PREGUNTAS A RESPONDER

Set up ambiente para testeo de los códigos

```
$ pip install -r requirements.txt
```

PREGUNTA 1:

Examine los datos de las 80 propiedades residenciales utilizando resúmenes estadísticos apropiados. ¿Los datos recopilados sobre las 80 propiedades residenciales son adecuados para el análisis? Si no es así, describa cualquier ajuste que deba realizarse.

RAZONAMIENTO:

> *Se hace un resumen estadístico de los datos, del se aprecia que los ouliers presentes los cuales corresponden a terrenos en los cuales no tienen casa y tienen una tasa de impuestos fija.*

- Se hace el filtro para eliminar aquellos TAX > 150, en la tarea de regresión no serán utilizados.

  > *Quedan aquellos en los que tienen casa y además presentan presentan un beneficio*

- Se opera sobre los casos que obtuvieron beneficio por vivir en el hogar y estar jubilado de manera de normalizar las tasas y que dependan únicamente de los pies de superficie. Se sigue la logica siguiente: 1) --> Si recibe ambos beneficios, lo que debería pagar sería TAX / 0.8 2) --> Si recibe solo un beneficio, lo que debería pagar sería TAX / 0.9 3) --> Si no recibe beneficios, lo que debería pagar sería TAX
- Con esto se logra un impuesto únicamente dependiente de los pies de superficie que contenga la casa

PREGUNTA 2:

Con base en los datos recopilados, ¿las cifras de pies cuadrados en la base de datos de RAD son razonablemente precisas? Explique su respuesta y respalde sus conclusiones a partir de los datos.

RAZONAMIENTO:

- Del enunciado: En el transcurso de los últimos dos meses, el tasador del personal ha examinado las casas en estas propiedades y sus planos para llegar a una estimación precisa de los pies. Entonces la columna ACT_SQFT presenta un medición muy precisa de los metros cuadrados
- Del enunciado la muestra es ALEATORIA (variables aleatorias independientes e idénticamente distribuidas) DIST Normal (podemos usar t. central del límite para intervalos de confianza)
- Dado los intervalos de confianza tanto de la columna DIFF como DIFF_ABS (diferencia entre BD y medición actual) como del error relativo son son muy bajas, utlizando un intervalo de confianza al 99%. Error absoluto = 1.606 (+-) 0.289 % --> máx error 3.779 % Error medio = 30.803 (+-) 6.068

- Considerando como un resultado imprciso aquellos con un eror absoluto mayor al 5 %, podemos decir que las mediciones son precisias.

PREGUNTA 3:

Realice un análisis estadístico apropiado para determinar si el promedio de pies cuadrados registrados en toda la base de datos RAD es significativamente diferente del promedio de pies cuadrados reales de todas las casas representadas en la base de datos. ¿Qué recomendación le haría a James Bradford basándose en sus respuestas a esta pregunta y a la pregunta anterior? Explique.

RAZONAMIENTO:

- Del enunciado: En el transcurso de los últimos dos meses, el tasador del personal ha examinado las casas en estas propiedades y sus planos para llegar a una estimación precisa de los pies. Entonces la columna ACT_SQFT presenta un medición muy precisa de los metros cuadrados
- Se estudia la distribución de las muestras, utilizando qqplot que muestra los cuantiles de la normal estimada vs los datos reales, vemos que en ambos casos, la distribución es muy cercana a la normal estimado, por lo que podemos hacer esta suposición para avanzar con un t-test sobre la columna de la base de datos.
- Al ser la medición actual muy precisa y una muestra aleatoria se estima como promedio esperado el de la variable ACT_SQFT
- Se realiza un ttest donde se tienen las siguientes hipotesis:

  > *H0: μ = promedio ACT_SFT (el promedio es este mostrado) HA: μ ≠promedio ACT_SFT (el promedio NO ES este mostrado)*

- Se considera un alpha < 0.05 para aceptar el test, lo cúal no es lo ocurrido.
- La conclusión es que no tenemos suficiente evidencia para aceptar la hipotesis de la muestra de base de datos su media es igual a la de la real
- Mi recomendación en base a lo anterior, sería utilizar solo la muestra de la variable ACT_SQFT y despreciar la estimación de la base de datos para construir los nuevos criterios de impuestos a las casas.
- Desde la mirada del negocio podría empezar un nuevo nuevo proceso de medición de los terrenos

PREGUNTA 4:

Usar una técnica estadística apropiada con los datos para determinar los valores de los componentes de impuestos fijos y variables (con respecto a los pies cuadrados) para propiedades mejoradas en el condado de Ryder. Explique cómo su enfoque para estimar estos componentes satisfará las inquietudes de James Bradford acerca de que el nuevo sistema tributario sea "neutral en cuanto a los ingresos" (es decir, que mantenga los impuestos a la propiedad casi iguales a los del sistema actual, tanto a nivel de propiedad individual como de base total)

RAZONAMIENTO, RESULTADOS E INQUETUDES:

- La regresión de lazo es un tipo popular de regresión lineal regularizada que incluye una penalización L1. Esto tiene el efecto de reducir los coeficientes para aquellas variables de entrada que no contribuyen mucho a la tarea de predicción. Esta penalización permite que algunos valores de coeficientes lleguen al valor de cero, lo que permite eliminar de manera efectiva las variables de entrada del modelo, en este caso es útil dado que nos servirá para encontrar mejor el mejor coeficiente para la cantidad fija, al hacer tender a cero la feature en el caso de necesitarlo

- Para determinar entre mejores modelos una vez realizado el hyper tunning se divide en dos conjuntos de datos (train, test)
- Del testing salen los mejores modelos
- Del script 4question_model.py salen los coeficientes que fueron encontrados por la regresión, los cuales son:

  > *[0.5436, 340.3796]: Donde 0.5436 es el coeficiente variable y 340.3796 la tasa fija*

- Resumen:

  > *Propiedad mejorada: 0.5436 * ACT_SQFT + 340.3796 Propiedad mejorada y vive ahí: (0.5436 * ACT_SQFT + 340.3796) * 0.9 Propiedad mejorada y esta jubilido ahí: (0.5436 * ACT_SQFT + 340.3796) * 0.9 Propiedad mejorada, vive ahí y esta jubilado: (0.5436 * ACT_SQFT + 340.3796) * 0.8 Impuesto fijo: 340.3796*

- Con respecto a la mantener el mismo balance se tiene:

  > *Suma de la predicción: 102489.34250000001 Suma de variable real FIXED_TAX: 102489.3425 Con un error porcentual de 1.419847e-14 %, puede estar tranquilo que no afectara el balance global*

- Con respecto a la mantener el mismo impuesto para cada uno:

  > *El error mae del modole fue de $ 135 La desviación típica del error fue de $ 48.2 El máximo error fue de $ 192 El mse fue de 7337.11 $^2 error absoluto medio de 5.41 %, con un máximo de 16% el que peor se verá afectado. No hay problemas con esta condición.*

PREGUNTA 5:

¿Dónde hubo datos inusuales que tuvieron un gran efecto en sus estimaciones de los componentes de impuestos fijos y variables? ¿Cómo manejó estos datos influyentes en su análisis? Proporcione la lógica de apoyo detrás de su enfoque para manejar estos datos influyentes.

RAZONAMIENTO

- Se resolvió en el proceso de cleaning de los datos, en los cuales se se eliminó aquellos casos que pagaban solo una tasa fija, dado que iban a ser un problema para la regresión.
- Aquellas personas que habían recibido beneficio, se debe saber cuanto deberían pagar sin el beneficio, para ellos, se normalizaron todos impuestos de la columna TAX, a una columna FIXED_TAX, en la cual aquellas casos que habían recibido un descuento por algún beneficio se colocó lo que deberían haber pagado si no hubieran recibido este beneficio y hibiesen pagado únicamente por propiedad mejorarda y los pies cuadrados de la construcción. Con esto teníamos el verdadero precio por metro cuadrado según la columna ACT_SQFT que es la (medición más certera sobre la propiedad) en la cual se paga un precio fijo + una cantidad variable por estos metros.
- De esta manera se logra un procedimiento justo.
- Dejar el proceso de optimización LASSO, más un hyper tunning de la regresión encontrar los coeficientes.

COMENTARIOS:

- Muchas gracias !
- Faltó deep learning en las preguntas xdd (comentario pasado a pelicula).