Hoja de Trabajo 6.
Modelos de Regresión Logística

INTRODUCCIÓN:

Kaggle

Kaggle es una comunidad en línea de científicos de datos, propiedad de Google LLC. Permite a los usuarios encontrar y publicar conjuntos de datos, explorar y construir modelos en un entorno de ciencia de datos basado en la web, trabajar con otros científicos de datos e ingenieros de aprendizaje automático, y participar en competencias para resolver los desafíos de la ciencia de datos. Tuvo su inicio al ofrecer competencias de aprendizaje automático y ahora también ofrece una plataforma pública de datos, una mesa de trabajo basada en la nube para la ciencia de la información y educación en IA de formato corto. El 8 de marzo de 2017, Google anunció que estaban adquiriendo Kaggle.

Conjunto de datos a utilizar

https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data

Notas:

- La hoja de trabajo se realizará en los mismos grupos.
- La hoja no se calificará si no pertenece a ningún grupo de los creados en canvas para esta hoja.

INSTRUCCIONES

- Utilice el data set House Prices: Advanced Regression Techniques que comparte Kaggle. Puede usar el análisis exploratorio que hizo en hojas anteriores. Genere un informe con las explicaciones de los pasos que llevó a cabo y los resultados obtenidos en la generación y aplicación de los modelos. Recuerde que la investigación debe ser reproducible por lo que debe guardar el código que ha utilizado para resolver los ejercicios y/o cada uno de los pasos llevados a cabo si utiliza una herramienta visual.

ACTIVIDADES

1. Cree una variable dicotómica por cada una de las categorías de la variable respuesta categórica que creó en hojas anteriores. Debería tener 3 variables dicotómicas (valores 0 y 1) una que diga si la vivienda es cara o no, media o no, económica o no.
2. Use los mismos conjuntos de entrenamiento y prueba que utilizó en las hojas anteriores.
3. Elabore un modelo de regresión logística para conocer si una vivienda es cara o no, utilizando el conjunto de entrenamiento y explique los resultados a los que llega. El experimento debe ser reproducible por lo que debe fijar que los conjuntos de entrenamiento y prueba sean los mismos siempre que se ejecute el código. Use validación cruzada.

4. Analice el modelo. Determine si hay multicolinealidad en las variables, y cuáles son las que aportan al modelo, por su valor de significación. Haga un análisis de correlación de las variables del modelo y especifique si el modelo se adapta bien a los datos.
5. Utilice el modelo con el conjunto de prueba y determine la eficiencia del algoritmo para clasificar.
6. Explique si hay sobreajuste (overfitting) o no (recuerde usar para esto los errores del conjunto de prueba y de entrenamiento). Muestre las curvas de aprendizaje usando los errores de los conjuntos de entrenamiento y prueba.
7. Haga otros dos modelos cambiando las variables predictoras de acuerdo con la significación de los coeficientes en el primer modelo. Explique por qué seleccionó las variables que uso para cada modelo.
8. Haga un análisis de la eficiencia del algoritmo usando una matriz de confusión. Tenga en cuenta la efectividad, donde el algoritmo se equivocó más, donde se equivocó menos y la importancia que tienen los errores, el tiempo y la memoria consumida. Para esto último puede usar “profvis” si trabaja con R y “cProfile” en Python.
9. Determine cual de todos los modelos es mejor, puede usar AIC y BIC para esto, además de los parámetros de la matriz de confusión y los del profiler.
10. Haga un modelo de árbol de decisión, uno de Random Forest y uno de Naive Bayes usando la misma variable respuesta y los mismos predictores que el mejor de los modelos de Regresión Logística.
11. Compare la eficiencia de los 3 modelos que creó en el punto anterior y el mejor de los de regresión logística ¿Cuál se demoró más en procesar?¿Cuál se equivocó más?¿Cuál se equivocó menos?¿por qué?

EVALUACIÓN

- **(25 puntos)** Análisis de los modelos generados. Verificar si están sobreajustados o no Recuerde explicar los razonamientos.
- **(25 puntos)** Análisis de las variables a incluir en los modelos. Pruebas de normalidad, correlación, etc. (Recuerde que las variables predictoras deberían ser las mismas para poder comparar)
- **(10 puntos)** Aplicación de los modelos al conjunto de prueba.
- **(20 puntos)** Matriz de confusión de cada modelo. Explicación de los resultados obtenidos
- **(20 puntos)** Comparación entre sí de los modelos generados.

MATERIAL A ENTREGAR

- Archivo .r o .py con el código y hallazgos comentados
- Link de Google docs con las conclusiones y hallazgos encontrados. Puede usar también Jupyter Notebooks o rmd.
- Vínculo del repositorio usado para trabajar la hoja de trabajo.