Tabal de contenido

Proyecto de Predicción de Lluvia en Australia

1. Descripción del Problema Predictivo:

Australia, conocida por su clima seco y desafiante, enfrenta limitaciones en el acceso a fuentes de agua. La gestión y regulación del agua son responsabilidades gubernamentales cruciales en este escenario. Los precios del agua fluctúan diariamente, siendo altamente influenciados por la demanda, especialmente por parte del sector agrícola, que utiliza el 80% de los recursos hídricos del país. La demanda de agua en agricultura varía significativamente según la lluvia; los días lluviosos muestran una menor demanda. Esta fluctuación, debido a la incertidumbre climática, dificulta la fijación de precios para este recurso vital.

El propósito de este proyecto es prever si lloverá al día siguiente en diversas regiones de Australia. Utilizaremos un extenso conjunto de datos que abarca aproximadamente 10 años de observaciones meteorológicas diarias de varias estaciones meteorológicas australianas. La columna central de interés, "RainTomorrow", representa si se anticipa una lluvia de más de 1 mm al día siguiente (Sí/No). Esta predicción tiene un impacto significativo en la toma de decisiones cotidianas de individuos y organizaciones, dada la fuerte influencia de la lluvia en múltiples actividades y sectores.

2. Dataset:

El conjunto de datos contiene información detallada sobre diversas variables meteorológicas, incluyendo la cantidad de lluvia en milímetros, velocidad del viento, temperatura, horas de sol, dirección del viento, entre otros. Estos datos se han registrado a lo largo de 10 años, resultando en un total de 145,460 filas que describen las condiciones climáticas de días específicos. A pesar de contar con 23 columnas, debido al análisis de 49 ubicaciones, se requiere transformar los datos en un formato donde se tendra un mínimo de 72 columnas. Adicionalmente, cada día contiene una columna con una respuesta de booleana (Sí/No) que indica si llovió al día siguiente, considerando la opción afirmativa cuando la precipitación excede 1mm.

3. Métricas de Desempeño:

En este caso, estamos abordando un problema de clasificación binaria, en el que buscamos predecir si lloverá al día siguiente en diferentes regiones de Australia.

Para analizar este tipo de problemas, es fundamental comprender qué resultados puede arrojar nuestro modelo predictivo. En este contexto, tenemos dos posibles respuestas: lluvia sí (Sí) o lluvia no (No). A partir de estas opciones, se desprenden cuatro escenarios de predicción del modelo:

- **Verdaderos Positivos (TP):** El modelo predice lluvia mañana (Sí) y realmente llueve al día siguiente.
- **Verdaderos Negativos (TN):** El modelo predice que no lloverá mañana (No) y efectivamente no llueve al día siguiente.
- **Falsos Positivos (FP):** El modelo predice lluvia mañana (Sí), pero no llueve al día siguiente.
- **Falsos Negativos (FN):** El modelo predice que no lloverá mañana (No), pero llueve al día siguiente.

Estos escenarios nos permiten evaluar la calidad y confiabilidad de nuestras predicciones, proporcionando información valiosa sobre el rendimiento de nuestro modelo en predecir la lluvia en Australia. Además, consideraremos métricas de negocio que cuantifican el costo asociado a predicciones incorrectas, siendo de gran relevancia en términos financieros y operativos.

Además, consideraremos métricas de negocio como el costo asociado con falsos positivos y falsos negativos, ya que estas predicciones incorrectas pueden tener implicaciones financieras y operativas importantes.

3.1 Matriz de Evaluación

Esta representación gráfica ilustra las predicciones generadas por el modelo. La matriz resume el número total de predicciones acertadas (verdaderos positivos), predicciones incorrectas de lluvia (falsos positivos), predicciones correctas de no lluvia (verdaderos negativos) y predicciones incorrectas de no lluvia (falsos negativos). Es un recurso ampliamente utilizado en diversos modelos, proporcionando información valiosa sobre el tipo de desafío que enfrenta el modelo y si este se encuentra balanceado o desbalanceado.

4.1 Tasa de Verdaderos Positivos (TPR) Sensibilidad

Esta métrica indica la proporción de predicciones positivas correctas sobre variables que son verdaderamente positivas. En nuestro caso, un verdadero positivo se registra si el modelo predice lluvia para el día siguiente y efectivamente llueve. La TPR se define como sigue:

$$TPR = \frac{TP}{TP+FN}$$

5.1 Tasa de Falsos Positivos (FPR)

Esta tasa señala la frecuencia de predicciones positivas incorrectas sobre variables que no son verdaderas. En nuestro problema, un falso positivo se presenta cuando se predice lluvia para el día siguiente pero, en realidad, no ocurre. La FPR se calcula de la siguiente forma:

$$TPR = \frac{FP}{FP + TN}$$

6.1 Curva de Características de Operación del Receptor (ROC)

La curva ROC (siglas en inglés de Receiver Operating Characteristic) es un gráfico que representa la compensación entre la Tasa de Verdaderos Positivos (TPR) y la Tasa de Falsos Positivos (FPR). Un modelo predictivo se considera mejor cuanto mayor sea la TPR y menor la FPR para cada umbral seleccionado. Así, los clasificadores con curvas cercanas a la esquina superior izquierda del gráfico se consideran más efectivos, como se muestra en la Figura 1.

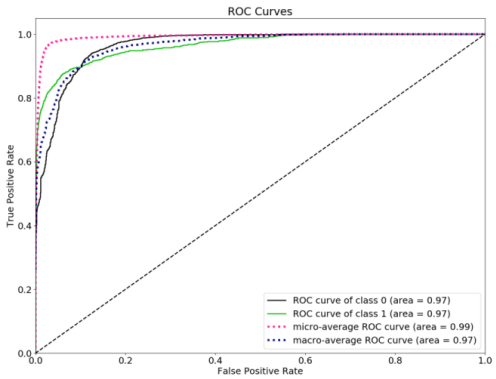

Figura 1: Ejemplo de curvas ROC para aprendizajes de máquinas

7.1 Área Bajo la Curva de Precisión-Recall (AUC)

El AUC se emplea como una métrica complementaria a la ROC para comparar el área bajo la curva. Ayuda a comparar diferentes métodos predictivos utilizados e identificar cuál tiene un mejor desempeño en el contexto del problema.

4. Criterio de Desempeño Deseado en Producción:

Buscamos que nuestro modelo de predicción del clima en Australia para un día específico logre una tasa de aciertos del 80%, esto se refiere a la capacidad del modelo de predicción del clima para hacer pronósticos precisos en un nivel mínimo del 80% de las ocasiones. Un alto nivel de tasa de acierto es importante porque demuestra la capacidad del modelo para predecir el clima de manera confiable. En el contexto de la gestión del recurso hídrico en Australia, una alta precisión en las predicciones es crucial para tomar decisiones informadas y evitar errores costosos. Adicionalmente, es fundamental mantener un índice de Falsos Positivos (False Positive Rate) por debajo del 15%.

Este último aspecto reviste gran importancia en el contexto de los mercados del agua en Australia. Predecir de manera errónea una lluvia que en realidad no ocurrirá (FP) conlleva una disminución en el precio del agua. Esta predicción genera una expectativa de alta oferta de agua, ya que se anticipa lluvia. Sin embargo, al tratarse de una predicción falsa (pues en realidad no lloverá), se subestima el precio del agua. Esto puede resultar en pérdidas financieras para las entidades responsables de la gestión del recurso hídrico, debido a una demanda mayor a la prevista.