Hoja de Trabajo 3.
Modelos de Regresión Lineal

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 grupos de las hojas de trabajo.
- La hoja no se calificará si no pertenece a ningún grupo de los creados en canvas para esta hoja.
- La nota individual se basará en los aportes de cada uno al trabajo grupal, por lo que deben versionar el código para poder revisar las contribuciones de cada uno.

INSTRUCCIONES

- Utilice el data set House Prices: Advanced Regression Techniques que comparte Kaggle. Debe hacer un análisis exploratorio para entender mejor los datos, sabiendo que el objetivo final es predecir los precios de las casas. Recuerde explicar bien cada uno de los hallazgos que haga. La forma más organizada de hacer un análisis exploratorio es generando ciertas preguntas de las líneas que le parece interesante investigar. Genere un informe con las explicaciones de los pasos que llevó a cabo y los resultados obtenidos. Recuerde que la investigación debe ser reproducible por lo que debe guardar el código que ha utilizado para resolver los ejercicios.

ACTIVIDADES

1. Descargue los conjuntos de datos de la plataforma kaggle.
2. Haga un análisis exploratorio extenso de los datos. Explique bien todos los hallazgos. No ponga solo gráficas y código. Debe llegar a conclusiones interesantes para poder predecir. Explique el preprocesamiento que necesitó hacer.
3. Incluya un análisis de grupos en el análisis exploratorio. Explique las características de los grupos.

4. Divida el set de datos preprocesados en dos conjuntos: Entrenamiento y prueba. Describa el criterio que usó para crear los conjuntos: número de filas de cada uno, estratificado o no, balanceado o no, etc. Si le proveen un conjunto de datos de prueba y tiene suficientes datos, tómelo como de validación, pero haga sus propios conjuntos de prueba.
5. Haga ingeniería de características, ¿qué variables cree que puedan ser mejores predictores para el precio de las casas? Explique en que basó la selección o no de las variables.
6. Todos los resultados deben ser reproducible por lo que debe fijar que los conjuntos de entrenamiento y prueba sean los mismos siempre que se ejecute el código.
7. Seleccione **una** de las variables y haga un modelo univariado de regresión lineal para predecir el precio de las casas. Analice el modelo (resumen, residuos, resultados de la predicción). Muéstrelo gráficamente.
8. Haga un modelo de regresión lineal con **todas** las variables numéricas para predecir el precio de las casas. Analice el modelo (resumen, residuos, resultados de la predicción). Muestre el modelo gráficamente.
9. Analice el modelo. Determine si hay multicolinealidad entre las variables, y cuáles son las que aportan al modelo, por su valor de significación. Haga un análisis de correlación de las características del modelo y especifique si el modelo se adapta bien a los datos. Explique si hay sobreajuste (overfitting) o no. En caso de existir sobreajuste, haga otro modelo que lo corrija.
10. Si tiene multicolinealidad o sobreajuste, haga un modelo con las variables que sean mejores predictoras del precio de las casas. Determine la calidad del modelo realizando un análisis de los residuos. Muéstrelo gráficamente.
11. Utilice cada modelo con el conjunto de prueba y determine la eficiencia del algoritmo para predecir el precio de las casas. ¿Qué tan bien lo hizo?
12. Discuta sobre la efectividad de los modelos. ¿Cuál lo hizo mejor? ¿Cuál es el mejor modelo para predecir el precio de las casas? Haga los gráficos que crea que le pueden ayudar en la discusión.

EVALUACIÓN

- **(25 puntos)** Análisis de los modelos generados, incluyendo los residuos. Recuerde explicar los razonamientos.
- **(25 puntos)** Análisis de las variables a incluir en el modelo. Pruebas de normalidad, correlación, etc.
- **(10 puntos)** Aplicación de los modelos al conjunto de prueba.
- **(20 puntos)** Explicación de los resultados obtenidos incluyendo el desempeño de los modelos.
- **(20 puntos)** Comparación entre los modelos elaborados y selección del mejor de todos para predecir el precio de las casas.

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.