When is that song?

J. Antonio García Ramírez

22 de noviembre, 2018

Esquema

- Objetivo
- Conjunto de datos
- EDA
- Resultado base:
  - Paper
  - Kaggle
- Resultados:
  - Regresión OLS y técnicas de contracción
  - Regresión con reducción de dimensión
  - Extra: regresión no lineal
- Conclusiones
- Anexo: Parallel analysis

Objetivo

- **Predecir el año de lanzamiento de una canción a partir de las características del audio, utilizando PLS Y PCR.**
- Producto secundario: contraste entre métodos basados en OLS, contracción, reducción de dimensionalidad y no lineales para realizar regresión en dimensiones *medianas*.

Conjunto de datos

- Canciones comerciales cuyo año de lanzamiento se encuentra en $[1922, 2011]$
- Subconjunto del famoso **Million Song Dataset**[1] (515,345 observaciones con 90 variables y etiqueta).
- Partición: 463,715 observaciones para el conjunto de entrenamiento, cerca del 89% de la muestra, y el resto 51,630 como conjunto de prueba.[2]

[1] Uno de sus fines es alentar la investigación en algoritmos a escalas comerciales

[2] Sugerencia del donador de los datos

Conjunto de datos, comentario

A pesar de no encontrarnos en una configuración HDLSS [3] sí tenemos un problema de muestreo respecto a la dimensionalidad.

[3] Hall, P., Marron, J. S. and Neeman, A. (2005) Geometric representation of high dimension low sample size data

Resultado base

- En un trabajo del donador de los datos[4] se reporta un error sobre el conjunto de prueba de 10.20 y 8.76 (medido como RMSE con las variables en su escala original) utilizando el método de 50 vecinos más cercanos y el algoritmo de Vowpal Wabbit.
- En la web Kaggle se reportan errrores de 8.86[5]

[4] T. Bertin-Mahieux, D. P.W. Ellis,B. Whitman y P. Lamere; *THE MILLION SONG DATASET*

[5] No existe evidencia del método utilizado y de si se utilizó el conjunto de datos completo

EDA, distribucón de y

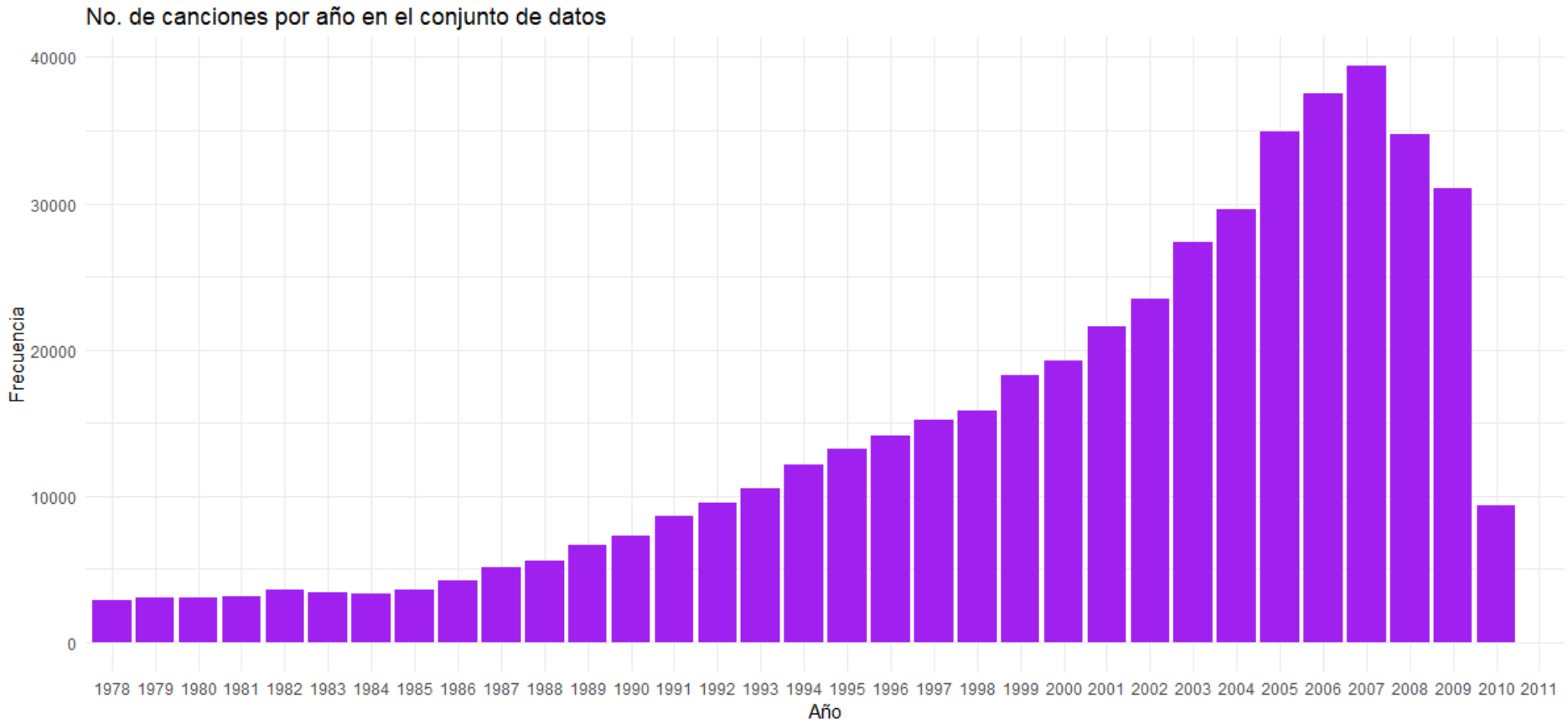

Figure 1:

EDA, MDS stress-1 inferior a 0.3

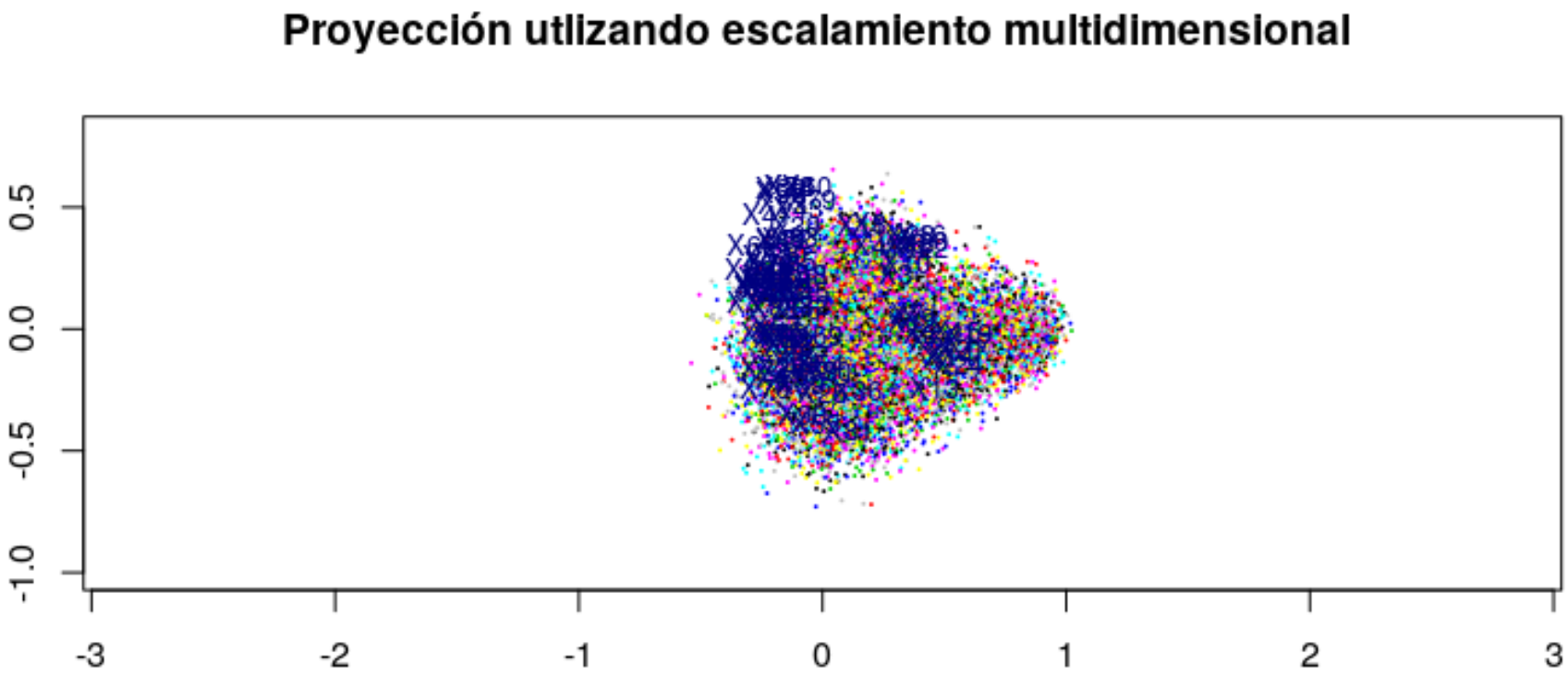

Figure 2:

EDA ¿Linealidad?

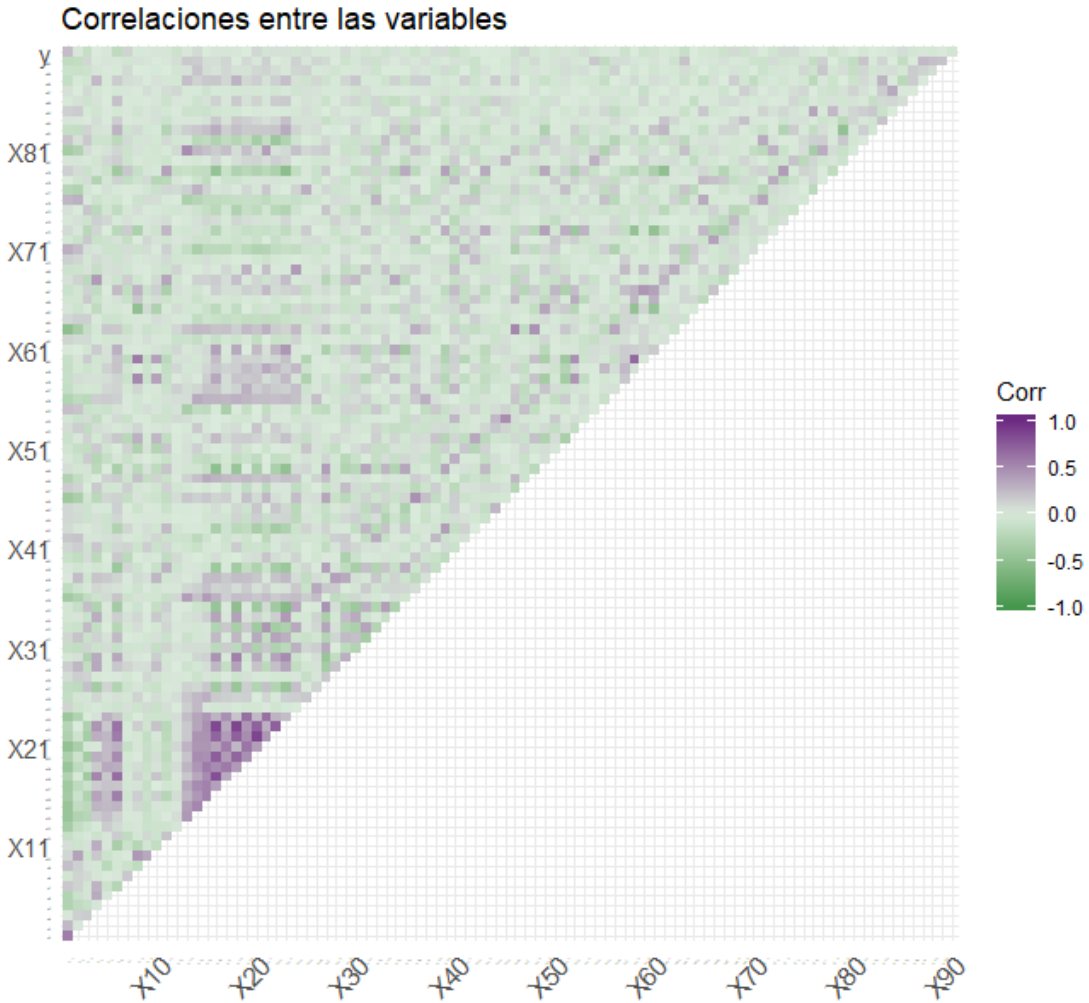

Figure 3:

Resultados, OLS y técnicas de contracción

Método	MSE entrenamiento	MSE prueba	Folds	Tiempo
OLS	0.7637302	0.7569239	2	2 segs
OLS, step	0.7637379	0.7569413	2	1.11 hrs
Ridge	1.001594	0.9856738	500*	26.26 mins
Lasso	1.001594	0.9856738	2500*	30.49 mins

La regularización se vuelve demasiado fuerte, por lo que las variables importantes pueden quedar fuera del modelo y los coeficientes se reducen excesivamente $\approx 1e{-}08$

Resultados: OLS estimación

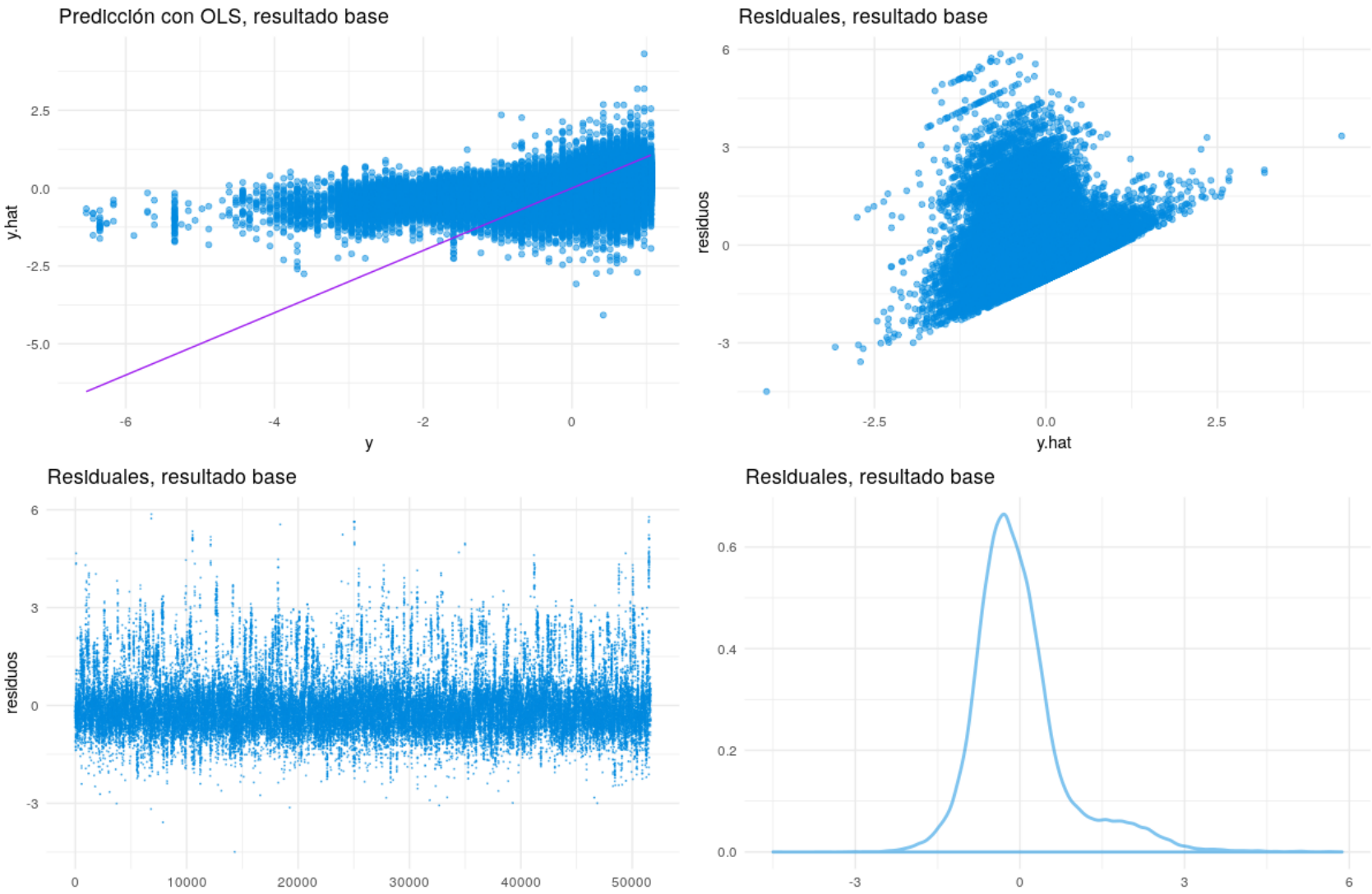

Figure 4:

Resultados, técnicas de reducción de dimensionalidad

Método	MSE entrenamiento	MSE prueba	Folds	Tiempo
PCR	0.8710691	0.8627718	50	20.18 mins
PLS	0.7637633	0.7568457	120*	20.34 mins

Resultados, regresión no lineal

Se realizó una búsqueda exhaustiva sobre diferentes grids utilizando SVM, KNN y el algoritmo de XGBoost.

Variables de entrada,			
XGBoost	MSE entrenamiento	MSE prueba	Tiempo
40 PC	0.7166808	0.8281409	3.15 hrs
16 PLS	0.7637633	0.7099578	20.34 mins
90 variables originales	1.3547	1.25721	7.25 hrs

Resultados

Usando la RMSE y la misma escala que el donador de los datos el error de PLS es de 9.509669, y usando los scores de PLS como entradas para XGB el error es de 9.210389

- Conclusiones
- Podríamos considerar que existen 12 variables latentes sin embargo encontramos una mejor predicción al estimar 16 componentes de PLS.
- En el espacio oblicuo de los scores de PLS, XBG muestra una ganancia marginal en comparación de los tiempos de cómputo que requiere.

Cierre

La agregación puede aportar grandes ganancias por encima de los componentes individuales. Funes era big data sin estadística [6]

[6] Stephen M. Stigler; *Los siete pilares de la sabiduría estadística*; Libros Grano de Sal, 1er edición 2017, pág. 23

Anexo: Parallel analysis

- Enfoque de PCR usando Parallel analysis: Imperativo una simulación eficaz y eficiente.
- Bootstrap y distribución apriori no informativa.

Anexo: Parallel analysis (simulaciones montecarlo ~5000)