Predicción de la percepción de sismos por la población a partir de datos básicos de terremotos

Trabajo final de la asignatura
"Taller de tesis"

Maestría en Explotación de Datos y Descubrimiento del Conocimiento

Facultad de Ciencias Exactas y Naturales
Facultad de Ingeniería

Autor:
Víctor A. Bettachini

Fecha:
9 de julio de 2024

Resumen

Una decena de terremotos se producen en el territorio nacional en forma diaria sin que sean detectados más que por instrumentos. Son muy pocos los casos en que las ondas sísmicas que estos producen son percibidos por la población. Se buscó desarrollar un modelo para predecir tal eventualidad utilizando usando solo datos básicos de terremotos. Limitaciones de los datos públicos utilizados limitaron el espacio geográficos del estudio a la provincia de San Juan. Se ensayaron modelos basado en regresión logística como uno basado en máquinas de potenciación de gradiente. Se logró ajustar modelos con alternativamente alta sensibilidad o precisión, pero no ambas simultáneamente, por lo que debe optarse entre estas alternativas según cual sea la prioridad para la aplicación que quiera dárseles.

Índice general

Capítulo 1

Introducción

1.1. Contexto y motivación científica

Un rápido desprendimiento entre dos facetas enfrentadas de sendas placas tectónicas que se traban mutuamente su desplazamiento relativo produce una rápida liberación de energía que se denomina terremoto. Esto sucede a cierta profundidad en la corteza terrestre en el punto denominado hipocentro a partir del cual parte de esta energía se encauza como ondas elásticas. El estudio de estas ondas es el área llamada sismología y de ahí el termino sismo para un evento particular detectado, pero que debiera aclararse si se produjo por un terremoto u otra fuente de ondas [1, sección 4.1.1].

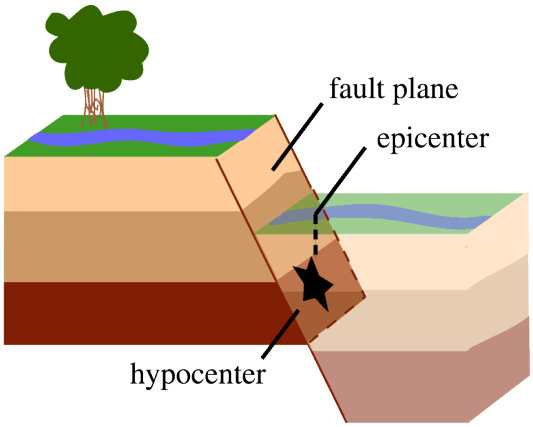

Figura 1.1: El epicentro es el punto de la superficie inmediatamente sobre el hipocentro que es el punto a cierta profundidad al que se adscribe un terremoto. Reproducido de [2].

Sea que las ondas de compresión longitudinal de la corteza, las tipo P, o las del tipo S transversales y más lentas, ambas arriban con mayor intensidad al punto de la superficial terrestre que se encuentra directamente sobre el hipocentro, que se denomina epicentro [1, sección 4.1.2] como ilustra la figura 1.1.

Cuanto más próximo es un punto en la superficie al hipocentro, la amplitud de las ondas sísmicas registradas en la superficie es mayor. Tanto esta amplitud como el período de oscilación son mucho mayores que el de otros desplazamientos de la corteza como los de las mareas solares y lunares de la corteza [1, sección 4.1.4]. Como resultado, estructuras artificiales pueden agitarse poniendo en riesgo su estabilidad estructural y haciendo caer elementos que no estaban fijados a esta o perdieron tal adhesión a causa de la agitación misma. Como consecuencia, los sismos más fuertes pueden generar graves daños, poniendo en riesgo la integridad física y la seguridad de las personas al generar daños en las viviendas y edificios, derrumbes de puentes, rompimiento de vidrios, entre otros [3].

La mayor parte de los sismos presentan ondas de pequeña amplitud y no generan daños materiales. Que sean o no detectados por la población es un factor relevante en su percepción de confianza vis-à-vís de los organismos de monitoreo y prevención de riesgos. Cuando informar a la población de la ocurrencia de un sismo es una decisión de política pública que debiera apuntar a no alertar innecesariamente sobre sismos menores imperceptibles [4]. El caso inverso es también problemático, en que ante una carencia de una comunicación oficial de la poca importancia de un evento sísmico llevó a la autoevacuación por parte de la población que lo percibió [5].

1.2. Objetivos del trabajo / Pregunta

Contar con una estimación rápida a partir de los datos sísmicos registrados por instrumental de si un dado evento será percibido por la población o no permitiría a las autoridades tomar decisiones informadas sobre la comunicación a la población. Este trabajo busca determinar el grado de certeza con que ciertos métodos de ciencias de datos pueden predecir si la población percibirá actividad sísmica producto de terremotos a partir de unos pocos datos básicos sobre los mismos publicados por el Instituto Nacional de Prevención Sísmica (INPRES) que es el organismo público de la República Argentina que realiza estudios e investigaciones básicas y aplicadas de sismología [6].

1.3. Estructura del documento

Se estructuró en los siguientes capítulos con temáticas diferenciadas

Introducción : se presentará el contexto y la motivación científica, los objetivos del trabajo y la estructura del documento.

Marco teórico : se revisarán los trabajos previos y relevantes, se presentarán los conceptos y técnicas de ciencia de datos utilizados en el trabajo.

Metodología : se describirán los datos utilizados, el preprocesamiento y limpieza de los mismos, el análisis exploratorio de los datos, las técnicas de análisis y modelado utilizadas, la selección de características, las métricas de evaluación de los modelos y los métodos estadísticos utilizados.

Resultados y discusión : se presentarán y analizarán los resultados obtenidos, se discutirán los resultados y su relevancia, se identificarán las limitaciones y posibles mejoras.

Conclusión : se resumirán los hallazgos principales, se presentarán las conclusiones generales y su relación con los objetivos del trabajo, se discutirán las aplicaciones y relevancia de los resultados.

Capítulo 2

Marco teórico

2.1. Relevamiento de trabajos previos y relevantes

En la prensa general se busca transmitir la importancia de un fenómeno sísmico informando un valor sea en la escala de Richter o de Mercalli como si fueran intercambiables. La primera se refiere a la magnitud, una característica física del terremoto (ver definición en la sección 3.5), en tanto que la segunda se refiere a una caracterización subjetiva del sismo, la llamada intensidad. Esta última cuantidad es función de fenómenos percibidos por el público como si hubo grietas en el suelo o vidrios o si temblaron edificios [1, sección 4.2.3]. Los organismos de monitoreo de sismos cuentan con portales en la world-wide web para que el público ofrezca tal información voluntariamente en formularios con preguntas de elección múltiple. Tal es el caso de los portales "Encuesta de sismos" del INPRES [7] y "Felt Report - Tell Us!" [8] parte del programa de riesgo de terremos del Geological Survey de los Estados Unidos de América (USGS) [9]. Este último aporta a la sistematización de reportes del público denominado "Did you fell it?" (DYFI) que ha nutrido en la último década y media a varios estudios relacionando la intensidad con diversos fenómenos citados en [10].

La intensidad expresada en la escala modificada de Mercalli (MMI) [11] en la base de datos DYFI mostró estar bien correlacionada con medidas de espectro de oscilación lo que impulsó a buscar un modelo que le relacione con parámetros físicos de terremoto. Una regresión múltiple basada en el método de máxima verosimilitud se encontraron los factores $c_{1\ldots7}$ de la [12, ecuación 1]

$$MMI = c_1 + c_2(M_w - 6) + c_3(M_w - 6)^2 + c_4 \log R + c_5 R + c_6 B + c_7 M_w \log R, \quad (2.1)$$

donde M_w es la magnitud del momento, según la definición discutida en la sección 3.1.1, $R = \sqrt{D + h}$ es la distancia entre observador e hipocentro en función de la distancia del primer al epicentro D y profundidad del segundo h, y B es nulo para $R < R_t$ una distancia de transición en la forma de atencuación de las ondas Una posterior actualización de este modelo por los mismos autores se basó en agrupar los valores promedio de MMI en cada código postal

$$MMI = c_1 + c_2 M_w + c_3 \log R + c_4 R + c_5 B + c_6 M_w \log R, \quad (2.2)$$

con $R = \sqrt{D + h + 14^2}$ donde el factor adicional, 14^2 da cuenta de la saturación a cortas distancias permitiendo simplificar el factor $B = \text{máx}(0, \log(R/50))$ que ya no requiere establecer un valor de transición R_t [13].

Lo anterior muestra que hay antecedentes de que un ajuste permite predecir la percepción de los mismos por parte de la población a partir datos físicos de los terremotos Para este trabajo lo que busca determinarse es una decisión binaria entre lo que correspondería a la intensidad 0 y 1 de la escala MMI, respectivamente una percepción solo con instrumentos o una por al menos una fracción de la población [11]. No es el enfoque a seguir discriminar con un umbral tras una predicción del valor MMI sino utilizar las técnicas de aprendizaje automático que generen una decisión binaria directamente desde los datos de parámetros físicos.

2.2. Conceptos y técnicas de ciencia de datos utilizados en el trabajo

Se utilizan en este trabajo dos técnicas con aplicación a la predicción pero en sendos extremos de complejidad y transparencia en cuanto al peso relativo que tienen las variables independientes en el resultado, el de regresión logística y el de máquinas de potenciación de gradiente, más conocidos por su nombre en inglés Gradient Boosting Machines (GBM).

2.2.1. Regresión logística para la predicción binaria

Lo siguiente es un resumen del material de la novena clase de la asignatura *Enfoque estadístico del aprendizaje* titulada *Regresión Logística* elaborado por Juan Barriola, Azul Villanueva y Franco Mastelli.

Un modelo de regresión lineal con coeficientes β_j para cada variable X_j que busque la probabilidad de una dependiente binaria $P(Y)$

$$P(Y) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j, \tag{2.3}$$

no presentaría un punto de corte claro para clasificar los datos en dos categorías. Para sortear esta dificulta se toma la salida de esta regresión como la variable dependiente de una regresión logística,

$$P(Y|X) = \frac{e^{\beta_0 + \sum_{j=1}^{p} \beta_j X_j}}{1 + e^{\beta_0 + \sum_{j=1}^{p} \beta_j X_j}}, \tag{2.4}$$

lo que asegura un valor entre 0 y 1. De esta expresión puede arribarse a

$$\log\left(\frac{P(x)}{1 - P(x)}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j, \tag{2.5}$$

cuyo lado izquierdo es la función *logit*, la inversa de lo que sería la logística de la probabilidad $\left(1 + e^{-P(x)}\right)^{-1}$.

Para implementar tal función en código en lenguaje *R*, el utilizado en este trabajo, se hace uso de la función *glm2* provista por la biblioteca homónima [14]. Esta tiene por argumentos *formula* y *data*, los mismo que la popular función *lm* incluida en el paquete base de R para modelos lineales. Pero como glm2 genera modelos lineales generalizados, requiere un argumento adicional, *family*, para indicar la distribución del error de la variable a predecir:

- Binomial: link=logit
- Poisson: link=log
- Gaussiana: link=identidad

La correspondiente función de enlace (link) relaciona el modelo lineal con función de probabilidad. Como se busca predecir un resultado booleano se indica la distribución binomial.

Aunque se parte de pocas variables independientes, se ensayará forzar una mayor simplificación del mismo mediante la técnica de regularización tipo Lasso (L1) que fuerza a que los coeficientes de las variables independientes que menos aportan a predicción vayan anulándose.

2.2.2. XGBoost para una predicción binaria

En otro extremo entre las herramientas de ciencia de datos por su complejidad se ensayará utilizar máquinas de potenciación de gradiente, más conocidos por su nombre en inglés *Gradient Boosting Machines* (GBM) para la predicción. Las distintas implementaciones de estos algoritmos, como *XGBoost*, *LightGBM* o *CatBoost* son capaces de producir un único modelo con fuerte poder predictivo a partir de la síntesis de resultados de modelos de predicción débiles, típicamente árboles de decisión. Para este trabajo se utiliza XGBoost implementado en lenguaje R por la biblioteca *xgboost* [15].

Capítulo 3

Metodología

3.1. Presentación y descripción de los datos utilizados

En el marco de los *Proyectos de Asistencia Estadística* del *Instituto de Cálculo* (IC) de la *Facultad de Ciencias Exactas y Naturales* (FCEyN) de la *Universidad de Buenos Aires* (UBA) se publicaron conjuntos de datos en un repositorio curado con el objeto de ser aplicados a la enseñanza de la estadística y la ciencia de datos por Daniela Parada, investigadora del IC [16]. De estos conjuntos el utilizado en este trabajo es el que se publica en el apartado "Visualización" que corresponden a datos de sismos de Argentina de la última década [17]. En este repositorio alojado por la firma GitHub, se provee un front-end html que da un contexto, hace una exploración inicial, un análisis para una provincia en particular, muestra una estimación de probabilidad y provee otra información sobre lo datos.

Los datos corresponden a detecciones por parte de estaciones de monitoreo sísmico en la República Argentina recopilados y publicados por el INPRES en su sitio web [18]. En el sitio de publicación de los datos se indica que el conjunto de datos comprende las fechas desde el 7 de enero de 2012 hasta el 18 de mayo de 2022 y fue realizado con datos *scrappeados* del buscador de sismos del INPRES por Gustavo Juantorena [17, sección 4.1].

Allí mismo se describe que el conjunto de datos reducido y curado denominado "sismos", el que se utilizó en este trabajo, es accesible a través de la importación de la biblioteca `datosIC` en lenguaje R [17, sección 5.1.1]. Este mismo conjunto reducido puede descargarse en formato de valores numéricos separado por comas (CSV) apuntando a su URL en el repositorio alojado en GitHub [19].

Las variables reportadas para cada sismo son:

- *Fecha*: en el formato `aaaa-mm-dd` de la norma ISO 8601 [20].
- *Hora*: una cadena de caracteres en formato `hh:mm:ss` con una exactitud al segundo.
- *Latitud, Longitud*: un número con una exactitud de un decimal con grados como unidad.

- *Provincia*: cadena de caracteres del nombre de la provincia donde se produjo el sismo (no donde se ubicó quién potencialmente lo percibiera) según se afirma en el sitio de publicación [17, pág. 5.1.1].
- *Magnitud*: un número función de un logaritmo de la amplitud de las ondas sísmicas, específicamente la escala de magnitud de momento (M_w) (leer discusión en el párrafo siguiente).
- *Profundidad*: un número entero con kilómetros como unidad que indico que tan bajo la superficie se ubicó el epicentro.
- *Percibido*: variable booleana de si hubo reportes de percepción del fenómeno por parte de la población,

Esta última variable es la que se busca predecir en este trabajo en función de las demás.

3.1.1. Unidad de magnitud en el conjunto de datos

Hay una multitud de especificaciones derivadas de la escala logarítmica originalmente propuesta por Richter en 1935 [1, sección 4.2.3]. Sin embargo no hay una especificación de en cual de estas se expresa la *magnitud* del conjunto de datos en [17]. Tampoco hay una nota adjunta a los datos que lo indique en el repositorio [19]. Asimismo, en la fuente original de donde se hizo el previamente mencionado *scrapping*, el buscador de sismos del INPRES [18], al solicitar datos solo se indica el término *magnitud* sin más detalle.

A pesar de carecer de una indicación explícita al respecto se opta por asumir que la magnitud en el conjunto de datos corresponde a la escala de magnitud de momento (M_w) que es la más comúnmente utilizada en la actualidad como se detalla en el párrafo siguiente. Avala tal suposición que en la página *Cálculo de la Magnitud* de la sección de educación del sitio del INPRES se la presenta como una escala englobadora [21]. Pero se considera un avala más crucial un documento de la *Comisión de trabajo de gestión de riesgo* en que el INPRES figura no solo como el *organismo con responsabilidad operativa* sino también entre los *organismos que generan información de base* en [22, anéxo X] detalla "actualmente la más utilizada es la magnitud momento, M_w".

Unidad de magnitud más corrientemente utilizada A medida que se fueron instalando más estaciones sismográficas en todo el mundo, se hizo evidente que el método desarrollado por Richter sólo era estrictamente válido para determinados rangos de frecuencia y distancia. Para aprovechar el creciente número de estaciones sismográficas distribuidas por todo el mundo, se desarrollaron nuevas escalas de magnitud que son una extensión de la idea original de Richter en adelante denominada M_L. Entre ellas se incluyen la magnitud de onda de cuerpo, m_b y la magnitud de onda de superficie M_s (ver sección 3.5). Cada una de ellas es válida para un rango de frecuencias y un tipo de señal sísmica concretos. En su rango de validez, cada una es equivalente a la magnitud Richter. Debido a las limitaciones de las tres escalas de magnitud, M_L, m_b y M_s, se desarrolló una nueva extensión de la escala de magnitud, conocida como magnitud de momento o M_w, de aplicación más uniforme. En particular, para los terremotos de gran magnitud, M_w ofrece la estimación más fiable de la importancia de mismo [23].

3.2. Adquisición y formateo de los datos

Puesto que el conjunto de datos es curado por un equipo de investigación de la UBA, se asume que los mismos son confiables y que no se requiere de un proceso de limpieza de los mismos. De todas formas se realizaron las verificaciones usuales cada vez que se utilizan datos tabulares en un estudio de estadístico y/o de ciencia de datos.

3.2.1. Carga y verificación de faltantes o duplicados

Tras descargar el archivo de datos en formato CSV se le importó en una estructura de datos *data.table* de en un entorno de trabajo en lenguaje R. Esta estructura de datos permite una consulta de los datos análoga a la del lenguaje SQL de bases de datos relacionales lo que le hace una herramienta versátil para el análisis de datos tabulares [24].

La presencia de valores faltantes indicados con el símbolo `NA` se descartó cuando la ejecución **sum**(**is.na**(sismos_arg)) arrojó un cero como resultado. Por el contrario ejecutar sismos_arg[**duplicated**(sismos_arg, fromLast = TRUE)] mostró unos 23 registros duplicadas sobre un total de 55 817 registros.

En una nueva tabla con nombre más corto, sismos, se copiaron los registros sin duplicados ejecutando sismos_arg[**!duplicated**(sismos_arg)].

3.2.2. Inspección y cambio de formato de datos

Ejecutar la función **colnames** con la *data.table* denominada sismos_arg como argumento permitió verificar que contuviera las columnas con los nombres anunciados en el sitio que publica los datos en su sección [17, Exploración inicial]. Las transformaciones o ingeniería de características que se detallan luego en esta sección se realizaron en función de los tipos de datos de cada columnas constatados con la función str.

Fecha y hora del sismo Si se quieren utilizar las variables *Fecha* y *Hora* para segmentar el día en franjas horarias, o comparar interanuales, es útil convertirles en funciones de continuas, días corridos en el año, para la primra, y segundos, para la segunda

La variable Fecha se reconocíó al importarse el conjunto de datos en formato ISO 8601 por lo que registrar una nueva variable para el año y otra para el día del año, fue sencillo con las funciones year y yday de la biblioteca base de R.

A partir de la varible Hora se escribió una función que genera otra continua contando el número entero de segundos transcurridos desde la medianoche del día en que se produjo el terremoto, *Segundos del día*, apoyandose para esto en la función strptime del paquete base de R. Esto habilita posibles análisis segmentando el día en franjas horarias.

Resumiendo las nuevas variable generadas fueron:

- *Segundos del día*: número entero de 0 a 86399
- *Día del año*: número entero de 1 a 366
  sismos[, ‘**D**ía del año‘ := yday(**as**.Date(Fecha, **format** = "%Y−%m−%d")),]
- *Año*: número entero de 2012 a 2022
  sismos[, Año := year(**as**.Date(Fecha, **format** = "%Y−%m−%d"))]

3.3. Delimitación del espacio geográfico

sec:geográfico

Distribución geográfica de los datos Como es esperable la mayor parte de los hipocentros se ubican en regiones con orografía elevada, como la cordillera de los Andes, producto de la subducción de la placa de Nazca bajo la placa Sudamericana, o las sierras cordobesas producto de procesos mucho más antiguos. Una ubicación de los mismos sobre un mapa físico lo ilustra en la figura 3.1. Esto conlleva a que la mayor parte de los terremotos estén alejados de las mayores urbanizaciones reduciendo la probabilidad de que sean percibidos por la población.

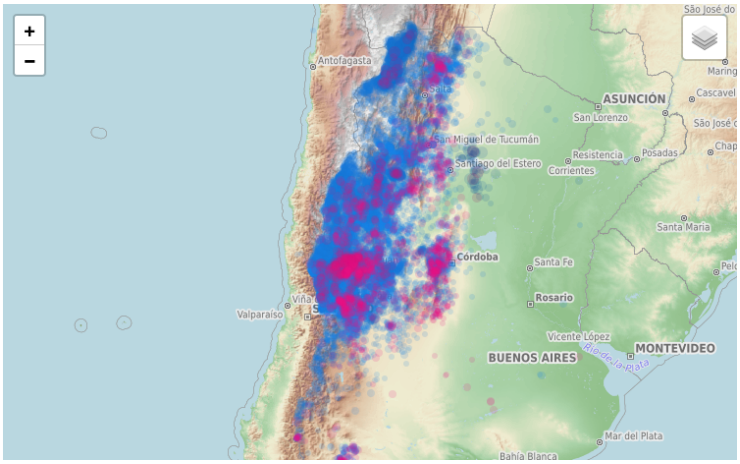

Figura 3.1: Superponer las ubicaciones de los terremotos en un mápa físico muestra que la mayoría de los terremotos se registran en zonas de montaña o cerros en su mayoría alejados de las mayores urbanizaciones lo que afecta negativamente la estadística de los percibidos. La escala de colores muestra los más superficiales en azul y profundos en rojo. Reproducido de [17]

La ubicación de los terremotos sobre un mapa político que muestra la figura 3.2 deja a las claras que no tienen una distribución homogénea aún en provincias con abundante actividad.

El problema de la distancia Evidentemente la proximidad de un epicentro a una población es un factor relevante en la percepción de un sismo. El INPRES tiene una metodología operativa para registrar la ubicación de los usuarios que reportan haber percibido un sismo a través de su página web [25]. Su buscador de sismos indica los sismos "sentidos" por usuarios con un color en sus resultados de búsqueda como ilustra la figura 3.3.

Lamentablemente en el conjunto de datos curados no figura la ubicación del sismografo o la población que hizo el registro lo que imposibilita incorporar la distancia al epicentro como una variable en el modelo.Puesto que no se tiene control posible sobre la posición de quienes potencialmente percibieron un sismo se decidió para reducir el impacto del parámetro distancia limitar el espacio

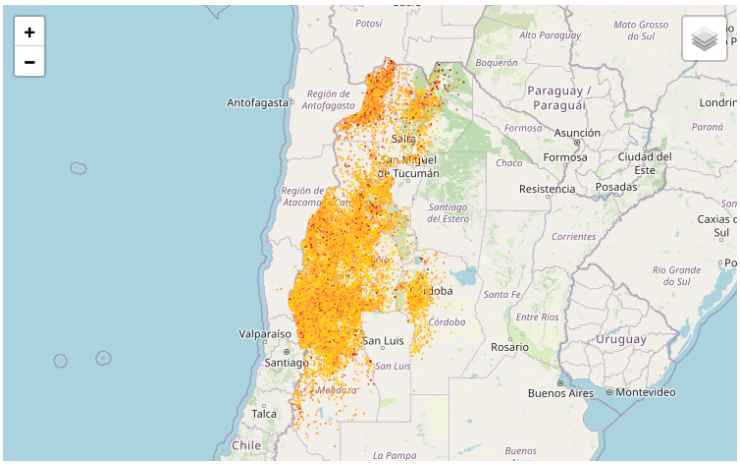

Figura 3.2: Sobre este mapa político se representan las ubicaciones de los terremotos como puntos. Esto permite apreciar que en algunas provincias, como San Luís, se omiten datas como evidencia el corte abrupto de reportes en su frontera. En otras se hacen más claros sesgos geográficos, como en Córdoba, Salta o Jujuy, Reproducido de [17]

geográfico de los terremotos. Las posibles fuentes de datos para realizar tal operación son su latitud, longitud y provincia.

Elección de la provincia de San Juan Respondiendo a la discutido en el párrafo anterior se buscó trabajar en un extensión geográfica relativamente limitada, pero cuidando que el un número de sismos restantes sea aún elevado. Tras una inspección visual de la figura 3.2 se determinó que la provincia de San Juan cumple con tales requisitos, por lo que se procedió a filtrar los datos con el comando sismos_SJ <—sismos[Provincia == "San_Juan"]. El subconjunto de datos de esta provincia representa un $\approx 54\,\%$ de los datos hasta aquí disponibles. Con 29 917 registros es aún un número suficiente para realizar análisis cono los métodos elegidos.

Latitud y longitud La distribución de terremotos en la provincia de San Juan dista de ser homogénea pues hay una preponderancia de las ubicaciones meridionales como puede apreciarse en la figura 3.4.

Lo anterior podría hacer sospechar de un sesgo con la latitud, pero la inspección visual de la figura muestra que la proporción de puntos rojos y amarillos, los de terremotos percibidos o no es similar con la que hay en el norte. En función de esto se decidió no eliminar las variable de latitud en este estadío y analizar más adelante su correlación con la percepción de los sismos. A fin de cuentas el objetivo del trabajo es determinar si un modelo puede predecir la percepción de un sismo por parte de la población y no determinar los factores que influyen en la percepción de los sismos por parte de la población. En caso de que haya factores geográficos regionales, estos serán capturados por el modelo de clasificación y no necesariamente por la variable de latitud.

Un razonamiento similar se aplica para no coartar la variable de longitud.

INSTITUTO NACIONAL DE PREVENCIÓN SÍSMICA

MINISTERIO DE ECONOMÍA - SECRETARÍA DE OBRAS PÚBLICAS

Sismología | Red Nacional de Estaciones Sismológicas | Red Nacional de Acelerógrafos | Ingeniería Sismorresistente | Laboratorio de Estructuras | INPRES-CIRSOC Reglamento | Publicaciones | Servicios

Buscador de sismos

Resultado de la búsqueda

12sísmos encontrados según las características especificadas. (Los sismos listados en color rojo han sido sentidos.)

Id	Fecha	Hora	Latitud	Longitud	Profund.	Magn.	Intensidad	Provincia	
1	14/06/2024	23:32:49	-31.860	-69.663	119 Km.	2.7		SAN JUAN	mapa
2	14/06/2024	22:12:32	-23.274	-66.653	233 Km.	3.5		JUJUY	mapa
4	14/06/2024	20:15:52	-31.897	-69.960	126 Km.	2.9		SAN JUAN	mapa
5	14/06/2024	18:51:09	-31.547	-69.410	125 Km.	2.8		SAN JUAN	mapa
6	14/06/2024	18:44:59	-28.521	-67.919	112 Km.	2.5		LA RIOJA	mapa
7	14/06/2024	16:12:04	-31.250	-68.664	108 Km.	3.8	II a III -Ciudad de San Juan, San Juan; II a III -Albardón, San Juan; II a III -Caucete, San Juan; II a III -Villa San Martín, San Juan	SAN JUAN	mapa
8	14/06/2024	14:41:57	-31.538	-69.343	102 Km.	2.8		SAN JUAN	mapa
9	14/06/2024	14:02:37	-31.898	-69.009	111 Km.	2.8		SAN JUAN	mapa
10	14/06/2024	12:38:36	-31.818	-69.715	100 Km.	2.5		SAN JUAN	mapa
13	14/06/2024	04:38:59	-31.297	-68.555	108 Km.	2.8		SAN JUAN	mapa
14	14/06/2024	04:15:16	-23.687	-66.592	227 Km.	3.2		JUJUY	mapa
15	14/06/2024	03:30:23	-31.210	-68.433	104 Km.	3.1		SAN JUAN	mapa

1

volver a Búsquedas

Figura 3.3: Resultados de una búsqueda manual de sismos en el sitio del INPRES. El indicado en rojo fue percibido por la población. En la columna *intensidad* se dan datos de ubicación de la población que lo percibió.

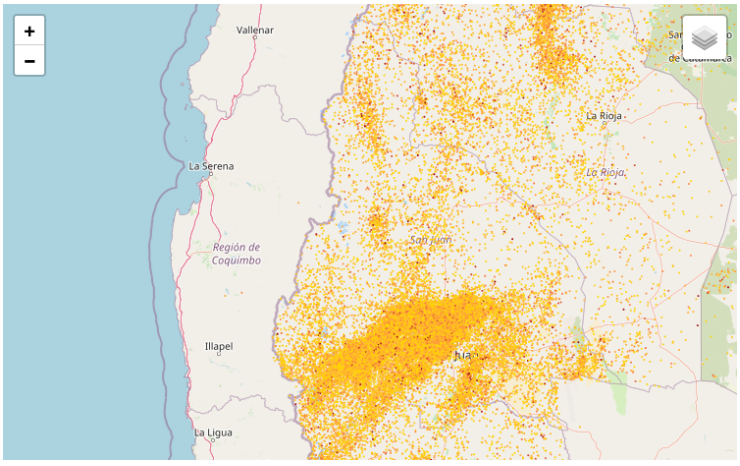

Figura 3.4: Recorte de la figura 3.2 en torno a la provincia de San Juan. Los terremotos se presentan mayoritarialmente en su región meridional. Los puntos rojos son los percibidos por personas. Reproducido de [17]

Como se comentó al presentar la figura 3.1, son esperables más terremotos y de mayor intensidad en las regiones occidentales de la provincia de San Juan. Nuevamente será una virtud de los modelos si pueden explotar tal información para mejorar la clasificación de los sismos percibidos por la población.

3.4. Análisis exploratorio de datos

Un primer vistazo sobre los datos con **summary**(sismos_SJ) permitió obtener un resumen de las variables numéricas y categóricas. Saltan a la vista que hay valores extremos en el extremo superior de la variable Magnitud bastante alejados de la mediana y que hay un fuerte desequilibrio en la variable "Percibido", la de clase de clasificación, en favor de los terremotos no percibidos.

Desequilibrio en la clase de clasificación Sobre el total de registros solo un valor cercano al 98 % fueron percibidos por el instrumental y no por la población. Restan tan solo unos 619 registros que efectivamente fueron percibidos por la población. Este desequilibrio llama al uso de técnicas de balanceo de clases en los modelos de clasificación a utilizar.

Distribución de la magnitud La magnitud de terremotos forzosamente presenta una distribución asimétrica por el hecho de que cuanto mayor es la energía liberada, más infrecuente es el fenómeno. La relación logarítmica de la frecuencia con la magnitud de los sismos que muestra la figura 3.5 para San Juan en los años 2012 a 2022 es universal es coincidente con lo enunciado en la ley empírica de Gutenberg-Richter [1, ec. 4.24].

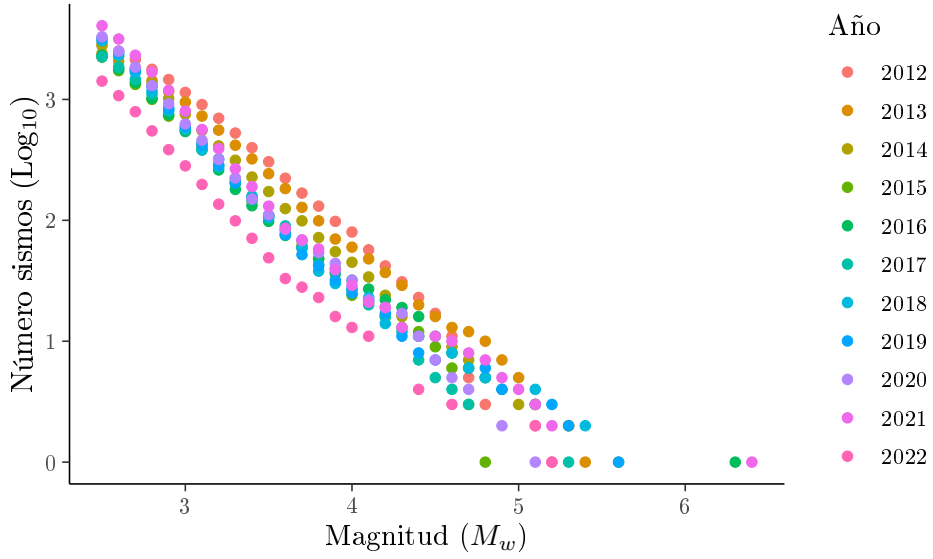

Figura 3.5: Los sismos de mayor magnitud son más infrecuentes. Generado con código provisto junto con los datos para la provincia de San Juan [17, sección 4.2.1]

Los sismos percibidos por la población no cumplen esta relación logarítmica, quedando relegados los de menor magnitud dando cuenta de la dificultad en percibirlos en ese caso como ilustra la figura 3.6.

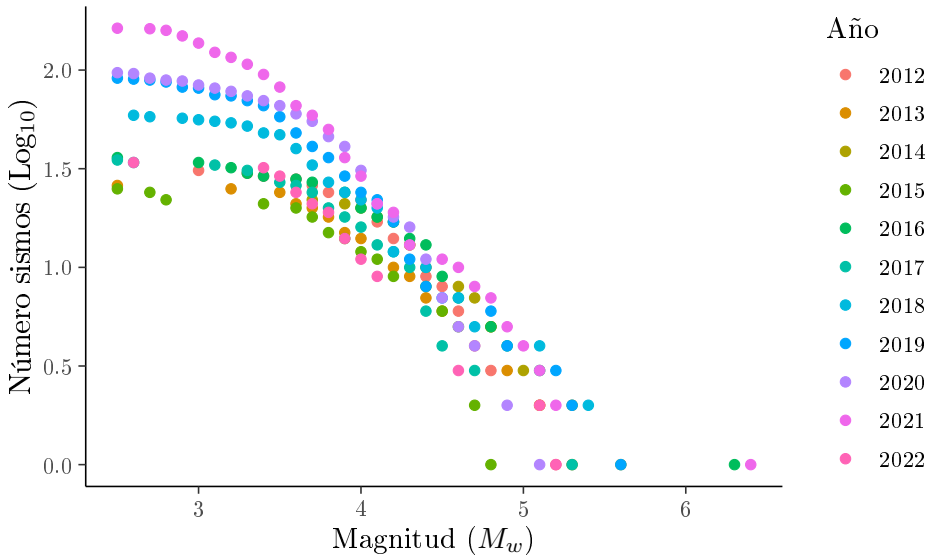

Figura 3.6: El apartamiento de la tendencia de la ley de Gutenberg-Richter en los sismos percibidos por la población muestra que la dificultad para que esto se produzca se incrementa con la baja de la magnitud.

Temporalidad de los percibidos Se puede hipotetizar que en los horarios de sueño nocturno sea menor la proporción de sismos percibidos. Segmentando los registros con la variable Segundos del día (ver sección 3.2.2) en intervalos de una hora se calculó la proporción entre casos percibidos o no. Contrariando la hipótesis, en la figura 3.7 no parece posible identificar una banda horaria, e.g. noche, en la que la proporción sea menor.

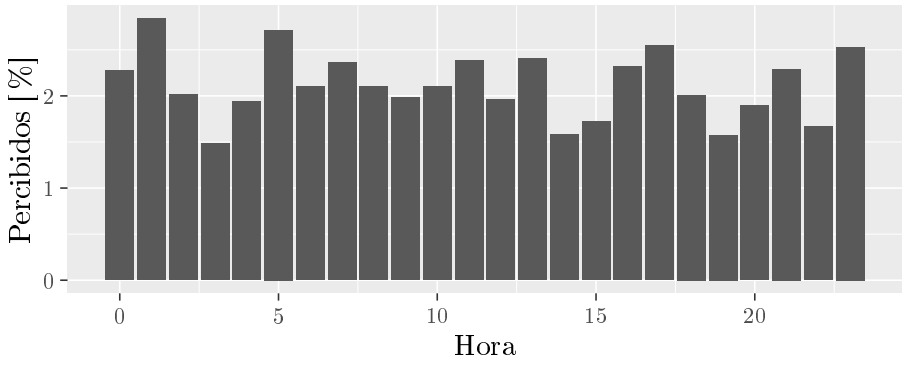

Figura 3.7: Proporción de sismos percibidos por la población en función de la hora del día.

Otra hipótesis que puede formularse es que en los meses de verano, enero y fe-

brero, sea menor la proporción de sismos percibidos asumiendo que la población tiene más horas de actividades al aire libre, y es en interiores de construcciones altas donde más fácilmente se perciben los sismos [11]. Sin embargo la figura 3.8 muestra que en el primer trimestre se halla la mayor proporción de sismos percibidos.

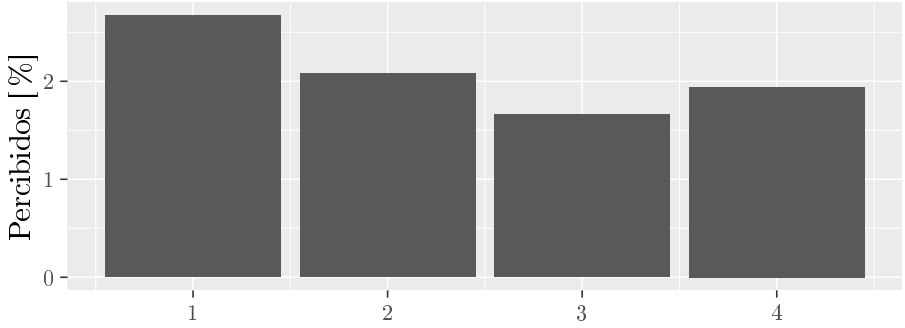

Figura 3.8: Proporción de sismos percibidos por trimestre.

Con el fin de explorar si los apartamientos trimestrales en las proporciones son signficativos se realizar un *ensayo binomial* de cada proporción trimestral contra la media. Este se alimenta con la tabla de contingencia de la figura 3.1 que muestra una cantidad de muestras que se consideran suficientes para proceder con el ensayo.

Trimestre	1.ero	2.do	3.ero	4.to
Percibido	215	154	112	138
No percibido	8046	7399	6745	7108

Cuadro 3.1: Tabla de contingencia de sismos percibidos por trimestre.

Para el primer y tercer cuatrimestre se obtuvieron valores p de 0,000452 y 0,016623 respectivamente, lo que indica que las proporciones de sismos percibidos en estos trimestres son significativamente distintas de la media. No se tiene una explicación para esta observación, que debe aclararse, debe tomarse con precaución pues se realiza sobre un número muy limitado de proporciones.

Dado que se son apreciables variaciones internanuales en el número total de terremotos, como lo evidencia la figura 3.5, tal vez debieran tomarse los trimestres de cada año en forma independientes años y analizar su medía y dispersión. Pero es posible que el número de muestras en cada trimestre sea insuficiente para obtener resultados representativos.

3.5. Ingeniería de características

3.5.1. Descarte de terremotos de poca profundidad

Hay sismos cuyo origen no son terremotos sino desplazamientos superficiales de tierra, explosiones para la minera o el fracturado hidráulico para la extracción

de hidrocarburos. Se busca omitir tales orígenes en los datos informados.

Siendo que la variable se informa como enteros de kilómetros, estos representarían los hipocentros hasta una profundidad de 500 m, compatibles con estas actividades artificiales. La omisión de estas fuentes a baja profundidad es una práctica usual en el análisis de datos orientados a sismos originados en terremotos [26]. El pequeño número que estos representan el conjunto de datos se aprecia en el histograma según profundidad en la figura 3.9. Son filtrados con la instrucción sismos_SJ <—sismos_SJ[Profundidad > 0].

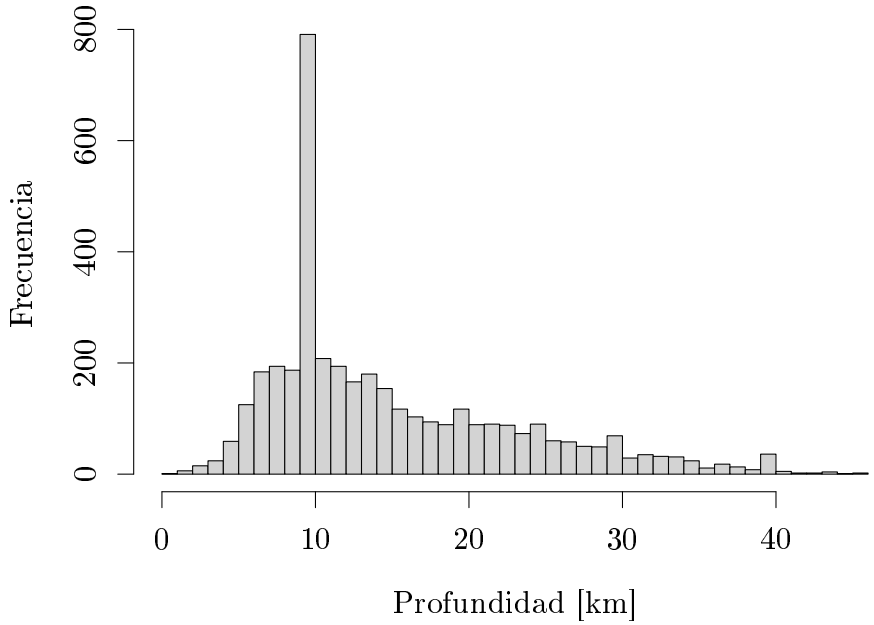

Figura 3.9: Los sismos con origen a profundidad de 0 km son pocos en el conjunto de datos de terremotos con hipocentros poco profundos, <50 km.

Este recorte tiene un efecto cuasi-irrelevante en la distribución de la profundidad de los terremotos en el conjunto de datos para San Juan. La figura 3.9 mostró que los hay con $h < 50\,\text{km}$, pero los hay hasta con $h = 750\,\text{km}$ siendo los de $h > 50\,\text{km} \approx 79\,\%$ del total como evidencia el histograma que reproduce la figura 3.10.

3.5.2. Linealización de la magnitud

/labelsec:linealización Todas las escalas de magnitud buscan dar cuenta de la energía liberada en el terremoto. La escala utilizada en los datos del INPRES es la de magnitud de momento (ver sección 3.1) definida como [1, ec. 4.23]

$$M_w = \frac{2}{3}\log_{10}(M_0) - 6{,}0 = \frac{2}{3}\log_{10}(\mu A u) - 6{,}0, \tag{3.1}$$

donde M_0 es el llamado *momento sísmico* a su vez función del módulo de cizalladura, μ, el área de falla involucrada, A, y su desplazamiento promedio, u, todas características del fenómeno en profundidad [1, sección 4.2.4].

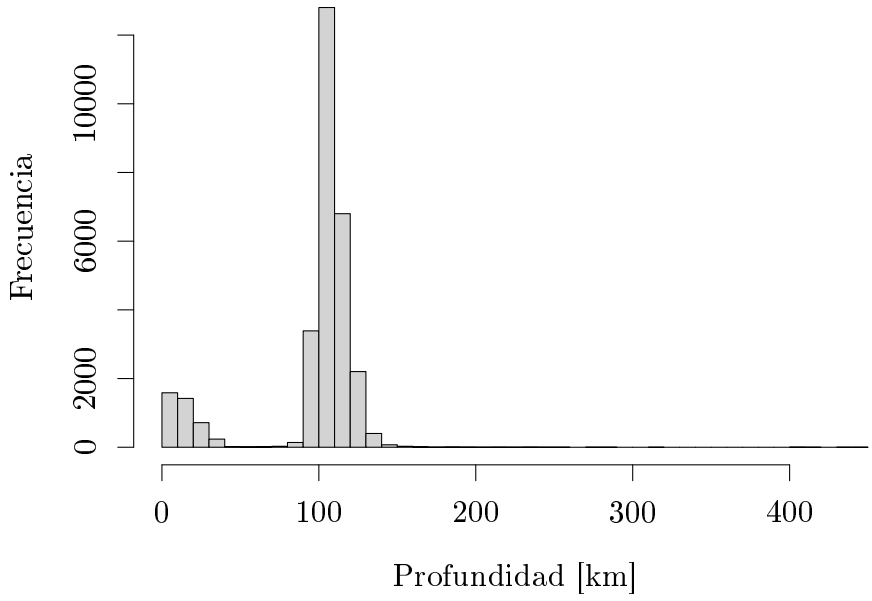

Figura 3.10: Los sismos con origen a mayor profundidad de 50 km son mayoritarios en el conjunto de datos.

Como se discutió en la sección 3.1.1 M_w fue precedida por otras escalas de magnitud ligadas a la amplitud de las ondas sísmicas percibidas en la superficie, como la de onda de cuerpo, m_b, y la de onda superficial, M_s, derivadas de la original de Richter de 1935. Puesto que este trabajo busca predecir la percepción por personas en la superficie, interesa relacionar los datos de M_w con estas escalas que tienen una forma genérica

$$M = \log_{10}\left(\frac{A}{T}\right) + q(\Delta, h) + a, \tag{3.2}$$

donde M es la magnitud, y aquí A es la amplitud de las ondas sismicas detectadas, T su período de oscilación, q es función de la profundidad del hipocentro, h, Δ el ángulo entre éste y el sismógrafo y la vertical, y a es una constante de ajuste [1, ecuación 4.13].

Dado que se carece del dato del punto de detección del sismo, no puede determinarse Δ y aunque se eligiera la escala que corresponde por h [1] no podría determinarse $q(\Delta, h)$. Frente a esto las diferencias entre las relaciones empíricas entre M_w con M_s o m_b[27], son irrelevantes a los fines de este trabajo. Se asume entonces por válida una fuerte aproximación

$$M_w \approx M_s \approx m_b. \tag{3.3}$$

Para las ondas de tipo S y P, las que contribuyen a los sismos percibidos en superficie (ver sección 1.1) se puede hacer uso de la escala propuesta origi-

[1] Si $h < 50$ km la mayor parte del aporte a la sismicidad la hacen ondas de propagación superficial por lo que se utiliza la *magnitud de onda superficial*, M_s, con distintos coeficientes de ajuste en la ecuación 3.2, que la *magnitud de onda de cuerpo*, m_b, que se usa para mayores profundidades [1, sección 4.2.3].

nalemente por Guntenberg en 1945 con la función de calibración para $q(\Delta, h)$ propuesta por *Gutemberg-Richter* en 1956 [1, ecuación 4.18] en la que se se omite de la expresión 3.2 la constante de ajuste a y para definir la razón $\frac{A}{T}$ se toma la mayor registrada por los sismógrafos,

$$m_b = \log_{10}\left(\frac{A}{T}\right)_{\text{máx}} + q(\Delta, h). \tag{3.4}$$

Para $q(\Delta, h)$ se usan valoren tabulados, e.g. para ondas P en $\Delta = 10$ a 110° corresponden $q(\Delta, h) \approx 6$ a 8 [28].

Para obtener un valor lineal A a partir de la escala 3.4, que es la utilizada en los datos de magnitud del INPRES, se consideró usar una valor medio del rango comentado en el párrafo anterior, $\overline{q}(\Delta, h)$, con lo que podría despejarse

$$\left(\frac{A}{T}\right)_{\text{máx}} = 10^{(m_b - \overline{q}(\Delta, h))}. \tag{3.5}$$

Pero la mínima magnitud, hallada con

$$\text{magnitud\_mínima} = \textbf{min}(\text{sismos\_SJ}[\,, \text{Magnitud}\,,\ ])$$

resultó ser menor que ese promedio, $\overline{q}(\Delta, h) = 7$, por lo que si se lo restara se obtendrían valores negativos para el valor a la izquierda de la ecuación y esto sería algo sin validez física (¡amplitudes negativas!). Se decidió utilizar entonces un ficticio $\tilde{q}(\Delta, h) = \text{magnitud\_mínima}$.

De querer despejar la amplitud de las ondas, A, debiera tenerse información sobre el período de las ondas, T, algo que no figura en el conjunto de datos. No se encontró otra alternativa que asumir que todos los fenómenos registrados tienen el mismo y en consecuencia asumir $T_{\text{constante}}$. Así realizando despejes a partir de la ecuación 3.4 y asumiendo tales condicionantes, se puede obtener un valor linealmente relacionado con amplitud registrada por un sismógrafo, $A_{\text{máx}}$,

$$\left(\frac{A_{\text{máx}}}{T_{\text{constante}}}\right) = 10^{(m_b - \tilde{q}(\Delta, h))}. \tag{3.6}$$

Con el comando

```
sismos_SJ[, Proxy\_amplitud := 10^(Magnitud- magnitud_mí
    nima)]
```

se generó una columna para este valor denominada Proxy amplitud, que por lo comentado anteriorme tiene por valor mínimo 1.

3.5.3. Percecpión sísmica, ¿de amplitud o energía?

Es posible también que que en el límite entre percpeción o no (I o II de la escala MMI [11]) pueda explicarse mejor en función de la energía de la onda sísmica que de la amplitud, por lo que podría esperarse una mejor correlación con el cuadrado de esta última variable derivada. Para ensayarlo se generó con

```
sismos_SJ[, `Proxy energía` := `Proxy amplitud`^2]
```

la nueva variable Proxy energía,

3.6. Variables a correlacionar con la percepción

De las variables originales, aquellas referidas al momento de detección del sismo, Fecha y Hora, y sus derivadas, Año, Día del año y Segundo del día, no se buscará explorar otra relación con la percepción que la presentado en el análisis exploratorio de datos, sección 3.4.

Restan, por un lado, las variables sobre la localización del terremoto, Latitud, Longitud y Profundidad, independientes entre sí. Por el otro, la única referida a la física del mismo, la Magnitud, y sus derivadas Proxy amplitud y Proxy energía, que se hipotetiza puede mostrar mejor relación con la percepción si ésta depende de la amplitud o energía de la oscilación sísmica. Como estas tres variables no tienen una relación lineal entre sí, se espera que puedan aportar información independiente al modelo de clasificación.

Como una medida preliminar de la valía de cada variable para los fines de clasificación se calcularon sus covarianzas con la variable de percepción. comando

```
cor(sismos_SJ[, .('Latitud', 'Longitud', 'Magnitud', '
    Profundidad', 'Proxy amplitud', 'Proxy energia',
    Percibido)])
```

que generó los valores que muestra el cuadro 3.2.

Magnitud	Proxy amplitud	Proxy energía	Longitud	Latitud	Profundidad
0.40571008	0.16537190	0.06037060	0.05548810	-0.01035141	-0.14235874

Cuadro 3.2: Covarianzas con la variable Percibido de las otras variables con las que buscará predecirse.

Respecto a la ubicación del terremoto se muestra el esperable resultado que cuanto más profundo menos percibido es. Menor impacto tiene donde está el epicentro. La latitud es casi irrelevante, siendo más relevante que cuanto al oeste se detecta un terremoto, esto es, con una mayor longitud, sean más percibidos. Cuanto más al oeste, se está más adentrado en la cordillera de los Andes, donde menos población se ubica, así que el incremento de la percepción no se debiera a una mayor proximidad entre terremoto y quién lo reporta. Lo que también cabe esperarse es que allí los terremotos sean de mayor magnitud, por lo que la explicación de la mayor percepción con mayor longitud tendría esta variable como mediadora en la relación. Sin embargo la correlación entre longitud y magnitud, calculada por

```
cor(sismos_SJ[, .('Longitud','Magnitud')], use = "
    complete.obs")
```

de 0,008 050 125, resulta ser despreciable, lo que lleva a abandonar esta explicación.

La magnitud es la variable que mayor correlación tiene con la percepción de los sismos. En esto aventaja apreciablemente a las variables sintetizadas a partir de ella, las aproximaciones a la amplitud y energía de las ondas sísmicas. Puesto que las relaciones de esta últimas con la primera no son lineales, no están perfectamente correlacionadas entre sí, lo que mostró la ejecución de

```
cor(sismos_SJ[, .('Magnitud', 'Proxy amplitud', 'Proxy
    energía')], use = "complete.obs")
```

arrojando que esta es 0,232 833 5 y 0,090 300 76 respectivamente. Por lo que estas variables no son redundantes entre sí y se espera que aporten información independiente al modelo de clasificación. Se ensayará retira de entre las variables utiilizadas estas variables derivadas para generar distintos modelos y evaluar el impacto de su inclusión.

3.7. Preprocesamiento

3.7.1. Escalamiento

Previo a la partición (splitting) se realiza un escalado uniforme sobre todo el conjunto de datos (scaling) de las variables numéricas, excluyendo las de fecha y hora con las que no se trabajará de aquí en adelante. Con esto variables de entrada a los modelos tendrán una media de cero y una desviación estándar de uno, lo que tiene consecuencias para interpretabilidad del modelo de regresión:

- los coeficientes serán comparables entre sí, no dependiendo de la escala de los datos,
- el intercepto del modelo será la predicción esperada para el caso en que los factores contemplados sean nulos, es decir, no tengan efecto,
- y finalmente si se hace uso de regularización se evita que los coeficientes de las variables de mayor escala tengan un peso desproporcionado en la función de pérdida [29, sección 3.4.1].

Para esto se hace uso de **scale**, función del conjunto base de R, para generar sismos_SJ_escalado con el comando

```
sismos_SJ_escalado <- scale(sismos_SJ[, .(Latitud,
    Longitud, Profundidad, Magnitud, 'Proxy amplitud', '
    Proxy energía')])
```

3.7.2. Partición con estratificación

Dado el fuerte desequilibrio de la clase Percibido comentado en la sección 3.4, ante una división de los datos en subconjuntos entrenamiento y prueba estocástica está el riesgo de que el subconjunto de prueba quede con muy pocos casos positivos y no sea representativo de la distribución de la clase en el conjunto de datos. Para evitar esto se realiza una división estratificada, esto es, manteniendo en los subconjuntos de entrenamiento (train) y ensayo (test) una proporción similar a la original de los casos de `Percibido`. Las evaluaciones sobre la calidad de los modelos de clasificación generados se realizarán sobre un subconjuntos de ensayo con el 20 % de los datos de la provincia de San Juan, el resto se utilizará para el entrenamiento. aplicando la función CreateDataPartion de la biblioteca caret en el comando

```
set.seed(123)
```

```
train_index <- caret::createDataPartition(sismos_SJ_
    escalado[ , Percibido], p = 0.8, list = FALSE)
entrenamiento_SJ <- sismos_SJ_escalado[train_index]
ensayo_SJ <- sismos_SJ_escalado[-train_index]
```

La función mencionada indicó los índices de los datos a incluir en el conjunto de entrenamiento, entrenamiento_SJ. Los datos restante se destinal al conjunto de ensayo, ensayo_SJ, este último con un número aún adecuado para su función de 5982 registros.

3.7.3. Desequilibrio en clase de clasificación

Para contrarrestar el desequilibrio en el conjunto de entreaniento sobre la clase de clasificación se utilizará la técnica de sobremuestreo de la clase minoritaria que genera nuevos casos sintéticos de la clase minoritaria a partir de los existentes. Para esto se utiliza la función ovun.**sample** de la biblioteca ROSE que genera un conjunto de datos de entrenamiento con un número de casos de la clase minoritaria igual al de la clase mayoritaria. Se generó así un nuevo conjunto de datos de entrenamiento entrenamiento_SJ_balanceado con 23 934 registros.

3.8. Métricas de evaluación de los modelos

Independientemente del modelo de clasificación que se utilice, la evaluación de su desempeño se realiza a partir de la comparación de las predicciones del modelo con los valores reales de la variable objetivo. Se reservará un subconjunto de los datos para evaluar el desempeño del modelo, el conjunto de prueba y generar una matriz de confusión.

Matriz de confusión La matriz de confusión es una tabla que muestra el número de verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos del modelo. A partir de esta matriz se pueden calcular como razones entre verdaderos y falso positivos y negativos la exactitud, precisión, sensibilidad, especificidad y el valor F1.

Exactitud (accuracy) La exactitud es la proporción de predicciones correctas sobre el total de casos. Se calcula como

$$\text{Exactitud} = \frac{\text{Verdaderos positivos} + \text{Verdaderos negativos}}{\text{Verd. pos.} + \text{Falsos pos.} + \text{Verd. neg} + \text{Falsos neg}}. \tag{3.7}$$

Precisión La precisión es la proporción de predicciones correctas sobre el total de predicciones realizadas. Se calcula como

$$\text{Precisión} = \frac{\text{Verdaderos positivos}}{\text{Verdaderos positivos} + \text{Falsos positivos}}. \tag{3.8}$$

En este trabajo cuantifica cuantos de las predicciones de percepción de sismos, efectivamente figuraban como tales en el conjunto de datos de ensayo. Esencialmente es una medida de que cuanto se equivoca la predicción en términos de reportar un evento que se percibirá cuando el publico no lo hizo según lo que figura en el conjunto de datos de ensayo.

Sensibilidad o exhaustividad (recall) La sensibilidad, o exhaustividad, es la proporción de verdaderos positivos sobre el total de casos positivos. Se calcula como

$$\text{Sensibilidad} = \frac{\text{Verdaderos positivos}}{\text{Verdaderos positivos} + \text{Falsos negativos}}. \tag{3.9}$$

En este trabajo cuantifica cuantos de los terremotos que tendrián que haber producido una predicción de percepción por parte de la población fueron efectivamente reportados como tales por el modelo al ensayarle en el conjunto de ensayo. Es esencialmente una medida de cuantos reportes falla en realizar.

Especificidad La especificidad es la proporción de verdaderos negativos sobre el total de casos negativos. Se calcula como

$$\text{Especificidad} = \frac{\text{Verdaderos negativos}}{\text{Verdaderos negativos} + \text{Falsos positivos}}. \tag{3.10}$$

F1-score El *F1-score* es la media armónica de la precisión y la sensibilidad (recall). Se calcula como

$$\text{F1-score} = 2 \times \frac{\text{Precisión} \times \text{Recall}}{\text{Precisión} + \text{Recall}}. \tag{3.11}$$

Capítulo 4

Resultados y discusión

4.1. Predictor por regresión logística múltiple

Como se indicó en la sección 2.2.1 una de las metodologías de predicción fue el de emplear modelos de predición logísticas. De acuerdo a cuales de las variables disponibles y que interecciones entre ellas se incluyen se obtuvieron distintas modelos de clasificación.

4.1.1. Todas las variables

El primer modelo ensayado tiene todas las variables como predictores sin interacción entre ellas,

$$\begin{aligned}\log\left(\frac{p}{1-p}\right) =& \beta_0 + \beta_1 \,\text{Magnitud} + \beta_2 \,\text{Proxy amplitud} \\ &+ \beta_3 \,\text{Latitud} + \beta_4 \,\text{Longitud} + \beta_5 \,\text{Profundidad},\end{aligned} \tag{4.1}$$

Este modelo se establece gracias a la función glm2 (ver sección 2.2.1) con el comando

```
logit_model <- glm2::glm2(
  Percibido ~ Magnitud + `Proxy_amplitud` + `Proxy_energía` + Longitud + Latitud + Profundidad,
   data = entrenamiento_SJ_balanceado,
  family = "binomial"
  )
```

donde la opción **family** = "binomial" indica que se ajustará un modelo de regresión logística de clasificación binaria sobre la variable Percibido que se muestra en función de las demás en el conjunto de dato indicado con **data** apuntando al conjunto de entrenamiento balanceado según se explicó en la sección 3.7.3. Los resultados del ajuste para los coeficientes los muestra una invocación de la función **summary** que muestra el cuadro 4.1.

Excepto los coeficientes para Proxy_amplitud y Proxy_energía, todos son significativos ya que sus probabilidades de que no tengan esos valores y se cumpla la hipótesis nula $Pr(>|z|)$ para el estadístico $z = \beta/\sigma_\beta$, se indican como

Variable	Coeficiente	Error estándar	z	$Pr(>\|z\|)$
Intercepto	-2.232632	0.033904	-65.851	<2e-16
Magnitud	1.575057	0.020722	76.010	<2e-16
Proxy amplitud	0.003223	0.008529	0.378	0.706
Proxy energía	-0.003303	0.008250	-0.400	0.689
Longitud	0.183377	0.021651	8.470	<2e-16
Latitud	-0.167088	0.021658	-7.715	1.21e-14
Profundidad	-0.987883	0.017962	-54.997	<2e-16

Cuadro 4.1: Coeficientes del modelo de regresión logística múltiple.

inferiores al umbral usual de 0,05. Para el modelo de predección a utilizar se optaría por obviar Proxy_amplitud y Proxy_energía. Antes de hacerlo y dejar definitivamente de lado las mismas se explorará si se puede mejorar el modelo con la inclusión de interacciones entre las variables. Estas debieran poder corregir sesgos introducidos al modelo por incluir entre sus variables aquellas supuestamente independientes pero que muestran correlación.

4.1.2. Interacción entre variables

Entre las 6 variables numéricas pueden formase pares de interacción, lo que lleva a preguntarse cuales incluir en un modelo. Puesto que se busca que incluirle aporte a mejorar la predección por corrección del sesgo que aporta su inclusión sin dar cuenta de la covarianza entre ellas se buscó determinar cuales pares presentan un mayor valor de la misma.

Las covarianzas entre las variables deben volver a calcularse pues estas no son las mismas en el subconjunto de entrenamiento que en el total de los datos. Aunque la modificación que causa la partición en si no es muy importante (ver sección 3.7.2), si es más significativa la causada por el posterior balanceo de la clase de clasificación (ver sección 3.7.3). Se generó una función basada en la ya mencionada **cor** para determinar los pares de variables con mayor valor absoluto de covarianza. Las que superan un corte arbitario de 0,03 se muestran en el cuadro 4.2, junto con el respectivo valor.

Variables	Covarianza
Proxy_energía - Proxy_amplitud	0.7432113838
Proxy_amplitud - Magnitud	0.3088276448
Profundidad - Longitud	0.2930435805
Proxy_energía - Magnitud	0.1609778094
Magnitud - Longitud	0.0427277428
Profundidad - Latitud	0.0382461548

Cuadro 4.2: Covarianza entre pares de variables numéricas.

No es de extrañar que figuren las covarianzan entre `Magnitud` y sus derivadas entre las de mayor covarianza. Pero también figuran unas no esperadas entre `Profundidad` y ubicación del epicentro, además de las más esperadasa entre `Magnitud` y `Longitud`, por las razones discutidas en la sección 3.6.

Variable	Coeficiente	Error estándar	z	$Pr(>\|z\|)$
Intercepto	-2.265e+00	3.527e-02	-64.221	<2e-16
Magnitud	1.598e+00	2.174e-02	73.503	<2e-16
Proxy amplitud	6.392e-04	1.203e-02	0.053	0.957625
Proxy energía	-2.872e-03	1.188e-02	-0.242	0.808996
Longitud	3.211e-01	3.213e-02	9.994	<2e-16
Latitud	-1.291e-01	2.301e-02	-5.612	2.00e-08
Profundidad	-1.004e+00	1.923e-02	-52.209	<2e-16
Proxy amplitud:Proxy energía	-4.126e-04	1.023e-03	-0.403	0.686825
Magnitud:Proxy amplitud	1.327e-03	5.903e-03	0.225	0.822178
Longitud:Profundidad	6.306e-02	1.636e-02	3.855	0.000116
Magnitud:Proxy energía	-1.257e-05	6.013e-03	-0.002	0.998332
Magnitud:Longitud	-1.135e-01	1.924e-02	-5.899	3.66e-09
Latitud:Profundidad	7.633e-02	1.326e-02	5.756	8.63e-09

Cuadro 4.3: Coeficientes del modelo de regresión logística múltiple incorporando la interacción entre variables.

Un modelo incorporando todas estas interacciones de pares a aquel propuesto en la sección 4.1.1 se generó con el comando

```
logit_model_interaccion <- glm2::glm2(
        Percibido ~ Magnitud * `Proxy_amplitud` +
            Longitud * Profundidad,
         data = entrenamiento_SJ_balanceado,
        family = "binomial"
        )
```

Nuevamente los p-values indican que las variables derivadas de `Magnitud` y las interacciones con las mismas no aportan significativamente al modelo, como muestra el cuadro 4.3, pero todas las demás interacciones restultan aportar significativamente al modelo.

Al parecer entonces se tendría un mejor modelo que ajusta mejor a los datos incorporando estas interacciones. Para afirmar que efectivamente es mejor se necesita utilizar alguna métrica de la bondad del ajuste.

4.1.3. Medida de la bondad del ajuste

La desviación (del inglés deviance) es una generalización de la suma de cuadrados de los residuos en el ajuste ordinario de caudrados mínimos a casos donde el mismo se realiza por máxima verosimilitud [29, sección 7.2]. Con este valor puede calcularse una especie de *pseudo valores de R cuadrado*, como medida de que tan bien el modelo explica la varianza en la variable dependiente. Para el modelo presentado en la sección 4.1.1, que denominaremos "sin interacción", se calcula con el comando

```
pseudo_r2 <- 1 - logit_model$deviance / logit_
    model$null.deviance
```

donde logit _**model$null.deviance** sería la del modelo con todos su coeficiente β_i nulos.

Asumiendo que cuanto menor es este pseudo R cuadrado se logró un mejor ajuste debe calcularse para las otras variantes de modelos de regresión logística múltiple hasta ahora propuestas. La primera es la variante del modelo presentado en la de la sección 4.1.1 "limpio" de las variables derivadas de `Magnitud`, que denominaremos "sin interacción, sin derivadas", generado con el comando

```
logit_model_limpio <- glm2::glm2(
  Percibido ~ Magnitud + Longitud + Latitud + Profundidad
    ,
   data = entrenamiento_SJ_balanceado,
  family = "binomial"
  )
```

El tercer modelo es el presentado en la sección 4.1.2 con las interacciones entre variables, que denominaremos "con interacción". Y el cuarto es el de esta misma sección pero sin las variables derivadas de `Magnitud`, que denominaremos "con interacción, sin derivadas", que se generó con el comando

```
        logit_pairwise_limpio <- glm2::glm2(
  Percibido ~
    Magnitud +
    Longitud +
    Latitud +
    Profundidad +
    Profundidad * Longitud +
    Magnitud * Longitud +
    Profundidad * Latitud
    ,
   data = entrenamiento_SJ_balanceado,
  family = "binomial"
  )
```

A los fines de comparar el pseudo R cuadrado para cada modelo se los resume en el cuadro 4.4.

Modelo	Pseudo R cuadrado
Sin interacción	0.5930901
Sin interacción, sin derivadas	0.5928425
Con interacción	0.5938916
Con interacción, sin derivadas	0.5935757

Cuadro 4.4: Pseudo R cuadrado para los distintos modelos de regresión logística múltiple.

Si bien el modelo "con interacción" es el que tiene el mayor pseudo R cuadrado, la diferencia con su variente "sin derivadas" que contiene coeficientes que no mostraron ser significativas (ver sección 4.1.2) llevan a optar entre ambos por el menos complejo.

Con lo anterior se tomó una decisión sobre cual modelo de regresión logística múltiple utilizar para predecir la percepción de los sismos en la provincia de San Juan, pero no se da una respuesta a si ajusta bien a los datos. Un ensayo

específico para la bondad de ajuste de modelos de predicción logísticas es la prueba de Hosmer-Lemeshow. Esta prueba compara la frecuencia observada de eventos en subgrupos de predicción con la frecuencia esperada. Se calculó con el comando

```
factor_forHL <- as.numeric(entrenamiento_SJ_
    balanceado$Percibido) - 1
hl_test <- ResourceSelection::hoslem.test(factor_
    forHL, fitted(logit_pairwise_limpio))
print(hl_test)
```

arrojando un resultado bastante negativo, un valor p $<2{,}2 \times 10^{-16}$.

4.1.4. Evaluación de la predicción del modelo logístico

El desempeño de los modelos se ensaya comparando la predicción con el conjunto de ensayo `ensayo_SJ` generado en la partición estratificada explicada en la sección 3.7.2. Como se discutió en la sección 2.2.1, los modelos de regresión logística múltiple generan una probabilidad de pertenencia a una clase, p. La predicción de la clase se realiza asignando a la instancia la clase con mayor probabilidad, es decir, si $p > 0{,}5$ se asigna a la clase positiva, y si $p \leq 0{,}5$ se asigna a la clase negativa. Debe entonces definirse este umbral, o punto de corte, entre las clases para clasificar las instancias. Para orientar tal decisión se armó una función basada a su vez en la función confusionMatrix de la biblioteca caret para calcular las métricas de evaluación de la predicción en todo un rango de puntos de corte. Las métricas calculadas son las presentadas en la sección 3.8: exactitud, especificidad, sensibilidad, especificidad y la media armónica de estás dos últimas, la métrica F_1.

Para el modelo favorecido por el análisis de bondad de ajuste (sección 4.1.3), el de regresión logística múltiple con interacción entre variables pero sin contemplar las derivadas de `Magnitud`, la figura 4.1 muestra las métricas de predicción en un rango 0,05 a 0,95 para el punto de corte.

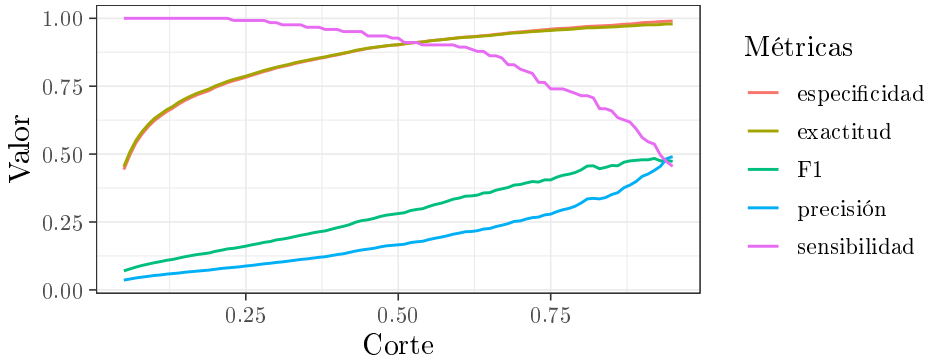

Figura 4.1: Métricas de evaluación para definir manualmente el punto de corte entre clases para el modelo de regresión logística múltiple con interacción entre variables, pero excluyendo las derivadas de la magnitud del terremoto.

No es evidente un punto óptimo, la presición es muy baja para todos los puntos de corte pero ascendente con el punto de corte, pero la sensibilidad

empieza a decaer muy rápido a partir de $\approx 0{,}6$. Se decidió entonces tomar el punto de corte en 0,6, en pos de mantener una alta sensibilidad, es decir que al predecir una predicción por la población hay una alta probabilidad de que esto sea así a costa de que no se generen reportes y luego la población perciban el sismo, lo que representa una baja precisión. Para este punto de corte de 0,6 los valores de las métricas se resumen en el cuadro 4.5.

exactitud	sensibilidad	especificidad	precisión	F_1
0,928 117 7	0,894 308 9	0,928 827 4	0,208 728 7	0,338 461 5

Cuadro 4.5: Métricas de predicción en el conjunto de ensayo del modelo de regresión logística múltiple con interacciones entre variablese excluyendo las derivadas de la magnitud del terremoto. Punto de corte en 0,6.

Se repite el análisis de las métricas de evaluación para el modelo con interacción entre variables y que no descarta las variables derivadas de la magnitud. La figura que muestra las métricas de la predicción en función del punto de corte praa este caso, figura 4.2, podría decirse que es una copia de la que realizada para el del modelo sin las derivadas de la magnitud, figura 4.1. Las diferencias son mínimas y esto es un testimonio del poco peso de los coeficientes correspondientes a las variables derivadas de la magnitud, a las relaciones con las otras, en este modelo.

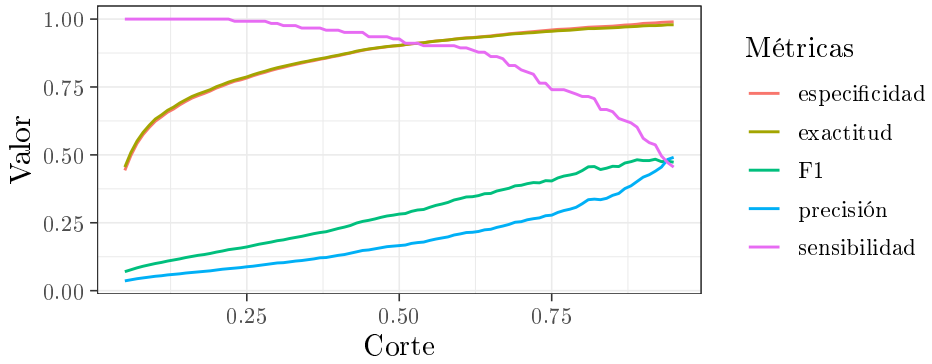

Figura 4.2: Métricas de evaluación para definir manualmente el punto de corte entre clases para el modelo de regresión logística múltiple con interacción entre variables, incluyendo las derivadas de la magnitud del terremoto. Esta figura es a primera vista una copia de la figura 4.2, pero los valores de las métricas son ligeramente distintos.

Por lo visto en la figura 4.2 se decidió tomar el mismo punto de corte en 0,6 para este modelo. Los valores de las métricas para este punto de corte también son muy similares a las del otro modelo, como muestra el cuadro 4.6. El desempeño de predicción es marginalmente mejor en este modelo en términos de sensibilidad y precisión, pero éste último sigue siendo muy bajo.

exactitud	sensibilidad	especificidad	precisión	F_1
0,928 284 9	0,894 308 9	0,928 998 1	0,209 125 5	0,338 983 1

Cuadro 4.6: Métricas de predicción en el conjunto de ensayo del modelo de regresión logística múltiple con interacciones entre variablese incluyendo las derivadas de la magnitud del terremoto. Punto de corte en 0,6.

4.2. Predictor por XGBoost

Como alternativa a los modelos de regresión logística múltiple, se ajustaron modelos de clasificación con el algoritmo XGBoost. Se entrenó en primer lugar un modelo con todas las variables disponibles, sin interacción entre ellas, con el comando

```
num_feat = c("Latitud", "Longitud", "Magnitud", "
    Profundidad", "Proxy_amplitud", "Proxy_energía")
entrenamiento_SJ_balanceado_matrix <- as.matrix(
    entrenamiento_SJ_balanceado[ ,.num_feat])
label_numeric <- as.numeric(entrenamiento_SJ_balanceado[,
    Percibido]) - 1
bstSparse <- xgboost(
  data = entrenamiento_SJ_balanceado_matrix,
  label = label_numeric,
  max.depth = 5, eta = 1,
  nthread = 2, nrounds = 5,
  objective = "binary:logistic",
  verbose = 2
  )
```

donde se indica que se ajustará un modelo de clasificación binaria con la función de pérdida logística, con un máximo de 5 niveles en los árboles y 5 iteraciones. Estos parámetros fueron elegidos a través de sucesivas pruebas manuales un busqueda de maximizar la métrica F1 (ver sección 3.8) al contrastar la predicción, con un umbral de 0,5 contra el conjunto de ensayo `ensayo_SJ` generado en la partición estratificada explicada en la sección 3.7.2 con el comando

```
ensayo_SJ_matrix <- as.matrix(ensayo_SJ[ ,.num_feat])
prediction_xgboost <- predict(bstSparse, ensayo_SJ_matrix
    )
umbral_p = 0.5
prediction_xgboost_binario <- ifelse(prediction_xgboost >
    umbral_p, 1, 0)
```

Para este modelo con todas las variables, incluso los anteriormente deprecadas derivadas de la `Magnitud`, las métricas de evaluación de la predicción del modelo con el punto de corte en 0,5 que se resumen en el cuadro 4.7.

El siguiente ensayo fue con el modelo de regresión logística múltiple sin las variables derivadas de la `Magnitud`, que se entrena con un comando en todo similar al anterior a excepcion de los cambios

```
num_feat = c("Latitud", "Longitud", "Magnitud", "
    Profundidad")
```

exactitud	sensibilidad	especificidad	precisión	F1
0,980 441 3	0,113 821 1	0,998 634 6	0,636 363 6	0,193 103 4

Cuadro 4.7: Métricas de evaluación para el mejor modelo de XGBoost hallado para todas las variables y punto de corte en 0,5.

```
...
  max.depth = 3, eta = 1,
```

donde el número de niveles en los árboles se redujo a 3 y se mantuvo el resto de los parámetros, lográndose un mejor desempeño en predecir en el conjunto de ensayo, como se muestra en el cuadro 4.8.

exactitud	sensibilidad	especificidad	precisión	F_1
0,922 266 8	0,894 308 9	0,922 853 7	0,195 729 5	0,321 167 9

Cuadro 4.8: Métricas de evaluación para el mejor modelo de XGBoost hallado solo con las variables originales y punto de corte en 0,5.

Se aprecia de comprarar los cuadros 4.5 y 4.8 que el modelo de regresión logística múltiple sin las variables derivadas de la `Magnitud` sacrifica la precisión por una mayor sensibilidad, es decir que al predicir que la poblaciño percibirá un sismo hay una alta probabilidad de que esto sea así a costa de que muchos reporten no se efectuen y la población perciba sismas, lo que representa una baja precisión. Se considera que una mayor sensibilidad es más importante a los fines de no emitir una alerta incorrecta a la población de que percibirán un sismo, por lo que se opta por este último modelo de XGBoost para predecir la percepción de sismos en la provincia de San Juan.

4.3. Relevancia de los resultados

De los cuatros modelos evaluados para la predicción de la percepción de los sismos en la provincia de San Juan, cual se elija depende de los objetivos de la predicción.

Si se considera que una mayor sensibilidad es más importante a los fines de no emitir una alerta incorrecta a la población de que percibirán un sismo, lo que se busca es una alta sensibilidad, aún a costa de que no se produzcan algunos reportes y aún así la población perciba sismos, lo que representa una baja precisión. En ese caso tanto el modelo de regresión logística múltiple con o sin las variables derivadas de la `Magnitud`, así como el de XGBoost sin estas variables resultan los más adecuados. Como resume el cuadro 4.9, todos estos tienen la misma alta sensibilidad, aunque al precio de una pobre precisión.

Si por el contrario se busca generar una alerta de que un terremoto será percibido como un sismo por la población, aunque esta resuelte luego en una "falsa alarma", debe optarse por un modelo que tenga una alta precisión, aún a costa de una baja sensibilidad. Entre los modelos evaluados, el de XGBoost con todas las variables, incluyendo las derivadas de la `Magnitud`, es el que tiene la

Modelo	exactitud	sensibilidad	especificidad	precisión	F_1
LOrig	0,928 117 7	0,894 308 9	0,928 827 4	0,208 728 7	0,338 461 5
LDer	0,928 284 9	0,894 308 9	0,928 998 1	0,209 125 5	0,338 983 1
XOrig	0,922 266 8	0,894 308 9	0,922 853 7	0,195 729 5	0,321 167 9
XDer	0,980 441 3	0,113 821 1	0,998 634 6	0,636 363 6	0,193 103 4

Cuadro 4.9: Métricas de predicción de los modelos seleccionados. Los modelos selecionados fueron los de regresión logística múltiple contemplando interacciones entre pares de variables incluyendo las variables derivadas de la magnitud del terremoto (LDer) o sin estas (LOrig), y los de XGBoost, también con (XDer) o sin estas variables derivadas (XOrig).

mayor precisión, como se muestra en el cuadro 4.9, aunque sacrifica para esto la sensibilidad.

4.4. Limitaciones y posibles mejoras

El mayor inconveniente enfrentado es la carencia de algún dato sobre la ubicación del reporte de percepción del sismo, ya discutido en la sección **??**, como "el problema de la distancia". Sumada esta ubicación para los reportados por la población, junto con la de un sismógrafo indicado como el que responsable primario para determinar magnitud y profundidad en los casos de terremotos percibidos o no, podrían hacerse inferencias en base a la distancia. Incluso sería posible calcular una aproximación al ángulo entre sismógrafo y terremoto y así aplicar la relación correcta entre magnitud y amplitud de onda (ver sección **??**).

Algo que resta estudiar es la distribución de la proporción de percecpión en función de la estación en el año, pues lo hallado en la sección 3.4 contradijo la hipótesis allí planteada. Se puede trabajar en la segmentación de los terremotos en trimestres por todos los años como se comenta allí.

Capítulo 5

Conclusión

5.1. Resumen de los hallazgos principales

Se hallaron dos familias de modelos que predicen la percepción por parte de la población de sismos en la provincia de San Juan, Argentina. Se utilizaró para esto datos públicos de sismos y reportes de percepción de la población publicados por el INPRES. No se pudo hallar un modelo que conjugue precisión y sensibilidad por lo que de aplicarse uno de ellos hay que optar a cuál de las métricas se le dá primacía.

5.2. Conclusiones generales y su relación con los objetivos del trabajo

En relación a los objetivos planteados en la sección 1.2, debe admitirse que si bien pudieron producirse modelos de predección de la percepción de sismos, esto se logró solo para una provincia en particular por limitación de los datos geográficos disponibles. Pero la mayor limitación a los modelos de predicción hallados es que no se logró un modelo de aplicación a múltiples problemas, que conjugue precisión y sensibilidad, por lo que de aplicarse uno de los modelos desarrolados hay que optar a cuál de las métricas se le dá primacía.

5.3. Aplicaciones y relevancia de los resultados

Los modelos desarrolados que mostraron alta sensibilidad, pero baja precisión, pueden ser útiles si las autoridades quieren mostrar a la población de que los terremotos son monitoreados y que se ponen a los sistemas gubernamentales en alerta. Pueden emitirse alertas de que un terremoto será percibido como un sismo por la población, y habrá una baja probabilidad de que esto resulte en una "falsa alarma".

Por el contrario el modelo con alta precisión y baja sensibilidad, al cometer menos errores sobre si un terremoto debiera ser percibido como sismo por la población puede ser útil, no para emitir alerta, pero sí para contrastar con los reportes de percepción de la población y así determinar si hay un faltante de

ellos. Con esto se puede determinar si hay un problema en los sistemas que reciben reportes de percepción de sismos.

Bibliografía

[1] C. M. R. Fowler. *The Solid Earth: An Introduction to Global Geophysics*. first. Cambridge University Press, 29 de jun. de 1990. 490 págs. ISBN: 978-0-521-37025-7. DOI: 10.1017/CBO9780511819643. URL: https://archive.org/details/solidearthintrod0000fowl (visitado 16-06-2024).

[2] Lisa Wald. *The Science of Earthquakes*. United States Geological Survey. URL: https://www.usgs.gov/programs/earthquake-hazards/science-earthquakes (visitado 25-06-2024).

[3] *¿Qué es un sismo?* Sistema Nacional para la Gestión Integral del Riesgo. 12 de nov. de 2018. URL: https://www.argentina.gob.ar/sinagir/riesgos-frecuentes/sismos (visitado 16-06-2024).

[4] Saunders, J. K., Minson, S. E., Cochran, E. S., Bunn, J., Baltay, A. S., Kilb, D. y O'Rourke, C. «A Twist of PLUM: Low-Magnitude Earthquakes and Ground-Motion-Based Early Warning». En: 2021 Southern California Earthquake Center Annual Meeting, SCEC Contribution #11360. URL: https://www.scec.org/publication/11360 (visitado 17-06-2024).

[5] Sandra Vaiciulyte, David A. Novelo-Casanova, Allen L. Husker y Ana B. Garduño-González. «Population response to earthquakes and earthquake early warnings in Mexico». En: *International Journal of Disaster Risk Reduction* 72 (1 de abr. de 2022), pág. 102854. ISSN: 2212-4209. DOI: 10.1016/j.ijdrr.2022.102854. URL: https://www.sciencedirect.com/science/article/pii/S2212420922000735 (visitado 17-06-2024).

[6] *Instituto Nacional de Prevención Sísmica*. Argentina.gob.ar. 28 de oct. de 2022. URL: https://www.argentina.gob.ar/inpres (visitado 17-06-2024).

[7] *Encuesta de sismos*. Instituto Nacional de Prevención Sísmica. URL: http://contenidos.inpres.gob.ar/encuesta/encuesta_sismo.php (visitado 30-06-2024).

[8] *Felt Report - Tell Us!* United States Geological Survey. URL: https://earthquake.usgs.gov/earthquakes/eventpage/tellus (visitado 30-06-2024).

[9] David Jay, Vincent Quitoriano, Charles Bruce, Margaret Hopper y James W. «USGS "Did You Feel It?" Internet-based macroseismic intensity maps». En: *Annals of Geophysics* 54.6 (14 de ene. de 2012). ISSN: 2037416X. DOI: 10.4401/ag-5354. URL: http://www.annalsofgeophysics.eu/index.php/annals/article/view/5354 (visitado 27-06-2024).

[10] *DYFI Scientific Background*. United States Geological Survey. URL: https://earthquake.usgs.gov/data/dyfi/background.php.

[11] *Intensidad y Magnitud*. Instituto Nacional de Prevención Sísmica. 7 de nov. de 2022. URL: https://www.argentina.gob.ar/inpres/docentes-y-alumnos/intensidad-y-magnitud (visitado 01-07-2024).

[12] G. M. Atkinson y D. J. Wald. «"Did You Feel It?Ïntensity Data: A Surprisingly Good Measure of Earthquake Ground Motion». En: *Seismological Research Letters* 78.3 (1 de mayo de 2007), págs. 362-368. ISSN: 0895-0695. DOI: 10.1785/gssrl.78.3.362. URL: https://pubs.geoscienceworld.org/srl/article/78/3/362-368/143359 (visitado 30-06-2024).

[13] G. M. Atkinson, C. B. Worden y D. J. Wald. «Intensity Prediction Equations for North America». En: *Bulletin of the Seismological Society of America* 104.6 (1 de dic. de 2014), págs. 3084-3093. ISSN: 0037-1106. DOI: 10.1785/0120140178. URL: https://pubs.geoscienceworld.org/bssa/article/104/6/3084-3093/332154 (visitado 30-06-2024).

[14] Ian Marschner C. «glm2: Fitting Generalized Linear Models with Convergence Problems». En: *The R Journal* 3.2 (2011), pág. 12. ISSN: 2073-4859. DOI: 10.32614/RJ-2011-012. URL: https://journal.r-project.org/archive/2011/RJ-2011-012/index.html (visitado 08-07-2024).

[15] Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, Rory Mitchell, Ignacio Cano, Tianyi Zhou, Mu Li, Junyuan Xie, Min Lin, Yifeng Geng, Yutian Li, Jiaming Yuan y XGBoost contributors (base XGBoost implementation). *xgboost: Extreme Gradient Boosting*. Ver. 1.7.7.1. 25 de ene. de 2024. URL: https://cran.r-project.org/web/packages/xgboost/index.html (visitado 22-06-2024).

[16] *IC-datasets-docencia*. URL: https://daniellaparada.github.io/IC-datasets-docencia/ (visitado 12-06-2024).

[17] Daniela Parada. *IC-datasets-docencia - 4 Visualización*. URL: https://daniellaparada.github.io/IC-datasets-docencia/04_visualizacion.html (visitado 12-06-2024).

[18] *Buscador de sismos*. Instituto Nacional de Prevención Sísmica. URL: http://contenidos.inpres.gob.ar/buscar_sismo (visitado 15-06-2024).

[19] Daniela Parada. *sismos-arg*. URL: https://github.com/daniellaparada/IC-datasets-docencia/blob/main/fuente/04_visualizacion/sismos-arg.csv (visitado 30-06-2024).

[20] *ISO 8601-1:2019(en), Date and time — Representations for information interchange — Part 1: Basic rules*. International Organization for Standardization. 2019. URL: https://www.iso.org/obp/ui/en/#iso:std:iso:8601:-1:ed-1:v1:en (visitado 16-06-2024).

[21] *Cálculo de la Magnitud*. Instituto Nacional de Prevención Sísmica. 7 de nov. de 2022. URL: https://www.argentina.gob.ar/inpres/docentes-y-alumnos/calculo-de-la-magnitud (visitado 01-07-2024).

[22] *Sismos en el territorio argentino*. 5 de mayo de 2015. URL: https://www.argentina.gob.ar/sites/default/files/sismos.pdf (visitado 01-07-2024).

[23] *Moment magnitude, Richter scale - what are the different magnitude scales, and why are there so many?* United States Geological Survey. URL: https://www.usgs.gov/faqs/moment-magnitude-richter-scale-what-are-different-magnitude-scales-and-why-are-there-so-many (visitado 30-06-2024).

[24] *Introduction to data.table*. The Comprehensive R Archive Network. 27 de mar. de 2024. URL: https://cran.r-project.org/web/packages/data.table/vignettes/datatable-intro.html#1.%20Basics (visitado 18-06-2024).

[25] *Acerca de tu ubicación*. Instituto Nacional de Prevención Sísmica. URL: https://www.inpres.gob.ar/desktop/conoce.html (visitado 17-06-2024).

[26] Yi Hu, Wentao Wang, Lei Li y Fangjun Wang. «Applying Machine Learning to Earthquake Engineering: A Scientometric Analysis of World Research». En: *Buildings* 14.5 (mayo de 2024). Number: 5 Publisher: Multidisciplinary Digital Publishing Institute, pág. 1393. ISSN: 2075-5309. DOI: 10.3390/buildings14051393. URL: https://www.mdpi.com/2075-5309/14/5/1393 (visitado 17-06-2024).

[27] Thomas C. Hanks e Hiroo Kanamori. «A moment magnitude scale». En: *Journal of Geophysical Research: Solid Earth* 84 (B5 1979). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/JB084iB05p02348, págs. 2348-2350. ISSN: 2156-2202. DOI: 10.1029/JB084iB05p02348. URL: https://onlinelibrary.wiley.com/doi/abs/10.1029/JB084iB05p02348 (visitado 04-07-2024).

[28] Willian L. Ellsworth. «Earthquake Magnitude: THE RICHTER SCALE (ML)». En: *The San Andreas Fault System, California*. Ed. por Robert E. Wallace. Vol. Professional Paper 15151. P. United States Geological Survey (USGS), 1991, pág. 177. URL: https://web.archive.org/web/20160425121745/http://www.johnmartin.com/earthquakes/eqsafs/safs_693.htm (visitado 14-10-2008).

[29] Trevor Hastie, Harry Friedman y Jerome Friedman. *The Elements of Statistical Learning Data Mining, Inference, and Prediction*. second. Springer Series in Statistics 0172-7397. New York, NY, USA: Springer, 26 de ago. de 2009. 745 págs. ISBN: 978-0-387-84858-7. URL: https://doi.org/10.1007/978-0-387-84858-7.

Anexos (opcionales)

5.4. Código fuente utilizado en el análisis

Enlace al repositorio en GitHub que aloja el código fuente utilizado en el análisis de los datos.

5.5. Tablas y gráficos adicionales

Esto es un placeholder para una sección aún vacía.

5.6. Otros materiales relevantes

Esto es un placeholder para una sección aún vacía.