Implementación de modelos lineales y random forests para la estimación de costos de hospitalización de accidentes.

Diego Herrera Malambo – malambod@utb.edu.co

Descripción del problema:

La necesidad de tener una estimación del costo promedio de la hospitalización radica en la proyección del presupuesto para la Secretaria de Salud del municipio, lo que permitirá estimar el valor del SOAT así como sus primas.

Descripción de los datos:

Link: https://www.datos.gov.co/Salud-y-Protecci-n-Social/Costos-de-la-atenci-n-hospitalaria-en-Bucaramanga-/g4vd-w4ip

Los datos son proporcionados por la Secretaria de Salud de Alcaldía Municipal de Bucaramanga

Los datos no están estructurados correctamente, tienen inconsistencias en su contenido, por ejemplo, palabras mal escritas o descripciones diferentes pero que hacen referencia a lo mismo (sinónimos).

Se dejaron columnas por fuera para este este estudio por la inconsistencia de información como la descrita anteriormente, así mismo se descartaron algunas filas que contenían nulos ya que por la naturaleza del evento podría ser que no haya sido medido o un error.

El dataset consta de 34.206 filas y 33 columnas, de los cuales después de la limpieza quedaron 23.305 filas y 13 columnas de las cuales 12 son atributos.

Metodología:

Se realizo preparación de la información de forma que sirviera para implementar modelos lineales y de random forest de machine learning en Python con las librerías Statsmodels y Sklearn.

En la preparación de la información se quitaron outliers (una sola pasada) quitando valores extremos, las variables categóricas se pasaron a dicotómicas por cada atributo, análisis de correlación de variables dejando una sola de las variables altamente correlacionadas, adicionalmente se estandarizo el dataset con el algoritmo MinMax, sobre el cual se implementaron los modelos lineales y de random forest para su posterior evaluación.

Resultados:

	r2_score	rms2
RandomForest_n=1000	0.77103	0.09270
RandomForest_n=500	0.77061	0.09278
RandomForest_n=300	0.77041	0.09283
RandomForest_n=100	0.76724	0.09346
RandomForest_n=10	0.72281	0.10200
LinearRegression(Statsmodels with k)	0.11043	0.18272
LinearRegression	0.11043	0.18272
LinearRegression(with k)	0.11043	0.18272
LinearRegression(Statsmodels with out k)	0.09706	0.18409
LinearRegression(Statsmodels best p-value)	0.09705	0.18409

El modelo que mejor se desempeño para predecir el costo fue el RandomForest con 1000 árboles. El desempeño de los modelos lineales fue muy pobre, lo que nos indica que la relación entre los factores y el valor a predecir o sigue una tendencia lineal.

A continuación, el link con todo el análisis realizado: https://github.com/dherrerambo/analisis-de-datos-avanzados/blob/master/Clase%207%20-%20Regressions/tarea/tarea-clase7-costo%20atencion%20hosp.ipynb

Propuesta:

- Hacer ingeniería sobre sobre alguna de las características presentes en la información
- Eliminar del instrumento de captura las variables altamente relacionadas
- Implementar análisis y modelos de machine learning que evalúen series de tiempo.
- Análisis mas exhaustivo sobre los outliers