Laboratorio #4

Ejercicio #1: Para el dataset house_prices.csv (visto en clase) adjunto a este laboratorio implemente todas las operaciones de ingeniería de características vistas en clase. En clase usamos algunas de las variables para ejemplificar como realizar las operaciones, ya que estamos usando el mismo dataset puede replicar dichas operaciones, pero deberá completar el dataset completo, es decir realizar ingeniería de características para las demás columnas que no tratamos en clase.

Para desarrollar este ejercicio deberá justificar por que seleccionó e implementó dicha operación, recuerde que esta decisión puede ir orientada a los siguientes puntos:

1) La operación no supone una deformación significativa en la distribución de densidad de la variable original.
2) La operación no perjudica la relación entre la variable modificada y la variable target.

Si usted considera necesario desechar alguna variable por que no genera beneficios o una relación adecuada puede hacerlo.

A continuación, se muestra el listado y orden de las operaciones que se deben realizar:

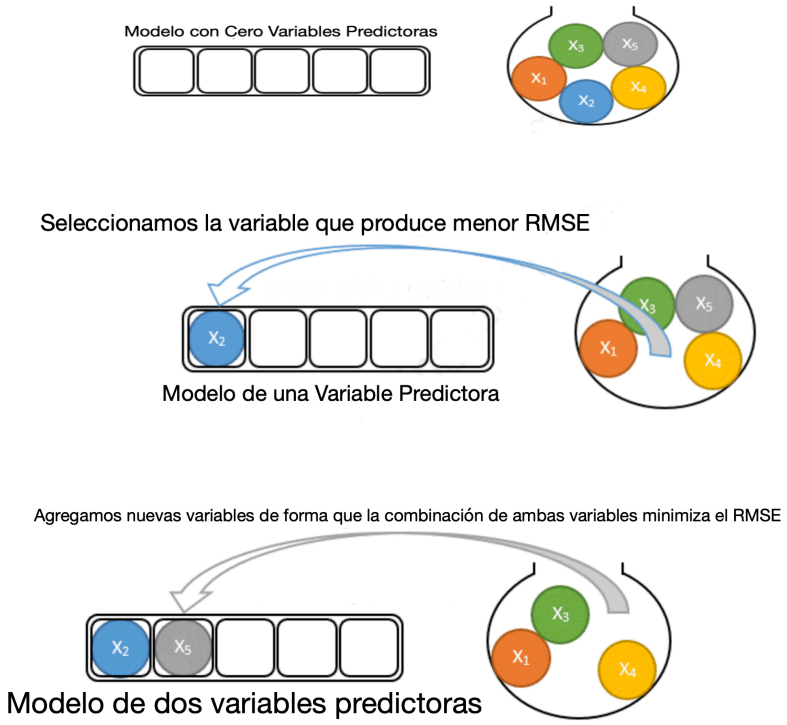
1) Imputación de data faltante,
2) Codificación de variables categóricas,
3) Manejo de Outliers,
4) Transformación de variables,
5) Feature scaling.

Su solución debe contener el notebook con todo el procedimiento y decisiones justificadas y el dataset resultante en formato .csv.

Ejercicio #2: Para este ejercicio deberá implementar una función que permita seleccionar un modelo de regresión de varias variables, utilizando el algoritmo Fordward Selection Regression. Considere los siguientes detalles para su implementación:

1) Considere M_0 como un modelo de cero variables predictoras,
2) Para $i = 1, \dots, p$ donde p denota la cantidad de variables disponibles en el dataset.
    a. Seleccione el modelo M_i como aquel modelo de i variables predictoras que producen el menor valor de RMSE (utilice K-Folds para hacer validación cruzada y seleccionar dicho modelo).
    b. Fije las variables del modelo M_i como base para producir el modelo M_{i+1} ,repita este procedimiento siempre y cuando el modelo M_{i+1} produzca un menor valor de RMSE.

Utilice la siguiente imagen como guía:



Para este ejercicio podrá utilizar el dataset que usted desee, recuerde que debe construir una función genérica que funcione con cualquier dataset.