Reporte de proyecto: Valuación de diamantes

# Abstract

Debido a un robo de diamantes en Gringotts Wizarding Bank se nos ha encomendado realizar una valuación de dichas gemas. Para realizar la valuación contaremos con información sobre las características físicas de dichos diamantes. Por lo tanto hemos construido un modelo de regresión a partir de la siguiente base de datos  [Datasets](https://nam04.safelinks.protection.outlook.com/?url=https%3A%2F%2Fs3.us-east-2.amazonaws.com%2Frecruiting-datasets%2Fdiamonds.zip&data=04%7C01%7Cdvelozso%40itam.mx%7C2e302c0df48148b6135f08d98531824a%7Cbc5eb99989e34663b72f13c503ec23cc%7C0%7C0%7C637687270550289439%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000&sdata=Z5qSPoMBlMrjFuUEiAYLAs9hcYAM4jhOMbIBwvoPDng%3D&reserved=0). Usando este modelo seremos capaces de otorgar una estimación de los valores perdidos.

# Desarrollo

Primero se hizo un análisis de nuestros datos disponibles. A continuación, enlistamos los descubrimientos más significativos.

* Contamos con variables tanto numéricas como categóricas.
* Algunos registros de las variables de dimensión (‘x’, ‘y’, ‘z’) tienen valores incoherentes (valores iguales o menores que 0).
* Existen variables de posición geográfica en nuestra base de datos. No encontramos una razón por la cual el posicionamiento geográfico influiría en el precio.

Posteriormente, realizamos el procesamiento y transformación de datos. A continuación, enlistamos los eventos importantes.

* Luego de un análisis de correlación decidimos tirar las variables geográficas.
* Los registros con datos faltantes fueron desechados debido a la pequeña dimensión de estos mismos.
* Se decidió desechar los registros con alguna dimensión igual a 0.
* Se tomo el valor absoluto de las dimensiones con valor negativo (suponemos un error al momento de introducir los datos).
* Se realizó una codificación “one-hot” de las variables categóricas
* Se hizo un escalamiento de las variables aleatorias.
* Se transformo mediante la fórmula ‘Box-Cox’ los precios de los diamantes (para calcular las predicciones reales, tenemos que .

Luego hemos seleccionado nuestro modelo y nuestro algoritmo de evaluación. Seleccionamos un árbol XGBoost y lo evaluamos mediante validación cruzada (K- folds) de 5 cortes. Utilizamos la métrica MSE (error cuadrado medio). En nuestra experiencia XGBoost es rápido y sencillo de implementar. El error cuadrado fue seleccionado para penalizar fuertemente errores grandes en la valuación

Posteriormente, realizamos una sencilla búsqueda aleatoria de los parámetros con mejor desempeño.

Finalmente, usando nuestro mejor estimador realizamos las predicciones de todo nuestro conjunto de datos y aplicamos la transformación inversa de ‘Box-Cox’ para revisar nuestro desempeño real.

# Conclusiones

Inicialmente nuestro modelo tuvo un error cuadrado medio de 0.11, tomando en cuenta que los valores transformados de precios se encuentran en el rango [4.79, 7.19] podemos concluir que tenemos un desempeño adecuado.

Después de una búsqueda de 15 minutos (20 iteraciones) encontramos un estimador con un MSE de 0.091. Una mejora considerable. Transformando mediante la inversa de ‘Box-Cox’ tenemos que

* MSE = 238,704.09
* MAE = 257.33

En general consideramos que el desempeño es más que aceptable.