Predicción de ganancias por venta de ropa de acuerdo con factores de sucursal

Métodos y aplicaciones de analítica I

Fabián Alexis Pallares Jaimes

Contexto de negocio:

Con el objetivo de conocer la influencia de la asignación de recursos a sus sucursales en la venta de ropa de mujer, el negocio de venta de ropa y joyería por catálogo XZ ha solicitado realizar un análisis predictivo que permita, basado en los datos recogidos de 100 de sus sucursales en 2019, estimar el valor de ventas de ropa de mujer para cada sucursal.

El negocio XZ desea modelar la demanda de las zonas, de tal manera que pueda llevar a cabo estrategias de negocio basado en las variables de mayor influencia en las ganancias por sucursal y optimizar sus recursos para generar más ganancia.

Objetivos de negocio y criterios de éxito:

El negocio XZ, basado en los datos provistos, quiere obtener un modelo predictivo que le permita optimizar sus recursos, esto es, determinar las variables que más influyen en la obtención de ganancias para las sucursales en las zonas determinadas.

El objetivo analítico para el problema planteado por el negocio XZ está relacionado con los indicadores de desempeño del negocio (KPI):

- Crecimiento de los ingresos de la compañía
- Índice de calidad del producto (Al buscar invertir en el desarrollo de estos)
- Tasa de retención de clientes
- Rentabilidad del cliente

Así pues, el negocio XZ buscar obtener un mayor crecimiento en sus ingresos mediante variables que se relacionan específicamente con el tratamiento a los clientes y la inversión de la compañía en la estimulación de compra de estos.

Obtener este modelo predictivo puede ayudar al negocio, por ejemplo, a saber en qué tipo de zonas o a qué mercados debería enfocar nuevas sucursales, así como los recursos que debe invertir en dichas sucursales de acuerdo a sus otras variables, para optimizar sus ganancias.

Contexto analítico: Para el desarrollo del modelo, el negocio entrega dos archivos de datos:

- Archivo de datos de 100 zonas y sus respectivas variables con el que se debe entrenar el modelo
- Archivo de datos de 20 zonas sin la variable objetivo para la evaluación del modelo

El modelo predictivo que se pretende realizar será una regresión lineal y debe hacerse después de los pasos de exploración y preparación de datos, obteniendo variables únicamente numéricas que permitan la ejecución del modelo.

Exploración y verificación de calidad de los datos

El conjunto de datos presentado por el negocio contiene información sobre 100 zonas ordenadas de manera contigua durante un año. Este conjunto de datos incluye valores correspondientes a la inversión de la compañía en variables de las sucursales de cada zona, así como el tipo de mercado que ataca cada zona, las promociones ofrecidas por estas y la nómina asignada para cada sucursal. Esto con el objetivo de determinar los factores más importantes a la hora de predecir las ganancias de cada zona.

Variable	Tipo de dato	Descripción
ropamujer	numérico	Ventas de prendas para mujer en millones de pesos durante el año
correo	numérico	Número de catálogos enviados durante el año.
páginas	numérico	Número de páginas del catálogo.
teléfono	numérico	Número de líneas para llamada abiertas para pedidos promedio.
impresa	numérico	Cantidad gastada en publicidad impresa.
servicio	numérico	Número de representantes del servicio al cliente.
idmercado	numérico (categórico)	Tipo de mercado. Se listan números de clases o tipos de mercado que se desea atacar.
tamañomer	factor (categórico)	Tamaño potencial del mercado, proyectado de acuerdo a cifras del área de marketing.
idloc	numérico	ID de la tienda.
edadloc	numérico	Años transcurridos desde la llegada a la zona.
promo	numérico (categórico)	Tipo de promoción que se llevó a cabo durante el año.
nomina	numérico	Valor total de la nómina durante el año.

Tabla 1. Descripción de datos

Sobre las variables numéricas del conjunto de datos se realizó una serie de estadísticas descriptivas para conocer la distribución de los datos y tener una noción básica de los datos que se presentan:

Variable	Min	Primer cuartil	Mediana	Media	Tercer cuartil	Max
ropamujer	16579	33010	40843	40818	47044	80246
correo	1147	9052	10200	10142	11250	15263
páginas	51.00	72.00	51.50	80.23	88.25	110.00
nómina	14337	18391	20918	21253	23876	30808
teléfono	17.00	28.75	34.50	35.08	41.00	59.00
impresa	18061	26790	28620	28511	30608	38739
servicio	15.00	28.00	36.00	36.19	44.00	68.00

Tabla 2. Estadística descriptiva del conjunto de datos

Variable	Asimetría	Curtosis
ropamujer	0.58	0.71
correo	-0.88	5.86
páginas	-0.1	-0.2
nómina	0.26	-0.7
teléfono	0.3	-0.29
impresa	-0.01	0.56
servicio	0.26	-0.47

Tabla 3. Asimetría y curtosis en las variables numéricas del conjunto de datos

Tal como observamos en las tablas 2 y 3, los datos parecen no presentar una distribución asimétrica. Asimismo, gracias a la cercanía entre los valores de media y mediana, y al valor provisto por la curtosis, podemos determinar que las variables no cuentan con valores extremos.

Este análisis puede ser comprobado al graficar los histogramas de dichas variables, los cuales se presentan a continuación:

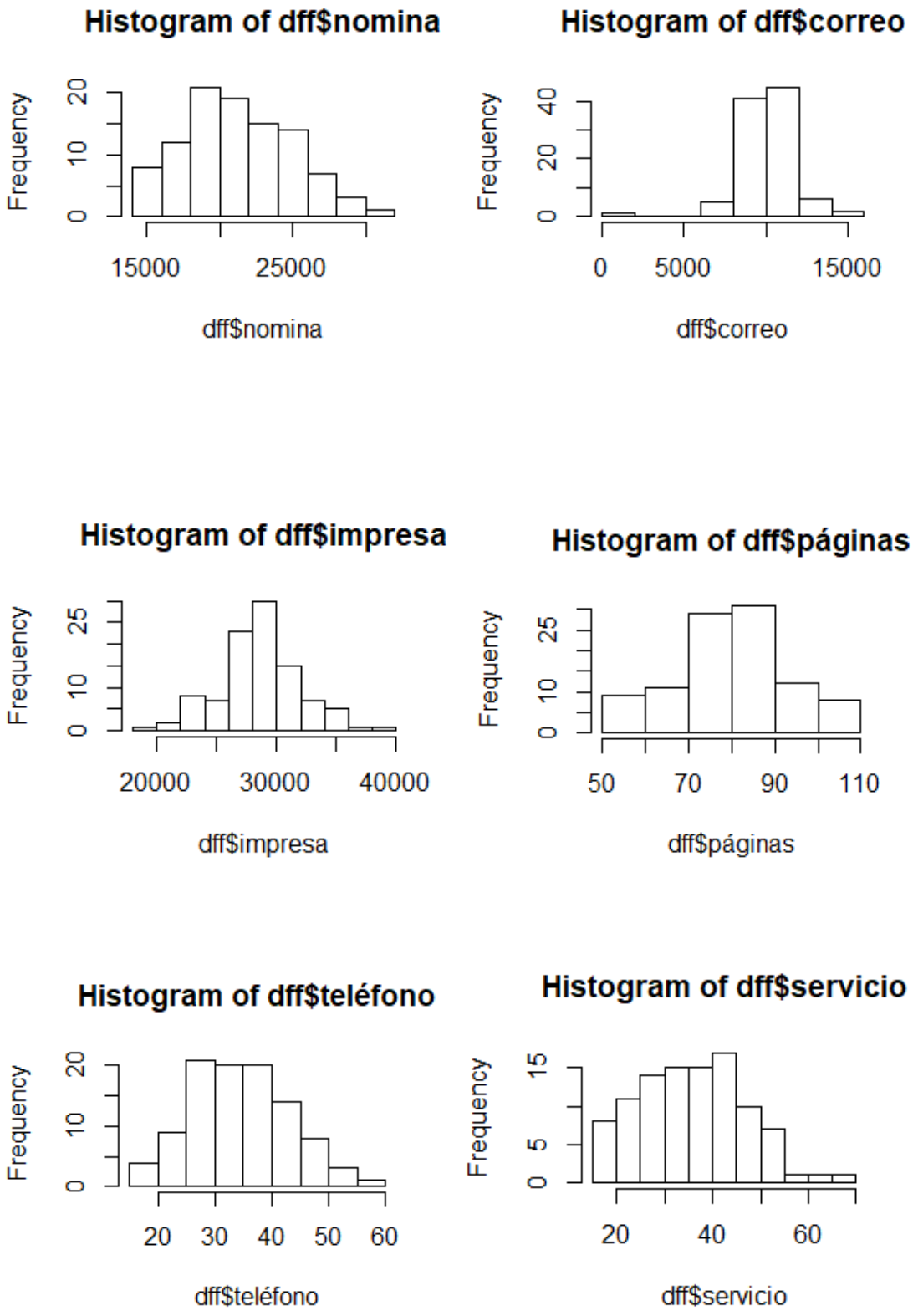

Ilustración 1. Histogramas de las variables numéricas del conjuntos de datos

Por otro lado, es posible determinar que nos encontramos con variables de tipo categórico, esto es, variables que representan categorías entre los datos.

Para este conjunto de datos encontramos las variables de tamaño del mercado, las promociones disponibles aplicadas por las sucursales y el tipo de mercado en el que se encuentra la sucursal.

Podemos describir estas variables observando el conteo de datos de cada categoría:

Niveles de tamañomer	Cantidad de datos
Pequeño	15
Mediano	21
Grande	64

Tabla 4. Variable categórica de tamaño de mercado

Niveles de promo	Cantidad de datos
32	32
34	34
34	34

Tabla 5. Variable categórica de promociones

Niveles de idmercado	Cantidad de datos
1	19
2	14
3	11
4	5
5	19
6	2
7	8
8	10
9	12

Tabla 6. Variable categórica de tipo de mercado

Preparación de la vista minable

Con el objetivo de correr un modelo de regresión lineal, cabe resaltar que nuestro conjunto de datos debe cumplir con unas condiciones básicas para funcionar como vista minable y/o para generar un modelo de regresión confiable:

- Las variables deben ser numéricas
- Las variables deben estar estandarizadas (solo para algunos casos)
- Las variables no deben contar con una correlación muy alta entre sí

Para la primera condición se trabajaron las variables categóricas como variables dummy, lo que quiere decir que se generaron columnas para las posiblescategorías de cada una de las variables. Esto permite que el modelo cuente únicamente con variables numéricas que permitan explicar la variable objetivo (ropamujer).

Al momento de trabajar las correlaciones, se hizo una gráfia de matriz de correlación a modo de *heatmap,* lo cual nos permite observar, tal como se ve en la ilustración 2, que sí hay variables altamente correlacionadas en nuestro conjunto de datos:

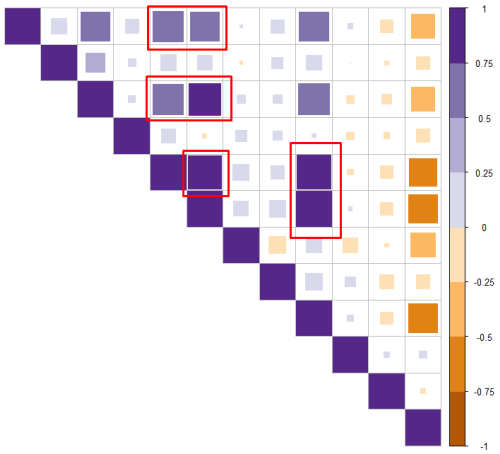

Ilustración 2. Matriz de correlaciones

Efectivamente se encontraron variables que pueden ayudar a explicarse entre sí, cosa que perjudica al modelo de regresión lineal. Es por esto que se buscaron a qué variables correspondía y se procedió a tomar decisiones sobre el uso de estas.

Se tomó la decisión de eliminar 2 variables del set de datos, esto por su alta correlación con otras variables o por su inutilidad en el modelo:

- La variable Idloc representa un id de la sucursal, por lo cual no aporta ningún valor al modelo y debe ser eliminada
- La variable servicio fue eliminada al tener una alta correlación con troas variables como correo y teléfono.

Modelado

Finalmente se realizó el modelo de regresión lineal con el siguiente código de R:

$$modeloaug < -lm(ropamujer \sim ., data = df)$$

$$modelocarstep < -step(modeloaug, direction = "both", trace = 0)$$

Posterior a esto el modelo se aplicó al conjunto de datos de prueba provisto por el negocio, lo cual arrojó las siguientes 20 predicciones para cada una de las sucursales de prueba:

idloc	ropamujer
101	*74604.62*
102	*35444.03*
103	*36957.58*
104	*33944.43*
105	*35863.17*
106	*45435.37*
107	*39822.43*
108	*40989.40*
109	*31373.54*
110	*25870.61*
111	*39916.52*
112	*42316.67*
113	*45918.91*
114	*52952.20*
115	*48536.37*
116	*39026.03*

117	*28989.02*
118	*42586.37*
119	*32379.50*
120	*43441.19*

Tabla 7. Predicciones del modelo sobre el conjunto de datos de prueba

Evaluación de resultados

Para evaluar los resultados del modelo de regresión lineal, se usó el criterio del error cuadrático medio (RMSE) con los datos reales de ganancias en las 20 sucursales indicadas (ganancias no provistas por el negocio).

Esta evaluación sobre el test de prueba se realizó dividiendo en conjunto de datos resultante en dos partes iguales y generando un error para cada conjunto:

Error del primer 50%	9658.81374
Error del segundo 50%	7975.46715

Tabla 8. RMSE calculado sobre el conjunto de datos real y el predicho dividido en dos partes iguales