FACULTAD DE INGENIERÍA

UNIVERSIDAD DE LA REPÚBLICA URUGUAY

INTRODUCCIÓN A LA CIENCIA DE DATOS

Tarea - Final

AÑO 2024

GRUPOS 2-10:

Lucía Coudet
Francisco Galletto

FECHA: 10 de julio de 2024

Tabla de contenidos

1. Planteo del problema

El objetivo es crear un modelo que permita predecir la probabilidad de que una persona que conduce un vehículo se vea involucrada en un accidente de tránsito. Esto es especialmente relevante en el campo de los seguros, donde entender estos riesgos es fundamental para definir políticas y precios.

Para lograr esta predicción, se debe calcular el valor de una variable de salida (si ocurre o no un accidente). Este tipo de problema pertenece al campo del aprendizaje supervisado, lo que significa que el modelo se entrena con datos donde el resultado es conocido. Matemáticamente, esto se expresa como:

$$Y = f(x) + \epsilon \tag{1}$$

Aquí, Y representa la variable de salida (ocurrencia de un accidente), X son las variables de entrada (factores que podrían influir en la probabilidad de un accidente, como la edad del conductor, la velocidad, el tipo de vehículo, etc.), $f(X)$ es la función que estamos tratando de aprender, y ϵ representa el error o la variabilidad que no puede ser explicada por el modelo.

Dado que la variable de salida Y es categórica y binaria (es decir, solo tiene dos posibles valores: ocurrió un accidente o no ocurrió un accidente), estamos tratando con un problema de clasificación supervisada. Esto significa que el modelo debe aprender a clasificar correctamente los datos en una de estas dos categorías.

2. Base de datos

2.1. Variables explicativas

Las variables que, a priori, se pueden considerar como influyentes en la probabilidad de que una persona tenga un choque automotor se agrupan en las siguientes categorías:

- **Datos personales:** Edad, género, estado civil, profesión, nivel de educación, número de hijos, etc.
- **Historial de siniestralidad e historial de conducción**: Número de accidentes anteriores, infracciones de tráfico, años de experiencia al volante, uso del cinturón de seguridad, hábitos de conducción (conducción nocturna, velocidad promedio), frecuencia de uso del vehículo, etc.
- **Datos del vehículo:** Tipo de vehículo, modelo, marca y año, antigüedad del vehículo, tipo de combustible, kilometraje, historial de mantenimiento, presencia de sistemas de asistencia a la conducción (como ABS, control de estabilidad), etc.
- **Zona de circulación:** Zona donde circula habitualmente la persona con ese vehículo, nivel de tráfico en la zona, índices de criminalidad, tipo de entorno (urbano o rural), presencia de intersecciones peligrosas, etc.

- **Datos contextuales:** Condiciones meteorológicas en la zona de circulación (frecuencia de lluvias, nieve, niebla), tipos de carretera (autopistas, carreteras secundarias, caminos rurales), iluminación de las vías, infraestructura vial (presencia de semáforos, señales de tránsito), eventos especiales en la zona (obras, festividades), etc.

2.2. Obtención de la base de datos

Base de datos relacional de pólizas de seguro automotor, del asegurado, e historia siniestral de una compañía aseguradora

2.3. Calidad de los datos: datos faltantes y completitud

- **Datos faltantes:** en caso de presencia de datos faltantes, si la no respuesta es generada al azar es posible proceder a un método de imputación por la media o la mediana. En caso contrario se puede recurrir a algún método más sofisticado de imputación como por ejemplo mediante modelado.
- **Datos incompletos o con errores:** para variables categóricas revisar que no haya datos inconsistentes. Para variables cuantitativas detección trabajando por ejemplo con percentiles.
- **Datos atípicos:** para variables categóricas revisar la frecuencia de los datos, datos con poca frecuencia podría ser un dato atípico. Para variables de tipo cuantitativa se pueden detectar los atípicos utilizando los percentiles, visualización mediante un boxplot o volin plot que además permite ver la densidad de los datos.

2.4. Análisis exploratorio de datos (EDA)

En el proceso de EDA el primer paso es identificar el tipo de variables que contiene la base de datos, si son cuantitativas (discretas o contínuas) o cualitativas (nominales o ordinales).

2.4.1. Análisis de variables cuantitativas

Para las variables cuantitativas se propone:

- **Resumen de Medidas Estadísticas:** Calcular mínimo, máximo, percentiles (25, 50, 75), media, mediana, varianza, desviación estándar y rango intercuartílico.
- **Visualizaciones:**
  - **Boxplot:** Para detectar valores atípicos (outliers) y visualizar la dispersión.
  - **Violin Plot:** Para identificar la densidad de los datos y su distribución.

- **Histogramas:** Utilizar variantes como histogramas apilados y apilados al 100 % para ver la distribución de los datos y comparaciones entre subgrupos.
- **Gráficos de dispersión:** Para analizar relaciones entre dos variables cuantitativas.

2.4.2. Análisis de variables cualitativas

Para las variables cualitativas se propone:

- **Frecuencia de cada Categoría:** Calcular las frecuencias y la moda (valor más frecuente).
- **Visualizaciones:**
  - **Gráfico de Barras:** Para visualizar la frecuencia de cada categoría.
  - **Gráficos de pastel (pie charts):** Para ver la proporción de cada categoría (aunque no siempre recomendados para muchas categorías).
- **Estudio sobre el Balance de los Datos:** Determinar si hay categorías con muy poca frecuencia en comparación a otras, y evaluar si es necesario agrupar o tratar estas categorías de manera especial.

2.4.3. Detección de outliers y anomalías

- **Boxplots y Violin Plots:** Ya mencionados, pero enfatizar su uso en la detección de outliers.
- **Gráficos de Dispersión con Diferentes Subgrupos:** Para detectar patrones anómalos en subgrupos específicos.

3. Estandarización

Dado que se tienen variables medidas en diferentes escalas, se propone la estandarización de los atributos, en particular la normalización (restar la media y dividir por el desvío).

4. Conjuntos para entrenamiento y testeo

Para poder evaluar la performance predictiva del modelo se particiona la base de datos en un conjunto de entrenamiento (70 %) y conjunto de testeo (30 %). A su vez, el error sobre el conjunto de entrenamiento se medirá utilizando un procedimiento de validación cruzada, estableciendo subconjuntos efectivamente de entrenamiento y de validación. A modo de resumen, la validación cruzada consiste en dividir el conjunto de entrenamiento en k partes, donde se utilizan (k-1) partes para entrenar, y la restante para evaluar el modelo. Este proceso se repite cambiando la parte elegida.

Se devuelve el promedio del valor de performance obtenido, y también la desviación estándar de los resultados.

5. Modelos a tener en cuenta

Una vez definido el problema y habiendo hecho el análisis descriptivo sobre la base de datos, se definen las técnicas estadísticas a utilizar para entrenar el modelo en el conjunto de entrenamiento. En primer lugar y a modo descriptivo, se propone implementar un árbol de decisión. Si bien la principal desventaja de este algoritmo es la inestabilidad, suelen ser útiles para saber cuáles son las variables más influyentes del análisis en caso que las haya (quedan en las primeras particiones). Luego se propone implementar los siguientes modelos:

- Regresión logística
- Random Forest
- Support vector machine
- Redes neuronales (con función de activación sigmoide)

6. Selección del mejor modelo

La selección del mejor modelo se realizará mediante validación cruzada. Por otra parte, para entrenar los modelos es necesario establecer valores para los hiperparámetros como la cantidad de árboles a ajustar en el algoritmo de Random Forest o la cantidad de variables seleccionadas al azar en cada partición. Se propone establecer un rango de valores para los mismos e ir viendo como impactan en el error del modelo (ver como cambia la predicción). Una vez seleccionado el mejor modelo según el error en el conjunto de validación de la muestra de entrenamiento, se vuelve a entrenar el modelo sobre todo el conjunto de entrenamiento (70La métricas utilizadas para evaluar al modelo son el Accuracy, la Precisión, el Recall y la Matriz de Confusión.

7. Interpretabilidad

Finalmente, si es posible, se le tratará de dar interpretabilidad al modelo, respondiendo a distintos tipos de preguntas.

7.1. Variables influyentes

- ¿Qué variables son las más influyentes a la hora de determinar la probabilidad de choque de un individuo?
- ¿Cómo contribuyen estas variables a la predicción? (por ejemplo, ¿incrementan o disminuyen la probabilidad de choque?)

7.2. Relación entre variables

- ¿Cómo se relacionan las covariables entre ellas?
- ¿Existen interacciones importantes entre algunas variables que afecten la probabilidad de choque?
- ¿Cómo varía la importancia de una variable en función de otra?

7.3. Peso de variables categóricas

- ¿Qué categorías específicas dentro de una variable (como tipo de vehículo o zona de circulación) están asociadas con una mayor o menor probabilidad de choque?
- ¿Hay diferencias significativas en la probabilidad de choque entre diferentes grupos demográficos?

8. Trabajo futuro

Se deberan identificar los problemas y limitaciones presentes en el trabajo actual para abordarlos y mejorar los modelos predictivos utilizados. Además, se deberá exporar posibles ampliaciones y profundizaciones de este estudio, considerando nuevas variables y técnicas avanzadas de análisis, como los los modelos de Redes Neuronales Convolucionales o Recurrentes. Este proceso no solo contribuirá a mejorar la precisión y la interpretabilidad del modelo actual, sino que también podría ser aplicable a otros problemas similares en diferentes contextos, ampliando su utilidad y relevancia.