Business understanding:

Tesla, Inc. es una empresa líder en innovación tecnológica y sostenible, reconocida por revolucionar la industria automotriz con vehículos eléctricos de alto rendimiento y soluciones de energía renovable.

En este proyecto se van a obtener datos de 3 fuentes de datos sobre distintos modelos de coches de la marca Tesla. La idea es obtener los datos mediante webscrapping y después usar esos datos para entrenar un modelo de regresión para intentar obtener el precio óptimo para un vehículo dependiendo del modelo, el color y el kilometraje.

Es importante que el modelo indique el mejor valor posible teniendo en cuenta las características para ajustar el precio lo mejor posible al precio de mercado del vehículo.

Requisitos:

Obtener los datos de necesarios de 3 fuentes distintas de datos.

El modelo tiene que obtener el precio de un vehículo dependiendo del modelo, el kilometraje y el color.

Los datos tienen que ser almacenados en una base de datos.

Restricciones:

Los datos tienen que obtenerse de fuentes abiertas de datos.

Los datos no pueden ser generados aleatoriamente.

Data understanding

Los datos que se van a utilizar para el proyecto son:

-Modelo: Modelo del vehículo tesla.

-Precio: Precio del vehículo.

-Color: Color del vehículo.

-Kilometraje: Km hechos por el vehículo. En caso de que no aparezca dicho dato se tomará 0 de referencia.

Las paginas de las que se van a obtener los datos son:

-Página oficial de Tesla.

-Página teslahunt.io

-Página Autocasion.com

-Página Autoscout24.com

Los algoritmos de regresión utilizados en el proyecto son:

-Decision tree regressor.

-Random forest regressor.

Preparación de los datos:

Para preparar los datos, se han obtenido todos los datos mediante webscrapping de las webs anteriormente indicadas. Y se han almacenado los datos en bruto en archivos .csv. Uno por cada página.

Posterior a la recogida de datos se procede a su correcta limpieza. Se cambian los datos de kilometraje y precio a valores enteros, puesto que al extraerlos de texto plano se encontraban con un formato y un tipo incorrecto. Posteriormente se estandariza el nombre de los modelos de vehículos obtenidos y se divide el nombre para obtener también el año de creación del vehículo.

Una vez realizadas las siguientes modificaciones se procede a estandarizar los colores de los vehículos. Puesto que dependiendo de la pagina los nombres pueden variar mucho entre variedades de color se ha optado por reducir los colores a los siguientes:

-Blanco.

-Negro.

-Gris.

-Rojo.

-Azul.

-Marrón.

Modelo:

-Decision tree regressor:

Mejor resultado:

'max\_depth': None

'min\_samples\_split': 2

'min\_samples\_leaf': 1

'Mean Squared Error': 40397427.61608545

'Root Mean Squared Error': 6355.897074063224

'Coeficiente de determinación': 0.9045229775671626

-Random forest regressor:

Mejor resultado:

n\_estimators': 50,

'max\_depth': None

'min\_samples\_split': 2

'min\_samples\_leaf': 1

'max\_features': 'sqrt'

'MSE': 36661374.55027225

'RMSE': 6054.863710297058

'R²': 0.9133529264877934

Conclusión:

Los datos arrojados con ambos modelos son bastante prometedores, pero aun así el root mean squared error sale relativamente alto. Aproximadamente un 15% del precio medio del vehículo. Entendemos que esa diferencia en el precio se basa en los distintos acabados interiores, extras que pueda tener el coche, colores específicos y estado de conservación general del vehiculo.