Universidad Francisco Marroquín

Machine Learning Models

José Alejandro Guzmán Zamora

Florida Used Cars Analysis

Introduction

Uno de los aspectos que generan más confusión al momento de realizar la compra de un carro es el precio de este. La manera en la que se valora un carro monetariamente depende de una cantidad considerable de factores incluyendo la relación subjetiva que cada individuo asigna al vehículo correspondiente. Sin embargo, se puede recolectar información específica de casos históricos para verificar qué relaciones existen y de qué manera se puede llegar a explicar este tema desde una perspectiva estadística y de computación. Tomando en cuenta lo mencionado, se puede aprovechar un dataset específico de información relacionada con la promoción y venta de carros usados. En este caso se está haciendo referencia al Dataset "Used Cars Dataset" y lo que se desea realizar a lo largo de este reporte es detallar el proceso por medio del cual se intentó predecir el precio de un carro en base a la información disponible. Debido a que en este tipo de datos hay cierta intuición de que se puede llegar a presentar una relación lineal entre alguna de las variables y el precio, se hicieron intentos principales sobre algoritmos básicos de regresión lineal. En pocas palabras lo que se intentó fue una regresión lineal simple, regresión lineal con múltiples variables involucradas, posteriormente se agregaron features polinómicas al modelo y por último se verificaron los beneficios proporcionados por la penalización de coeficientes mayores bajo el método de regularización de Ridge.

Data

El dataset de carros usados contiene información obtenida del sitio craiglist.org; este sitio web se utiliza para una variedad de cosas, sin embargo, la compra y venta de diferentes artículos es una de las actividades que ocurren con mayor frecuencia en la plataforma. El autor del dataset hizo una exploración de la página por medio de web scrapping para obtener información de ventas de carros usados. Se debe de mencionar, además, que la información recolectada abarca las ventas dentro de los Estados Unidos. El dataset tiene un total de 25 columnas, entre estas algunas de las que se pueden considerar más importantes (antes de comenzar con el análisis directo de selección) incluyen:

- Año
- Región
- Precio
- Manufacturero
- Modelo
- Condición

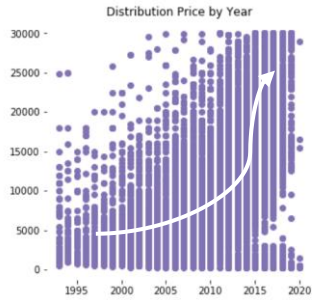
Para un listado completo de las variables referirse al anexo No. 1. Del total de variables, 7 son de tipo numérico y las demás son categóricas. No obstante, se debe de aclarar que de las variables numéricas una es el identificador y dos más describen la posición de cada anunció por medio de longitud y latitud. Estas tres variables mencionadas se descartaron desde el inicio porque no van a ser útiles en el análisis de relación lineal que se desea encontrar. Inicialmente el dataset contiene un total de 509,577 observaciones. Con el objetivo de simplificar el análisis, y con una dirección de interés personal se decidió utilizar un subconjunto del dataset. Por lo tanto, se seleccionó la data específica para el estado de Florida con un total de 35,244 observaciones y la misma cantidad de variables inicialmente.

Methods

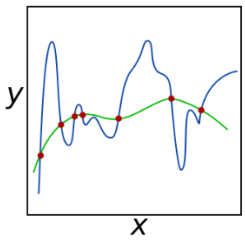
Después de hacer la limpieza y el análisis exploratorio inicial, se definió que las variables sobre las cuales es más conveniente trabajar son “year”, “condition”, “cylinders” y “odometer”. Todas estas variables son númericas, sin embargo, las variables de número de cilindros y de condición eran originalmente variables categóricas que se tuvieron que trasladar al tipo de dato numérico para ingresar adecuadamente a los algoritmos utilizados. Tal como se mencionó en la introducción, la naturaleza de este registro de datos abre las puertas para que se desarrolle una intuición de que hay una relación lineal entre los mismos. El primer paso que se tomó fue utilizar una regresión lineal simple. La regresión lineal es un algoritmo que intenta definir una relación lineal entre diferentes variables de un dataset; esto lo hace calculando coeficientes que se utilizan para graficar o evaluar la relación lineal obtenida. Vale la pena mencionar que la regresión lineal utiliza la minimización del costo de cada modelo de predicción en el que el error es la diferencia entre la predicción y el valor real. En este caso se le denominó una regresión lineal simple porque se intentó predecir la variable de precio solamente utilizando una de las demás variables disponibles. Con el objetivo de realizar una mejor predicción de la variable, se evaluó el modelo con cada una de las variables independientes. Por lo tanto, se llevaron a cabo 4 regresiones lineales que produjeron 4 modelos.

Posteriormente se quiso aumentar la capacidad de descripción del modelo por medio de un análisis de distintas variables, en otras palabras, se realizó una regresión lineal múltiple. En este caso se utiliza la palabra múltiple para indicar que se utiliza más de una variable para intentar describir o predecir la variable de respuesta. El algoritmo de regresión lineal múltiple es exactamente el mismo que el de la regresión lineal simple, la diferencia principal es la interpretación de este. Al momento de tener más de 1 variable independiente, el modelo pasa de estar en el campo vectorial bidimensional a estar en cantidades mayores de dimensiones. Por ejemplo, si se quiere empezar a predecir el precio en base a 2 variables, se proporcionarían la cantidad de coeficientes necesarios para la representación de un plano en 3 dimensiones. Esto se replica para cualquier cantidad de dimensiones, sin importar que el humano no pueda pensar o visualizarlas. La manera en la que se realizó la evaluación de estos modelos fue un poco distinta, esto es porque al momento de poder tener más de una variable se pueden empezar a crear distintas combinaciones que proporcionen mejores descripciones. Para no dejar ningún espacio de duda, lo que se hizo fue que se generaron todas las posibles combinaciones de estas variables descriptivas. Posteriormente se corrió el modelo de regresión lineal y se seleccionó el de mejor rendimiento en base a la métrica de R cuadrado. Esta métrica utilizada describe la proporción de la varianza que una o muchas variables independientes describen de la dependiente o de respuesta.

Una vez seleccionado el mejor modelo, se pudo observar que en algunas distribuciones de las variables había cierto crecimiento cuadrático o de mayor grado:



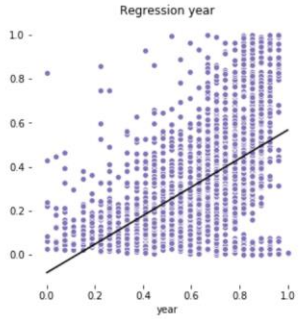
En base a estas observaciones, se procedió a utilizar features polinómicas. Una variable polinómica es una variable calculada que intenta agregar cierto nivel de descripción de mayor grado al dataset. Lo más interesante de la regresión lineal es que a pesar de ser "lineal" como lo dice su nombre, la adición de estas variables polinómicas le dan la habilidad de hacer predicciones de mayor grado que se ajusten tanto como se desee a una nube de datos. En base a la considerable dispersión que se observó en el dataset, se concluyó que se podría aumentar el grado en una cantidad relativamente grande, pero con una clara presencia de overfitting. A causa de esto se utilizó un grado cuadrático para aumentar la capacidad de predicción del modelo múltiple. Por último, se consideró aplicar regularización enfocada en la penalización de coeficientes de valores altos. Lo que la regularización hace es realizar una suma adicional a la función de costo en base a los coeficientes, esto penaliza coeficientes "complicados" o altos y esto resulta en un modelo más simple.

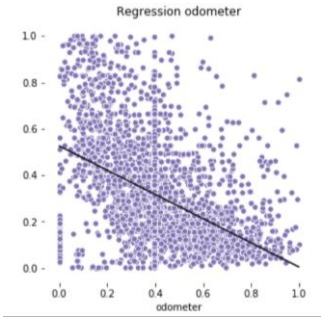


En este caso se utilizó el método de Ridge regresión, este suma el cuadrado de la suma de los coeficientes y no se enfoca directamente en la reducción de variables como lo hace el método de Lasso. Se aplicó esta regularización con un parámetro alpha de fuerza de 10.

Results

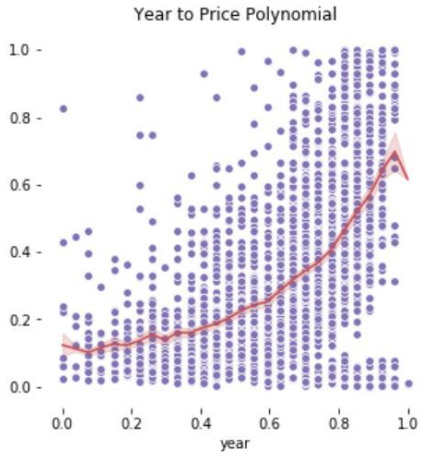
En el segmento anterior se explicó que se utilizó una métrica de rendimiento conocida como R cuadrado, por lo tanto, se presentarán los resultados en base a la misma. Como primer punto se evaluó el modelo de regresión lineal simple, la variable que logró una mejor descripción del precio fue la de "year" con un **R2 de 0.310** en comparación con el segundo mejor modelo, este utiliza la variable de "odometer" y resulta en un **R2 de 0.190**:





En cuanto al resultado de las combinaciones para el modelo de regresión lineal múltiple, se obtuvo que las variables de “years”, “cylinders” y “odometer” en conjunto describen a la variable de precio de una mejor manera. Este modelo obtuvo un valor de **R2 de 0.475 y un valor de R2 ajustado de 0.474**. La métrica de R2 ajustado es muy similar a la de R2 convencional, sin embargo, esta penaliza la inclusión de mayor cantidad de variables que no necesariamente aportan al modelo. En este caso se puede ver que las dos métricas son muy similares, por lo que no se puede establecer que la cantidad de variables involucradas sea mayor de lo óptimo.

Lo siguiente que se evaluó fue la diferencia de añadir las variables polinómicas con grado cuadrático, afortunadamente mejoró el modelo con un aumento de la métrica de **R2 a 0.519**:



El modelo que incluye las tres variables más descriptivas en conjunto con las features polinomiales de grado 2 fue el que mejor rendimiento tuvo para la predicción del precio en este dataset específico. No obstante, se puede declarar que la métrica a la que se llegó no es particularmente la deseada. Esto es porque la métrica de R2 representa un mejor modelo al acercarse al valor de 1 o 100%, sin embargo, no se logró llegar a un valor considerablemente grande.

Conclusions

A lo largo del presente reporte se describieron los métodos y procesos involucrados en el análisis de un dataset de ventas de carros usados en la plataforma de craiglist.org; además, se decretó que se deseaba predecir el precio de carros usados en base a información básica de los mismos. En base a la utilización de distintos métodos de predicción, todos basados en la regresión lineal, se descubrió que sí hay cierta

relación entre algunas variables y específicamente con la variable de precio de los carros. Sin embargo, por la manera en la que se distribuyen las observaciones y las métricas obtenidas, no se puede especificar directamente que alguna de estas relaciones sea explícitamente lineal. Respecto a los resultados obtenidos, el modelo ganador fue el que utilizó las variables de “year”, “cylinders” y “odometer” en conjunto con variables polinomiales de grado 2. Este grado de los features calculados se escogió arbitrariamente después de identificar la forma en las visualizaciones, además, cualquier grado mayor pudo haber aumentado la métrica o disminuido el error, pero esto estaría acompañado de una menor capacidad de generalización para datos desconocidos. Vale la pena mencionar que en base al mismo argumento de la dispersión de la nube de datos y el modelo polinomial que se evaluó, la regularización no aporta un beneficio lógico. Esto es porque al querer penalizar los coeficientes de mayor valor, se estaría manipulando el modelo hacia la representación de la regresión lineal múltiple de grado 1.

En cuanto al análisis y las predicciones futuras (para este y otros datasets), se recomienda recolectar la información de manera automatizada. Si bien es cierto el scrapper es un método automático de recolección, la fuente de datos es puramente manual y esto introduce una cantidad considerable de error humano que puede llegar a afectar el dataset y sus características primordiales. Asimismo, para el análisis de este dataset específico, se recomienda tomar en cuenta los demás estados del país. En este caso solo se obtuvieron datos de Florida, sin embargo, podría ser interesante obtener una muestra de cada uno de los estados y hacer el análisis a nivel del país. Al momento de hacer la selección de variables para la predicción, se le dio la prioridad a las variables numéricas y a las categóricas ordinales, para el futuro podría ser útil usar las demás variables e incluso tomar en cuenta otros modelos para la predicción del precio.

Appendix

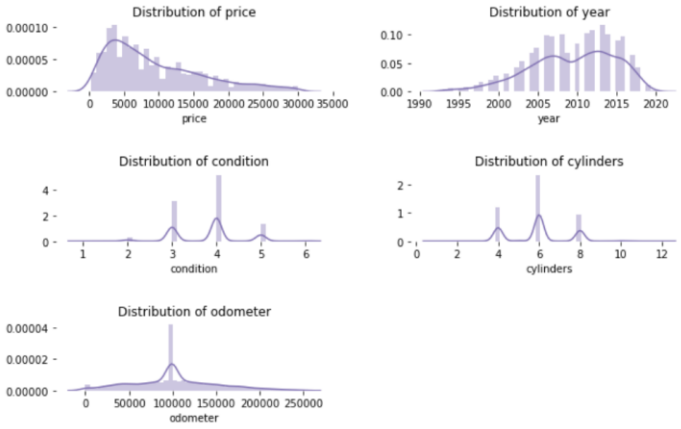
Anexo No. 1

Información Básica del Total de Variables

```
Data columns (total 25 columns):
id              509577 non-null int64
url             509577 non-null object
region          509577 non-null object
region_url      509577 non-null object
price           509577 non-null int64
year            508050 non-null float64
manufacturer    486813 non-null object
model           501588 non-null object
condition       277643 non-null object
cylinders       309894 non-null object
fuel            505592 non-null object
odometer        417253 non-null float64
title_status    506515 non-null object
transmission    505858 non-null object
vin             302152 non-null object
drive           365434 non-null object
size            167574 non-null object
type            368046 non-null object
paint_color     344871 non-null object
image_url       509563 non-null object
description     509561 non-null object
county          0 non-null float64
state           509577 non-null object
lat             499285 non-null float64
long            499285 non-null float64
dtypes: float64(5), int64(2), object(18)
```

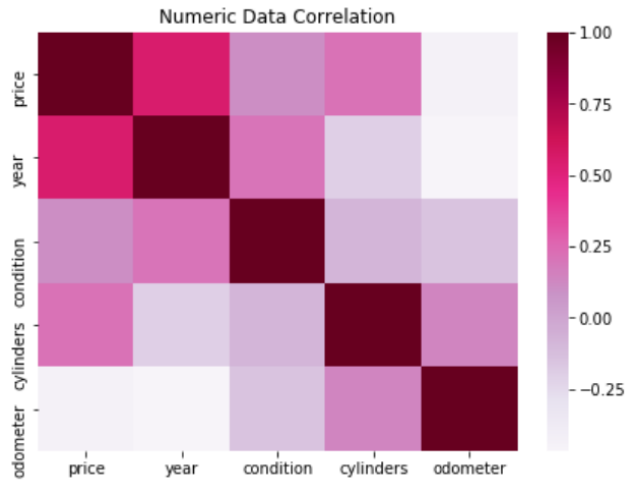
Anexo No. 2

Distribución de Variables Numéricas



Anexo No. 3

Correlación de Variables Seleccionadas



Anexo No. 4

Dataset Original

https://www.kaggle.com/austinreese/craigslist-carstrucks-data