II MAPI – Segunda Conferencia Colombiana de Matemáticas Aplicadas e Industriales
Medellín, Antioquia, Colombia, 8 al 10 de Junio de 2022

Sistema de recomendación de inmuebles Bancolombia basado en Simulación Bayesiana y redes neuronales

Recommendation system for estate Bancolombia with bayesian simulation and neural networks

Juan Camilo Díaz Zapata[1,a]

[1]Gerencia de Modelos Analíticos, Centro de Excelencia en Analítica, Bancolombia, Medellín, Colombia

Resumen

Un sistema de recomendación de inmuebles para Bancolombia hace que la experiencia del cliente al conseguir su próxima vivienda sea memorable, ya que con este nuevo sistema el banco te facilita y acompaña desde la etapa de búsqueda de la vivienda hasta la legalización del inmueble (Bancolombia 2021). Bancolombia día a día cuenta con millones de datos, sin embargo, no cuenta con la información suficiente para recomendar viviendas de acuerdo a las afinidades y composición familiar de los usuarios, por esto se desarrolló un sistema de recomendación en tiempo real basado en cinco preguntas fundamentales hacia cualquier persona: (1) zona geográfica de interés, (2) rango de precio de búsqueda, (3) presupuesto para cuota inicial, (4) gustos y afinidades y (5) la composición familiar del usuario. El sistema de recomendación se basa en como integrar las respuestas de estas cinco preguntas a un conjunto de inmuebles que satisfaga de la mejor forma las respuestas de cada usuario.

La forma en cómo se manipulan las respuestas a estas cinco preguntas es clave para obtener una buena recomendación a los usuarios, por esto las tres primeras preguntas se tratan como un conjunto de filtros sobre la base total de inmuebles en la oferta existente, las dos últimas preguntas se tratarán como variables de clasificación que permitirán darle un orden a la base de ofertas resultante después de aplicar el filtro mediante las tres primeras preguntas, dando como resultado para cada cliente una lista de recomendaciones donde en primer lugar aparecerá la oferta más recomendada de compra.

La composición familiar es una de las variables que tiene mayor preponderancia a la hora de la búsqueda de una vivienda según grandes portales inmobiliarios, como lo describe por ejemplo Planeatubien.com (2021) y por supuesto es la variable de mayor influencia sobre el tipo de configuración del inmueble (tamaño del inmueble, número de habitaciones y de baños, etc) conllevando a que el tratamiento de esta respuesta sea ideal para el éxito del sistema de recomendación propuesto. Sin embargo, dada la importancia de esta variable y la ausencia de información de parte del banco que contenga una relación entre el número de personas a vivir en un inmueble y las características de un inmueble, se optó por simular una base de datos de modelado que permitirá el ajuste una red neuronal a partir de la elicitación de una distribución a priori basada en la distribución beta de acuerdo con el conocimiento experto de varias personas.

Una elicitación es el proceso de hallar una distribución de probabilidad de una variable aleatoria donde dichas probabilidades son basadas en el conocimiento de uno o varios expertos acerca de un punto de interés (Finerri 1937). Para la extracción de este conocimiento se diseña un cuestionario de elicitación tomando consideraciones desde el campo de estudio, en este caso el campo inmobiliario.

[a]Cientifico de Datos lider, Bancolombia.. E-mail: jucadiaz@Bancolombia.com.co

El proceso de la obtención de una base de datos simulada a partir de una elicitación, para la estimación ha sido estudiada y aplicada por distintos autores como Tangarife & Correa (2017) quienes estimaron los parámetros de un modelo logístico a partir de un elicitación de una distribución a priori beta y normal truncada. Correa & Barrera (2021) estimaron los parámetros de un modelo lineal múltiple a partir de una elicitación bayesiana a varios expertos, el algoritmo propuesto en este articulo para la obtención de la base de datos simulada de la cual se estimaran los parámetros de la red neuronal se basó en la metodología desarrollada por Tangarife & Correa (2017).

A continuación, se dará la explicación del algoritmo aplicado para la obtención de una base simulada que dará pie a la estimación de los parámetros de la red neuronal.

1. Se deben fijar los niveles de las covariables de forma adecuada donde $X_1, X_2, \ldots, X_k$ son puntos de diseño que se eligen en consenso con el experto en este caso, evaluadores de viviendas y asesores comerciales inmobiliarios. Donde:

   $$X_i = [x_{i1}, x_{i2}, x_{i3}, x_{i4}]$$

   Donde: x_{i1} : Número de personas en habitar el inmueble, x_{i2} : Tamaño del inmueble (mts), x_{i3} : Número de habitaciones del inmueble, x_{i4} : Número de baños del inmueble;
   en la tabla 1 se describe los puntos de diseño seleccionados para las variables x_{i2}, x_{i3}, x_{i4}.

x_{i2}	x_{i3}	x_{i4}
46	1	1
63	2	2
70	3	2
85	3	2
100	3	2
120	4	3
180	4	3

Tabla 1: Puntos de diseño para las variables x_{i2}, x_{i3}, x_{i4}

   Cada uno de los puntos descritos en la tabla 1 es combinado con los valores de la variable $x_{i1} = 1, 2, 3, 4, 5$; obteniendo un total de 35 $(k = 35)$ puntos de diseño.
   La variable respuesta esta definida como:

   $$y_j = \begin{cases} 1 & \text{si j-ésima persona con una composición familiar } x_{i1} \text{ adquiere el inmueble} \\ 0 & \text{en otros casos} \end{cases}$$

   $j = 1, 2, \ldots, n$

2. Para cada punto de diseño se procede a:

   - Se fija un n y se pide al experto dar el número personas con alguna composición familiar x_{i1} que comprarían en una muestra hipotética de tamaño n, dígase Y_0, calcule $E(\pi) = Y_0/n$.
   - Para el mismo n se pide al experto dar el número mínimo de éxitos que él esperaría aceptable, Y_I, calcule $\hat{\pi_I} = Y_I/n$.
   - Para el mismo n se pide al experto dar el número máximo de éxitos que él esperaría aceptable, Y_S, calcule $\hat{\pi_S} = Y_S/n$.

   Este paso se hizo con un tamaño n de 100 y una muestra de 16 expertos para cada punto de diseño.

3. A los valores elicitados en el punto dos, se ajusta una distribución beta para estimar los parámetros α y β, sean:
   $E(\pi) = y_0/n$

$P(\pi \leq Y_I/n) = 0.05$
$P(\pi \geq Y_S/n) = 0.05$

los valores de α y β se obtienen de minimizar la siguiente función:

$$f(\alpha, \beta) = (\pi_I - qbeta(0.05, \alpha, \beta))^2 + (\pi_S - qbeta(0.95, \alpha, \beta))^2 + (\pi + (\alpha/(\alpha + \beta)))^2$$

4. Calcule el N equivalente, esto permite cuantificar el conocimiento del experto en términos de tamaño muestral, este tamaño representa realmente el nivel de conocimiento que los expertos tienen en la clasificación adecuada de inmuebles a una composición familiar determinada.

$$N = \frac{4Z^2_{\alpha/2}\hat{\pi}(1-\hat{\pi})}{(\hat{\pi_S} - \hat{\pi_I})^2}$$

5. Para cada nivel se repiten los siguientes pasos m veces:

   - Genere un valor de la beta con α_i y β_i hallados en el punto tres.

   $$\begin{bmatrix} \pi_1 \\ \pi_2 \\ \vdots \\ \pi_k \end{bmatrix}$$

   En este paso se obtiene un vector de tamaño k, en nuestro caso $k = 35$
   - Genere una muestra de valores y a partir de la distribución conjugada de la beta, la distribución binomial.

   $$\begin{bmatrix} y_1^{(1)}, & X_1 \\ y_2^{(1)}, & X_1 \\ \vdots & \vdots \\ y_{n_{eq}}^{(1)}, & X_1 \\ y_1^{(2)}, & X_2 \\ y_2^{(2)}, & X_2 \\ \vdots & \vdots \\ y_{n_{eq}}^{(2)}, & X_2 \\ \vdots & \vdots \\ y_1^{(k)}, & X_k \\ y_2^{(k)}, & X_k \\ \vdots & \vdots \\ y_{n_{eq}}^{(k)}, & X_k \end{bmatrix}$$

   En este paso se obtiene una matriz de tamaño $(n_{eq} * k)$x5, donde n_{eq} es el N equivalente hallado en el punto cuatro y X_k es una variable que representa los niveles definidos en el punto uno de tamaño cuatro.

En nuestro caso el algoritmo fue construido en el sistema computacional R (R Core Team 2021) y obtuvimos una base de datos con un total de 770 mil registros entre todas las combinaciones de puntos de diseño y composiciones familiares con su respectiva variable respuesta. A partir de esta base de datos generada, se procedió a estimador una red neuronal por medio del paquete **Keras** en el sistema computacional Python (Van Rossum & Drake 2009) con cuatro neuronas de entrada dadas por x_{i1}, x_{i2}, x_{i3} y x_{i4} y una capa oculta de seis neuronas con función de activación sigmoide, obteniendo un $AUC = 80\,\%$ en base de validación y prueba.

Con este modelo puesto en producción los clientes de Bancolombia han podido obtener una mejor experiencia a la hora de adquirir una vivienda nueva y al modelo ir obteniendo información real de

los mismos clientes que le permitirá retroalimentar el modelo inicialmente desarrollado por una base de datos simulada.

***Palabras clave*:** Sistema de Recomendación, Inmuebles, Estadística Bayesiana, Redes Neuronales.

Referencias

Bancolombia (2021), 'Bancolombia recomendados'.
*https://tu360.grupobancolombia.com/inmobiliario/onboarding

Correa, J. & Barrera, C. (2021), 'Elicitation of the parameters of multiple linear models', *Revista Colombiana de Estadística - Applied Statistics* **44**(1), 159–170.

Finerri, B. D. (1937), 'La prevision: ses lois logigues, ses sources subjectives', *Annal es de I'Institut Henri Poincard* **7**, 1–68.

Planeatubien.com (2021), 'La dimensión ideal para tu departamento'.
*https://planeatubien.com/metros-cuadrados-la-dimension-ideal-para-tu-departamento/

R Core Team (2021), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
*http://www.R-project.org/

Tangarife, J. & Correa, J. (2017), 'Elicitación de una distribución a priori para el modelo logístico', *Comunicaciones en Estadística* **10**(2), 225–246.

Van Rossum, G. & Drake, F. L. (2009), *Python 3 Reference Manual*, CreateSpace, Scotts Valley, CA.