Universidad Nacional Autónoma de México

Instituto de Investigaciones en Matemáticas Aplicadas y Sistemas

Análisis de bienes raíces en EEUU

Minería de datos

Autor:

Aguilar Rodríguez José Antonio

Moran Peraza Hugo Carlos

Introducción

A lo largo de los años la industria de bienes raíces ha demostrado ser una fuente constante de ingresos para muchos inversionistas. En general el comportamiento que ha tenido este mercado ha sido estable, salvo en algunos períodos excepcionales (Crisis 2007-2008), lo cual convierte a este mercado en uno de los más atractivos para invertir. No es por nada que la riqueza de una buena parte de los millonarios en Estados Unidos provenga de esta industria. Por esta y otras razones, hacer un buen análisis de este mercado es bastante fructífero en muchos sentidos.

Si bien desarrollar un buen análisis puede ser muy rentable, no es para nada sencillo, ya que se tiene que tener en cuenta muchos factores microeconómicos y macroeconómicos que afectan al precio de los inmuebles.

Por estos motivos, nos ha contratado una de las empresas con mayor presencia en Estados Unidos en materia de bienes raíces, solicitando la ayuda de dos científicos de datos que les puedan generar un modelo que estime el precio de las casas, tomando en cuenta variables sociodemográficas.

Preliminares

En esta sección hablaremos un poco de los preliminares del proyecto, estos puntos preliminares servirán como pilar para el resto del proyecto.

Objetivo

Nuestro objetivo va orientado a brindar un modelo de valuación de casas para una empresa que se dedica a comprar y vender inmuebles en Estados Unidos. De esta manera aprovechar las oportunidades en las que las casas se encuentren subvaluadas o sobrevaluadas para generar un rendimiento atractivo para la empresa.

Alcance

El alcance de este proyecto se limita a trabajar con datos que se obtuvieron de forma gratuita por diversas instituciones, la ventana de tiempo de estos datos se encuentra entre el año 1998 y el año 2018, dándonos un periodo de veinte años para analizar. No contamos con datos actuales, por tanto, nuestra meta se enfoca más en encontrar características sociodemográficas que ayuden a predecir el precio de un inmueble, trazando así un camino que pueda generar una metodología que pueda ser aplicada a datos actualizados y pueda ser puesta en producción.

Marco estratégico

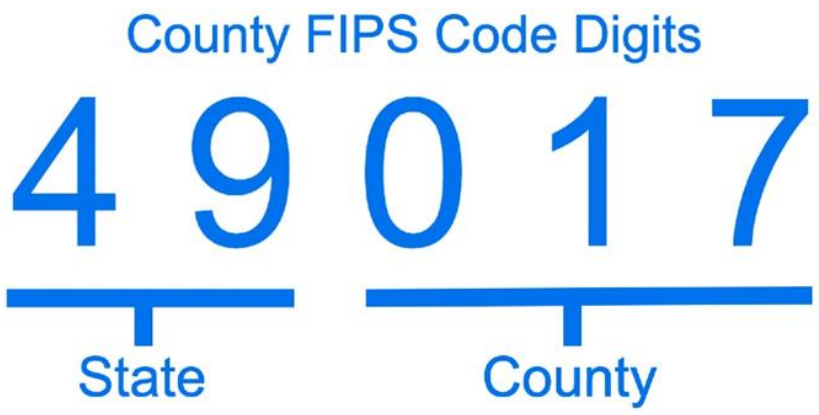
Primero: Crearemos una base de datos con las características que creemos que afectan de manera directa el valor de los inmuebles. Integraremos en una sola base todos los datos que obtengamos de las diferentes fuentes que consideremos.

Segundo: Una vez obtenida la base de datos, realizaremos una limpieza sobre ellos, luego haremos un análisis exploratorio para realizar un estudio del mercado en el que se desenvuelve la empresa interesada. Posteriormente aplicaremos un preprocesamiento en donde, entre otras cosas, trataremos de reducir las dimensiones de los datos.

Tercero: Por último, utilizaremos la base de datos para probar varios modelos (de regresión: lineal, árboles de decisión y potenciación del gradiente) y de este modo generar una forma eficiente de estimar el precio de una casa dada.

Definiciones

FIPS: El código FIPS tiene como finalidad la de representar numéricamente los condados de los estados en Estados Unidos, veamos un ejemplo:



Los primeros dos dígitos hacen referencia al estado y los últimos tres hacen referencia al condado en cuestión, todos los códigos son de 5 dígitos.

Marco Teórico

Antes de adentrarnos en el contenido de nuestro proyecto, conviene realizar un breve repaso de algunos conceptos matemáticos importantes. A lo largo de esta sección daremos una breve introducción a los modelos que utilizaremos más adelante en el proyecto: regresión lineal, árboles de decisión y XGboost. Así mismo, definiremos un par de conceptos matemáticos conocidos como métricas. Estas métricas, como su nombre lo sugiere, nos ayudarán en la parte final del proyecto a "medir" qué tan bien se desempeñan nuestros modelos. Existe una gran variedad de métricas e indicadores que tienen este propósito, sin embargo, dada la naturaleza de nuestro proyecto y para no saturar de información este documento, lo ideal es exponer dos: la R^2 y el error cuadrático medio (MSE por sus siglas en inglés).

El problema de la Regresión

El objetivo principal cuando se trata de resolver un problema de regresión es predecir una etiqueta, comúnmente llamada variable objetivo, a una observación que no posee una. El punto clave de la regresión es que dicha etiqueta es en realidad una variable que puede tomar cualquier valor en los números reales, es decir, puede tomar valores como 1, 2, 3.5, 2.66, 1000. Un problema de regresión que es muy famoso, y que de hecho guarda mucha relación con el objetivo general de nuestro proyecto es el de estimar el precio de una casa con base en características de ésta como el área, número de habitaciones, ubicación y otras más.

Un problema de regresión puede ser resuelto mediante un algoritmo de aprendizaje de regresión, el cual utiliza una colección de ejemplos u observaciones como entrada y produce, de alguna manera, un modelo que tiene la capacidad de recibir ejemplos sin etiqueta como entrada y devolver los valores de la variable objetivo correspondientes a éstos.

Como ya se mencionó en párrafos anteriores, el problema que deseamos solucionar mediante este proyecto se trata de uno de regresión. Por esta última razón fue conveniente exponer una breve introducción a estos problemas. Lo que sigue es dar una muy pequeña definición de algunos algoritmos de aprendizaje de regresión, los cuales son los que más adelante utilizaremos, evaluaremos y compararemos.

Regresión Lineal

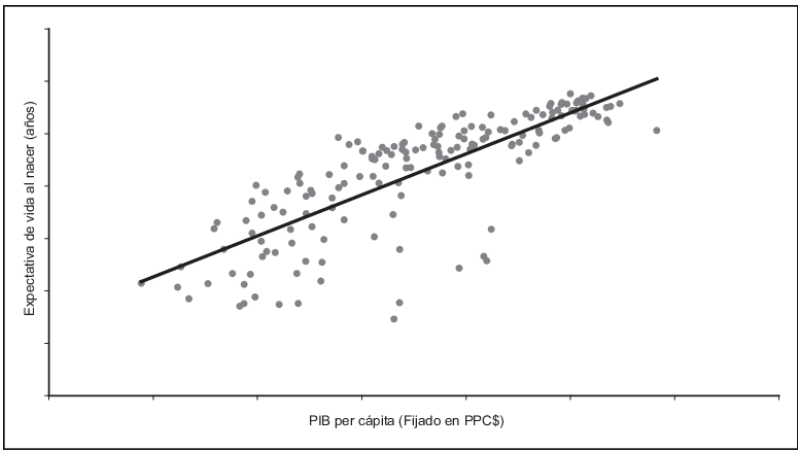
La regresión lineal es uno de los algoritmos de aprendizaje de regresión más populares y que, en general, tienen muy buen desempeño. Otro par de características importantes que posee la regresión lineal, y por lo cual fue un algoritmo que decidimos utilizar, es su simplicidad tanto de implementación como de interpretación.

El algoritmo de regresión lineal ajusta un modelo que se trata de una simple combinación lineal de los valores de las características (recordemos que con características nos referimos a propiedades de las observaciones como el peso, las dimensiones de alto, ancho y largo, etc.), esto significa que a cada característica le corresponderá un peso que indica qué tan importante es dicha característica a la hora de predecir un valor de la variable objetivo.

En términos un poco más matemáticos, si se tiene una colección de ejemplos etiquetados $\{(x_i, y_i)\}_{i=1}^{N}$, donde N es el tamaño de la colección, x_i e y_i son el vector de características y su correspondiente etiqueta para la observación $i = 1, \ldots, N$. El objetivo final de la regresión lineal es encontrar un modelo tal que se cumpla lo que sigue:

$$f_{w,b}(x) = wx + b$$

donde w es un vector de pesos y b es un número real, es decir, se desea encontrar los valores de w y b para que la ecuación anterior funcione lo "suficientemente bien". Más adelante en esta misma sección mencionaremos con más detalle a qué nos referimos con "suficientemente bien". Finalizamos la introducción a la regresión lineal con la siguiente figura que muestra un ejemplo de una regresión lineal simple:



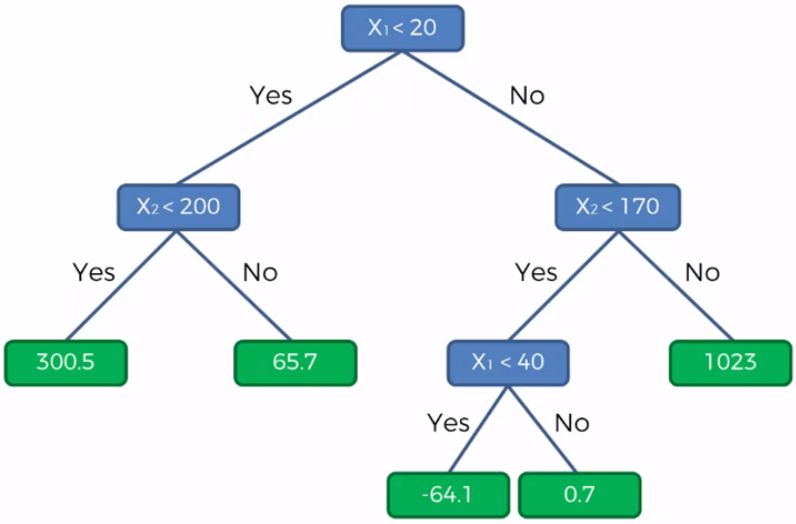
Árboles de decisión

Un árbol de decisión es, esencialmente, un diagrama que fluye en una única dirección con varias bifurcaciones, la forma en la que es posible interpretar éstas últimas es considerarlas como decisiones. En cada una de las bifurcaciones se examina una característica específica del vector de características. Si el valor de la característica examinada se encuentra por debajo de un número específico, entonces se toma el camino del lado izquierdo; de otro modo, el camino que se debe seguir es el derecho. Esta estrategia se sigue al pie de la letra hasta que se llegue a un nodo en el diagrama en el cual ya no haya opción de tomar una

decisión, a este nodo se le suele llamar “hoja” (leaf en inglés). Cabe destacar que el valor que se encuentre en el nodo hoja será el valor de la variable objetivo que corresponda.

Como los árboles de decisión son útiles para intentar resolver un problema de regresión, lo que se busca es hallar una estructura del diagrama tal que, para todos los ejemplos etiquetados en conjunto, los valores en los nodos hoja que resulten de seguir la estructura del diagrama se “asemejen lo suficiente” a sus respectivas etiquetas. Conviene hacer énfasis en la importancia que ahora tiene el definir qué tanto es suficiente. Esta incógnita representa uno de los problemas más grandes en el aprendizaje de máquina ya que se busca optimizar el modelo.

Para el caso de los árboles de decisión no daremos una definición en un lenguaje matemático ya que consideramos que explicar con vigor matemático el modelo de árboles no resulta sencillo y resulta mucho más difícil entenderlo al lector en caso de que éste no posea suficiente formación matemática. Sin embargo, así como con la regresión lineal, consideramos pertinente la exposición de un ejemplo de una estructura de un árbol de decisión aplicado a un problema ejemplar:



XGBoost

El XGBoost (eXtreme Gradient Boosting) es una implementación de código abierto popular y eficiente del algoritmo de árboles “boosteados” por gradientes. El “boost” por gradiente es un algoritmo de aprendizaje que intenta predecir con precisión una variable objetivo mediante la combinación de un conjunto de estimaciones de un conjunto de modelos más simples y débiles.

El algoritmo XGBoost funciona bien en las competiciones de aprendizaje automático debido a su manejo robusto de una variedad de tipos de datos, relaciones, distribuciones y la variedad de hiperparámetros que puede ajustar.

A muy grandes rasgos, el objetivo del algoritmo XGBoost es utilizar diferentes modelos para que cada uno de éstos “aporte su opinión”. Cuando todos los modelos hayan opinado, se determinará el valor de la variable objetivo considerando estas opiniones de alguna manera. Una analogía útil podría ser que un paciente desea saber si padece cáncer, para esto decide ir a consulta con diferentes médicos (cada uno de los médicos representa un modelo) y recibe sus respectivos diagnósticos (valores de la variable objetivo). Para que el paciente pueda determinar si tiene cáncer o no con mayor seguridad, decide concluir que el diagnóstico que más haya aparecido es el verdadero. La idea del XGBoost es esencialmente esta.

Métricas de evaluación de modelos

Las métricas de evaluación de modelos, como cualquiera podría intuir, nos ayudan a determinar qué tan bueno resulta ser un modelo. Además, pueden servir como punto de referencia para comparar dos o más modelos y, de esta forma, determinar qué modelo es mejor. Por esta razón, una buena selección de métricas de evaluación de los modelos puede traducirse en una buena selección de modelos y, por consiguiente, un trabajo mejor echo.

Uno de los principales problemas relacionados con las métricas de evaluación de modelos es que hay demasiadas y que en los problemas donde el uso de un conjunto de métricas es ideal, el uso de otro conjunto de métricas puede ser engañoso y, por lo tanto, peligroso. Sin embargo, como la base de nuestro proyecto es un problema de regresión, existe un par de métricas que son generalmente buenos puntos de referencia en la evaluación de los modelos producto de los algoritmos de aprendizaje de regresión expuestos hasta este momento: la R^2 y el error cuadrático medio (MSE).

El coeficiente de determinación, mejor conocido como R^2, es una métrica estadística que representa la proporción de la variabilidad de la variable objetivo explicada por las características consideradas en un modelo de regresión. Por esto mismo, la R^2 puede tomar valores entre 0 y 1. Si su valor es igual a 0.5, significa que aproximadamente la mitad de la variabilidad observada en la variable objetivo puede ser explicada por las variables de entrada (características consideradas) del modelo. Podemos concluir fácilmente que el caso ideal es que la R^2 tome un valor cercano a 1.

Por el lado del error cuadrático medio, su definición es un poco más fácil y rápida de entender. Mide el promedio de los errores al cuadrado entre el valor de la variable objetivo que arroja un modelo para una observación y su verdadero valor. Con esto es relativamente fácil deducir que un buen modelo debería tener un MSE lo más cercano posible a 0. Esto indicaría que el modelo propuesto predijo correctamente todos los valores de la variable objetivo para todos los ejemplos.

Con esto creemos que es suficiente introducción a las herramientas que utilizaremos a lo largo del proyecto para que cualquier lector, incluso si este no posee una sólida formación en matemáticas, estadística y computación, pueda entender cada detalle de nuestro proyecto.

Integración de la información:

Variables totales

En esta sección describiremos un poco el proceso con el cual se consiguió los datos necesarios para el proyecto, así como la forma en que se limpiaron y se integraron para su uso en la modelación. Cabe mencionar que la información recabada tiene una venta de tiempo del año 1998 al año 2018.

Los datos usados así como sus bases de datos respectivas, fueron:

- Ratio de desempleo, la base de datos original fue obtenida por Geographical Economic Data de ST. LOUIS FED.
- Hospitales en Estados Unidos, la base de datos original fue obtenida por Homeland Infrastructure Foundation.
- Rating de hospitales, la base de datos original fue creada por Centers for Medicare & Medicaid Services.
- Precio de inmuebles empresa Zillow. (county_time_series) y el dataset countycrosswalk.
- Escuelas públicas del país creado por Homeland Infrastructure Foundation.
- Ratio de crímenes obtenido por ICPSR.

Una vez descargados estos datasets, se hizo uso del lenguaje de programación Python, con el cual se desecharon las columnas que no eran necesarias. Al querer juntar ya todas las tablas, hubo un pequeño error al querer juntar la tabla de crimines, para la cual usamos una herramienta de perfilado y limpieza de datos, llamada Talend. Con esta herramienta exploramos de forma visual los datos, y nos dimos cuenta de ciertos detalles en algunas variables.

Limpieza de la información

Como ya se mencionó anteriormente en esta parte usamos la herramienta Talend Preparation, la cual es bastante fácil de usar y contiene bastantes técnicas de limpieza y transformación, que nos servirán para el proyecto.

Una vez cargado los datos a Talend nos dimos cuenta que una variable tenía entradas inválidas, después de investigar un poco detectamos el error, el cual era básicamente que en algunos números hacía falta un digito, generando todo el problema anterior.

Posteriormente, notamos que los valores flotantes tenían muchos decimales, así que decidimos dejarlos hasta dos puntos decimales. Luego también observamos que cierta columna (Rating de hospitales) tenía valores nulos, los cuales completamos con la moda de cada condado.

Analytical Base Table (ABT)

Una vez terminada la limpieza, nos pasamos otra vez a Python, donde ya tendremos la base de datos completa, con la cual se hará la modelación. Aunque primero haremos una pequeña exploración de unos estadísticos relevantes de nuestros datos.

Veamos la tabla final, después de hacer todo este proceso:

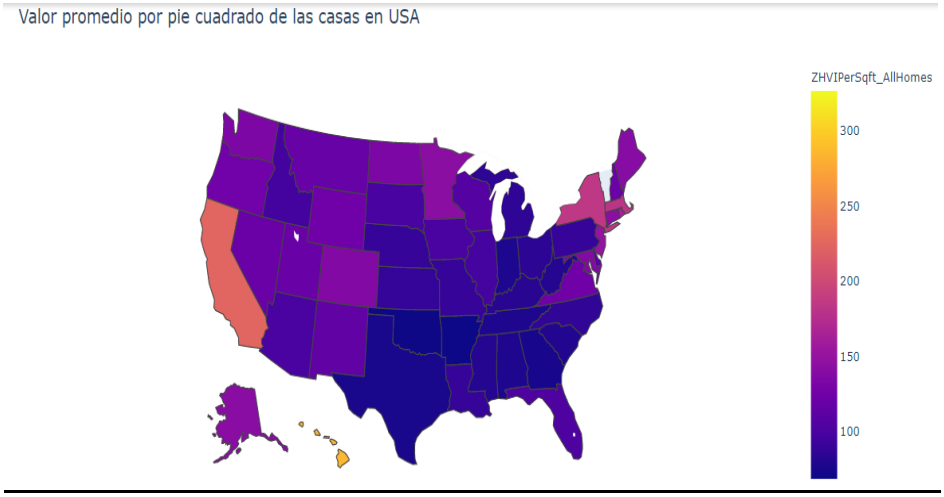
	FIPS	CountyName	StateName	NumberOfSchools	NumberOfHospitals	AverageHospitalRating	UnemploymentRate	crime_rate_per_100000
0	1001	Autauga	Alabama	15	1	4.00	3.6	251.601926
1	1003	Baldwin	Alabama	47	4	3.00	3.6	228.086325
2	1069	Houston	Alabama	31	5	3.50	4.1	401.281012
3	1073	Jefferson	Alabama	226	14	2.75	3.7	798.357491
4	1081	Lee	Alabama	37	1	4.00	3.6	246.466975

AverageHousePrice	ZHVIPerSqft_AllHomes	MedianRentalPrice_AllHomes
114483.67	72.56	1255.21
164861.69	90.39	1165.94
114941.00	72.94	975.63
99759.77	68.25	855.86
148163.22	74.79	1143.89

Exploración

En esta parte visualizaremos algunos estadísticos interesantes.

Para tener una idea general del valor de las casas, sin tomar en cuenta el tamaño de ellas, tenemos una variable que nos dice el valor por pie cuadrado de los inmuebles, dándonos información intrínseca del valor de las casas.

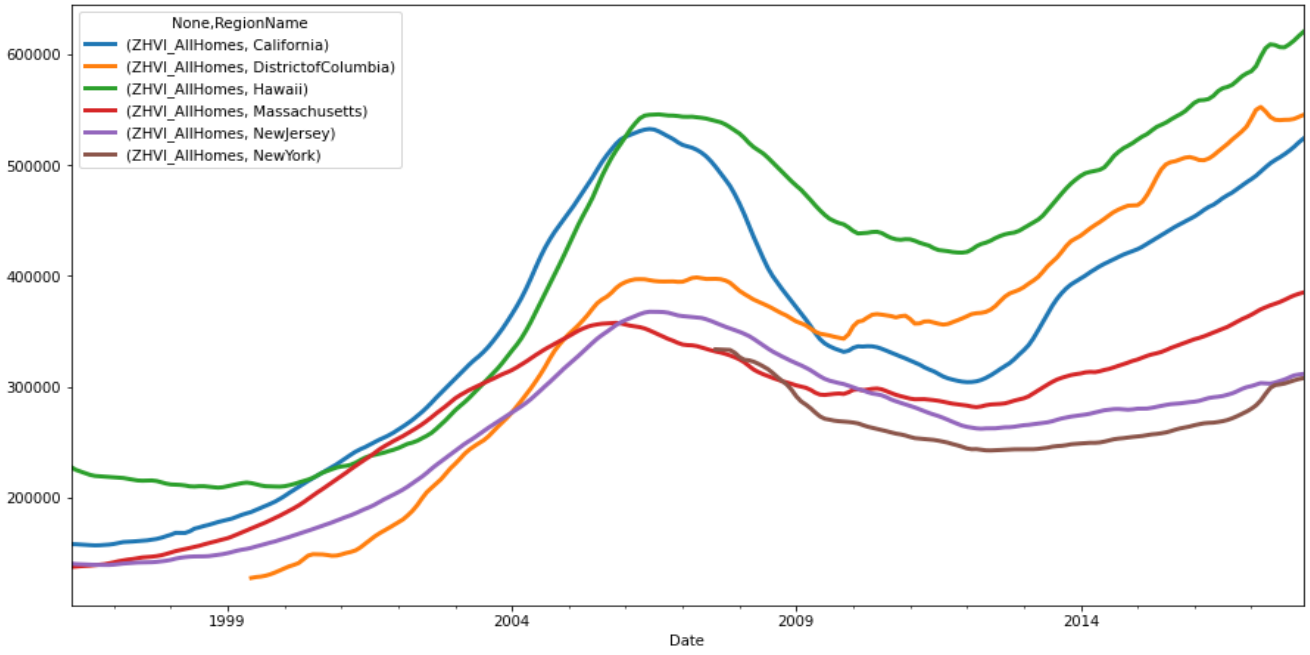


Observaciones de los resultados:

- Los estados al sureste del país tienen menor valor comparados con los estados del noreste.
- Hawái tiene el valor más grande de todos los estados, seguido de california.
- En general, los estados del norte tienen un mayor valor que los del sur.

Veamos ahora como se ha desarrollado el precio de las casas a lo largo de los años para el top de estados:

Serie de tiempo, precio de inmuebles
(En USD)

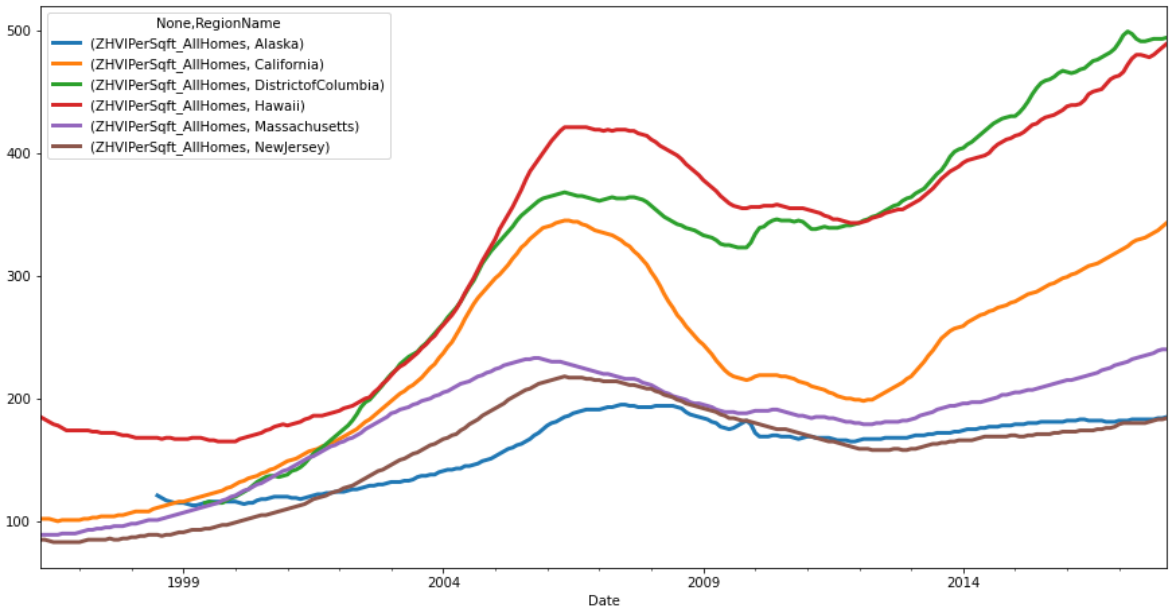


Observaciones de los resultados:

- Se puede ver cómo quedó marcada la crisis inmobiliaria (2007-2008) en el precio de los inmuebles a lo largo de esos años. Después del 2012 se empieza a recuperar los precios para posteriormente iniciar una tendencia alcista (hasta la próxima crisis).
- Al estado de New Jersey le ha costado más recuperarse de esa crisis, vemos como antes de la crisis tenía una muy buena tendencia alcista pero después del 2008 su tendencia ha sido plana.

A continuación veremos algo muy parecido al gráfico anterior, pero en esta ocasión veremos reflejado el valor por pie cuadrado

Serie de tiempo, precio por pie cuadrado
(EN USD)

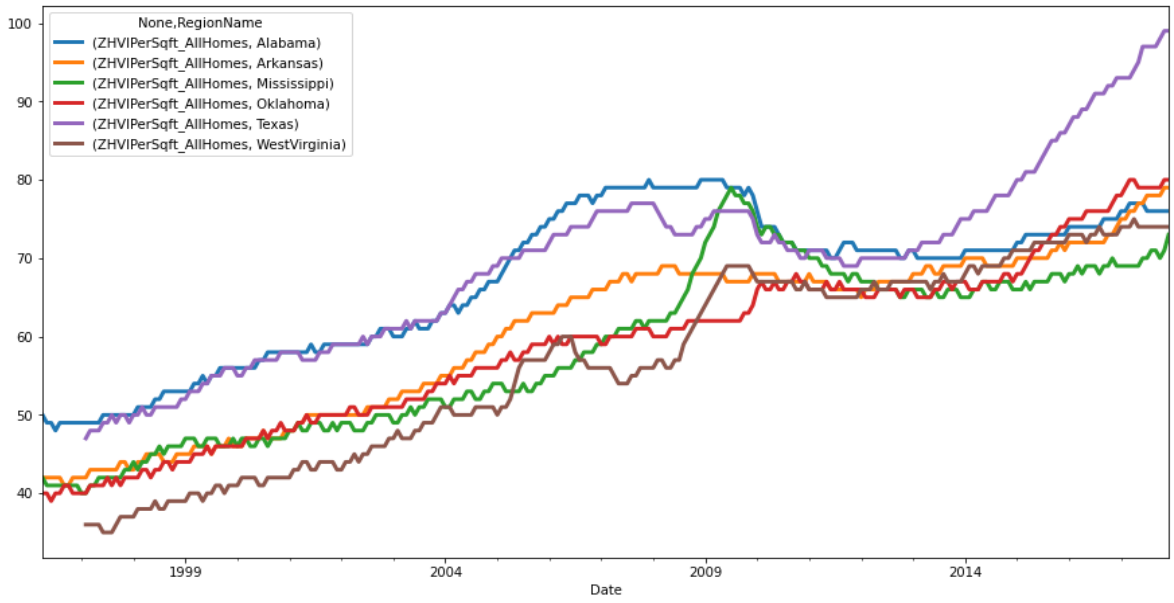


Observaciones de los resultados:

- Notamos un patrón similar al gráfico anterior.
- El caso de New Jersey nos llamó la atención, otra vez. Antes de la crisis tenía toda la pinta de que se iba a posicionar en los primero lugares, pero una vez pasada la crisis no se pudo recuperar de ella.
- También notamos que California fue bastante castigada por la crisis.

Por último veamos la tendencia del valor por pie cuadrado para el top de los estados más baratos.

Serie de tiempo, precio por pie cuadrado
(EN USD)



Observaciones de los resultados:

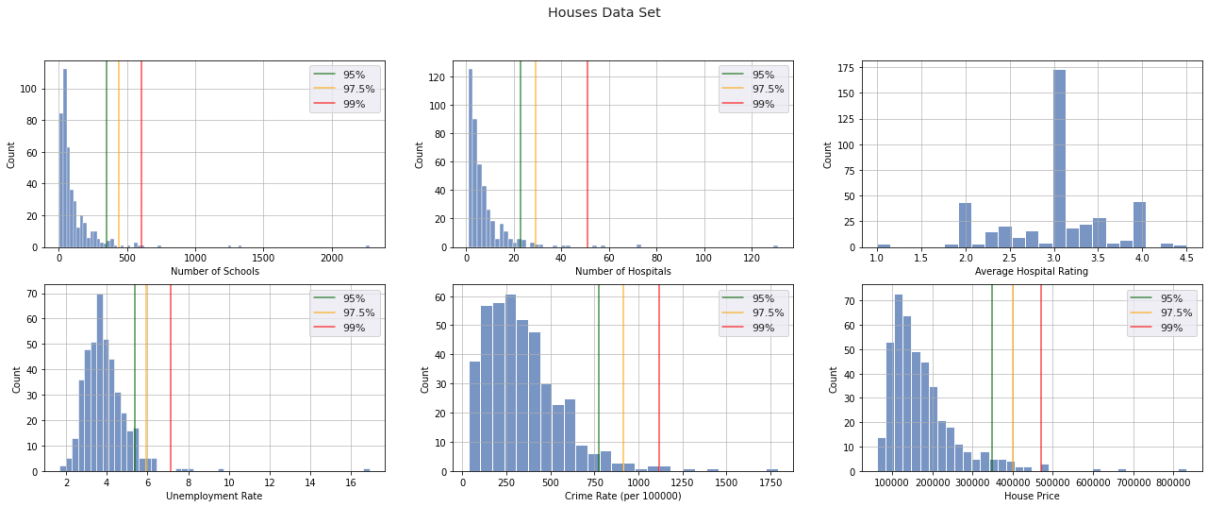
- En este gráfico ocurre todo lo contario, en vez de bajar por la crisis tienen un considerable aumento. Esto se debe a que las personas dejaron de comprar en los estados caros y se fueron a vivir a los estados más “económicos”.

Modelación

Como ya hemos mencionado en secciones anteriores, el objetivo particular de este proyecto es hallar una metodología que sea útil para valuar las casas dentro del mercado estadounidense. Lo que proponemos para completar este objetivo es ajustar un modelo cuya variable objetivo sea el precio de las casas y las variables candidatas a ser consideradas como predictores, es decir aquellas variables que formaran parte de los vectores de características para nuestras observaciones, sean las variables sociodemográficas que conforman la base de datos que logramos conseguir, limpiar e integrar en la sección anterior.

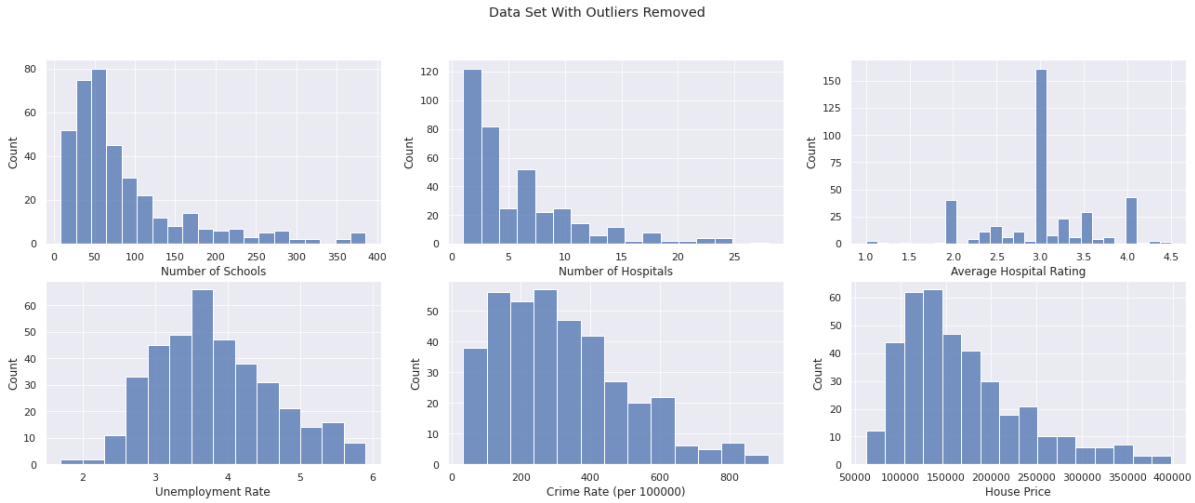
Una vez que tenemos lista nuestra base de datos, lo que siempre es recomendable hacer es un análisis exploratorio preliminar. En este análisis exploratorio preliminar, nuestro objetivo principal será seleccionar aquellas variables que contengan la mayor cantidad de información respecto al precio de las casas. La forma de determinar qué tanta información nos aportan estas variables candidatas depende del tipo de la variable, si es numérica y ordinal, o si es categórica.

El primer paso en el análisis de las variables numéricas y ordinales suele ser el analizar los histogramas de cada una de éstas. Haciendo esto, es posible darse cuenta de 4 características, no necesariamente independientes una de otra, importantes correspondientes a las variables en cuestión: la distribución (si es normal, exponencial, uniforme, etc.), el rango (los valores que puede llegar a tomar la variable), los cuantiles (los cuales nos indican dónde se concentra cierta cantidad de los datos) y los valores atípicos (aquellas observaciones que presuntamente se alejan de un escenario normal). Por esta razón, decidimos hacer uso de esta poderosa herramienta, los resultados se pueden apreciar en la siguiente imagen:

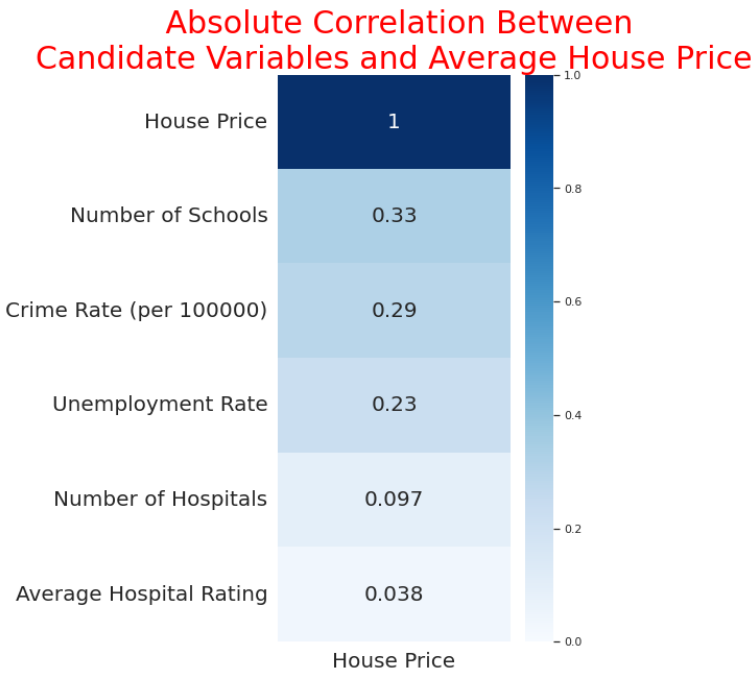


Las líneas verticales de color verde, amarillo y rojo representan los cuantiles al **95%**, **97.5%** y **99%** respectivamente. Lo cual se puede interpretar como que el **95%** de los ejemplos dentro de nuestro data set se encuentran concentrados hasta la línea verde, el **97.5%** antes de la línea amarilla y el **99%** de los ejemplos se encuentran del lado izquierdo de la línea roja. Otra observación importante que debería hacerse de las gráficas es que muchas veces las líneas se encuentran en la mitad izquierda de las gráficas, esto quiere decir que existen valores que podrían considerarse atípicos. No es recomendable considerar valores atípicos al momento de modelar o sacar conclusiones respecto de las relaciones que guardan las diferentes variables, pues éstos únicamente representan ruido (representan información falsa o basura). Lo que decidimos hacer es considerar

exclusivamente aquellos ejemplos que se encuentren del lado izquierdo de la línea amarilla para cada una de las variables. El resultado es el siguiente:



Con las distribuciones más estables, podemos continuar con nuestro análisis. El siguiente paso consta de encontrar las relaciones que existen entre las variables numéricas y ordinales candidatas y nuestra variable objetivo. El método clásico para estudiar estas relaciones es mediante la correlación y ésta puede observarse claramente con la ayuda de la siguiente visualización:



Lo ideal es que la correlación sea lo más cercana a la unidad. Mientras más cercano a 1 se encuentre el valor de la correlación implica que más información aporta la variable correspondiente respecto de nuestra variable objetivo. Por esta

razón utilizaremos aquellas características que más correlación tengan con el precio de las casas. Consideraremos las variables: del número de escuelas dentro del condado, el ratio de crímenes, el ratio de desempleo del condado y el número de hospitales.

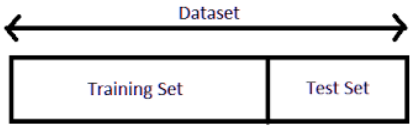
Ahora, para analizar la única variable categórica que tenemos (el estado en donde se encuentra la casa), nos apoyaremos de una prueba estadística que recibe el nombre de prueba de Kruskal. El objetivo de la prueba de Kruskal es determinar si una variable categórica aporta una cantidad significativa de información respecto de una variable numérica y ordinal. La hipótesis nula de esta prueba es que la variable categórica NO aporta información relevante. El resultado obtenido al aplicar dicha prueba para la variable correspondiente al estado en donde se encuentra la casa se encuentra en la tabla que se encuentra a continuación:

Estadístico	P-valor
172.25	8.8478 e-17

El valor más importante es el P-valor, el cual nos indica que, como dicho valor es menor a 0.05, podemos rechazar la hipótesis nula de la prueba con un 95% de confianza. Con esto concluimos que la variable correspondiente al estado donde se ubica la casa es una variable importante si se desea predecir el precio de la misma. Lo único que falta es encontrar una buena representación de dicha variable. En nuestro caso haremos uso del código FIPS, el cual en sus primeros dos digitos tiene la información del estado al que pertenece el condado, y además a condados cercanos se le asocia códigos cercanos, así habrá una relación de cercanía entre condados, esto nos puede ayudar ya que si un condado tiene valores de casas muy altos entonces esperaríamos que los condados vecinos tengan valores cercanos.

Ahora que hemos seleccionado las variables que utilizaremos para el modelado, conviene hablar sobre los modelos que utilizaremos. Como nuestro problema es uno de regresión, lo consecuente es usar algoritmos de aprendizaje de regresión. Los modelos candidatos que seleccionamos para intentar ajustarlos, probarlos, evaluarlos y compararlos son: regresión lineal, árbol de decisión y XGBoost. Respecto de las métricas que utilizaremos para evaluar y comparar los métodos anteriores son la R^2 y el MSE, esto debido a la naturaleza del problema y de los modelos que ajustaremos.

Para realizar todo este procedimiento de ajuste, prueba, evaluación y comparación de los modelos. Lo primero que debemos hacer es dividir nuestro conjunto de datos en dos subconjuntos: uno de entrenamiento y otro de prueba. El conjunto de entrenamiento constará del 80% del conjunto total mientras que el conjunto de prueba representará el sobrante 20%. La manera gráfica de verlo es la siguiente:



Posteriormente eliminamos el ruido de nuestro conjunto de entrenamiento, es decir, consideraremos únicamente las observaciones cuyas variables se encuentren antes de todas las líneas amarillas de la primera figura presentada en esta sección. Luego ajustamos los modelos sobre este conjunto preprocesado y, con estos modelos ajustados, estimamos los valores para el conjunto de ejemplos de prueba. Con estas estimaciones podemos calcular las métricas mencionadas algunos párrafos atrás.

Resultados

Los resultados finales se pueden resumir en la siguiente tabla:

Modelo	R^2	MSE
Regresión Lineal	0.34	0.158
Árbol de decisión	0.18	0.180
XGBoost	0.554	0.102

Conclusiones de los modelos empleados

Como podemos notar en la tabla anterior, los resultados obtenidos no son tan satisfactorios como se esperaría, el r^2-score no es tan alto como nos gustaría, aún así podemos notar cosas interesantes:

- El modelo de árbol de decisión tuvo el peor desemepeño con respecto al r^2-score, teneniendo así un error medio cuadratico mayor.
- Aunque el XGBoost se basa en arboles de decisión, gracias a sus parámetros y de su función objetivo, obtiene un mejor resultados con sus dos contricantes.
- Por último la regresión lineal se lleva el segundo lugar, encontrandose justo a la mitad de los otros dos modelos.

Conclusiones

En este trabajo se exploró una metodología para tratar de predecir los precios de inmuebles en EEUU, se analizó de forma profunda algunas características sociodemograficas que sirvieron para predecir dichos precios, si bien no se obtuvieron grandes resultados, se trazó el camino para seguir explorando más características que puedan aportar información a los modelos.

Unas de las ideas que tenemos para trabajo a futuro, es conseguir la información del ingreso promedio por condado, esto nos ayudará a comprender que tanto “presupuesto” tienen las personas de ese lugar para invertir en una casa.

Referencias

Burkow Andriy, “The Hundred-Page Machine Learning Book”, Enero 2019, Andriy.

Jerome H. Friedman, Robert Tibshirani, Trevor Hastie, “The Elements of Statistical Learning”, 2001, Springer

Jake Morgan, “Classification and Regression Tree Analysis”, 2014 [online]