Manuel Soto Romero

Informe de Análisis del Problema y Resultados de los Modelos de Regresión

Análisis del Problema

El problema planteado se centra en la predicción del precio de vehículos en función de varias características que incluyen:

- **Year** (Año de fabricación)
- **Mileage** (Kilometraje)
- **EngineSize** (Tamaño del motor)
- **FuelType** (Tipo de combustible)
- **Transmission** (Tipo de transmisión)
- **Horsepower** (Potencia del motor)
- **Price** (Precio del vehículo, que es la variable objetivo)

El conjunto de datos utilizado contiene 15,000 registros y 8 columnas. Estas columnas incluyen tanto variables numéricas como categóricas. El objetivo es predecir el **Precio** del vehículo (variable dependiente) basándose en las otras características (variables independientes).

Análisis de los Datos

Antes de aplicar los modelos, se realizó un análisis exploratorio de los datos para obtener una visión general de las características. Las estadísticas descriptivas del conjunto de datos son las siguientes:

- **Año (Year)**: Los vehículos varían desde el año 2000 hasta el 2022, con una media alrededor de 2011.
- **Kilometraje (Mileage)**: El kilometraje promedio es de 129,700 km, con un rango que va desde 10,016 km hasta 249,990 km.
- **Tamaño del Motor (EngineSize)**: El tamaño promedio del motor es de 3.5 litros, con un rango de 1 a 6 litros.
- **Potencia (Horsepower)**: La potencia promedio de los vehículos es de 274 caballos de fuerza, con un rango que va desde 50 hasta 499 caballos de fuerza.
- **Precio (Price)**: El precio promedio de los vehículos es de aproximadamente 52,283, con un mínimo de 5,002 y un máximo de 99,998.

Se observó que algunas variables, como el **FuelType** y **Transmission**, son categóricas, lo que requirió una codificación para su uso en los modelos de regresión.

Modelos Aplicados

Se aplicaron cuatro modelos de regresión para predecir el precio de los vehículos:

1. **Regresión Lineal**: La regresión lineal es el modelo básico de predicción que asume una relación lineal entre las variables predictoras y la variable objetivo. Los resultados obtenidos fueron:

    - **MSE (Error Cuadrático Medio)**: 0.4944
    - **MAE (Error Absoluto Medio)**: 0.5708
    - **R² (Coeficiente de Determinación)**: -0.0011
2. **Regresión Polinómica (grado 2)**: La regresión polinómica intenta capturar relaciones no lineales entre las variables, añadiendo términos de grado superior en el modelo. Los resultados obtenidos fueron:
    - **MSE**: 0.4955
    - **MAE**: 0.5697
    - **R²**: -0.0032
3. **Regresión Ridge**: La regresión Ridge es una variante de la regresión lineal que utiliza regularización para reducir el sobreajuste (overfitting). Los resultados obtenidos fueron:
    - **MSE**: 0.4944
    - **MAE**: 0.5708
    - **R²**: -0.0011
4. **Regresión Lasso**: La regresión Lasso, al igual que la regresión Ridge, también utiliza regularización, pero además realiza una selección de características, estableciendo algunos coeficientes de variables a cero. Los resultados obtenidos fueron:
    - **MSE**: 0.4941
    - **MAE**: 0.5705
    - **R²**: -0.0004

Interpretación de los Resultados

- **MSE (Error Cuadrático Medio)**: Este valor indica qué tan cerca están las predicciones de los valores reales. Cuanto más bajo sea el MSE, mejor es el modelo en términos de precisión de predicción. En todos los modelos, los valores de MSE son similares, indicando que ninguno de los modelos tiene una ventaja significativa en términos de precisión.
- **MAE (Error Absoluto Medio)**: Similar al MSE, pero con menos penalización para los errores grandes. Los valores de MAE también son muy cercanos entre sí, lo que sugiere que los modelos no son capaces de predecir con gran exactitud el precio de los vehículos.
- **R² (Coeficiente de Determinación)**: Este valor mide la proporción de la varianza en el precio que es explicada por el modelo. Un valor de R² cercano a 1 indica un buen modelo, mientras que valores negativos o cercanos a cero indican un mal modelo. Todos los modelos mostraron un R² muy bajo o negativo, lo que significa que ninguno de los modelos es capaz de capturar adecuadamente la relación entre las variables predictoras y el precio de los vehículos.

Conclusión

Los resultados obtenidos sugieren que los modelos de regresión aplicados no logran predecir de manera efectiva el precio de los vehículos. El valor de **R²** para todos los modelos es bajo (incluso negativo), lo que indica que no hay una relación clara entre las variables utilizadas y el precio de los vehículos.

Manuel Soto Romero

El **mejor modelo** fue la **Regresión Lasso**, con un R^2 de **-0.0004**, que es ligeramente mejor que los otros modelos, pero aún así presenta un rendimiento muy bajo. Este bajo rendimiento puede deberse a la falta de características relevantes en los datos, la presencia de valores atípicos o la complejidad inherente del problema, lo que hace que los modelos lineales y polinomiales no sean suficientes para capturar la variabilidad del precio de los vehículos.

Es recomendable explorar nuevas características, ajustar el preprocesamiento de los datos (por ejemplo, manejando valores atípicos y realizando transformaciones adecuadas), y considerar modelos más complejos o técnicas de aprendizaje automático no lineales para mejorar el rendimiento en la predicción del precio de los vehículos.