Proyecto 2. Entrega 1
Modelos de Regresión Lineal

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

Descripción de la Consultoría:

La consultoría se dividirá en varias etapas. Cada semana deberá entregar resultados de la aplicación de un algoritmo de predicción y/o clasificación. La conclusión será la elección del mejor algoritmo predictivo para estimar el valor de una vivienda.

Resultados esperados en la Primera Entrega Consultoría:

Se espera la entrega de un informe detallado donde incluya el análisis exploratorio del conjunto de datos (incluyendo agrupamiento) y la evaluación de varios modelos de regresión lineal simple y múltiple. Debe seleccionar el mejor modelo de los construidos para predecir el valor de la vivienda.

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.

INSTRUCCIONES

- Utilice el data set House Prices: Advanced Regression Techniques. 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.
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 cada uno.
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. Use el conjunto de datos llamado "**train.csv**". Extraiga de ahí su subconjunto 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 reproducibles 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. 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?¿Qué medidas usó para determinar la calidad de la predicción?
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

Nota: 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 poder ser evaluado en la entrega. Utilice una herramienta que permita registrar los aportes de cada uno.

- **(25 puntos)** Análisis exploratorio de los datos. Se hizo un análisis exploratorio lo suficientemente extenso que permita explicar las variables que pudieran estar influyendo en la variable respuesta, que es el precio de las casas. Análisis de las variables a incluir en el modelo. Pruebas de normalidad, correlación, etc.
- **(25 puntos)** Análisis de los modelos generados, incluyendo los residuos. Recuerde explicar los razonamientos.
- **(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 .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 de controlador de versiones utilizado.

FECHAS DE ENTREGA

- **AVANCE:** Descripción de las variables, análisis exploratorio, análisis de relaciones con la variable respuesta: viernes 28 de febrero 23:59.
- **ENTREGA FINAL:** Domingo 2 de marzo a las 23:59

NOTA: Solo se calificará el Documento Final si está entregado el avance con todo lo que se pide.