Caso de estudio
Temperaturas mínimas diarias en Melbourne, Australia, 1981-1990
Introducción a los métodos de aprendizaje automático

Noviembre 2019

Autores:

- Santiago Casás
- Agustín Betancor
- Martín Rose
- Mauricio Coniglio
- Gianni Iaquinta

ÍNDICE

Introducción

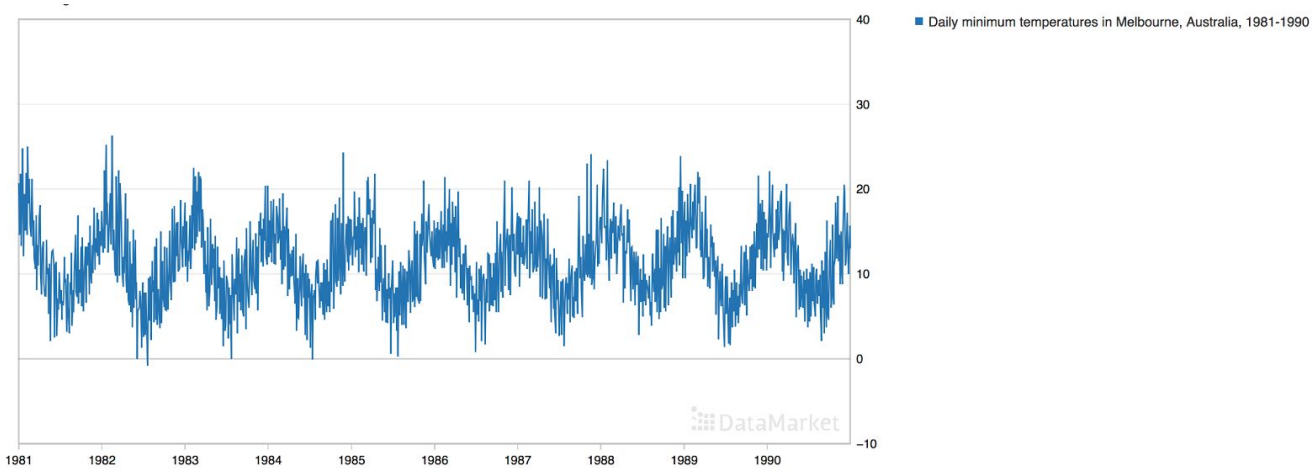
El caso de estudio se basa en la técnica Time series forecasting la cual es una de las técnicas de análisis predictivo más antigua conocida. La misma se basa en una secuencia de datos, observaciones o valores, medidos en determinados momentos y ordenados cronológicamente para luego con dicha informacion historica hacer pronósticos sobre el valor de los mismos datos en el futuro.
Este tipo de técnica tiene dos diferencias importantes con otros modelos predictivos supervisados:

1. En este tipo de modelos se focaliza en pronosticar una variable específica, dado que se sabe como esta variable ha cambiado con el tiempo anteriormente. Mientras que en otras técnicas el componente del tiempo en los datos no era importante o no estaba disponible.
2. En este tipo de modelos no interesan los datos de otros atributos que pueden influir en la variable objetivo, las variables independientes o predictoras no son estrictamente necesarias para predecir en este tipo de modelos.

Los métodos de Time series forecasting se denominan métodos de pronóstico basados en datos, donde no hay diferencia entre un predictor y un objetivo. Dado que el predictor es también la variable objetivo. El mas simple de estos métodos es la regresión lineal con el tipo de función $y(t) = a + b * t$. Siendo y(t) el valor de la variable en el tiempo a predecir y a,b siendo los valores de los coeficientes para poder realizar la predicción. Pero también se pueden utilizar otro tipo de funciones como exponencial.

Preparación de los datos

El dataset utilizado para este estudio fue el dataset “Minimum Daily Temperature Dataset” el cual fue recomendado por el creador de Mastering Machine Learning en el siguiente blogpost: https://machinelearningmastery.com/time-series-datasets-for-machine-learning/
Dicho dataset se compone de dos atributos, uno correspondiente a la fecha en la cual se hizo la medición y el otro a la temperatura medida.
Las mediciones se hicieron diariamente dentro de un periodo de 10 años (1981-1990). Por lo tanto el dataset contiene 3650 instancias. Cabe destacar que no hay faltantes en los datos y que tampoco se encontraron outliers.
En la siguiente gráfica puede apreciarse la distribución de los datos en el periodo de 1981-1990:



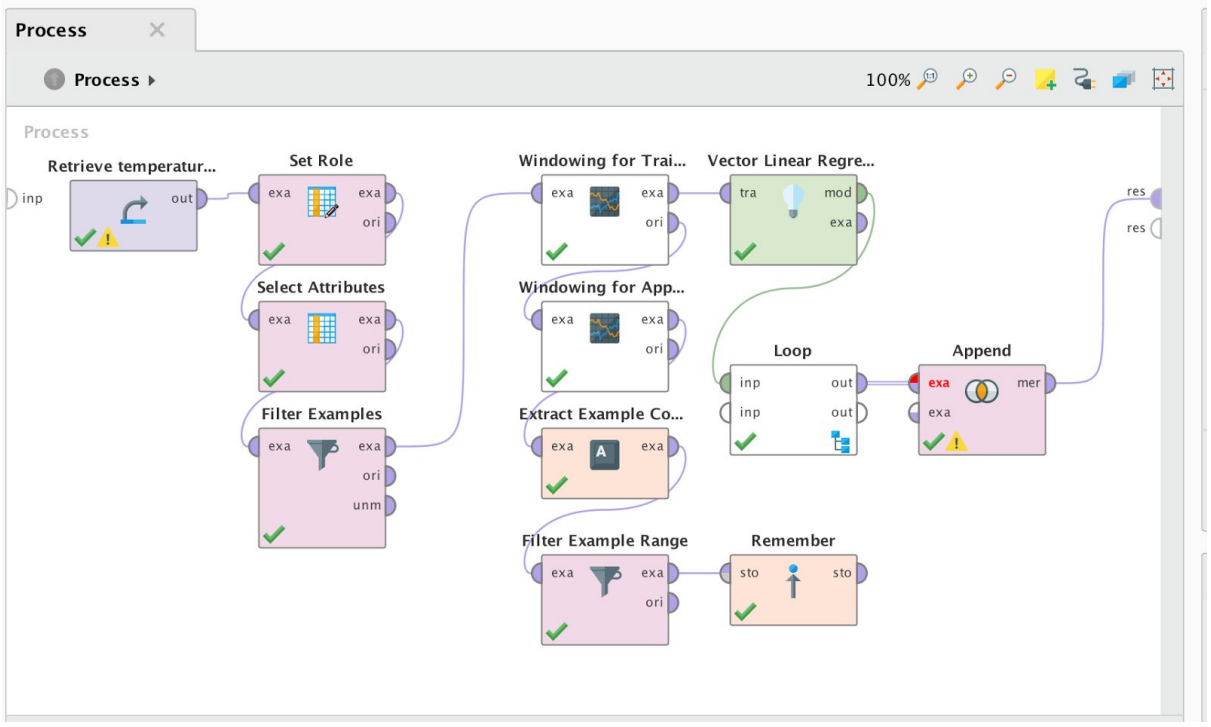
Definición del problema

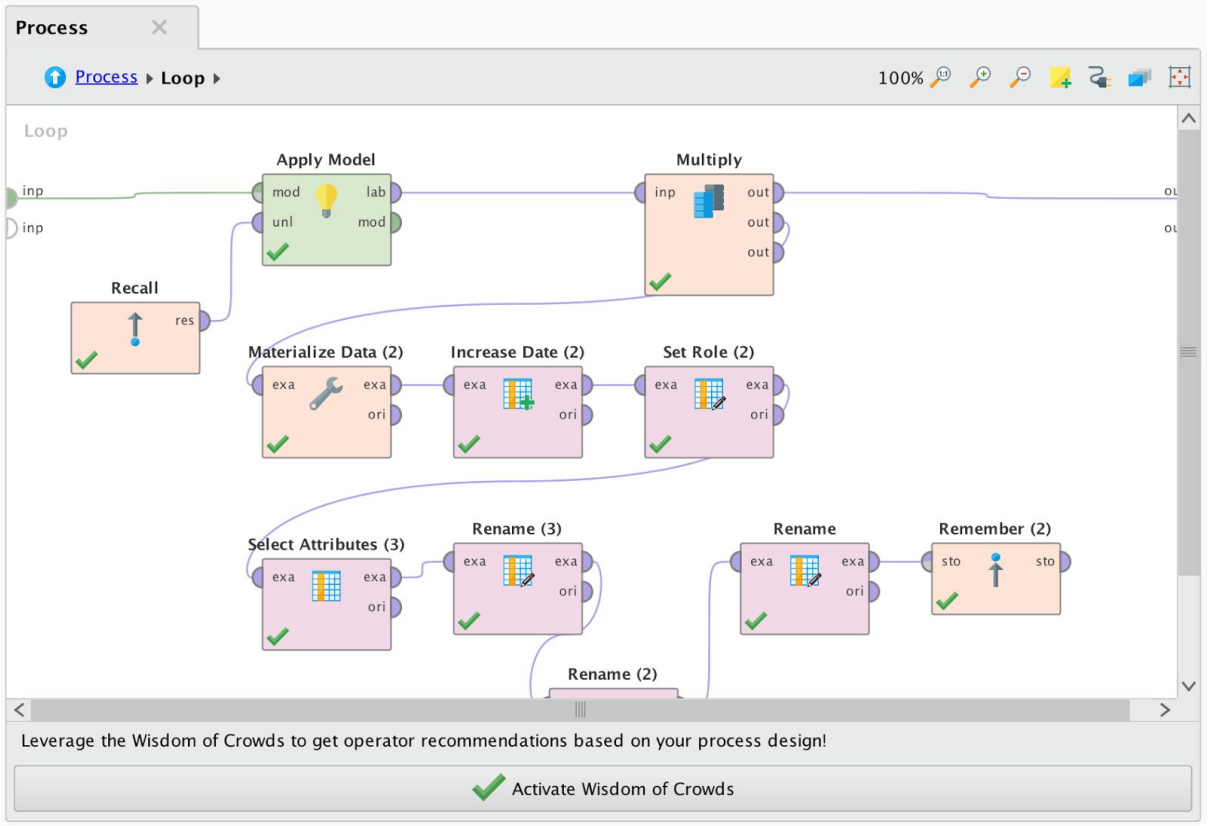
El problema planteado en este caso de estudio es poder crear y entrenar a partir de los datos históricos correspondientes al periodo de 1981-1990 de las temperaturas mínimas diarias de Melbourne, Australia, para poder luego aplicarlo a datos actuales similares para poder predecir la salida, en este caso la temperatura mínima para ese dia.
Dada la estructura intrínseca del problema es ideal para aplicar la técnica de Time Series Forecasting. A continuación se explicara como se aplicó dicho tecnica a la solución del problema.

Solución planteada

Como solución a dicho problema se presentó el siguiente modelo creado en RapidMiner. En el mismo lo que se hizo fue, enseguida de cargar el dataset en la plataforma, utilizar un operador Set Role para indicar como id al atributo Date. Los siguientes operadores que se encuentran en el diagrama era para manejar los datos faltantes, pero en el caso de este dataset no había ninguno. A continuación utilizamos el operador Windowing para dado una tupla agregar componentes a esta utilizando el parámetro window para definir la cuantos registros anteriores se quiere agregar a cada tupla y para marcar cuál de los atributos se va a querer predecir, en este caso escogimos 365 por lo que a cada tupla se le agrega 365 columnas correspondientes a los 365 días previos a esa observación y el atributo a predecir es la temperatura.
En paralelo se utiliza otro operador windowing junto con un operador extract example y un filter example para sacar la última tupla del dataset. Esta última tupla se guarda en memoria para luego ser sustituida por la tupla predecida posterior y guardar acá el dataset predecido.
Luego de generado el modelo y guardado la última tupla del dataset se utiliza un operador loop repetidas veces, 365 veces, para predecir las temperaturas de cada uno de los días del año siguiente.
Dentro del loop es donde se realiza la predicción de las temperaturas para los siguientes días. Para realizar esto comienza levantando la tupla generada anteriormente, y luego se le aplica el modelo. Con la tupla ya cargado con la predicción hecha se pasa a renombrar las

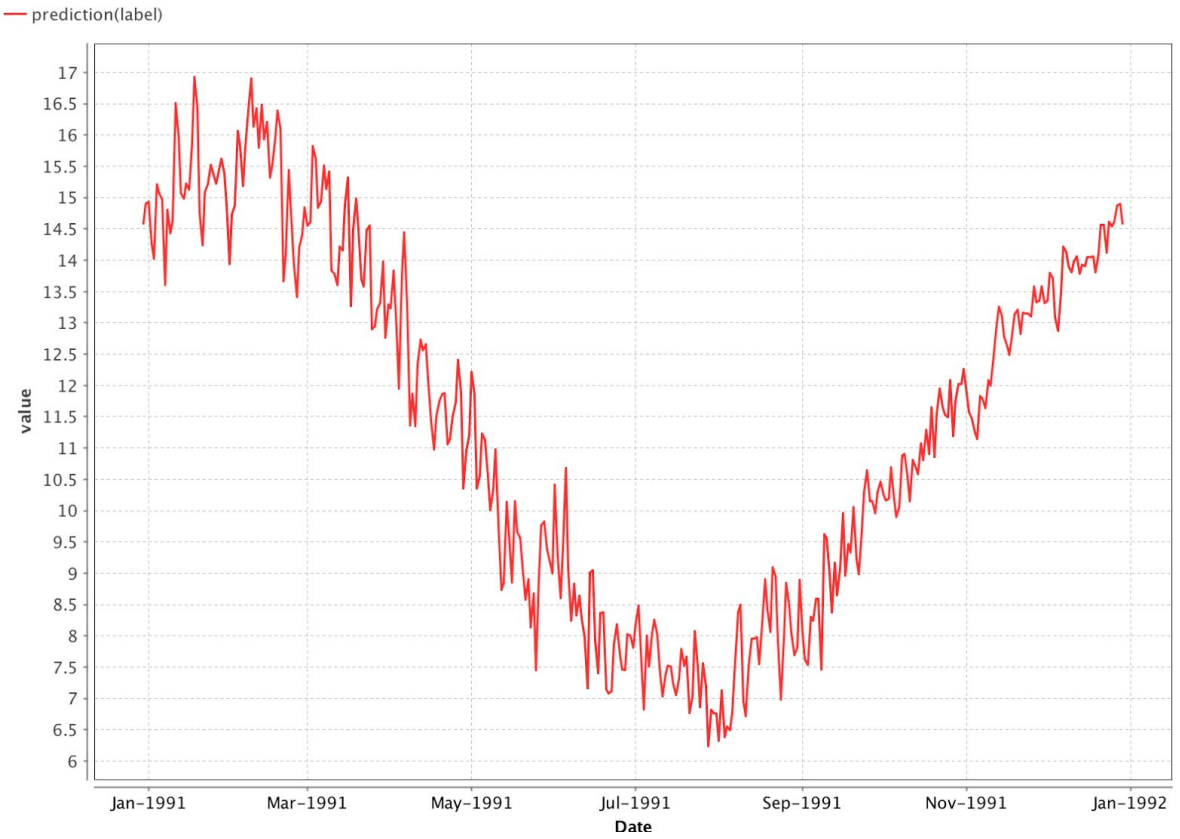
columnas para que se pueda utilizar esta nueva tupla para predecir el siguiente valor. Se elimina la columna asociada al registro más viejo, y se cambian todos los nombres(restandoles un dia al nombre de la columna). Luego de esto se guarda la tupla con el operador remember.
Luego que se completa el loop se agrupan todas las tuplas predichas dentro del loop para su extracción.

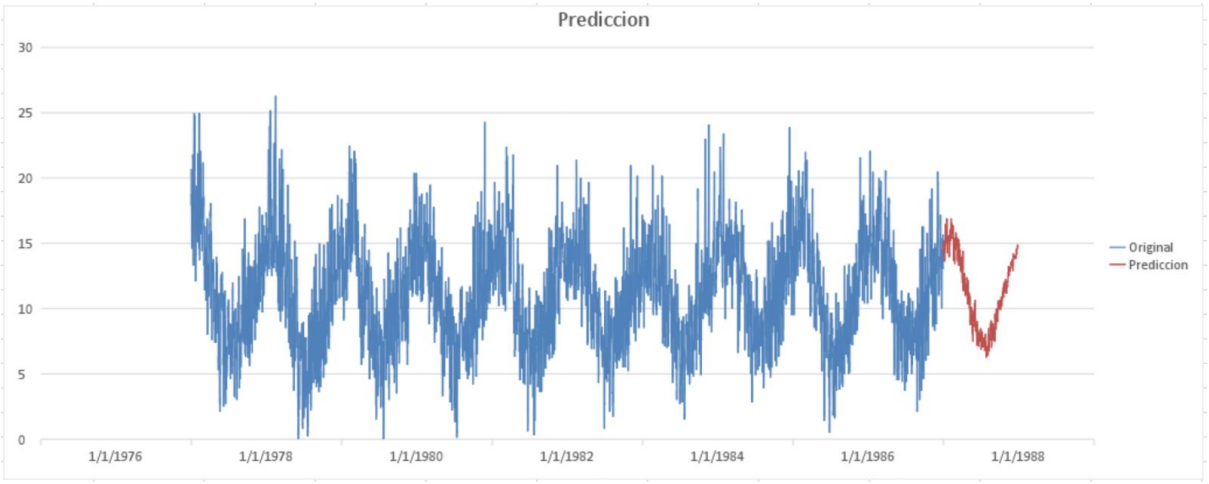




Resultados obtenidos

Como se puede apreciar en la primera gráfica vemos como se comporta la predicción de la temperatura mínima para el año 1991 (el siguiente al último del dataset). La curva de la misma se comporta de manera razonable, dado que, como se puede ver en la segunda gráfica, en la cual se compara la predicción con los datos que teníamos de los 10 años anteriores, la curva se comporta de manera similar, lo cual nos da a entender de que el resultado fue correcto. El RMSE obtenido de nuestro modelo fue root_mean_squared_error: 2.409 +/- 0.085 grados centígrados, de lo cual se desprende que el modelo creado performa bastante bien.





Conclusiones

- Técnica muy utilizada actualmente en la industria, estadistica, finanzas, etc.
- Se precisa un dataset confiable y grande para poder realizar una buena predicción. Si esto no se cumple pueden haber predicciones erroneas.
- En rapidminer realizar este caso de estudio fue tedioso al tener que renombrar 365 veces las variables, dado que para generar el dataset de predicción en la ventana de 365 días se debe agregar la nueva predicción y quitar la mas vieja de todas.

Referencias

- https://machinelearningmastery.com/time-series-forecasting/
- https://machinelearningmastery.com/time-series-datasets-for-machine-learning/
- https://docs.rapidminer.com/studio/operators/
- Libro: “Predictive Analytics and Data Mining, Kotu & Desphande”