Proyecto 2. Entrega 5.
Modelos de Regresión Logística

INTRODUCCIÓN:

InmoValor S.A. es una empresa innovadora del sector inmobiliario que ha apostado por la transformación digital para ofrecer valoraciones precisas y objetivas de propiedades. Ante un mercado dinámico y competitivo, la compañía ha adoptado técnicas avanzadas de análisis y modelos de regresión para estimar el valor de inmuebles basándose en un amplio conjunto de datos que recopila información detallada de viviendas. Este dataset incluye variables clave como ubicación, tamaño, calidad constructiva y otros factores determinantes, lo que permite desarrollar modelos predictivos capaces de reflejar con mayor exactitud las condiciones del mercado.

La empresa ha decidido incorporar un equipo de analistas de datos con la finalidad de trabajar con el conjunto de datos "House Prices: Advanced Regression Techniques" para desarrollar modelos predictivos que proyecten de manera precisa el precio de las viviendas. Mediante el análisis de variables clave como la ubicación, el tamaño y la calidad de las propiedades, el equipo utilizará técnicas avanzadas de regresión para mejorar la estimación de valores inmobiliarios y facilitar la toma de decisiones estratégicas en el mercado de bienes raíces.

Vínculo del conjunto de datos a utilizar

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

Resultados esperados en la Quinta Entrega:

Se espera la entrega de un informe detallado donde incluya:

- Los modelos de Regresión Logística para clasificación usando la variable dicotómica y la variable categórica de los precios que creó.
- Una comparación para estimar la categoría de los precios de las viviendas donde determine qué algoritmo funcionó mejor (Árboles de regresión, Random Forest, Bayes Ingenuo, KNN, Regresión Logística). Compare los resultados del mejor modelo de cada uno de los algoritmos. Recuerde que para establecer una comparación válida, es necesario comparar bajo las mismas condiciones. Mismo conjunto de datos de entrenamiento y de prueba.
- 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.
- Como la vez anterior, le han pedido un informe formal, que enlace los temas usando subtítulos explicativos que mantenga al lector en un hilo conductual de los temas que aborda la consultoría. No incluya los enunciados de las actividades de esta guía como subtítulos.

Notas:

- La consultoría es en grupo, por lo que solo se tendrán en cuenta los grupos conformados por más de un especialista.
- Cada individuo será evaluado de forma individual basado en sus aportes al trabajo grupal, por lo que deben versionar el código para poder revisar las contribuciones de cada uno.

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 un tuneo del modelo para determinar los mejores parámetros, recuerde que los modelos de regresión logística se pueden regularizar como los de regresión lineal.
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 regresión logística para la variable categórica para el precio de las casas (categorías: barata, media y cara). Asegúrese de tunearlo para obtener el mejor modelo posible.
11. Compare la eficiencia del modelo anterior con los de clasificación de las entregas anteriores ¿Cuál se demoró más en procesar?¿Cuál se equivocó más?¿Cuál se equivocó menos?¿por qué?

EVALUACIÓN

Notas: Tiene que poderse comprobar su aporte al trabajo grupal a través de commits. Si no existen al menos 3 commits con su aporte significativo no va a tener nota de la hoja de trabajo. Utilice una herramienta que permita registrar los aportes de cada uno.

- **(20 puntos)** Análisis de los modelos generados. Explique todos los razonamientos. Uso de las métricas correctas
- **(10 puntos).** Tuneo de parámetros. Se probó con varios valores para los hiperparámetros. Se explican los resultados.
- **(20 puntos)** Análisis de los modelos generados para la variable dicotómica. Explicación de resultados

- **(20 puntos)** Matriz de confusión de cada modelo de clasificación. Explicación de los resultados obtenidos. Selección del mejor modelo usando las métricas adecuadas. Análisis de otras métricas además del accuracy para los modelos generados.
- **(20 puntos)** Comparación del modelo de clasificación de Regresión Logística con el mejor modelo de clasificación de todas las entregas pasadas (KNN, Árbol de Decisión, Random Forest, Naive Bayes). Elección del mejor modelo de clasificación basado en las métricas adecuadas.
- **(10 puntos)** El informe de resultados está bien redactado, se establece una continuidad lógica en las explicaciones y el análisis de los modelos. Tiene un hilo conductor claro que no pierde al lector y no tiene las actividades de esta guía como subtítulos.

MATERIAL A ENTREGAR

- Archivo .rmd, .ipynb o Google docs con el informe con lo solicitado en las instrucciones
- Script de R o de Python que utilizó debidamente organizado y comentado (Si utilizó .rmd debe añadir el .html que se genera)
- Link del repositorio utilizado para la hoja de trabajo.

FECHAS DE ENTREGA

- **AVANCE:** Puntos del 1 al 6 de la sección de actividades: viernes 11 de abril
- **ENTREGA FINAL:** domingo, 13 de abril a las 23:59