



UNIVERSIDAD
SERGIO
ARBOLEDA

TRABAJO FINAL - MODELO GLM

MAESTRÍA EN ANALÍTICA Y GERENCIA DE DATOS II

Casos de Estudios Aplicados al Sector Seguros y Financieros en
Colombia

Por:

Yessica Paola García

Docentes:

Julian Caro
Ximena Quiroga

Fecha:

Miércoles, 27 de
noviembre de 2024

ÍNDICE

PARTE 1 - LIMPIEZA DE LOS DATOS

PARTE 2 - ELECCIÓN DE VARIABLES

PARTE 3 - MODELO GLM

PARTE 4 - MEJOR MODELO GLM

LIMPIEZA DE LOS DATOS

1

Carga y exploración de datos

- Se cargaron los datos desde el archivo 'AGUA_V2_40.parquet' en un DataFrame llamado df.
- También se cargó un diccionario de datos (dict) para entender la naturaleza de cada variable y definir el plan de limpieza.
- Comprender los datos antes de cualquier limpieza asegura que se tomen decisiones informadas sobre cómo manejar los valores faltantes y definir el tipo de cada variable.

2

Muestreo y reducción de tamaño

- Se realizó un muestreo aleatorio (sample()) del dataset completo, con una semilla fija (42) para garantizar la reproducibilidad.
- Reducir el tamaño del DataFrame ayuda a evitar problemas de memoria durante el procesamiento y facilita el análisis inicial de los datos.

3

Manejo de valores faltantes y transformaciones

- Filtro de peso: Se eliminó cualquier registro cuyo valor de peso fuera cero, ya que estos registros podrían sesgar el análisis.
- Cálculo de la variable de respuesta: Se definió la variable de respuesta como el cociente de la variable de interés sobre el peso.
- Se utilizaron distintos métodos para tratar los valores faltantes (NaN):
 - Relleno con ceros, en los casos en los que era apropiado asumir un valor nulo.
 - Relleno con la moda, cuando los valores eran categóricos y frecuentes.
 - Eliminación de variables irrelevantes.
- Tratamiento de antigüedad: Para los registros sin fecha de reforma del edificio, se usó la fecha de construcción. Luego, se agruparon los años en intervalos de 10 años para simplificar la variable.
- Estas decisiones se basaron en la naturaleza de los datos y el análisis del diccionario para garantizar la coherencia y evitar problemas de colinealidad.

ELECCIÓN DE VARIABLES

1

2

3

Conversión y agrupación de variables

- Se convirtieron variables de antigüedad en categorías (por grupos de 10 años) y se crearon variables dummy para las variables categóricas.
- Las variables codpos y codin se eliminaron tras comprobar que no tenían influencia en los resultados mediante pruebas previas.

Modelo de selección de variables con Random Forest

- Se usó un modelo de RandomForestRegressor con 100 estimadores, utilizando toda la capacidad de los procesadores disponibles (-1), para identificar la importancia de cada variable.
- La selección de características mediante Random Forest permite evaluar de manera eficiente la relevancia de las variables, reduciendo la dimensionalidad del modelo.
- Solo se seleccionaron aquellas variables con una importancia superior al 0.05%, lo cual permitió centrarse en aquellas características con un mayor impacto en la predicción.

Resultado final de selección

Las variables seleccionadas fueron aquellas que superaron el umbral de importancia en Random Forest, junto con algunas variables adicionales relacionadas con el peso de los edificios, ya que se consideró que estas podían influir en el comportamiento del seguro.

ELECCIÓN DE VARIABLES

4

- La lógica del negocio, como lo entiendo, se enfoca en cómo el estado físico de los edificios, la calidad de las instalaciones hidráulicas y el contexto histórico de las reformas influyen en el riesgo de sufrir daños relacionados con el agua. En términos generales:
 - Las variables de antigüedad (antiguedif_2, antigref_2, ANTIGUEDAD_VIVIENDA_2) representan la vulnerabilidad inherente de las instalaciones debido al paso del tiempo.
 - Las variables como CUPD_CAP_Corr_aguaacagbc y K_ACAGBC reflejan la capacidad operativa y la condición física de las instalaciones, lo cual afecta directamente la probabilidad de daños.
 - EXPOSICION describe el contexto de riesgo externo del edificio, mientras que anualidad_seguro puede relacionarse con el historial de siniestros y la capacidad para hacer frente a los daños.
- Cada una de estas variables fue seleccionada porque aporta una perspectiva clave sobre los factores de riesgo que afectan a los daños por agua en los conjuntos residenciales, lo cual ayuda a construir un modelo predictivo robusto para evaluar el riesgo de averías hidráulicas.

MODELO GLM

1

Modelo de Regresión Lineal generalizada (GLM)

- Utilizando las variables previamente seleccionadas, se construyó un modelo GLM.
- Se eligió un modelo GLM por su capacidad para modelar relaciones lineales entre variables en situaciones en las que no necesariamente se cumplen los supuestos de normalidad o linealidad. Además, facilita la interpretación de los efectos de las variables independientes sobre la variable dependiente.

2

Evaluación inicial del Modelo

- Se observaron los valores p para determinar la significancia estadística de cada variable.
- Ninguna de las variables incluidas tenía un valor p mayor a 0.05, lo cual indica que todas eran estadísticamente significativas.
- Esto sugiere que todas las variables contribuyen de manera significativa al modelo, pero aún se podía optimizar el ajuste para obtener el mejor criterio de información bayesiano (BIC).

Optimización del Modelo utilizando Stepwise

MEJOR MODELO GLM



- Se utilizó un enfoque de selección Stepwise (hacia adelante y hacia atrás) para optimizar el modelo, minimizando el criterio de información bayesiano (BIC).
- Minimizar el BIC es crucial para obtener un modelo parsimonioso, es decir, un modelo que sea lo suficientemente sencillo sin perder capacidad predictiva.



Resultado del mejor Modelo

- Durante el proceso de optimización, se determinó que eliminar la variable antigüedad_ref2 (que representaba reformas entre 46 a 51 años) mejoraba el valor del BIC.
- La eliminación de esta variable simplificó el modelo sin afectar su rendimiento predictivo, mostrando que no tenía una contribución significativa a la calidad del ajuste del modelo final.

Muchas Gracias

Email:

jessicap8010@gmail.com

