1

Buenos Días, hoy les presentaremos nuestra solución al caso ¿Seguro para su vehículo?

En este caso nos piden encontrar un algoritmo que nos muestre que clientes que actualmente tienen un seguro de vida desean comprar un seguro para su vehículo

Para resolver esto nos presentaron una base de datos en la cual indica muchas características de personas a las cuales ya se les ofreció este seguro y cómo fue su respuesta

2

Este caso nos presenta las siguientes variables, las más importantes son el Genero de la persona, su edad, si cuentan con licencia de conducir, el código de la región en la que se encuentran, si su vehículo estuvo previamente asegurado, la edad del vehículo, si el vehículo está dañado, cuanto pagaría de prima anual por su vehículo, el canal por el cual se comunicaron con el cliente, la antigüedad dentro de la empresa para este cliente y si adquirió o no el seguro de vehículo

Con estos datos se encontraron las siguientes relaciones interesantes

3

Se encontró que en un mayor número de hombres que mujeres decidieron tomar el seguro, que más vehículos de entre 1 y 2 años adquirieron el seguro que el resto de vehículos, pero que es más grande la proporción los que obtuvieron el seguro con un vehículo mayor de 2 años que el resto menores a ellos, y que

4

Al observar la matríz de correlación encontramos que las relaciones más importantes con la variable de respuesta son la edad, si tenían seguro previamente, el canal de ventas, el género, la edad del vehículo y si este esta previamente dañado

5

Con esto definimos como hipótesis que Las variables **Age, Previoulsy Ensured, Sale Channel, Vehicle Age, Vehicle Damage y Gender** tienen una relación que explica la variable **Response** mejor que todas las variables del dataset.

Trataremos de corroborar o negar esta hipótesis para encontrar el mejor modelo

6

A continuación planteamos unas preguntas para encontrar las relaciones entre las previas variables.

Aquí vemos que que las personas que son más precavidas y adquieren el seguro son, en promedio, de una edad mayor

7

Con esta gráfica podemos ver que los vehículos de edad entre 1 y 2 años adquieren más el seguro, pero una mayor proporción de los vehículos mayores de 2 años adquieren el seguro en total

8

Aquí podemos ver que en su mayoría las personas que ya han tenido daños a su vehículo son más propensas a tomar el seguro

9

Con esta gráfica encontramos que no, que las mujeres no son más propensas a tomar este seguro, sino los hombres

10

Con estas relaciones y variables escogidas realizamos nuestros modelos y decidimos que el Modelo 1 tendrá todas las variables del dataset, mientras que el Modelo 2 tendrá las variables que definimos en nuestra hipótesis

Con esto encontramos los siguientes valores de Accuracy. Esto indica que tan propensos son nuestro modelos en predecir la decisión de la persona.

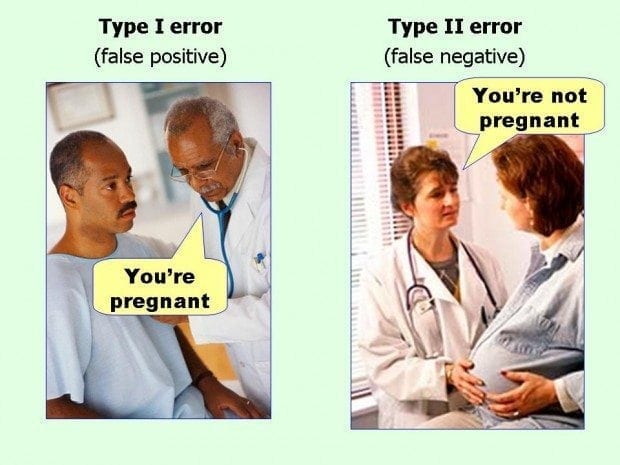
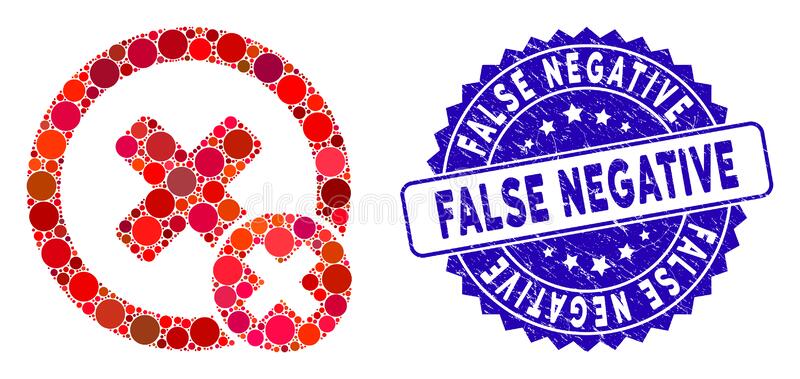
11

El precisión indica un error de falso positivo, es decir que le ofreceremos el seguro a una persona que nuestro modelo indica que lo desea, cuando en realidad no lo desea

12

El recall indica un error de falso negativo, es decir que el modelo nos indica que una persona no desea el seguro, cuando en realidad si lo desea

13

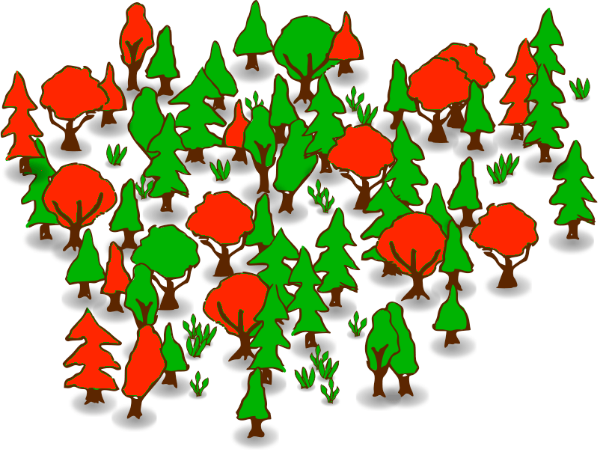
Para seleccionar nuestro buscamos el modelo con una mejor Accuracy, ya que esto muestra que tan acertado será el modelo. Como segundo punto buscamos un modelo con un el Recall más alto ya que con poco porcentaje de respuesta positiva no queremos perder a ningún cliente. Por último el Precision no es tan influyente ya que si una persona no desea adquirir el seguro actualmente es posible que alguien que conozca o en el futuro si lo desee y consideramos que compartir esta información no está de mas.  
  
  

14

Nuestro mejor modelo es el modelo Random Forest 1, utilizando todas la variables del modelo. Este modelo es:  
#2 en Accuracy

#1 en Recall

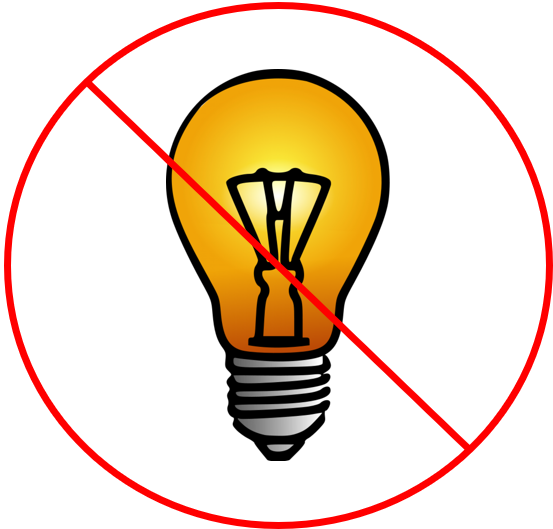
#1 en Precision



15

Conclusiones:  
Nuestra hipótesis que las variables Age, Previoulsy Ensured, Sale Channel, Vehicle Age, Vehicle Damage y Gender son más predictores que todas las variables del dataset fue refutada, por lo que alguna relación la cual no tomamos en cuenta está en juego entre estas variables

Se utilizáron correctamente los modelos de Naive Bayes, Arbol de Decision, Random Forest y Regresión Logística

Se encontró un modelo, Random Forest 1 con todas las variables, que con un alto Accuracy y Recall estima correctamente el comportamiento de las personas que desean el seguro  


16

Recomendaciones:  
Obtener un listado de canales de venta sin cifrado numérico para poder agrupar los canales similares y poder presentar un mejor modelo

Compartir el costo por unidad de los canales de venta y la ganancia promedio de la venta de un seguro para poder evaluar cual de los canales de venta no son viables por la cantidad de seguros vendidos

Utilizar técnicas de Big Data para realizar mejores modelos de regresión logística, arbol de decisión y Random Forest  
  


17

Muchas gracias por su atención, este proyecto fue presentado por Mario Morales, Estuardo Umaña y Alejandro Cortes