Trabajo final integrador Análisis de Datos

June 24, 2021

Fernando Emir Garade

0.1 Introducción

Para la realización del trabajo integrador se utilizó un conjunto de datos relativos al estado del tiempo en distintas ciudades de Australia.

El objetivo final es crear un modelo que permita predecir, a partir de información del día actual, si el día siguiente va a llover o no.

0.2 Estructura del dataset

El dataset está compuesto por 23 variables y 145460 registros. Una de estas variables es RainTomorrow y es la variable que tenemos intención de predecir a partir del resto de la información.

En las 22 variables restantes tenemos datos acerca de la fecha, la ciudad, la humedad, la temperatura, la presión, direcciones y velocidades de vientos y ráfagas, cantidad de lluvia de ese día, nubosidad e información acerca de la evaporación y la luz solar para cada registro.

0.3 Características de las variables

0.3.1 Variables numéricas

Algunas de las variables como MinTemp, MaxTemp, WindGustSpeed, WindSpeed3pm, Humidity9am, Humidity3pm, Pressure9am, Pressure3pm, Temp9am y Temp3pm tienen una distribución de valores que se asemejan a una distribución normal. Por supuesto, hay ciertas diferencias en el sesgo y apuntamiento respecto de una normal pero podemos decir que se asemejan bastante.

Las demás variables tienen distribuciones más sesgadas y con distintos grados de apuntamiento.

En general, las variables que miden una misma magnitud en distintos puntos del día se encuentran muy correlacionadas entre sí.

0.3.2 Variables categóricas

Los valores que toman las variables categóricas tienen sentido y no hay grandes desbalances en cuanto a cantidad de registros para cada valor posible de cada variable, con excepción de la variable RainToday la cual se encuentra muy desbalanceada. Esta variable nos indica que solo el 28% de los registros están asociados a días de lluvia y el resto a días que no llovió.

0.4 Esquema de validación de resultados

En principio el dataset completo fue dividido en 22 columnas como variables predictoras (X) y una variable a predecir (y). Luego, cada uno de estos conjuntos se separó en grupos de entrenamiento y de testeo del 80% y 20% de los datos respectivamente.

La variable objetivo RainTomorrow fue encodeada a un valor de 1 cuando el valor era 'Yes' y 0 cuando el valor era 'No'.

0.5 Limpieza y preparación de datos

0.5.1 Valores faltantes

Los registros con valores faltantes en la variable respuesta fueron eliminados porque de mantenerlos, no tendríamos cómo saber si la predicción para esos registros fue buena o mala.

De las 23 columnas en el conjunto de variables predictoras, 22 tenían algún porcentaje de datos faltantes.

Las variables más afectadas son Sunshine, Evaporation, Cloud3pm y Cloud9am cada una con más del 37% de datos faltantes. Se verificó que esta información faltante se explicaba por las ciudades de los registros, ya que existían ciudades con todos los datos faltantes para estas variables. Por esta razón y la gran cantidad que significaban, se decidió eliminar estas columnas y no considerarlas para el análisis.

Se eliminaron los registros que contenían datos faltantes de variables que representaban menos del 5% del total de datos.

Para las demás variables numéricas se imputó por la media para evitar la distorsión que pueda introducir la media (por valores outliers o por el sesgo).

Para las variables categóricas se imputó por la moda o valor más frecuente.

0.5.2 Ingeniería de Features

La variable Date la convertimos a tres variables numéricas Year, Month y Day para almacenar información relevante a estacionalidad. Particularmente habíamos notado que la mayor estacionalidad se daba de forma anual, coincidiendo con las distintas estaciones del año.

Para las ciudades se decidió utilizar la técnica de OHE. Esta técnica es buena porque es sencilla de realizar y de interpretar, pero tiene la desventaja de generar muchas variables nuevas.

Para las variables categóricas asociadas a la dirección de los vientos se decidió encodearlas a partir del ángulo al que corresponden en la rosa de los vientos.

0.5.3 Detección y tratamiento de outliers

Para la detección y tratamiento de valores extremos o outliers se ensayaron dos técnicas de las vistas en clase. Una de ellas la técnica del rango intercuartílico y la otra llamada LOF (Local Outlier Factor). La primera reducía en gran medida el tamaño del dataset sin conseguir ninguna mejora en las métricas de performance del modelo de regresión logística (uno de los utilizados). La segunda tampoco conseguía una mejoría en estos indicadores, pero la cantidad de registros que se descartaban era mucho menor por lo que la preferimos.

0.5.4 Relación entre variables de entrada

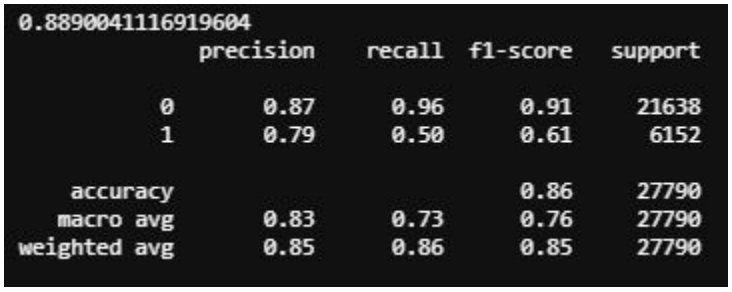
Se estudió la relación de las variables de entrada del modelo utilizando la correlación de Pearson, la correlación de Spearman, la Información Mutua y el test ANOVA. Tanto la primera como la segunda evidencian altas correlaciones entre variables referentes a la misma magnitud (MaxTemp, MinTemp, Temp9am, Temp3pm o Pressure9am con Pressure3pm, entre otras). Luego, con los valores de Información Mutua y de ANOVA pudimos identificar, entre las variables más correlacionadas, cuáles eran más relevantes para nuestra variable objetivo. Se probó eliminar variables con alta correlación y poca información mutua / F ratio, pero esto no aportó ninguna mejora a la performance del modelo.

0.5.5 Balanceo de clases variable objetivo

Dado que las muestras que pertenecían a la clase 1 ('Yes') de la variable objetivo se encontraban subrepresentadas en el dataset (igual de subrepresentadas en todos los conjuntos), se exploró la posibilidad de balancear la cantidad de muestras pertenecientes a cada clase en el conjunto de entrenamiento. Esto lo hicimos duplicando las muestras existentes de esa clase y no trajo mayores beneficios mas que aumentar un poco el recall para la clase en cuestión.

0.6 Entrenamiento y evaluación de modelos

Con los datasets descriptos y debidamente estandarizados se entrenaron 3 modelos: Uno de regresión logística, otro de árboles de decisión y un tercero de random forest. El modelo de regresión logística y el de random forest superaron en performance al de árboles de decisión siendo el mejor de los 3 el random forest. Se hizo una pequeña optimización de hiperparámetros y se reentrenó este último obteniendo los siguientes resultados:

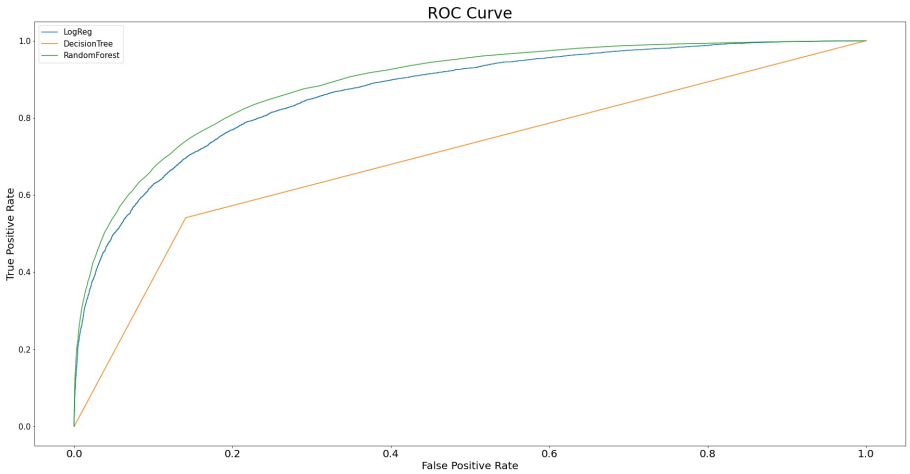


0.8890041116919604

	precision	recall	f1-score	support
0	0.87	0.96	0.91	21638
1	0.79	0.50	0.61	6152
accuracy			0.86	27790
macro avg	0.83	0.73	0.76	27790
weighted avg	0.85	0.86	0.85	27790

El primer valor corresponde al AUC del modelo.

También podemos ver la comparación de las curvas ROC con los otros dos modelos aquí:



Y por último las 10 features más importantes para el modelo resultaron: