Trabajo practico integrados - Análisis de Datos

Santiago Rivier

18 de junio de 2021

1. Introducción

El presente trabajo practico integrador consiste en el procesado de un dataset correspondiente al clima en Australia para luego predecir si lloverá al día siguiente.

2. Desarrollo

2.1. Análisis exploratorio

2.1.1. Variables de Entrada (Numéricas)

En esta primera etapa se analizo el dataset para ver la cantidad de categorías que contenía, cantidad de NaN y algunos valores estadísticos de cada categoría. Luego se procedió a realizar un análisis mas exhaustivo de las variables de entrada, inicialmente con las variables numéricas. Como primer paso se intuyo que tipo de variables podían estar fuertemente correlacioadas con otras y se imprimieron gráficos de caja para verificar. El primer análisis, fue ver que tan correlacionado podían llegar a estar la temperatura mínima, temperatura máxima, temperatura a las 9am y temperatura a las 3pm. Esto arrojó como resultado que la temperatura máxima y la temperatura a las 3pm están fuertemente correlacionados, como así también, la temperatura mínima y la temperatura a las 9am.

Luego se analizo la incidencia solar, la nubosidad a las 9am y la nubosidad a las 3pm. Entre la incidencia solar y las nubosidades se puede ver en gráfico de histogramas que tienen una pequeña correlación pero no dice mucha información. Lo que si se puede observar es que los histogramas de nubosidad presentan picos a la izquierda y derecha y puede llegar a representar días nublados y días no nublados, Dicho histograma se puede trabajar con "Gausian Mixture Model"para extraer sus componentes. Por ultimo, para asegurarse de que realmente están correlacionados, se realizo un gráfico scatter y se logro comprobar que están inversamente correlacionados la incidencia solar con la nubosidad.

Por otro lado, se analizo la lluvia con respecto a la humedad a las 3pm y la humedad a las 9am. De este gráfico scatter sacamos que dichas variables están fuertemente correlacionados y que cerca de un 100 % de humedad se presentan algunos outliers. Esto hay que tenerlo en cuenta al modelo de entrenar el modelo o realizar la imputación de las variables.

2.1.2. Variables de Entrada (Categóricas)

Para el caso de las variables categóricas, se analizo la cantidad de categorías que contenía la variable Locación y se analizo si en alguna de esas locaciones había una mayor cantidad de días lluvioso el cual, no se logro inferir nada por lo que esta variable se decidió eliminar del dataset por el momento. Una posible solución habría sido analizar la ubicación de la localidad y la dirección del viento para así analizar la probabilidad de que llueva dado que el viento provenga del mar.

Por otro lado, tenemos como variable categorica la direccion del viento el cual se pudo obtener que contiene 16 categorías o direcciones de viento (17 si se consideran los NaN).

Además, se analizo la variable RainToday, dicha variable tiene una fuerte correlación con la predicción de si lloverá al día siguiente. La cantidad de NaN que posee no es demasiado elevada y se puede observar en el grafico de lluvias mensuales que hay una mayor cantidad de lluvias para los meses de Junio, Julio y Agosto.

2.1.3. Variables de salida

Como variable de salida tenemos RainTomorrow (llueve mañana) lo que primero se analiza la cantidad de NaN obteniendo como resultado un 2,24 % de NaN respecto a la salida. Además, se puede observar que el dataset esta fuertemente des balanceado hacia el lado de los días que no llueven.

2.2. Esquema de validación de resultados

En esta etapa se eliminan las muestras con salidas NaN y se particiona el dataset en un 70 % para entrenamiento y un 30 % para test. En esta etapa me surgió una complicación que como luego se realiza la limpieza y preparación de los datos sobre los datos de entrenamiento, cuando llegaba la hora de implementar el modelo las filas y columnas de mis datos de entrenamiento no coincidían con los de test. Es por esto que para solventar este problema se realizo la separación del dataset luego de aplicar el algoritmo de imputación MICE y antes de realizar el balanceo del dataset por oversampling para no introducir valores ficticios a la salida.

2.3. Limpieza y preparación de datos/ingeniería de features

Como primera instancia se analizo cual variable contenía mayor cantidad de valores NaN dando como resultado la incidencia solar y la nubosidad.

Luego se tomaron las variables categóricas y se transformaron a variables numéricas. Primero se tomo la variable RainToday, se eliminaron las filas que contenían NaN ya que no son demasiados y esta variable tiene gran correlación con RainTomorrow. Luego se utilizo el método de LabelEncoder para transformar de variable categórica a variable numérico.

Por otro lado, para el caso de las direcciones del viento, los valores de NaN se imputaron por la moda y se realizo un OneHotEncoding para convertirlas a variables numéricas.

La variable Date se separo en Años, meses y días y luego Date y Locación se eliminaron del dataset en este caso. De este modo nuestro dataset ya estaría limpio de cualquier variable categórica.

Para el caso de la imputación de las variables numéricas, se realizo la imputación por MICE. Se estudiaron para 4 casos (None, 1, 2, 3) y el que mejor resultados dio para el caso nuestro resulto ser None por lo que la imputación se realizo con ese parámetro.

Luego, con todo el dataset sin NaN y totalmente numérico, se realizo la separación de datos de entrenamiento y test.

Por ultimo, para el caso del dataset de entrenamiento se realizo un balanceo del mismo por el método de oversampled y se eliminaron las categorías que estaban fuertemente correlacionadas para no repetir los datos.

2.4. Entrenamiento de modelos

Para el caso del entrenamiento del modelo se realizo por dos métodos, el primero Random Forest y el segundo Logistic Regresion dando como mejor resultado el modelo de Random Forest.

3. Conclusiones

Se lograron aplicar los conceptos aprendidos durante el cursado de la materia a un caso practico real y se obtuvieron resultados muy buenos.

Dicho desarrollo se puede continuar y mejorar en muchos aspectos pero por temas de complejidad y tiempos se llego hasta esta instancia.