicados satisfactoriamente. Los pasos serán reportados para poder llevar un control de la información. Se realizarán de nuevo los gráficos del primer paso para ver visualmente los cambios realizados en las representaciones y se extraerá la información estadística básica del set de datos.

Siguientes pasos

Una vez terminada esta fase del proyecto se pasará con la etapa de Modelado en la cual se evaluarán las posibles elecciones de modelos para aplicar en esta problemática, se generarán los diseños de las pruebas y se procederá a construir los modelos que cumplan los objetivos, para luego ser ejecutados y entrenado.

Finalmente, evaluaremos los datos obtenidos con los modelos, se determinará si estos pueden responder las preguntas de interés y evaluar si es necesario realizar un proceso iterativo con la etapa anterior hasta conseguir un resultado adecuado en esta.