Tecnológico de Monterrey

Generación de los modelos de análisis de datos (micro-retailers)

Escuela: Instituto Tecnológico de Estudios Superiores de Monterrey

Materia: Desarrollo de proyectos y análisis de datos

Profesor: Alfredo García Suárez

Nivel Académico: Profesional

Ciudad: Puebla

Autores

Omar Eduardo Pelcastre Reyes	A01735985@tec.mx
Saúl Jesús Cuervo Méndez	A01735937@tec.mx
Juan José Lara García	A01736667@tec.mx
Cristian Marino Gutiérrez Jiménez	A01736337@tec.mx
Kevin Vergara Lara	A01735970@tec.mx
Marco Ivan Olalde Gonzalez	A01733378@tec.mx

Etapa 3. Generación de modelos de análisis de datos.

Para conocer cuáles son los casos de correlación de las variables del dataframe es necesario generar modelos de regresión múltiple y regresión logística donde podamos ver si hay una influencia entre las distintas variables independientes con una variable designada dependiente.
A continuación se analizan 5 casos de correlación por los dos métodos tomando las distintas variables de nuestro archivo de datos constando de valores numéricos y dicotómicos.

Caso 1:

El primer caso de correlación que se analizó fue si la variable de "2_current_permanent_employees" se ve afectada por las variables independientes "145_number_direct_competitors" y "4_number_permanent_employees_last_year".
Posteriormente, en un código declaramos las variables dependientes e independientes y se procede a hacer el código para la regresión lineal.
Con esto, obtenemos el coeficiente de determinación que nos indica qué tan fiable es nuestro modelo.

En este caso, el modelo de regresión múltiple tiene una R^2 de 0.35 por lo tanto vemos que el número de competidores directos y el número de empleados del año pasado no afectan al número de empleados actuales del negocio debido a que a nivel de las micro empresas o tienditas de la esquina realmente no hay una competencia por personal ya que generalmente son negocios supervisados por los mismos familiares o por sólo una persona.

Caso 2:

El segundo caso analizado es si la variable "24_burnout" depende o se ve afectada por las variables independientes "78_number_home_deliveries_week" y "317_home_deliveries".
De igual forma se hizo uso de un código para declarar las variables y obtener una regresión lineal múltiple.

El código indicó que el modelo tiene una R^2 de 0.0029 lo cuál nos indica que no hay absolutamente ninguna relación entre las variables dependientes e independiente.
Con esto podemos concluir que el cansancio que pueda experimentar el empleado o la persona que atiende la tienda no tiene relación con el número de entregas a domicilio a la semana y de las entregas a domicilio totales. Esto se debe a que una gran parte de los negocios no contaba con un sistema de repartidores o acceso a aplicaciones para llevarlo a cabo tales como Rappi, Uber Eats, Didi, etc.

Caso 3:

El tercer caso cuenta con las variables dependientes de "272_card_days_receive_money", "24_burnout" y "103_number_own_fridges". La variable independiente a la que se le buscó relación fue a "97_number_of_customers_in_store". Quiere decir, queremos analizar si los

días en que se espera que la tienda reciba dinero fiado tiene que ver con el cansancio y el número de refrigeradores que poseen.
Para esto, volvimos a hacer uso del código de Colab para hacer la regresión lineal y el resultado de esta regresión múltiple fue que el coeficiente de determinación tiene un valor de 0.01.
Esta R^2 nos dice como el caso anterior que no hay relación entre nuestras variables dependientes e independiente. Esto se debe a que desde un análisis lógico, se puede afirmar que el número de clientes que pueda tener un negocio al momento de visitarlo para las entrevistas, no depende de si la tienda cuenta con refrigeradores propios o por el tiempo que tardan los clientes en regresar el dinero fiado ya que la mayoría de estos no recurren a ese compromiso con el encargado de la tienda.

Caso 4:

En el caso 4 encontramos nuestra primera regresión logística ya que la variable "108_does_the_micro_retailer_has_a_barred_window_" tiene valores dicotómicos de si la tienda sí cuenta con una ventana de protección o no lo hace.
Las variables dependientes seleccionadas fueron las siguientes "145_number_direct_competitors", "272_card_days_receive_money" y "276_expected_days_informal_credit".
Se procedió a declarar variables y a realizar y entrenar al código para que pudiera hacer una predicción.
De esta predicción se obtuvo la siguiente matriz de confusión.

```
Matriz de Confusión:
[[106   0]
 [ 26   2]]
```

Esta nos dice que el modelo predijo que las tiendas sí tienen protección 106 veces y realmente sí la tenían. No hay falsos positivos pero sí falsos negativos ya que el modelo predijo 26 veces que la tienda no tenía protección cuando realmente sí la tenía. Y por último, 2 veces predijo que no tenía protección y efectivamente no la tenía.
Con esta matriz se obtienen los siguientes valores de precisión, exactitud y sensibilidad.

```
Precision del modelo:
1.0

Exactitud del modelo
0.8059701492537313

Sensibilidad del modelo
0.07142857142857142
```

Estos valores nos dicen que el modelo tiene altos valores de precisión y exactitud pero un valor de sensibilidad bajo. Estos resultados se deben a las variables independientes escogidas y al número de datos con los que contamos para realizar el modelo.

Caso 5:

La segunda regresión logística tiene las variables independientes “268_number_fridges”, “104_how_many_shelves_does_the_micro_retailer_have” y ”2_current_permanent_employees”. La variable dicotómica es “99_does_the_micro_retailer_exhibits_products_outside_”.
Esta variable dicotómica nos dice si los negocios exponen o no sus productos fuera del establecimiento.
Como en el caso anterior se declararon variables y se entrenó al código para llevarlo a cabo.
De la predicción se obtuvo la siguiente matriz de confusión.

```
Matriz de Confusión:
[[96  0]
 [38  0]]
```

Esta nos dice que el modelo predijo que 96 veces la respuesta era sí y realmente fue sí, no hay falsos positivos, hay 38 falsos negativos que predijo que la respuesta sería no per era sí y finalmente no contamos con respuestas negativas.
Finalmente tenemos los valores de precisión, exactitud y sensibilidad.

```
Precision del modelo:
0.0
```

```
Exactitud del modelo
0.7164179104477612
```

```
Sensibilidad del modelo
0.0
```

Estos valores nos dicen que la regresión logística generada no tiene precisión ni sensibilidad pero tiene un alto grado de exactitud. Esto nuevamente debido a las variables independientes usadas y a las respuestas que se generaron en la columna ya que no obedecía un carácter dicotómico.