Primera línea del título del trabajo
Segunda línea del título

Trabajo Fin de Grado

Autor:

Autor del trabajo

Tutor:

Tutor 1 del trabajo
Tutor 2 del trabajo

Grado en Matemáticas

Junio, 2015
Universidad de Almería

Índice general

Abstract in English

Resumen en español

1 Objetivos

Los objetivos que nos proponemos obtener en este trabajo son los siguientes:

1. Creación de un algoritmo, haciendo uso de un entorno de desarrollo Java como es Intellij IDEA, el cual nos permitirá la creación de un nuevo modelo mediante redes bayesianas y con la que podremos asignar nuevos valores a los datos faltantes que encontraremos en nuestro dataset.
2. Comparar los resultados obtenidos mediante un grupo de control que se ha creado previamente.
3. Estudiar diferentes modelos para una mejora de los resultados.

1.1 dudas

¿Cual es nuestro problema? ¿Como podemos solucionarlo? ¿A que solución hemos llegado?

- Punto tres objetivo es por si sale mal el modelo y dar posibles soluciones (merece la pena?)

¿Se intercala parte teorica y parte práctica?

2 Introducción

En esta sección, trataremos de explicar, de manera escueta, como está la situación actual de una de las ramas más influyentes de las matemáticas como es el "análisis de Ddatos".

El análisis de datos es un proceso en el cual se examinan un conjunto de datos para extraer conclusiones sobre la información que contienen dichos datos. Con el paso de los años, estas conclusiones se han ido mejorando con el avance de las nuevas tecnologías. Hoy en día las técnicas y tecnologías de análisis se aplican considerablemente en el mundo comercial ya que gracias a ellas se pueden tomar mejores decisiones tanto en el ámbito comercial como en el ámbito social. Estas nuevas técnicas son llevadas a cabo por "*Data Scientist*". Un "*Data Scientist*" es aquella persona que posee las habilidades y técnicas necesarias para resolver problemas complejos y además, tenga las aptitudes necesarias para poder gestionarla. De este modo, se han unificado dos mundos que hasta este momento se encontraban separados como es el de la gestión de datos y el análisis de datos.

En este trabajo, nos introduciremos en la piel de un "*Data Scientist*" con lo que para ello, deberemos de seguir los siguientes pasos:

1. **Extracción de datos**: Seremos capaces de poder extraer la información independientemente del origen (*csv* , *txt* , *arff* . . .) y del volumen de esta (*Big Datat* o *Small Data*)

2. **Limpieza de datos**: Tendremos que ser capaces de eliminar toda aquella información que nos puedan ocasionar un perjuicio en el resultado final ó una distorsión de la información.

3. **Procesamiento de datos**: Tendremos que ser capaces de poder aplicar diferentes métodos estadísticos para obtener una respuesta lo más óptima posible. Dentro de estos métodos podremos hacer uso de los modeles de regresión, pruebas de hipótesis, inferencia estadística...

4. **Diseño**: Tendremos que ser capaces de poder diseñar o experimentar nuevos test en caso de que lo veamos necesario.

5. **Visualización**: Tendremos que ser capaces de poder mostrar esta información lo más nítida e inteligible posible para que pueda estar al alcance del mayor número posible de sujetos.

Comenzaremos partiendo de una base de datos la cual contiene información acerca de las viviendas (tipo de vivienda, número de metros cuadrados construidos, número de baños..) que posee en propiedad el Grupo Cooperativo Cajamar (GCC) y que es comercializado a través de la operadora Haya Real Estate.

3

Creación del algoritmo

En este capítulo se expondrá de forma detallada el proceso seguido para poder obtener el principal objetivo del presente trabajo.

3.1 Pre-procesamiento

Como dijo *Dhanurjay Patil*, ex-anlista de datos de Google, el 80% de un proyecto consiste en la limpieza de datos. Es por ello, que gran parte de este trabajo se lo dedicaremos al pre-procesamiento de los datos.

Como hemos dicho al final del capitulo anterior, hemos seleccionado una base de datos en la que encontramos múltiples variables acerca de las características y morfología de las viviendas que tienen en propiedad el GCC.

Haciendo uso de una sentencia SQL realizaremos una llamada a la tabla que contiene toda la información de la que disponemos de los inmuebles. En total, encontramos 237 variables de las cuales nos quedaremos con 7 variables. Estas variables son:

1. **Tipo vivienda**.
2. **Número de baños**.
3. **Número de habitaciones**.
4. **Número de ascensore**.
5. **Número de metros cuadrados construidos**.
6. **Número de metros cuadrados útiles**.
7. **Importe de tasación**.

El motivo de la selección de esta variables es debido a que presenta un alto grado de datos faltantes y un alto grado de datos erróneos por culpa de una mala introducción por parte del operario. En total, nuestra base de datos constan de 27.945 registros. A continuación procederemos a guardar toda esta información en un libro de Excel (.csv) para así poder realizar una limpieza de los datos.

Tipo vivienda

Según nuestro dataset, podemos encontrar los siguientes tipos de vivienda, tal y como se muestra en el siguiente gráfico.

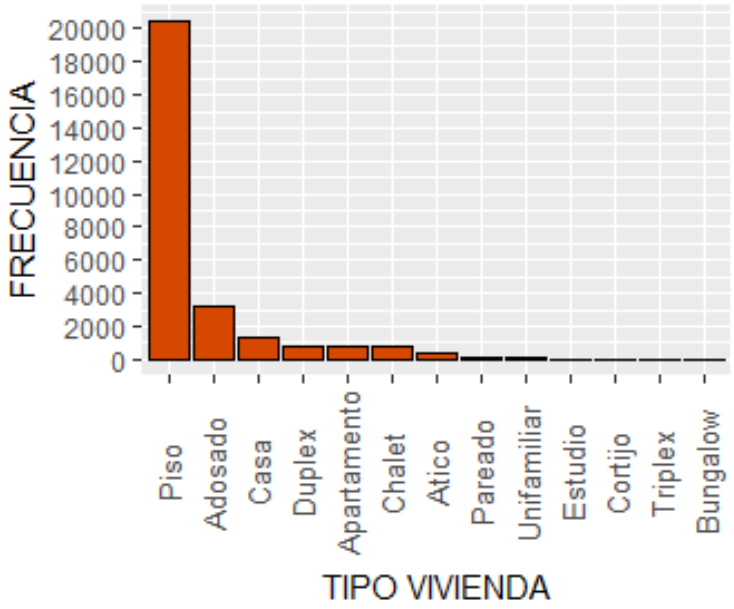

Figura 3.1: Tipología vivienda

Ademas, podemos observar como se distribuyen el número de registros en función de la vivienda.

Tipo vivienda	Total
Piso	20.421
Adosado	3.163
Casa	1.319
Dúplex	839
Apartamento	825
Chalet	753
Ático	416
Pareado	90
Unifamiliar	84
Estudio	16
Cortijo	12
Tríplex	5
Bungalow	2
Total	27.945

Cuadro 3.1: Frecuencia viviendas

Debido a que el tipo vivienda Pareado , Unifamiliar , Estudio , Cortijo , Tríplex y Bungalow no tiene suficiente representatividad en la muestra, procedemos a la eli-

minación de ello. Con lo que en total, obtenemos un dataset compuesto por 27.736 registros y 7 tipos de viviendas.

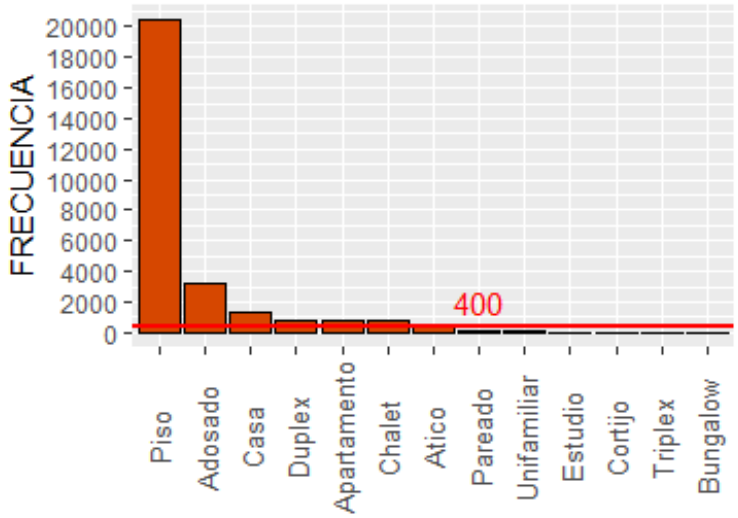

Figura 3.2: Tipología vivienda

Número de baños

En la siguiente imagen, se puede observar un histograma del número de baños

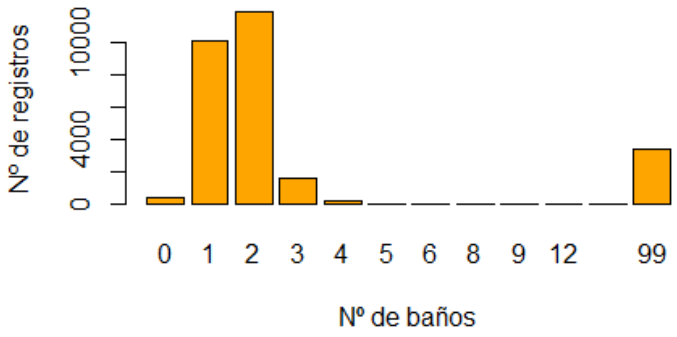

Figura 3.3: Nº de baños

Hay que tener en cuenta que, para que una vivienda consiga la célula de habitabilidad debe de contar con al menos un baño y una habitación. Teniendo en cuentas esto y observando la Figura 3.3 se procederá a pasar como *missing* aquellos valores que sean 0 ó que tengan más de 6 baños.

Prueba Prueba Prueba Prueba Prueba PruebaPrueba Prueba PruebaPrueba Prueba PruebaPrueba Prueba PruebaPrueba Prueba PruebaPrueba Prueba PruebaPrue-

ba Prueba PruebaPrueba Prueba Prueba

Bibliografía

[1] M. Abramowitz, I.A. Stegun, *Pocketbook of Mathematical Functions*, Verlag Harri Deutsch, 1984.

[2] D. Dominici, *Mehler–Heine type formulas for Charlier and Meixner polynomials*, arXiv:1406.6193v1.

[3] S.F. Khwaja, A.B. Olde Daalhuis, *Uniform asymptotic approximations for the Meixner–Sobolev polynomials*, Anal. Appl., **10** (2012), 345–361.

[4] Página web de los usuarios de TEX: http://www.tug.org.